Chương 4

Thống kê. Ước lượng tham số

BÀI 11 (2 tiết)

4.1 Lý thuyết mẫu

"Thống kê là một khoa học đồng thời là một công nghệ cung cấp cho ta những phương pháp, công cụ để thu thập và tạo dữ liệu, trình bày và phân tích dữ liệu để hiểu nội dung ẩn chứa trong dữ liệu. Từ đó rút ra những thông tin, tri thức hữu ích và đưa ra những quyết định, chính sách thích hợp".[1]

Thống kê toán là bộ môn toán học nghiên cứu quy luật của các hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu thập và xử lý số liệu thống kê các kết quả quan sát về những hiện tượng ngẫu nhiên này. Nếu ta thu thập được các số liệu liên quan đến tất cả đối tượng cần nghiên cứu thì ta có thể biết được đối tượng này (phương pháp toàn bộ). Tuy nhiên trong thực tế điều đó không thể thực hiện được vì quy mô của các đối tượng cần nghiên cứu quá lớn hoặc trong quá trình nghiên cứu đối tượng nghiên cứu bị phá hủy. Vì vậy cần lấy mẫu để nghiên cứu.

Mục này giới thiệu về phương pháp lấy mẫu ngẫu nhiên và các thống kê thường gặp của mẫu ngẫu nhiên.

4.1.1 Tổng thể và mẫu

Khái niệm tổng thể

Khi nghiên cứu các vấn đề về kinh tế - xã hội, cũng như nhiều vấn đề thuộc các lĩnh vực vật lý, sinh vật, quân sự . . . thường dẫn đến khảo sát một hay nhiều dấu hiệu (định tính hoặc định lượng) thể hiện bằng số lượng trên nhiều phần tử. Tập hợp tất cả các phần tử này gọi là tổng thể hay đám đông (population). Số phần tử trong tổng thể có thể là hữu hạn hoặc vô hạn. Cần

[1] Đặng Hùng Thắng, Trần Mạnh Cường (2019), *Thống kê cho Khoa học xã hội và Khoa học sự sống*, NXB Đại học Quốc gia Hà Nội.

nhấn mạnh rằng ta không nghiên cứu trực tiếp bản thân tổng thể mà chỉ nghiên cứu dấu hiệu nào đó của nó.

Ký hiệu N là số phần tử của tổng thể; $\mathcal{X}$ là dấu hiệu cần khảo sát.

Ví dụ 4.1. (a) Muốn điều tra thu nhập bình quân của các hộ gia đình ở Hà Nội thì tập hợp cần nghiên cứu là các hộ gia đình ở Hà Nội, dấu hiệu nghiên cứu là thu nhập của từng hộ gia đình (dấu hiệu định lượng).

(b) Một doanh nghiệp muốn nghiên cứu các khách hàng của mình về dấu hiệu định tính có thể là mức độ hài lòng của khách hàng đối với sản phẩm hoặc dịch vụ của doanh nghiệp, còn dấu hiệu định lượng là số lượng sản phẩm của doanh nghiệp mà khách hàng có nhu cầu được đáp ứng.

Một số lý do không thể khảo sát toàn bộ tổng thể

(a) Do quy mô của tập hợp cần nghiên cứu quá lớn nên việc nghiên cứu toàn bộ sẽ đòi hỏi nhiều chi phí về vật chất và thời gian, có thể không kiểm soát được dẫn đến bị chồng chéo hoặc bỏ sót.

(b) Trong nhiều trường hợp không thể nắm được toàn bộ các phần tử của tập hợp cần nghiên cứu, do đó không thể tiến hành toàn bộ được.

(c) Có thể trong quá trình điều tra sẽ phá hủy đối tượng nghiên cứu...

Do đó thay vì khảo sát tổng thể, ta chỉ cần chọn ra một tập nhỏ để khảo sát và đưa ra quyết định.

Khái niệm tập mẫu

Thông thường quy mô của một tổng thể là rất lớn. Vì thế người ta thường chọn ra một tập hợp con các cá thể để nghiên cứu. Việc chọn ra từ tổng thể một tập hợp con nào đó được gọi là phép lấy mẫu. Tập hợp con được chọn được gọi là mẫu (sample). Số cá thể trong mẫu được gọi là kích thước mẫu, ký hiệu là n.

Ví dụ 4.2. Ta muốn đánh giá số giờ trong một ngày mà một sinh viên đại học sử dụng Facebook. Vì số sinh viên đại học rất lớn, nên ta không thể điều tra trên tất cả các sinh viên được. Ta chọn ngẫu nhiên một mẫu gồm 50 sinh viên để khảo sát và tìm được số giờ trung bình dùng Facebook của 50 sinh viên này là 4,7 giờ. Con số này cho ta một hình ảnh về việc sử dụng Facebook của các sinh viên đại học.

Ví dụ 4.3. Ta muốn đánh giá tỷ lệ phế phẩm trong các sản phẩm của nhà máy A. Giả sử nhà máy chế tạo được 400000 sản phẩm. Ta không đủ thời gian và tiền bạc để xem xét được toàn bộ 400000 sản phẩm. Ta chọn ra một mẫu gồm 300 sản phẩm để kiểm tra và phát hiện ra có 18

sản phẩm mắc lỗi. Tỷ lệ phế phẩm trong mẫu kiểm tra là $18/300 = 6\%$. Từ đó ta nhận định tỷ lệ phế phẩm của nhà máy A khoảng 6%.

Chương 4 và Chương 5 sẽ nghiên cứu tổng thể thông qua mẫu. Nói nghiên cứu tổng thể có nghĩa là nghiên cứu một hoặc một số đặc trưng nào đó của tổng thể. Khi đó, ta không thể đem tất cả các phần tử trong tổng thể ra nghiên cứu mà chỉ lấy một số phần tử trong tổng thể ra nghiên cứu và làm sao qua việc nghiên cứu này có thể kết luận được về một hoặc một số đặc trưng của tổng thể mà ta quan tâm ban đầu.

Một số cách chọn mẫu cơ bản

Các kết luận suy diễn từ mẫu có đáng tin cậy không? Câu nói nổi tiếng của Mark Twain, nhà văn Anh "Có ba kiểu nói dối: Nói dối, nói dối trắng trợn và thống kê" ("Thera are three kinds of lies: Lies, Damned lies and Statistics"). Tuy nhiên, thống kê không nói dối. Kết quả sai (mà ta gọi là dối trá) do thống kê đưa ra là do phương pháp lấy mẫu không đúng:

1. Việc lấy mẫu đã được tiến hành không khách quan, theo hướng có lợi cho người nghiên cứu.

2. Mẫu được chọn không đại diện.

Ví dụ 4.4. Để điều tra mức thu nhập trung bình của sinh viên tốt nghiệp đại học mới ra trường, nếu mẫu được chọn trong số các sinh viên tốt nghiệp ngành Công nghệ thông tin thì rõ ràng mức lương trung bình trong mẫu không phản ánh trung thực mức lương trung bình của sinh viên mới ra trường nói chung.

Các kết luận suy diễn từ mẫu có đáng tin cậy chỉ đạt được nếu mẫu được chọn phản ánh trung thực, thực sự đại diện cho tổng thể. Do đó vấn đề chọn mẫu là một vấn đề rất quan trọng và phong phú của thống kê. Các kỹ thuật chọn mẫu đúng đắn sẽ giúp ta đảm bảo được tính đại diện trung thực cho tổng thể. Để trả lời cho câu hỏi đặt ra là làm sao chọn được tập mẫu có tính chất tương tự như tổng thể để các kết luận của tập mẫu có thể dùng cho tổng thể, ta sử dụng một trong những cách chọn mẫu sau:

(a) Lấy mẫu ngẫu nhiên: mỗi cá thể của tổng thể được chọn một cách độc lập với xác suất như nhau.

(b) Lấy mẫu theo khối: Tổng thể được chia làm N khối, mỗi khối xem là một tổng thể con. Chọn ngẫu nhiên ra m khối trong N khối đó. Tập hợp tất cả các cá thể của m khối được chọn sẽ được lập thành một mẫu để khảo sát.

Phương pháp này được áp dụng khi ta không liệt kê danh sách tất cả các cá thể trong tổng thể.

(c) Lấy mẫu phân tầng: Chia tổng thể ra một số tầng, sao cho các cá thể trong mỗi tầng khác nhau càng ít càng tốt. Mỗi tầng được coi là một tổng thể con. Trong mỗi tầng ta sẽ thực hiện việc lấy mẫu ngẫu nhiên.

Phương pháp này được sử dụng khi các cá thể quá khác nhau về vấn đề mà nhà nghiên cứu đang quan tâm khảo sát.

Ví dụ 4.5. Tại một trường Đại học có 20000 sinh viên với 5 hệ đào tạo khác nhau: 10000 sinh viên hệ chính quy; 2000 sinh viên hệ liên thông; 2000 sinh viên hệ văn bằng hai; 5000 sinh viên hệ tại chức và 1000 học viên hệ sau đại học. Bộ phận đảm bảo chất lượng tiến hành cuộc khảo sát về chất lượng và mức độ hài lòng của người học. Chọn ngẫu nhiên 1000 sinh viên để khảo sát. Xem mỗi hệ đào tạo là một tầng, số sinh viên ở mỗi tầng được chọn như sau:

Hệ đào tạo	**Số SV**	**% SV**	**Số SV được chọn**
Chính quy	10000	50	500
Liên thông	2000	10	200
Văn bằng hai	2000	10	200
Tại chức	2000	10	200
Sau đại học	1000	5	50
Tổng	**20000**	**100**	**1000**

4.1.2 Mẫu ngẫu nhiên

Biến ngẫu nhiên và quy luật phân phối gốc

Giả sử ta cần nghiên cứu dấu hiệu $\mathcal{X}$ của tổng thể có $E(\mathcal{X}) = \mu$ và $V(\mathcal{X}) = \sigma^2$ (μ và σ chưa biết). Ta có thể mô hình hóa dấu hiệu $\mathcal{X}$ bằng một biến ngẫu nhiên. Thật vậy, nếu lấy ngẫu nhiên từ tổng thể ra một phần tử và gọi X là giá trị của dấu hiệu $\mathcal{X}$ đo được trên phần tử lấy ra thì X là biến ngẫu nhiên có bảng phân phối xác suất là

X	x_1	x_2	$\dots$	x_n
P	$P(X = x_1)$	$P(X = x_2)$	$\dots$	$P(X = x_n)$

Như vậy dấu hiệu $\mathcal{X}$ mà ta nghiên cứu được mô hình hóa bởi biến ngẫu nhiên X, còn cơ cấu của tổng thể theo dấu hiệu $\mathcal{X}$ (tập hợp các xác suất) chính là quy luật phân phối xác suất của X.

Biến ngẫu nhiên X được gọi là biến ngẫu nhiên gốc. Quy luật phân phối xác suất của X là quy luật phân phối gốc, đồng thời $E(X) = \mu$, $V(X) = \sigma^2$.

Các đặc trưng của tổng thể

(a) Xét tổng thể về mặt định lượng : tổng thể được đặc trưng bởi dấu hiệu $\mathcal{X}$ được mô hình hóa bởi biến ngẫu nhiên X. Ta có các tham số đặc trưng sau đây:

1. Trung bình tổng thể: $E(X) = \mu$.
2. Phương sai tổng thể: $V(X) = \sigma^2$.
3. Độ lệch chuẩn của tổng thể: $\sigma(X) = \sigma$.

(b) Xét tổng thể về mặt định tính : tổng thể có kích thước N, trong đó có M phần tử có tính chất A. Khi đó $p = \dfrac{M}{N}$ gọi là tỷ lệ tính chất A của tổng thể.

Khái niệm mẫu ngẫu nhiên

Giả sử tiến hành n phép thử độc lập. Gọi X_i là "giá trị của dấu hiệu $\mathcal{X}$ đo lường được trên phần tử thứ i của mẫu" $i = 1, 2, \ldots, n$. Khi đó, $X_1, X_2, \ldots, X_n$ là n biến ngẫu nhiên độc lập có cùng quy luật phân phối xác suất với X.

Định nghĩa 4.1 (Mẫu ngẫu nhiên)**.** Cho biến ngẫu nhiên gốc X có quy luật phân phối xác suất $F_X(x)$ nào đó. Một mẫu ngẫu nhiên kích thước n được thành lập từ biến ngẫu nhiên X là n biến ngẫu nhiên độc lập có cùng quy luật phân phối xác suất $F_X(x)$ với biến ngẫu nhiên X.

Ký hiệu mẫu ngẫu nhiên: $W_X = (X_1, X_2, \ldots, X_n)$.

Thực hiện một phép thử đối với mẫu ngẫu nhiên W_X tức là thực hiện một phép thử đối với mỗi thành phần X_i của mẫu. Giả sử X_1 nhận giá trị x_1, X_2 nhận giá trị x_2, …, X_n nhận giá trị x_n ta thu được một mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$.

Ví dụ 4.6. Gọi X là "số chấm xuất hiện khi gieo một con xúc xắc". X là biến ngẫu nhiên có bảng phân phối xác suất

X	1	2	3	4	5	6
p	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Nếu gieo con xúc xắc 3 lần và gọi X_i là "số chấm xuất hiện ở lần gieo thứ i", $i = 1, 2, 3$ thì ta có 3 biến ngẫu nhiên độc lập có cùng quy luật phân phối xác suất với X. Vậy ta có một mẫu ngẫu nhiên $W_X = (X_1, X_2, X_3)$ cỡ $n = 3$ được xây dựng từ biến ngẫu nhiên gốc X. Thực hiện một phép thử đối với mẫu ngẫu nhiên này (tức là gieo 3 lần một con xúc xắc). Giả sử lần thứ nhất xuất hiện mặt 6, lần thứ hai xuất hiện mặt 2, lần thứ ba xuất hiện mặt 1 thì ta có một giá trị của mẫu ngẫu nhiên $W_x = (6, 3, 1)$.

Mẫu ngẫu nhiên có thể phản ánh được kết quả điều tra, thực nghiệm bởi vì những kết quả này được coi là một giá trị của nó. Mặt khác mẫu ngẫu nhiên là tập hợp các biến ngẫu nhiên. Do vậy ta có thể nghiên cứu quy luật phân phối xác suất của nó, tức là khái quát được thực nghiệm. Quan hệ giữa mẫu ngẫu nhiên và mẫu cụ thể (hay một giá trị của nó) tương tự quan hệ giữa biến ngẫu nhiên và một giá trị có thể có của nó.

4.1.3 Mô tả giá trị của mẫu ngẫu nhiên

Phân loại dữ liệu

Từ tổng thể ta trích ra tập mẫu có n phần tử. Ta có n số liệu.

(a) Dạng liệt kê: Các số liệu thu được được ghi lại thành dãy $x_1, x_2, \ldots, x_n$.

(b) Dạng rút gọn: Số liệu thu được có sự lặp đi lặp lại một số giá trị thì ta có dạng rút gọn sau:

(b1) Dạng tần số: $(n_1 + n_2 + \ldots + n_k = n)$

Giá trị	x_1	x_2	$\ldots$	x_k
Tần số	n_1	n_2	$\ldots$	n_k

(b2) Dạng tần suất: $(f_k = n_k/n)$

Giá trị	x_1	x_2	$\ldots$	x_k
Tần suất	f_1	f_2	$\ldots$	f_k

(c) Dạng khoảng: Dữ liệu thu được nhận giá trị trong (a, b). Ta chia (a, b) thành k miền con bởi các điểm chia: $a_0 = a < a_1 < a_2 < \cdots < a_{k-1} < a_k = b$.

(c1) Dạng tần số: $(n_1 + n_2 + \ldots + n_k = n)$

Giá trị	$(a_0 - a_1]$	$(a_1 - a_2]$	$\ldots$	$(a_{k-1} - a_k]$
Tần số	n_1	n_2	$\ldots$	n_k

(c2) Dạng tần suất: $(f_k = n_k/n)$

Giá trị	$(a_0, a_1]$	$(a_1, a_2]$	$\ldots$	$(a_{k-1}, a_k]$
Tần suất	f_1	f_2	$\ldots$	f_k

Chú ý, thông thường, độ dài các khoảng chia bằng nhau. Khi đó ta có thể chuyển về dạng rút gọn:

Giá trị	x_1	x_2	$\ldots$	x_k
Tần số	n_1	n_2	$\ldots$	n_k

trong đó x_i là điểm đại diện cho $(a_{i-1}, a_i]$ thường được xác định là trung điểm của đoạn đó: $x_i = \dfrac{1}{2}(a_{i-1} + a_i)$.

Đặt w_i là tần số tích lũy của x_i và $F_n(x_i)$ là tần suất tích lũy của x_i, ta sẽ có

$$w_i = \sum_{x_j < x_i} n_j; \quad F_n(x_i) = \frac{w_i}{n} = \sum_{x_j < x_i} f_j$$

thì $F_n(x_i)$ là một hàm của x_i và được gọi là hàm phân phối thực nghiệm của mẫu hay hàm phân phối mẫu. Chú ý rằng theo luật số lớn (Định lý Béc-nu-li) $F_n(x)$ hội tụ theo xác suất về $F_X(x) = P(X < x)$, trong đó X là biến ngẫu nhiên gốc cảm sinh ra tổng thể (và cả tập mẫu). Như vậy hàm phân phối mẫu có thể dùng để xấp xỉ luật phân phối của tổng thể.

Biểu diễn dữ liệu

Một câu ngạn ngữ Trung Hoa "Một hình ảnh có tác dụng bằng một nghìn lời nói". Để có được một hình ảnh rõ ràng và dễ nhớ về mẫu các giá trị của biến ngẫu nhiên X, ta dùng các đồ thị và các biểu đồ để thể hiện chúng.

(a) Biểu đồ hình cột (bar chart): là biểu đồ nhằm biểu diễn cho dữ liệu được phân nhóm (thường dùng cho dữ liệu định tính) như các tháng trong năm, các nhóm tuổi...Các nhóm được biểu diễn thường xuất hiện theo trục hoành, trục tung là chiều cao của các hình chữ nhật tỷ lệ với giá trị được biểu diễn. Mục tiêu của việc dùng biểu đồ hình cột là đưa ra so sánh giữa các nhóm.

(b) Biểu đồ hình quạt (pie chart): cũng được dùng để biểu diễn dữ liệu được phân nhóm, nhưng các nhóm được biểu diễn bằng các hình quạt trong hình tròn. Số lượng hoặc tỷ lệ của mỗi hạng mục (mỗi nhóm) tỷ lệ với diện tích hình quạt biểu diễn nó. Biểu đồ này thường dùng để phân tích hoặc so sánh ở mức độ tổng thể.

(c) Tổ chức đồ (histogram): thường được dùng để biểu thị tần số hay tần suất các giá trị trong mỗi khoảng giá trị.

1. Nếu độ rộng các khoảng bằng nhau, thì chiều cao của hình chữ nhật dựng trên mỗi khoảng chính là tần số hay tần suất tương ứng của khoảng.
2. Nếu độ rộng các khoảng không bằng nhau, chiều cao của hình chữ nhật dựng trên mỗi khoảng được tính toán sao cho diện tích mỗi hình chữ nhật tỷ lệ với tần số hoặc tần suất của khoảng đó.

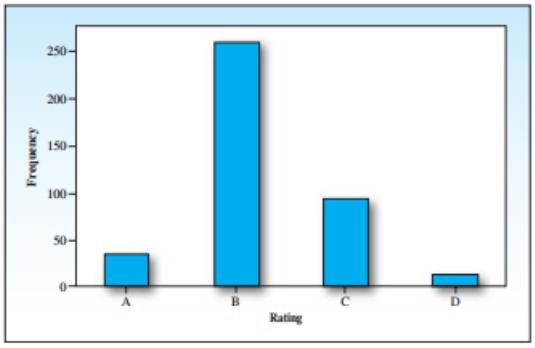
(d) Đa giác tần số, tần suất: dùng khi dữ liệu là liên tục và khoảng dữ liệu rất rộng. Tại mỗi giá trị của dữ liệu x_i và tần số n_i ta chấm một điểm có tọa độ (x_i, n_i). Nối các điểm này với nhau ta được đa giác tần số. Nếu muốn có đa giác tần suất ta thay n_i bằng $f_i = n_i/n$.

Ví dụ 4.7. Khảo sát 400 nhà quản lý giáo dục về đánh giá chất lượng giáo dục công ở Hoa Kỳ, ta nhận được bảng dữ liệu

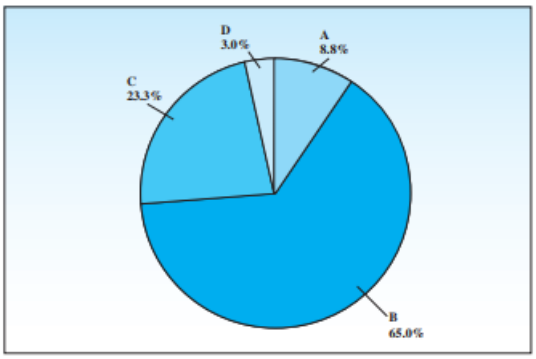
Xếp hạng	A	B	C	D
Tần số	35	260	93	12

- Tổng số nhà quản lý giáo dục được khảo sát $n = 400$.

- 35 người xếp hạng A chiếm 9%; 260 người xếp hạng B chiếm 65%; 93 người xếp hạng C chiếm 23%; 12 người xếp loại C chiếm 3%.
- Biểu đồ hình cột cho tập dữ liệu này biểu diễn ở Hình 4.1
- Biểu đồ hình quạt cho tập dữ liệu này biểu diễn ở Hình 4.2



Hình 4.1: Biểu đồ hình cột cho dữ liệu trong Ví dụ 4.7



Hình 4.2: Biểu đồ hình quạt cho dữ liệu trong Ví dụ 4.7

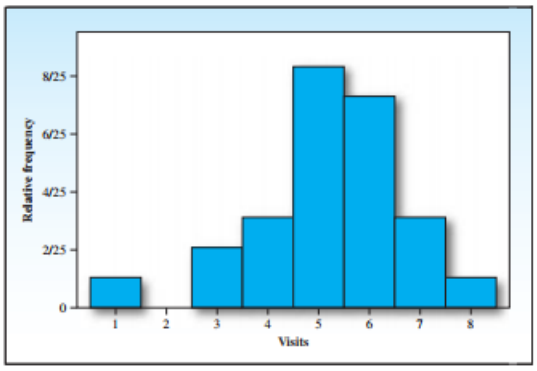
Ví dụ 4.8 (Ví dụ về tổ chức đồ)**.** 25 khách hàng của Starbucks được thăm dò ý kiến trong một cuộc khảo sát tiếp thị "Trong một tuần bạn đến Starbucks bao nhiêu lần?". Số liệu được cho trong bảng sau:

6 7 1 5 6 4 6 4 6 8 6 5
6 3 4 5 5 5 7 6 3 5 7 5 5

Biến được đo lường là "số lần đến Starbucks", một biến rời rạc chỉ nhận các giá trị nguyên. Trong trường hợp này, cách đơn giản nhất là chọn các lớp hoặc khoảng con dưới dạng giá trị nguyên trên phạm vi giá trị quan sát: 1, 2, 3, 4, 5, 6, 7 và 8. Bảng dưới đây cho thấy các lớp và tần số tương ứng của chúng cùng tần số.

Số lượt đến Starbucks	Tần số	Tần suất
1	1	0,04
2	0	0,00
3	2	0,08
4	3	0,12
5	8	0,32
6	7	0,28
7	3	0,12
8	1	0,04

Biểu đồ tần suất tương đối được thể hiện trong Hình 4.3.



Hình 4.3: Biểu đồ tổ chức đồ cho dữ liệu trong Ví dụ 4.8

4.1.4 Đại lượng thống kê và một số thống kê thông dụng

Để nghiên cứu mẫu ngẫu nhiên gốc X, nếu dừng lại ở mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ thì rõ ràng chưa giải quyết được vấn đề gì, bởi các biến ngẫu nhiên X_i có cùng quy luật phân phối xác suất với X mà ta chưa biết hoàn toàn. Vì vậy ta phải liên kết hay tổng hợp các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ lại sao cho biến ngẫu nhiên mới thu được có những tính chất mới, có thể đáp ứng được yêu cầu giải những bài toán khác nhau về biến ngẫu nhiên gốc X.

Đại lượng thống kê

Định nghĩa 4.2 (Thống kê)**.** Trong thống kê toán việc tổng hợp mẫu $W_X = (X_1, X_2, \ldots, X_n)$ được thực hiện dưới dạng hàm của các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$. Ký hiệu

$$\boxed{G = f(X_1, X_2, \ldots, X_n)} \tag{4.1}$$

ở đây f là một hàm nào đó và G được gọi là một thống kê.

Ví dụ 4.9. Cho một mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ kích thước n. Một ví dụ về thống kê dạng (4.1) là

$$G = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Nhận xét 4.1. **(a)** Thống kê G là một hàm của các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ nên cũng là một biến ngẫu nhiên. Do đó ta có thể tìm ra các "tính chất mới" thông qua việc khảo sát quy luật phân phối xác suất của G và các tham số $E(G)$, $V(G)\ldots$

(b) Nếu mẫu ngẫu nhiên có giá trị $W_x = (x_1, x_2, \ldots, x_2)$ (mẫu cụ thể), ta tính được giá trị cụ thể của G, ký hiệu là $g = f(x_1, x_2, \ldots, x_n)$ hay g_{qs} còn gọi là giá trị quan sát của thống kê G.

Sau đây ta xét một số thống kê thông dụng.

Một số thống kê thông dụng

(a) Trung bình mẫu ngẫu nhiên: Cho mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ kích thước n được xây dựng từ biến ngẫu nhiên gốc X. Trung bình của nó là một thống kê, ký hiệu là $\overline{X}$ và được định nghĩa bởi hàm sau đây:

$$\boxed{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i} \tag{4.2}$$

Do $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên nên $\overline{X}$ cùng là biến ngẫu nhiên. Nếu biến ngẫu nhiên gốc X có kỳ vọng $E(X) = \mu$, phương sai $V(X) = \sigma^2$ thì thống kê $\overline{X}$ có kỳ vọng $E(\overline{X}) = \mu$ và phương sai $V(\overline{X}) = \dfrac{\sigma^2}{n}$ nhỏ hơn phương sai của biến ngẫu nhiên gốc n lần, nghĩa là các giá trị có thể có của $\overline{X}$ ổn định quanh kỳ vọng μ hơn các giá trị có thể có của X. Điều này thể hiện "chất lượng mới" của thống kê $\overline{X}$ so với biến ngẫu nhiên gốc X.

(b) Phương sai mẫu ngẫu nhiên và phương sai hiệu chỉnh mẫu ngẫu nhiên: Cho mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ kích thước n được xây dựng từ biến ngẫu nhiên gốc X. Phương sai của nó là một thống kê, ký hiệu là $\hat{S}^2$ và được xác định bởi hàm

$$\boxed{\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4.3}$$

trong đó $\overline{X}$ là trung bình của mẫu ngẫu nhiên W_X.

Do $\hat{S}^2$ là biến ngẫu nhiên, nên có thể tính được $E(\hat{S}^2)$ bởi công thức

$$E(\hat{S}^2) = \frac{n-1}{n}\sigma^2, \tag{4.4}$$

trong đó $\sigma^2 = V(X)$.

Để kỳ vọng của phương sai mẫu ngẫu nhiên $\hat{S}^2$ trùng với phương sai của biến ngẫu nhiên gốc X ta cần một sự hiệu chỉnh. Từ (4.4) suy ra

$$E\left(\frac{n}{n-1}\hat{S}^2\right) = \sigma^2.$$

Đặt

$$\boxed{S^2 = \frac{n}{n-1}\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \tag{4.5}$$

và gọi S^2 là phương sai hiệu chỉnh của mẫu ngẫu nhiên (vì nó bằng phương sai mẫu ngẫu nhiên nhân thêm hệ số $\frac{n}{n-1}$). Đồng thời ta có $E(S^2) = \sigma^2$.

(c) Độ lệch tiêu chuẩn và độ lệch tiêu chuẩn hiệu chỉnh mẫu ngẫu nhiên: 1. Độ lệch tiêu chuẩn của mẫu ngẫu nhiên được ký hiệu và xác định bởi

$$\boxed{\hat{S} = \sqrt{\hat{S}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}} \tag{4.6}$$

2. Độ lệch tiêu chuẩn hiệu chỉnh của mẫu ngẫu nhiên được ký hiệu và xác định bởi

$$\boxed{S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}} \tag{4.7}$$

(d) Tần suất mẫu ngẫu nhiên: Trường hợp cần nghiên cứu một dấu hiệu định tính A nào đó mà mỗi cá thể của tổng thể có thể có hoặc không, giả sử p là tần suất có dấu hiệu A của tổng thể. Nếu cá thể có dấu hiệu A ta cho nhận giá trị 1, trường hợp ngược lại ta cho nhận giá trị 0. Lúc đó dấu hiệu nghiên cứu có thể xem là biến ngẫu nhiên X có phân phối Béc-nu-li tham số p có kỳ vọng $E(X) = p$ và phương sai $V(X) = p(1-p)$.

Lấy mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ trong đó X_1, X_2, $\ldots X_n$ là các biến ngẫu nhiên độc lập có cùng phân phối Béc-nu-li với tham số p. Tần số xuất hiện A trong mẫu là $m = \sum_{i=1}^{n} X_i$. Khi đó tần xuất mẫu là một thống kê ký hiệu và xác định bởi

$$\boxed{f = \frac{m}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}} \tag{4.8}$$

Như vậy tần suất mẫu là trung bình mẫu của biến ngẫu nhiên X có phân phối Béc-nu-li $\mathcal{B}(p)$ tham số p. Ngoài ra,

$$\boxed{E(f) = p, \quad V(f) = \frac{p(1-p)}{n}} \tag{4.9}$$

Dưới đây là một số quy luật phân phối xác suất của một số thống kê thông dụng.

Quy luật phân phối xác suất của một số thống kê thông dụng

Giả sử dấu hiệu nghiên cứu trong tổng thể có thể xem như một biến ngẫu nhiên X có phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$ với kỳ vọng $E(X) = \mu$ và phương sai $V(X) = \sigma^2$. Các tham số này có thể đã biết hoặc chưa biết. Từ tổng thể rút ra một mẫu ngẫu nhiên cỡ n: $W_X = (X_1, X_2, \ldots, X_n)$. Các biến ngẫu nhiên thành phần X_i, $i = 1, \ldots, n$, độc lập có cùng quy luật phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$ như X.

Chú ý rằng mọi tổ hợp tuyến tính của các biến ngẫu nhiên có phân phối chuẩn là biến ngẫu nhiên có phân phối chuẩn. Vì vậy ta có các kết quả sau.

(a) Quy luật phân phối xác suất của trung bình mẫu $\overline{X}$: Thống kê trung bình mẫu $\overline{X}$ có phân phối chuẩn và thống kê $U = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n}$ có phân phối chuẩn tắc

$$\boxed{\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)} \tag{4.10}$$

Phân vị mức α của U được ký hiệu là $u_{1-\alpha}$. Chẳng hạn phân vị mức $\alpha = 5\%$ của U là $u_{0,95} = 1,645$ vì $\Phi(1,645) = 0,95$; phân vị mức $\alpha = 2,5\%$ của U là $u_{0,975} = 1,96$ vì $\Phi(1,96) = 0,975$.

(b) Quy luật phân phối xác suất của phương sai hiệu chỉnh mẫu ngẫu nhiên: 1. Thống kê $S = \dfrac{n\hat{S}^2}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2}$ có phân phối khi bình phương với $n-1$ bậc tự do

$$\boxed{S = \frac{n\hat{S}^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2_{(n-1)}} \tag{4.11}$$

2. Nếu trong (4.11) ta thay $\overline{X}$ bằng μ thì thống kê

$$\boxed{S = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)} \tag{4.12}$$

Phân vị mức α của S ký hiệu là $\chi^2_{1-\alpha}(n)$.

3. Thống kê $T = \dfrac{\overline{X} - \mu}{S}\sqrt{n} = \dfrac{\overline{X} - \mu}{\hat{S}}\sqrt{n-1}$ có phân phối Student với $n-1$ bậc tự do:

$$\boxed{T = \frac{\overline{X} - \mu}{S}\sqrt{n} = \frac{\overline{X} - \mu}{\hat{S}}\sqrt{n-1} \sim t(n-1)} \tag{4.13}$$

Phân vị mức α của T ký hiệu là $t_{1-\alpha}(n)$.

Chú ý 4.1. Trong thực hành khi $n \geq 30$ ta có thể không cần đến giả thiết chuẩn của biến ngẫu nhiên gốc, thống kê $T = \dfrac{\overline{X} - \mu}{S}\sqrt{n}$ xấp xỉ phân phối chuẩn tắc $\mathcal{N}(0,1)$.

(c) Phân phối của thống kê tần suất mẫu: Khi n đủ lớn ($np \geq 5$ và $n(1-p) \geq 5$) thì thống kê $U = \dfrac{f-p}{\sqrt{p(1-p)}}\sqrt{n}$ có phân phối xấp xỉ phân phối chuẩn tắc

$$\boxed{U = \frac{f-p}{\sqrt{p(1-p)}}\sqrt{n} \sim \mathcal{N}(0,1)} \tag{4.14}$$

(d) Phân phối Fisher: Nếu hai mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_{n_1})$, $W_Y = (Y_1, Y_2, \ldots, Y_{n_2})$ kích thước n_1, n_2 được xây dựng từ biến ngẫu nhiên X và Y có phân phối chuẩn $\mathcal{N}(\mu_1; \sigma_1^2)$, $\mathcal{N}(\mu_2; \sigma_2^2)$ tương ứng thì người ta đã chứng minh được rằng:

1. Nếu $\sigma_1^2 = \sigma_2^2$ thì thống kê $F = \dfrac{S_1^2}{S_2^2}$ có phân phối Fisher với $(n_1 - 1, n_2 - 1)$ bậc tự do:

$$\boxed{F = \frac{S_1^2}{S_2^2} \sim \mathcal{F}(n_1 - 1; n_2 - 1)} \tag{4.15}$$

2. Nếu $\sigma_1^2 \neq \sigma_2^2$,

$$\boxed{F = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim \mathcal{F}(n_1 - 1; n_2 - 1)} \tag{4.16}$$

Phân vị mức α của F được ký hiệu là $\mathcal{F}_{1-\alpha}(n_1 - 1; n_2 - 1)$.

4.1.5 Cách tính giá trị cụ thể của một số thống kê thông dụng

Bên cạnh việc nghiên cứu quy luật phân phối xác suất của các thống kê, còn cần phải tính toán các giá trị của chúng. Giả sử mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ có một giá trị là $W_x = (x_1, x_2, \ldots, x_n)$. Mẫu cụ thể này có thể cho ở các dạng khác nhau.

(a) Mẫu cho dưới dạng liệt kê. (Tần số của các x_i bằng 1)

(a1) Trung bình mẫu:

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i} \tag{4.17}$$

(a2) Phương sai mẫu:

$$\boxed{\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \tag{4.18}$$

(a3) Phương sai hiệu chỉnh mẫu:

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}\hat{s}^2} \tag{4.19}$$

(a4) Các độ lệch chuẩn:

$$\boxed{\hat{s} = \sqrt{\hat{s}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}; \quad s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{4.20}$$

Để tính các công thức (4.17)–(4.20), ta lập bảng tính toán

x_i	x_i^2
x_1	x_1^2
x_2	x_2^2
$\dots$	$\dots$
x_n	x_n^2
$\sum_{i=1}^n x_i$	$\sum_{i=1}^n x_i^2$

Ta thấy trung bình mẫu $\overline{x}$ và phương sai mẫu $\hat{s}^2$ là hai giá trị cơ bản nhất đối với mẫu cụ thể này, còn các giá trị s^2, $\hat{s}$, s có thể tính trực tiếp từ $\hat{s}^2$; giá trị của nhiều thống kê khác cũng được tính trên cơ sở đã có $\overline{x}$ và $\hat{s}^2$. Do đó cần cải tiến các công thức tính $\overline{x}$ và $\hat{s}^2$ phù hợp với từng trường hợp số liệu.

(b) Mẫu cho ở dạng rút gọn. (Tần số của các x_i là $n_i > 1$, $\sum_{i=1}^k n_i = n$)

(b1) Trung bình mẫu:

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i} \tag{4.21}$$

(b2) Phương sai mẫu:

$$\boxed{\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{k} n_i x_i\right)^2} \tag{4.22}$$

(b3) Phương sai hiệu chỉnh mẫu:

$$\boxed{s^2 = \frac{n}{n-1}\hat{s}^2} \tag{4.23}$$

(b4) Các độ lệch chuẩn:

$$\boxed{\hat{s} = \sqrt{\hat{s}^2}; \quad s = \sqrt{s^2}} \tag{4.24}$$

Để tính các công thức (4.21)–(4.24), ta lập bảng tính toán

x_i	n_i	$n_i x_i$	$n_i x_i^2$
x_1	n_1	$n_1 x_1$	$n_1 x_1^2$
x_2	n_2	$n_2 x_2$	$n_2 x_2^2$
...	...	...	...
x_k	n_k	$n_k x_k$	$n_k x_k^2$
	$\sum_{i=1}^k n_i = n$	$\sum_{i=1}^k n_i x_i$	$\sum_{i=1}^k n_i x_i^2$

(c) Phương pháp đổi biến. (Trong trường hợp độ dài các khoảng bằng nhau)

(c1) Trung bình mẫu:

$$\boxed{\overline{x} = x_0 + h\overline{u} = x_0 + \frac{h}{n}\sum_{i=1}^{k} n_i u_i} \tag{4.25}$$

(c2) Phương sai mẫu:

$$\boxed{\hat{s}^2 = h^2\left[\frac{1}{n}\sum_{i=1}^{k} n_i u_i^2 - \left(\frac{1}{n}\sum_{i=1}^{k} n_i u_i\right)^2\right] = h^2 \hat{s}_u^2} \tag{4.26}$$

trong đó

x_i là điểm giữa của khoảng thứ i, $i = 1, 2, \ldots, k$;
$u_i = \dfrac{x_i - x_0}{h}$, h là độ dài các khoảng;
$x_0 = x_i$ ứng với n_i lớn nhất.

Để tính các công thức (4.25)–(4.26), ta lập bảng tính toán

x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$
x_1	n_1	u_1	$n_1 u_1$	$n_1 u_1^2$
x_2	n_2	u_2	$n_2 u_2$	$n_2 u_2^2$
...	...	...	...	...
x_k	n_k	u_k	$n_k u_k$	$n_k u_k^2$
	$\sum_{i=1}^k n_i = n$		$\sum_{i=1}^k n_i u_i$	$\sum_{i=1}^k n_i u_i^2$

Tính tham số đặc trưng mẫu trên máy tính CASIO FX570VN PLUS

Bước 1 Chuyển đổi máy tính về chương trình thống kê **MODE** → **3** → **AC**

Bước 2 Bật chức năng cột tần số/tần suất **SHIFT** → **MODE** → **Mũi tên đi xuống** → **4(STAT)** → **1(ON)**

Bước 3 Bật chế độ màn hình để nhập dữ liệu, Nhập số liệu **SHIFT** → **1** → **1(TYPE)** → **1(1-VAR)**

Chú ý nhập xong số liệu thì bấm **AC** để thoát.

Bước 4 Xem kết quả:

- Trung bình mẫu ($\overline{x}$): **SHIFT** $\rightarrow$ **1** $\rightarrow$ **4(VAR)** $\rightarrow$ **2**
- Độ lệch tiêu chuẩn mẫu hiệu chỉnh (s): **SHIFT** $\rightarrow$ **1** $\rightarrow$ **4** $\rightarrow$ **4**

Ví dụ 4.10. Ở một địa điểm thu mua vải, kiểm tra một số vải thấy kết quả sau

Số khuyết tật ở mỗi đơn vị	0	1	2	3	4	5	6
Số đơn vị kiểm tra (10m)	8	20	12	40	30	25	15

Hãy tính kỳ vọng mẫu và độ lệch chuẩn hiệu chỉnh mẫu của mẫu trên.

Lời giải Ví dụ 4.10
Cách 1: Gọi X là số khuyết tật ở mỗi đơn vị. Lập bảng tính toán

x_i	n_i	$n_i x_i$	$n_i x_i^2$
0	8	0	0
1	20	20	20
2	12	24	48
3	40	120	360
4	30	120	480
5	25	125	625
6	15	90	540
$\sum$	$n = 150$	$\sum_i n_i x_i = 499$	$\sum_i n_i x_i^2 = 2073$

Suy ra $\overline{x} = \dfrac{499}{150} = 3,3267$; $\overline{x^2} = \dfrac{2073}{150} = 13,82$; $\hat{s}^2 = \overline{x^2} - (\overline{x})^2 = 13,82 - (3,3267)^2 = 2,7531$; $s^2 = \dfrac{150}{149} \times 2,7531 = 2,7715$; $s = \sqrt{2,7715} = 1,6648$.

Cách 2: Sử dụng máy tính CASIO FX570VN PLUS tính được $\overline{x} = 3,3267$; $s = 1,6648$.

4.2 Ước lượng điểm

Như đã biết, các tham số của dấu hiệu nghiên cứu $\mathcal{X}$ như trung bình, phương sai, cơ cấu của tổng thể theo dấu hiệu $\mathcal{X}$ được sử dụng rất rộng rãi trong phân tích kinh tế - xã hội và nhiều lĩnh vực khác. Song các tham số này thông thường lại chưa biết. Vì vậy đặt ra vấn đề ước lượng chúng nhờ phương pháp mẫu.

Sau khi đã mô hình hóa dấu hiệu $\mathcal{X}$ và cơ cấu tổng thể bằng biến ngẫu nhiên X và quy luật phân phối xác suất của nó, ta có thể phát biểu vấn đề thực tế nêu trên dưới dạng toán học như sau: "Cho biến ngẫu nhiên X, có thể đã biết hoặc chưa biết quy luật phân phối xác suất dạng tổng quát, nhưng chưa biết tham số θ của nó. Hãy ước lượng θ bằng phương pháp mẫu". Đây là một trong những bài toán cơ bản của thống kê toán học.

Dưới đây ta sẽ nghiên cứu các phương pháp tìm ra một số hay một khoảng số để ước lượng θ. Các phương pháp này xuất phát từ cơ sở hợp lý nào đó để tìm ước lượng của θ, chứ không phải là sự chứng minh chặt chẽ.

Phương pháp ước lượng điểm chủ trương dùng giá trị quan sát của một thống kê để ước lượng một tham số (véc tơ tham số) nào đó theo các tiêu chuẩn: vững, không chệch, hiệu quả.

4.2.1 Phương pháp hàm ước lượng

Mô tả phương pháp

Giả sử cần ước lượng tham số θ của biến ngẫu nhiên X. Thông thường X là một biến ngẫu nhiên mà ta muốn biết phân phối xác suất của X. Trong xác suất, biết phân phối của X nghĩa là ta đã có thông tin "đầy đủ" về nó, nói khác đi ta có thể tính được xác suất để biến ngẫu nhiên nhận giá trị trong một miền bất kỳ. Tuy nhiên trên thực tế phân phối xác suất của X thường rất khó nắm bắt. Chính bì vậy ta mong muốn biết được những thông tin chính về X như giá trị trung bình, độ lệch chuẩn, trung vị, mốt, mômen...của X.

Từ X ta lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ kích thước n. Chọn lập thống kê $G = f(X_1, X_2, \ldots, X_n)$. Một trong những cách chọn dạng hàm f là tương ứng thống kê đặc trưng của mẫu ngẫu nhiên với tham số cần ước lượng của biến ngẫu nhiên. Phương pháp này gọi là phương pháp mô-men (moment estimation). Chẳng hạn $G = f(X_1, X_2, \ldots, X_n) = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ nếu là ước lượng cho kỳ vọng $E(X)$, còn $G = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ nếu ước lượng phương sai...

Tiến hành lập mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$. Tính giá trị cụ thể của G ứng với mẫu này, tức là $g = f(x_1, x_2, \ldots, x_n)$. Đây là ước lượng điểm của θ.

Thống kê $G = f(X_1, X_2, \ldots, X_n)$, viết gọn là G, là hàm ước lượng của θ.

Chú ý rằng θ là một số chưa biết, còn $G = f(X_1, X_2, \ldots, X_n)$ là một biến ngẫu nhiên. Như vậy, ở đây ta đã lấy một biến ngẫu nhiên để xấp xỉ cho một số. Câu hỏi đặt ra là

1. Ước lượng đưa ra có "tốt" không?
2. "Ước lượng tốt" được hiểu theo nghĩa nào?

Dưới đây là một số tiêu chuẩn cho ước lượng điểm.

Một số tiêu chuẩn lựa chọn hàm ước lượng

Cùng một mẫu ngẫu nhiên có thể xây dựng nhiều thống kê G khác nhau để ước lượng cho tham số θ. Vì vậy ta cần lựa chọn thống kê tốt nhất để ước lượng cho tham số θ dựa vào các tiêu chuẩn sau.

(a) Ước lượng không chệch (unbiased estimator): Ước lượng G của θ được gọi là ước lượng không chệch của θ nếu

$$\boxed{E(G) = \theta \quad \text{hay} \quad E(G-\theta) = 0} \tag{4.27}$$

nghĩa là về trung bình "độ chệch" bằng 0. Tính chất này có nghĩa là ước lượng G không có sai số hệ thống mà chỉ có sai số ngẫu nhiên.

Điều kiện (4.27) của ước lượng không chệch có nghĩa là trung bình các giá trị của G bằng θ. Tuy nhiên, không có nghĩa là mọi giá trị của G đều trùng khít với θ mà từng giá trị của G có thể sai lệch rất lớn so với θ. Vì vậy ta tìm ước lượng không chệch sao cho độ sai lệch trung bình là bé nhất.

(b) Ước lượng vững (consistent estimator): Ước lượng G của θ được gọi là ước lượng vững nếu với mọi $\varepsilon > 0$ ta có

$$\lim_{n\to+\infty} P\{|G-\theta| < \varepsilon\} = 1$$

hay

$$\lim_{n\to+\infty} P\{\theta-\varepsilon < G < \theta+\varepsilon\} = 1.$$

Tính chất này đảm bảo cho ước lượng G gần θ tùy ý với xác suất cao khi cỡ mẫu đủ lớn.

(c) Ước lượng hiệu quả (efficient estimator): Đại lượng $G-\theta$ được gọi là sai số (nó là một biến ngẫu nhiên) còn giá trị trung bình của sai số $E(G-\theta) = E(G)-\theta$ được gọi là độ chệch. Ta mong muốn tìm được ước lượng sao cho sai số bình phương trung bình

$$MSE(G) = E(G-\theta)^2$$

nhỏ nhất có thể. Tuy nhiên trong nhiều trường hợp tìm ước lượng thỏa mãn điều kiện này khá khó. Lưu ý rằng

$$E(G-\theta)^2 = [E(G)-\theta]^2 + V(G) = (\text{Độ chệch})^2 + V(G).$$

Vì vậy ta thường chỉ tìm ước lượng tốt nhất trong các ước lượng không chệch tức là ước lượng không chệch có phương sai $V(G)$ nhỏ nhất trong các ước lượng không chệch. Ước lượng đó gọi là ước lượng hiệu quả.

Khi ta không biết ước lượng hiệu quả có tồn tại hay không thì để so sánh các ước lượng không chệch ta sẽ so sánh độ lệch tiêu chuẩn hay phương sai của chúng. Ước lượng không chệch có độ lệch tiêu chuẩn hay phương sai nhỏ hơn sẽ "tốt hơn". Độ lệch tiêu chuẩn của ước lượng điểm G, ký hiệu là σ_G được gọi là sai số tiêu chuẩn (standard error). Ước lượng điểm của sai số tiêu chuẩn được ký hiệu là $\hat{\sigma}_G$. Tổng quát hơn để so sánh hai ước lượng điểm G_1 và G_2 cho tham số θ bất kỳ, ta so sánh sai số bình phương trung bình,

ước lượng nào có sai số bình phương trung bình bé hơn là ước lượng tốt hơn. Tức là nếu độ hiệu quả tương đối

$$\frac{MSE(G_1)}{MSE(G_2)}$$

nhỏ hơn 1 thì ta kết luận rằng ước lượng G_1 hiệu quả hơn ước lượng G_2.

Để xét xem ước lượng không chệch G có phải là ước lượng hiệu quả của θ hay không ta cần phải tìm một cận dưới của phương sai của các ước lượng không chệch và so sánh phương sai của G với cận dưới này. Điều này được giải quyết bằng bất đẳng thức Cramer–Rao phát biểu như sau.

Định lý 4.1. Cho mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ được lấy từ tổng thể có dấu hiệu nghiên cứu được mô hình hóa bởi biến ngẫu nhiên X mà hàm mật độ xác suất $f(X, \theta)$ hay hàm phân phối xác suất $F(X, \theta)$ thỏa mãn một số điều kiện nhất định (thường được thỏa mãn trong thực tế) và $\hat{\theta}$ là ước lượng không chệch bất kỳ của θ thì

$$\boxed{V(G) \geq \frac{1}{nE\left(\frac{\partial(\ln f(X,\theta))}{\partial\theta}\right)^2}} \tag{4.28}$$

4.2.2 Ước lượng điểm cho một số tham số thông dụng

(a) Ước lượng điểm cho kỳ vọng hay giá trị trung bình: Giả sử X là biến ngẫu nhiên với kỳ vọng $E(X) = \mu$ chưa biết, μ được xem là trung bình của tổng thể. Từ X ta lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ cỡ n. Chọn

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

làm ước lượng điểm cho kỳ vọng $E(X) = \mu$. Ước lượng điểm $\overline{X}$ thỏa mãn cả ba tính chất tốt đã nêu ở trên: không chệch, vững và hiệu quả.

Khi ta có một mẫu cụ thể $W_X = (x_1, x_2, \ldots, x_n)$ thì

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

là một ước lượng điểm của μ.

(b) Ước lượng điểm cho phương sai: Giả sử X là biến ngẫu nhiên với phương sai $V(X) = \sigma^2$ chưa biết, σ^2 được xem là phương sai của tổng thể. Nếu ta có một mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ cỡ n của X thì xuất phát từ công thức tính phương sai, đại lượng

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

được xem xét để dùng làm ước lượng cho σ^2. Tuy nhiên không khó để chỉ ra rằng

$$E(\hat{S}^2) = \frac{n-1}{n}\sigma^2$$

nghĩa là $\hat{S}^2$ là một ước lượng chệch của σ^2. Để thu được ước lượng không chệch cho σ^2 ta "hiệu chỉnh" đại lượng này một chút bằng cách đặt

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{n}{n-1}\hat{S}^2.$$

Đại lượng S^2 (chính là phương sai hiệu chỉnh mẫu ngẫu nhiên) là ước lượng không chệch cho σ^2. Người ta đã chứng minh được rằng cả $\hat{S}^2$ và S^2 đều là ước lượng vững cho σ^2. Như vậy ước lượng tốt cho σ^2 là S^2.

Khi có mẫu cụ thể $W_X = (x_1, x_2, \ldots, x_n)$ ta tính được các giá trị cụ thể của $\hat{S}^2$ và S^2, ký hiệu là $\hat{s}^2$ và s^2:

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2; \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{n}{n-1}\hat{s}^2.$$

đây là các ước lượng điểm của σ^2.

Ví dụ 4.11. Chỉ số IQ của 10 sinh viên được cho như sau:

$$87, \quad 81, \quad 88, \quad 85, \quad 100, \quad 90, \quad 114, \quad 93, \quad 86, \quad 98$$

1. Ước lượng điểm cho chỉ số IQ trung bình là trung bình mẫu $\overline{x} = 92,2$.
2. Ước lượng điểm cho độ lệch tiêu chuẩn σ của chỉ số IQ là độ lệch tiêu chuẩn hiệu chỉnh mẫu $s = 9,64$.
3. Sai số tiêu chuẩn của trung bình mẫu là $\sigma_{\overline{x}} = \sigma/\sqrt{n}$, trong đó $n = 10$ là cỡ mẫu. Ước lượng điểm cho sai số tiêu chuẩn của trung bình mẫu là

$$\hat{\sigma}_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{9,64}{\sqrt{10}} = 3,05.$$

Nếu ta dùng $\hat{\mu} = \dfrac{x_1 + x_2}{2}$ như một ước lượng điểm cho chỉ số IQ trung bình thì sai số tiêu chuẩn là $\sigma_{\hat{\mu}} = \sigma/\sqrt{2}$ và ước lượng điểm cho sai số tiêu chuẩn này là

$$\hat{\sigma}_{\overline{\mu}} = \frac{s}{\sqrt{2}} = \frac{9,64}{\sqrt{2}} = 6,81.$$

Rõ ràng ước lượng $\hat{\mu}$ không hiệu quả bằng ước lượng $\overline{x}$.

(c) Ước lượng điểm cho tỷ lệ: Cho p là một tỷ lệ hay xác suất của một sự kiện A trong tổng thể chưa biết. Ta thực hiện n quan sát độc lập và gọi m là số lần xuất hiện A. Khi đó tần suất mẫu

$$f = \frac{m}{n}$$

là ước lượng điểm cho p. Người ta chứng minh được rằng ước lượng này có cả ba tính chất tốt đã nêu ở trên: không chệch, vững và hiệu quả.

Ví dụ 4.12. Trong đợt vận động bầu cử tổng thống người ta phỏng vấn ngẫu nhiên 1600 cử tri thì được biết 960 người sẽ bỏ phiếu cho ứng cử viên A. Hãy chỉ ra ước lượng điểm cho tỷ lệ phiếu thực mà ứng cử viên A sẽ thu được.

Lời giải Ví dụ 4.12 Ước lượng điểm cần tìm là $f = \dfrac{960}{1600} = 0,6 = 60\%$.

4.2.3 Phương pháp ước lượng hợp lý cực đại (maximum-likelihood estimation)

Giả sử ta đã biết quy luật phân phối xác suất dạng tổng quát của biến ngẫu nhiên X, chẳng hạn hàm mật độ xác suất $f_X(x,\theta)$ (có thể hiểu $f_X(x,\theta)$ là công thức xác suất nếu X rời rạc). Cần ước lượng tham số θ của X ta lập mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$.

Hàm của đối số θ

$$L\Big(x_1, x_2, \ldots, x_n, \theta\Big) = f_X(x_1,\theta) f_X(x_2,\theta) \ldots f_X(x_n,\theta) \tag{4.29}$$

và gọi là hàm số hợp lý của tham số θ. Giá trị của hàm hợp lý chính là xác suất hay mật độ xác suất tại điểm $W_x = (x_1, x_2, \ldots, x_n)$. Giá trị $g = g(x_1, x_2, \ldots, x_n)$ được gọi là ước lượng hợp lý cực đại của θ nếu ứng với giá trị này của θ hàm hợp lý đạt cực đại.

Do hàm L và hàm $\ln L$ đạt cực đại tại cùng một giá trị của θ nên ta có thể tìm giá trị của θ để $\ln L$ đạt cực đại theo các bước sau.

1. Tìm đạo hàm bậc nhất của hàm $\ln L$ theo θ.
2. Lập phương trình $\dfrac{d \ln L}{d\theta} = 0$. Phương trình này gọi là phương trình hợp lý. Giả sử nó có nghiệm $\theta = g = g(x_1, x_2, \ldots, x_n)$.
3. Tìm đạo hàm bậc hai $\dfrac{d^2 \ln L}{d\theta^2}$. Nếu tại điểm $\theta = g$ đạo hàm bậc hai âm thì tại điểm này hàm $\ln L$ đạt cực đại. Do đó $g = g(x_1, x_2, \ldots, x_n)$ là ước lượng điểm hợp lý tối đa cần tìm.

Ví dụ 4.13. Bằng phương pháp hợp lý cực đại ta tìm được ước lượng của tham số p trong quy luật phân phối nhị thức $\mathcal{B}(n;p)$ là $\frac{\overline{x}}{n}$ và ước lượng của tham số λ trong quy luật phân phối Poisson là $\frac{1}{\overline{x}}$. (Phần chứng minh xem như bài tập).

Ví dụ 4.14. Tìm ước lượng hợp lý cực đại của các tham số μ và σ của biến ngẫu nhiên tuân theo quy luật phân phối chuẩn $\mathcal{N}(\mu;\sigma^2)$.

Lời giải Ví dụ 4.14 Dễ thấy hàm hợp lý có dạng

$$L = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Lập hệ phương trình hợp lý

$$\begin{cases} \dfrac{d\ln L}{d\mu} = 0, \\ \dfrac{d\ln L}{d\sigma} = 0. \end{cases}$$

Giải hệ này ta nhận được $\mu = \overline{x}$ và $\sigma = s$.

Chú ý 4.2. Đối số của hàm hợp lý là θ chứ không phải $x_1, x_2, \ldots, x_n$. Do vậy nếu thay giá trị mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$ bằng mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ thì kết quả và chứng minh trên vẫn đúng. Do đó ta thu được kết quả tổng quát hơn. Chẳng hạn, nếu X tuân theo quy luật phân phối chuẩn thì ước lượng hợp lý của μ là $\overline{X}$ và ước lượng hợp lý của σ là S.

Ngoài ra còn các phương pháp Bayes, phương pháp minimax, phương pháp bootstrap ...Các phương pháp này tìm ước lượng điểm cũng như các tiêu chuẩn để kiểm tra các tính chất tốt của một ước lượng điểm nằm ngoài phạm vi của bài học.

BÀI 12 (2 tiết)

4.3 Ước lượng khoảng

Phương pháp ước lượng điểm nói trên có nhược điểm là khi kích thước mẫu bé thì ước lượng điểm có thể sai lệch rất nhiều so với giá trị của tham số cần ước lượng. Mặt khác phương pháp trên cũng không thể đánh giá được khả năng mắc sai lầm khi ước lượng là bao nhiêu. Do đó khi kích thước mẫu bé người ta thường dùng phương pháp ước lượng khoảng tin cậy cho trường hợp một tham số.

Khái niệm ước lượng khoảng

Giả sử chưa biết đặc trưng θ nào đó của biến ngẫu nhiên X. Ước lượng khoảng của θ là chỉ ra một khoảng số (g_1, g_2) nào đó chứa θ, tức là có thể ước lượng $g_1 < \theta < g_2$.

Phương pháp khoảng ước lượng tin cậy

Để ước lượng tham số θ của biến ngẫu nhiên X, từ biến ngẫu nhiên này ta lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ cỡ n. Chọn thống kê $G(X, \theta)$ sao cho mặc dù chưa biết giá trị của θ, quy luật phân phối xác suất của G vẫn hoàn toàn xác định. Do đó, với xác suất α khá bé ta tìm được $P(G_1 < \theta < G_2) = 1 - \alpha$. Vì α khá bé, nên $\gamma = 1 - \alpha$ khá lớn (thông thường yêu cầu $1 - \alpha = \gamma \geq 0,95$ để có thể áp dụng nguyên lý xác suất lớn cho sự kiện $(G_1 < \theta < G_2)$). Khi đó, sự kiện $(G_1 < \theta < G_2)$ hầu như chắc chắn xảy ra trong một phép thử. Thực hiện một phép thử đối với mẫu ngẫu nhiên W_X ta thu được mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$, từ đó tính được các giá trị của G_1, G_2, ký hiệu là g_1, g_2. Như vậy có thể kết luận: với độ tin cậy $1 - \alpha = \gamma$ tham số θ nằm trong khoảng (g_1, g_2).

(a) (G_1, G_2) được gọi là khoảng tin cậy của θ với độ tin cậy $\gamma = 1 - \alpha$.

(b) $1 - \alpha = \gamma$ được gọi là độ tin cậy của ước lượng.

(c) $I = G_2 - G_1$ được gọi là độ dài khoảng tin cậy.

Các cận G_1 và G_2 phụ thuộc vào mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ nên chúng là các biến ngẫu nhiên. Ta thường mong muốn tìm được các khoảng tin cậy có hai tính chất sau:

1. Có độ tin cậy cao.
2. Độ dài khoảng tin cậy nhỏ, tức là hiệu $G_2 - G_1$ bé.

Đôi khi ta chỉ quan tâm đến giá trị lớn nhất hay giá trị nhỏ nhất của tham số. Ví dụ ta xét hai kết luận sau:

1. Với xác suất 95% thì chiều cao trung bình của người Việt Nam nhỏ hơn 168cm.

2. Với xác suất 95%, chiều cao trung bình của sinh viên Đại học Bách khoa Hà Nội lớn hơn 165cm.

Ở kết luận thứ nhất, với độ tin cậy 95% thì chiều cao trung bình của người Việt Nam nằm trong khoảng $(0; 168)$ và ta chỉ quan tâm đến giá trị lớn nhất. Ở kết luận thứ hai, với độ tin cậy 95% thì chiều cao trung bình của sinh viên Đại học Bách khoa Hà Nội thuộc khoảng $(165; +\infty)$ và ta chỉ quan tâm đến giá trị nhỏ nhất. Các khoảng tin cậy dạng này được gọi là khoảng tin cậy một phía. Khoảng tin cậy mà chỉ quan tâm đến cận trên được gọi là khoảng tin cậy lớn nhất hay khoảng tin cậy trái còn khoảng tin cậy mà chỉ quan tâm tới cận dưới được gọi là khoảng tin cậy nhỏ nhất hay khoảng tin cậy phải. Khoảng tin cậy mà ta quan tâm tới cả hai cận (trên và dưới) được gọi là khoảng tin cậy hai phía (đôi khi gọi là khoảng tin cậy đối xứng vì các cận trên, cận dưới thường đối xứng qua ước lượng điểm của tham số).

Dưới đây là các khoảng tin cậy cho kỳ vọng hay giá trị trung bình, phương sai và tỷ lệ hay xác suất.

4.3.1 Khoảng tin cậy cho kỳ vọng

Bài toán 4.1. Giả sử biến ngẫu nhiên X có kỳ vọng $E(X) = \mu$ và phương sai $V(X) = \sigma^2$, trong đó $E(X) = \mu$ chưa biết. Bài toán đặt ra là tìm ước lượng khoảng cho μ dựa trên các quan sát $W_x = (x_1, x_2, \ldots, x_n)$.

Từ tổng thể, ta lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ và xét các trường hợp sau.

Trường hợp phân phối chuẩn, phương sai đã biết

Giả sử $X \sim \mathcal{N}(\mu; \sigma^2)$ với σ^2 đã biết. Khi đó $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ là biến ngẫu nhiên có phân phối chuẩn với kỳ vọng $E(\overline{X}) = \mu$ và phương sai $V(\overline{X}) = \dfrac{\sigma^2}{n}$.Do đó

$$\boxed{U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}} \tag{4.30}$$

có phân phối chuẩn tắc $\mathcal{N}(0; 1)$.

Chọn cặp số không âm α_1, α_2 thỏa mãn $\alpha_1 + \alpha_2 = \alpha$, tìm các phân vị chuẩn tắc $u_{\alpha_1}, u_{1-\alpha_2}$ sao cho $P(U < u_{\alpha_1}) = \alpha_1$; $P(U < u_{1-\alpha_2}) = 1 - \alpha_2$. Do tính chất của phân phối chuẩn tắc $u_{\alpha_1} = -u_{1-\alpha_1}$, suy ra

$$\begin{aligned} P(-u_{1-\alpha_1} < U < u_{1-\alpha_2}) &= P(u_{\alpha_1} < U < u_{1-\alpha_2}) \\ &= P(U < u_{1-\alpha_2}) - P(U < u_{\alpha_1}) = 1 - \alpha_2 - \alpha_1 = 1 - \alpha. \end{aligned}$$

Như vậy,

$$1-\alpha = P(-u_{1-\alpha_1} < U < u_{1-\alpha_2}) = P\bigg(-u_{1-\alpha_1} < \frac{\overline{X}-\mu}{\sigma}\sqrt{n} < u_{1-\alpha_2}\bigg).$$

Hay

$$1-\alpha = P\bigg(\overline{X} - u_{1-\alpha_2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\alpha_1}\frac{\sigma}{\sqrt{n}}\bigg).$$

Khi có mẫu cụ thể $W_X = (x_1, x_2, \ldots, x_n)$, tính được giá trị cụ thể $\overline{x}$ của $\overline{X}$, khi đó khoảng tin cậy cho μ với độ tin cậy $\gamma = 1-\alpha$ là:

$$\boxed{\bigg(\overline{x} - u_{1-\alpha_2}\frac{\sigma}{\sqrt{n}} \quad ; \quad \overline{x} + u_{1-\alpha_1}\frac{\sigma}{\sqrt{n}}\bigg)} \tag{4.31}$$

Như vậy, với độ tin cậy $\gamma = 1-\alpha$ cho trước, có vô số khoảng tin cậy cho μ vì có vô số cặp α_1, α_2 thỏa mãn $\alpha_1 + \alpha_2 = \alpha$. Ở đây ta chỉ xét một số trường hợp đặc biệt.

(a) Khoảng tin cậy hai phía (đối xứng) $(\alpha_1 = \alpha_2 = \alpha/2)$

$$\boxed{\bigg(\overline{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad ; \quad \overline{x} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\bigg)} \tag{4.32}$$

trong đó $u_{1-\frac{\alpha}{2}}$ được tra từ bảng giá trị hàm phân phối chuẩn tắc (Phụ lục 3) từ hệ thức

$$\boxed{\Phi(u_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}} \tag{4.33}$$

hoặc tra từ bảng giá trị hàm Laplace (Phụ lục 2) từ hệ thức

$$\boxed{\phi(u_{1-\frac{\alpha}{2}}) = \frac{1-\alpha}{2}} \tag{4.34}$$

(b) Khoảng tin cậy trái $(\alpha_1 = \alpha, \alpha_2 = 0)$:

$$\boxed{\bigg(-\infty \quad ; \quad \overline{x} + u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\bigg)} \tag{4.35}$$

trong đó $u_{1-\alpha}$ được tra từ bảng giá trị hàm phân phối chuẩn tắc (Phụ lục 3) từ hệ thức

$$\boxed{\Phi(u_{1-\alpha}) = 1-\alpha}. \tag{4.36}$$

hoặc tra từ bảng giá trị hàm Laplace (Phụ lục 2) từ hệ thức

$$\boxed{\phi(u_{1-\alpha}) = \frac{1-2\alpha}{2}}. \tag{4.37}$$

(c) Khoảng tin cậy phải $(\alpha_1 = 0, \alpha_2 = \alpha)$:

$$\boxed{\bigg(\overline{x} - u_{1-\alpha}\frac{\sigma}{\sqrt{n}} \quad ; \quad +\infty\bigg)} \tag{4.38}$$

Ví dụ 4.15. Trọng lượng của một loại sản phẩm là biến ngẫu nhiên tuân theo luật phân phối chuẩn với độ lệch tiêu chuẩn là 6 gam. Cân thử 36 sản phẩm loại này ta thu được kết quả sau:

177	165	152	174	159	166	160	152	162	175	158	169
166	162	181	168	170	150	173	164	167	177	167	175
165	182	166	158	166	170	168	165	160	160	169	166

1. Với độ tin cậy $1-\alpha = 95\%$, hãy tìm khoảng tin cậy đối xứng của trọng lượng trung bình của loại sản phẩm nói trên.

2. Có thể khẳng định trọng lượng trung bình của sản phẩm ít nhất là bao nhiêu với độ tin cậy 99%.

3. Với độ tin cậy 99% hãy tìm khoảng tin cậy cho trọng lượng trung bình của sản phẩm.

Lời giải Ví dụ 4.15 Gọi X là trọng lượng sản phẩm, $X \sim \mathcal{N}(\mu, \sigma^2)$ với $\sigma = 6$. Trọng lượng trung bình của sản phẩm là $E(X) = \mu$ chưa biết cần ước lượng.

1. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng cho kỳ vọng $E(X) = \mu$ của tổng thể có phân phối chuẩn khi đã biết phương sai với $\sigma = 6$.

   - Chọn thống kê $U = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n}$. Thống kê $U \sim \mathcal{N}(0;1)$.
   - Với $\alpha = 0,05$, $\Phi(u_{1-\frac{\alpha}{2}}) = \Phi(u_{0,975}) = 0,975$, tra bảng giá trị hàm phân phối chuẩn tắc nhận được $u_{0,975} = 1,96$.
   - Từ số liệu đã cho ta có cỡ mẫu $n = 36$, trung bình mẫu $\overline{x} = 166,22$ (gam), suy ra khoảng tin cậy 95% cho trọng lượng trung bình là

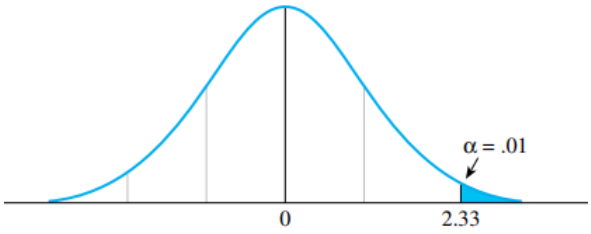
     $$\left(\overline{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\ ;\ \overline{x} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = \left(166,22 - 1,96 \times \frac{6}{\sqrt{36}}\ \ 166,22 + 1,96 \times \frac{6}{\sqrt{36}}\right)$$

     hay $(164,26 < \mu < 168,18)$.
   - Vậy với độ tin cậy 95%, trọng lượng trung bình của loại sản phẩm nói trên từ 164,26 gam đến 168,18 gam.

2. Với độ tin cậy $\gamma = 99\%$ suy ra $\alpha = 0,01$, $\Phi(u_{1-\alpha}) = \Phi(u_{0,99}) = 0,99$, tra bảng giá trị hàm phân phối chuẩn tắc nhận được $u_{0,99} = 2,33$ (xem Hình 4.4). Giá trị bé nhất cho trọng lượng trung bình là

   $$\overline{x} - u_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 166,22 - 2,33\frac{6}{\sqrt{36}} = 163,89.$$

   Tức là với xác suất 99% ta có thể khẳng định trọng lượng trung bình của sản phẩm lớn hơn 163,89 gam.

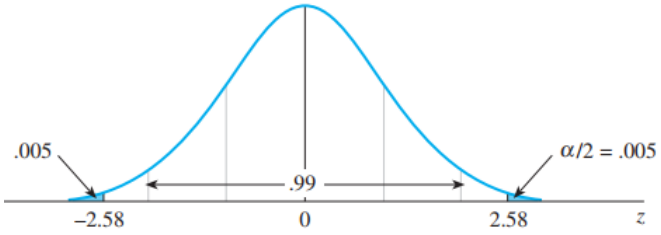


Hình 4.4: Giá trị của $u_{1-\alpha}$ với độ tin cậy 99%

3. Ta có $u_{1-\frac{\alpha}{2}} = 2,58$ (xem Hình 4.5) nên khoảng tin cậy cho trọng lượng trung bình của sản phẩm với độ tin cậy 99% là

$$= \left(166,22 - 2,58 \times \frac{6}{\sqrt{36}}\ \ 166,22 + 2,58 \times \frac{6}{\sqrt{36}}\right) = (163,64\ ;\ 168,80).$$

Như vậy, trên cùng một mẫu nếu độ tin cậy càng lớn thì độ dài của khoảng tin cậy càng lớn.



Hình 4.5: Giá trị của $u_{1-\frac{\alpha}{2}}$ với độ tin cậy 99%

Chú ý 4.3. 1. Chú ý rằng không thể viết $P(164,26 < \mu < 168,18) = 0,95$ vì độ tin cậy gắn với khoảng tin cậy ngẫu nhiên chứ không gắn với mẫu cụ thể. Hơn nữa vì μ là một hằng số nên nó chỉ có thể thuộc hoặc không thuộc khoảng (164,26; 168,18) nên $(164,26 < \mu < 168,18)$ không phải là sự kiện ngẫu nhiên.

2. Ta có thể xác định $u_{1-\frac{\alpha}{2}} = 1,96$ ở ý Ví dụ 4.15(1) từ bảng giá trị hàm Laplace (Phụ lục 2) từ hệ thức $\phi(u_{1-\alpha}) = \dfrac{1-\alpha}{2}$.

Trong phần tiếp theo ta chỉ đưa ra công thức áp dụng.

Trường hợp phân phối chuẩn, phương sai chưa biết

Trong nhiều bài toán thực tế ta không biết phương sai của tổng thể. Với giả thiết dữ liệu tuân theo phân phối chuẩn thì người ta đã chứng minh được rằng thống kê

$$\boxed{T = \frac{\overline{X} - \mu}{S}\sqrt{n}} \tag{4.39}$$

có phân phối Student với $n-1$ bậc tự do, $T \sim t(n-1)$. Bằng phương pháp tương tự như đã trình bày ở trên, ta có thể tìm được các khoảng tin cậy cho kỳ vọng $E(X) = \mu$ với độ tin cậy $\gamma = 1 - \alpha$.

(a) Khoảng tin cậy hai phía (đối xứng)

$$\boxed{\left(\overline{x} - t^{(n-1)}_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \quad ; \quad \overline{x} + t^{(n-1)}_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)} \tag{4.40}$$

(b) Khoảng tin cậy trái

$$\boxed{\left(-\infty \quad ; \quad \overline{x} + t^{(n-1)}_{1-\alpha} \frac{s}{\sqrt{n}}\right)} \tag{4.41}$$

(c) Khoảng tin cậy phải

$$\boxed{\left(\overline{x} - t^{(n-1)}_{1-\alpha} \frac{s}{\sqrt{n}} \quad ; \quad +\infty\right)} \tag{4.42}$$

trong đó $t^{(n-1)}_{1-\frac{\alpha}{2}}$, $t^{(n-1)}_{1-\alpha}$ được tra từ bảng phân phối Student với $n-1$ bậc tự do (Phụ lục 4).

Ví dụ 4.16. Để kiểm tra sự chính xác của hệ thống đóng gói tự động các bao gạo khi xuất khẩu tại một nhà máy, người ta đã chọn ngẫu nhiên 16 bao và tính được trọng lượng trung bình là 49,75 kg và độ lệch tiêu chuẩn hiệu chỉnh mẫu là 0,5 kg. Giả thiết rằng trọng lượng của những bao gạo có phân phối chuẩn.

1. Tìm khoảng tin cậy cho trọng lượng trung bình của bao gạo với độ tin cậy 95%.
2. Với độ tin cậy 95% có thể khẳng định trọng lượng trung bình của bao gạo cao nhất là bao nhiêu? Dựa trên kết quả thu được có thể khẳng định về trung bình các bao gạo đã bị đóng thiếu hay không nếu biết rằng trọng lượng chuẩn mỗi bao là 50 kg.

Lời giải Ví dụ 4.16 Gọi X là trọng lượng các bao gạo được đóng gói tự động, $X \sim \mathcal{N}(\mu, \sigma^2)$. Trọng lượng trung bình là $E(X) = \mu$ chưa biết, cần ước lượng.

1. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng cho kỳ vọng $E(X) = \mu$ của tổng thể có phân phối chuẩn khi chưa biết phương sai.
   - Chọn thống kê $T = \dfrac{\overline{X} - \mu}{S}\sqrt{n}$. Thống kê T có phân phối Student, $T \sim t^{(n-1)}$.
   - Với độ tin cậy 99% suy ra $\alpha = 0,05$, $t^{(n-1)}_{1-\frac{\alpha}{2}} = t^{(15)}_{0,975} = 2,13$, tra bảng phân phối Student (Phụ lục 4).

- Ta có $n = 16, \overline{x} = 49,75, s = 0,5$, suy ra khoảng tin cậy 95% cho trọng lượng trung bình là

$$\left(\overline{x} - t_{1-\alpha}^{(n-1)} \frac{s}{\sqrt{n}} \quad ; \quad +\infty\right) = \left(49,75 - 2,13 \times \frac{0,5}{\sqrt{16}} \;\; 49,75 + 2,13 \times \frac{0,5}{\sqrt{16}}\right)$$

  hay $(49,48 < \mu < 50,02)$.

- Vậy với độ tin cậy 95%, ta có thể khẳng định trọng lượng trung bình của các bao gạo nói trên từ 49,48 kg đến 50,02 kg.

2. Đây là bài toán ước lượng bằng khoảng tin cậy một phía cho kỳ vọng $E(X) = \mu$ của tổng thể có phân phối chuẩn khi chưa biết phương sai.

   - Với độ tin cậy 95% suy ra $t_{1-\alpha}^{(n-1)} = t_{0,95}^{(15)} = 1,75$, tra bảng phân phối Student (Phụ lục 4).

   - Khoảng tin cậy lớn nhất cho lượng trung bình là

$$\left(0; \overline{x} + t_{1-\alpha}^{(n-1)} \frac{s}{\sqrt{n}}\right) = (0; 49,75 + 1,75 \frac{0,5}{\sqrt{16}}) = (0; 49,97).$$

   - Với độ tin cậy 95% giá trị lớn nhất cho trọng lượng trung bình của các bao gạo là 49,97 kg.

     Ta có thể khẳng định về trung bình các bao gạo đã bị đóng thiếu.

Trường hợp mẫu cỡ lớn ($n \geq 30$)

Khi mẫu cỡ n lớn, trung bình mẫu ngẫu nhiên $\overline{X}$ có phân phối xấp xỉ phân phối chuẩn với trung bình μ và phương sai σ^2/n do đó thống kê

$$\boxed{U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)} \tag{4.43}$$

Cũng do n lớn nên ta có thể xấp xỉ σ (chưa biết) bởi S do đó thống kê

$$\boxed{U = \frac{\overline{X} - \mu}{S}\sqrt{n} \sim \mathcal{N}(0,1)} \tag{4.44}$$

Trong thực hành cho phép vận dụng với $n \geq 30$.

Các lập luận như đã trình bày ở trên, các khoảng tin cậy cho μ là:

(a) Khoảng tin cậy hai phía (đối xứng)

$$\boxed{\left(\overline{x} - u_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \quad ; \quad \overline{x} + u_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)} \tag{4.45}$$

(b) Khoảng tin cậy trái

$$\boxed{\left(-\infty \quad ; \quad \overline{x} + u_{1-\alpha}\frac{s}{\sqrt{n}}\right)} \tag{4.46}$$

(c) Khoảng tin cậy phải

$$\boxed{\left(\overline{x} - u_{1-\alpha}\frac{s}{\sqrt{n}} \quad ; \quad +\infty\right)} \tag{4.47}$$

Ví dụ 4.17. Để ước lượng trọng lượng trung bình của loại trái cây A tại một vùng, người ta thu hoạch ngẫu nhiên 100 trái cây A của vùng đó và thu được kết quả sau

Trọng lượng (gam)	40-42	42-44	44-46	46-48	48-50	50-52
Số trái	7	13	25	35	15	5

Hãy ước lượng trọng lượng trung bình của loại trái cây A trong vùng bằng khoảng tin cậy đối xứng với độ tin cậy 95%. Cho biết trọng lượng loại trái cây A là biến ngẫu nhiên tuân theo luật phân phối chuẩn.

Lời giải Ví dụ 4.17 Gọi X là trọng lượng loại trái cây A, $X \sim \mathcal{N}(\mu, \sigma^2)$. Trọng lượng trung bình của loại trái cây A là $E(X) = \mu$ chưa biết, cần ước lượng. Đây là bài toán ước lượng khoảng của kỳ vọng của biến ngẫu nhiên phân phối chuẩn trường hợp chưa biết phương sai, mẫu cỡ $n = 100 > 30$.

- Chọn thống kê $U = \dfrac{\overline{X} - \mu}{S}\sqrt{n}$. Vì $n = 100 > 30$ nên thống kê $U \sim \mathcal{N}(0, 1)$.
- Khoảng tin cậy đối xứng cho $E(X) = \mu$ là $\left(\overline{x} - u_{1-\frac{\alpha}{2}}\dfrac{s}{\sqrt{n}} \; ; \; \overline{x} + u_{1-\frac{\alpha}{2}}\dfrac{s}{\sqrt{n}}\right)$ trong đó, với $\alpha = 0,05$, $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$ được tra từ bảng giá trị hàm phân phối chuẩn tắc (Phụ lục 3).
- Từ số liệu đã cho tính được $n = 100$, $\overline{x} = 46,06$, $s = 2,48$. Suy ra khoảng tin cậy đối xứng của μ là $\left(46,06 - 1,96 \times \dfrac{2,48}{\sqrt{100}} \; ; \; 46,06 + 1,96 \times \dfrac{2,48}{\sqrt{100}}\right)$ hay $(45,573 \; ; \; 46,546)$.
- Vậy với độ tin cậy 95%, trọng lượng trung bình của loại trái cây A ở vùng trên từ 45,573 gam đến 46,546 gam.

4.3.2 Khoảng tin cậy cho phương sai (đọc thêm)

Bài toán 4.2. Giả sử X là biến ngẫu nhiên có phân phối chuẩn, $X \sim \mathcal{N}(\mu; \sigma^2)$, với phương sai $V(X) = \sigma^2$ chưa biết. Hãy ước lượng khoảng cho phương sai σ^2 dựa trên số liệu quan sát $W_x = (x_1, x_2, \ldots, x_n)$.

Lập một mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ kích thước n và xét các trường hợp sau.

Trường hợp kỳ vọng chưa biết

Người ta đã chứng minh được rằng thống kê

$$\boxed{\chi^2 = \frac{(n-1)S^2}{\sigma^2}} \tag{4.48}$$

là biến ngẫu nhiên có phân phối khi bình phương với $(n-1)$ bậc tự do, $\chi^2(n-1)$.

Chọn cặp số không âm α_1, α_2 thỏa mãn $\alpha_1 + \alpha_2 = \alpha$, tìm các phân vị $P(\chi^2 < \chi^2_{(n-1,\alpha_1)}) = \alpha_1$; $P(\chi^2 < \chi^2_{(n-1,1-\alpha_2)}) = 1-\alpha_2$. Từ đó suy ra

$$P\left(\chi^2_{(n-1,\alpha_1)} < \chi^2 < \chi^2_{(n-1,1-\alpha_2)}\right) = P\left(\chi^2_{(n-1,\alpha_1)} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{(n-1,1-\alpha_2)}\right) = 1-\alpha,$$

hay tương đương với

$$P\left(\frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha_2)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1,\alpha_1)}}\right) = 1-\alpha.$$

Khi có mẫu cụ thể $W_X = (x_1, x_2, \ldots, x_n)$, tính được giá trị cụ thể s^2 của S^2, khi đó khoảng tin cậy cho σ^2 với độ tin cậy $\gamma = 1-\alpha$ là:

$$\boxed{\left(\frac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha_2)}} \quad ; \quad \frac{(n-1)s^2}{\chi^2_{(n-1,\alpha_1)}}\right)} \tag{4.49}$$

Ta xét một số trường hợp cụ thể của (4.49).

(a) Khoảng tin cậy hai phía $(\alpha_1 = \alpha_2 = \alpha/2)$

$$\boxed{\left(\frac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha/2)}} \quad ; \quad \frac{(n-1)s^2}{\chi^2_{(n-1,\alpha/2)}}\right)} \tag{4.50}$$

(b) Khoảng tin cậy trái $(\alpha_1 = \alpha, \alpha_2 = 0)$:

$$\boxed{\left(0 \quad ; \quad \frac{(n-1)s^2}{\chi^2_{(n-1,\alpha)}}\right)} \tag{4.51}$$

(c) Khoảng tin cậy phải $(\alpha_1 = 0, \alpha_2 = \alpha)$:

$$\boxed{\left(\frac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha)}} \quad ; \quad +\infty\right)} \tag{4.52}$$

Chú ý 4.4.

1. Giá trị $\chi^2_{(n-1,1-\alpha/2)}$, $\chi^2_{(n-1,\alpha/2)}$, $\chi^2_{(n-1,\alpha)}$ và $\chi^2_{(n-1,1-\alpha)}$ được tra từ bảng phân phối khi bình phương với $n-1$ bậc tự do (Phụ lục 6).

2. Lấy căn bậc hai các cận trong các khoảng tin cậy cho phương sai ta sẽ thu được các khoảng tin cậy cho các độ lệch chuẩn.

Trường hợp kỳ vọng đã biết

Nếu kỳ vọng $E(X) = \mu_0$ đã biết, người ta đã chứng minh được rằng thống kê

$$\boxed{\chi^2 = \frac{\sum_{i=1}^{n}\left(X_i - \mu_0\right)^2}{\sigma^2}} \tag{4.53}$$

là biến ngẫu nhiên có phân phối khi bình phương với n bậc tự do, $\chi^2(n)$. Làm giống như trường hợp trên ta nhận được:

(a) Khoảng tin cậy hai phía

$$\boxed{\left(\frac{\sum_{i=1}^{n}\left(x_i - \mu_0\right)^2}{\chi^2_{(n,1-\alpha/2)}} \quad ; \quad \frac{\sum_{i=1}^{n}\left(x_i - \mu_0\right)^2}{\chi^2_{(n,\alpha/2)}}\right)} \tag{4.54}$$

(b) Khoảng tin cậy trái

$$\boxed{\left(0 \quad ; \quad \frac{\sum_{i=1}^{n}\left(x_i - \mu_0\right)^2}{\chi^2_{(n,\alpha)}}\right)} \tag{4.55}$$

(c) Khoảng tin cậy phải

$$\boxed{\left(\frac{\sum_{i=1}^{n}\left(x_i - \mu_0\right)^2}{\chi^2_{(n,1-\alpha)}} \quad ; \quad +\infty\right)} \tag{4.56}$$

Ví dụ 4.18. Trọng lượng của một loại sản phẩm tuân theo luật phân phối chuẩn. Cân thử từng sản phẩm của một mẫu ngẫu nhiên gồm 25 đơn vị, ta nhận được kết quả sau:

Trọng lượng sản phẩm (gam)	29,3	29,7	30	30,5	30,75
Số sản phẩm	4	5	8	5	3

Với độ tin cậy 95% hãy tìm khoảng tin cậy cho phương sai của trọng lượng sản phẩm trong hai trường hợp

1. Đã biết kỳ vọng $E(X) = 30$.
2. Không biết kỳ vọng.

Lời giải Ví dụ 4.18 Gọi X là trọng lượng sản phẩm, $X \sim \mathcal{N}(\mu, \sigma^2)$. Phương sai của trọng lượng sản phẩm là $V(X) = \sigma^2$ chưa biết, cần ước lượng. Đây là bài toán ước lượng khoảng của phương sai của biến ngẫu nhiên phân phối chuẩn.

1. Trường hợp $E(X) = 30$ đã biết.

- Chọn thống kê $\chi^2 = \dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}$. Thống kê χ^2 có phân phối khi bình phương với n bậc tự do.
- Với độ tin cậy 95% suy ra $\alpha = 0,05$, tra bảng phân phối khi bình phương (Phụ lục 5), $\chi^2_{(n,1-\alpha/2)} = \chi^2_{(25;0,975)} = 40,65$, $\chi^2_{(n,\alpha/2)} = \chi^2_{(25;0,025)} = 13,12$.
- Từ số liệu đã cho tính được $n = 25$, $\sum_{i=1}^{5} n_i(x_i-30)^2 = 5,13$. Suy ra khoảng tin cậy cần tìm là

$$\left(\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\chi^2_{(n,1-\alpha/2)}}\,;\,\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\chi^2_{(n,\alpha/2)}}\right) = \left(\frac{5,13}{40,65}\,;\,\frac{5,13}{13,12}\right)$$

  hay $(0,1262 < \sigma^2 < 0,3910)$.
- Vậy với độ tin cậy 95%, phương sai của trọng lượng sản phẩm từ 0,1262 đến 0,3910.

2. Trường hợp kỳ vọng chưa biết ta sử dụng thống kê

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

Tra bảng phân phối khi bình phương với độ tin cậy 95% và bậc tư do $n-1 = 24$ ta được $\chi^2_{(n-1,1-\alpha/2)} = \chi^2_{(24;0,975)} = 39,36$, $\chi^2_{(n-1,\alpha/2)} = \chi^2_{(24;0,025)} = 12,40$.

Tính $\overline{x} = \dfrac{1}{25}\sum_{i=1}^{5} n_i x_i = 30,012$, $(n-1)s^2 = \sum_{i=1}^{5} n_i(x_i-30,012)^2 = 5,1264$.

Vậy khoảng tin cậy cần tìm là

$$\left(\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\chi^2_{(n-1,1-\alpha/2)}}\,;\,\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\chi^2_{(n-1,\alpha/2)}}\right) = \left(\frac{5,1264}{39,36}\,;\,\frac{5,1264}{12,40}\right)$$

hay $(0,1302 < \sigma^2 < 0,4134)$.

Vậy với độ tin cậy 95% phương sai σ^2 lớn hơn 0,1302 và nhỏ hơn 0,4134 hay độ lệch chuẩn σ nằm từ 0,3608 gam đến 0,6430 gam.

4.3.3 Khoảng tin cậy cho xác suất

Bài toán 4.3. Ta xét một tổng thể mà mỗi cá thể hoặc có tính chất A hoặc không có tính chất A nào đó. Gọi p là tỷ lệ cá thể có tính chất A trong tổng thể. Thông thường p chưa biết. Dựa trên một mẫu được chọn ngẫu nhiên, hãy tìm khoảng tin cậy cho p với độ tin cậy $\gamma = 1-\alpha$ cho trước.

Ta thực hiện n phép thử độc lập, cùng điều kiện. Gọi m là số cá thể có tính chất A trong n cá thể được chọn. Khi đó m là biến ngẫu nhiên có phân phối nhị thức $\mathcal{B}(n;p)$. Khi n lớn thì m có phân phối xấp xỉ phân phối chuẩn và do đó tần suất $f = \dfrac{m}{n}$ cũng có phân phối xấp xỉ phân phối chuẩn với trung bình $E(f) = p$ và phương sai $V(f) = \dfrac{p(1-p)}{n}$. Xấp xỉ này "tốt"

nếu $np > 5$ và $n(1-p) > 5$. Tuy nhiên vì p chưa biết nên phương sai $V(f)$ chưa biết. Do f là ước lượng điểm tốt cho p nên phương sai $V(f)$ có thể được xấp xỉ bằng $\frac{f(1-f)}{n}$. Khi đó thống kê

$$\boxed{U = \frac{f-p}{\sqrt{f(1-f)}}\sqrt{n}} \tag{4.57}$$

có phân phối xấp xỉ phân phối chuẩn tắc $\mathcal{N}(0;1)$. Từ đó ta tìm được các khoảng tin cậy cho p khi có mẫu cụ thể là:

(a) Khoảng tin cậy hai phía (đối xứng)

$$\boxed{\left(f - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} \quad ; \quad f + u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right)} \tag{4.58}$$

(b) Khoảng tin cậy trái

$$\boxed{\left(0 \quad ; \quad f + u_{1-\alpha}\sqrt{\frac{f(1-f)}{n}}\right)} \tag{4.59}$$

(c) Khoảng tin cậy phải

$$\boxed{\left(f - u_{1-\alpha}\sqrt{\frac{f(1-f)}{n}} \quad ; \quad 1\right)} \tag{4.60}$$

Chú ý 4.5. Vì p chưa biết nên ta không kiểm tra được điều kiện $np > 5$ và $n(1-p) > 5$. Trong thực hành ta dùng các điều kiện $nf > 5$ và $n(1-f) > 5$.

Ví dụ 4.19. Điều tra nhu cầu tiêu dùng loại hàng A trong 100 hộ gia đình ở khu dân cư B thấy 60 hộ gia đình có nhu cầu loại hàng trên. Với độ tin cậy $1-\alpha = 95\%$ hãy tìm khoảng tin cậy đối xứng của tỷ lệ hộ gia đình có nhu cầu loại hàng đó. Giả thích kết quả thu được.

Lời giải Ví dụ 4.19 Gọi p là tỷ lệ hộ gia đình ở khu dân cư B có nhu cầu mặt hàng A. Kiểm tra điều kiện $nf = 100 \times 0,6 = 60 > 5$ và $n(1-f) = 100 \times 0,4 = 40 > 5$. Đây là bài toán ước lượng khoảng của tỷ lệ trường hợp mẫu cỡ n đủ lớn.

- Thống kê $U = \frac{f-p}{\sqrt{f(1-f)}}\sqrt{n}$ có phân phối xấp xỉ phân phối chuẩn tắc $\mathcal{N}(0;1)$.
- Khoảng tin cậy đối xứng của xác suất p là

$$\left(f - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\,;\, f + u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right)$$

trong đó $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$ được tra từ bảng giá trị hàm phân phối chuẩn tắc.

- Với $n = 100$, $m = 60$, $f = \dfrac{m}{n} = 0,6$, suy ra khoảng tin cậy cần tìm là

$$\left(0,6 - 1,96\sqrt{\frac{0,6 \times 0,4}{100}}\ ;\ 0,6 - 1,96\sqrt{\frac{0,6 \times 0,4}{100}}\right) = (0,504\ ;\ 0,696).$$

- Vậy tỷ lệ hộ gia đình ở khu dân cư B có nhu cầu loại hàng A là từ 50,4% đến 69,6% với độ tin cậy 95%.

Vì mọi giá trị trong khoảng tin cậy này đều nhỏ hơn 70% nên ta cũng có thể khẳng định rằng có ít hơn 70% hộ gia đình có nhu cầu mặt hàng A.

4.3.4 Xác định kích thước mẫu

Bài toán 4.4. Giả sử θ là một tham số của tổng thể và $G = G_n(X_1, X_2, \ldots, X_n)$ là một ước lượng cho θ dựa trên mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ có kích thước n. Cho trước số $\varepsilon > 0$ và $\gamma \in (0;1)$. Ta nói rằng G_n có độ chính xác (hay sai số) ε với độ tin cậy γ nếu

$$\boxed{P\Big(|\theta - G_n| \le \varepsilon\Big) \ge \gamma} \tag{4.61}$$

Nếu kích thước mẫu n càng lớn thì độ chính xác của ước lượng càng cao, sai số càng nhỏ. Tuy nhiên kích thước mẫu càng lớn thì nhà nghiên cứu càng tốn nhiều thời gian, tiền bạc và công sức cho việc thu thập dữ liệu. Bài toán đặt ra là cần chọn kích thước mẫu tối thiểu là bao nhiêu để đủ đạt được độ chính xác mong muốn.

Ta xét các trường hợp sau đây.

(a) Trường hợp ước lượng cho giá trị trung bình

Giả sử với độ tin cậy γ cho trước, ta muốn có ước lượng cho giá trị trung bình μ với sai số không quá ε. Trong trường hợp X có phân phối chuẩn và phương sai σ^2 đã biết thì

$$\boxed{P\left(|\overline{x} - \mu| \le u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \gamma} \tag{4.62}$$

ở đây $\alpha = 1 - \gamma$. Do đó nếu n thỏa mãn

$$u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \varepsilon$$

hay tương đương với

$$\boxed{n \ge \frac{\sigma^2 (u_{1-\alpha/2})^2}{\varepsilon^2}} \tag{4.63}$$

thì

$$P\Big(|\overline{x} - \mu| \le \varepsilon\Big) \ge P\left(|\overline{x} - \mu| \le u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \gamma.$$

Ta sẽ chọn kích thước mẫu n là số nguyên dương bé nhất thỏa mãn điều kiện (4.63) thì sẽ đạt được độ chính xác (hay sai số) với độ tin cậy γ mong muốn.

Tuy nhiên công thức trên chỉ áp dụng được khi σ đã biết. Trong trường hợp X có phân phối chuẩn và phương sai chưa biết thì

$$P\left(|\overline{x} - \mu| \leq t^{(n-1)}_{1-\alpha/2}\frac{s}{\sqrt{n}}\right) = \gamma.$$

Do đó nếu n thỏa mãn

$$t^{(n-1)}_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \varepsilon$$

hay tương đương với

$$\boxed{n \geq \frac{s^2(t^{(n-1)}_{1-\alpha/2})^2}{\varepsilon^2}} \tag{4.64}$$

thì ta sẽ có

$$P\left(|\overline{x} - \mu| \leq \varepsilon\right) \geq \gamma.$$

Ta không thể tìm n thỏa mãn (4.64) vì cả s và $t^{(n-1)}_{1-\alpha/2}$ đều phụ thuộc vào n. Tuy nhiên, nhìn vào bảng phân phối Student ta thấy khi số bậc tự do lớn thì các phân vị của phân phối Student $t^{(n-1)}_{1-\alpha/2}$ và các phân vị của phân phối chuẩn tắc $u_{1-\alpha/2}$ gần như nhau, vì vậy ta có thể chọn n thỏa mãn

$$\boxed{n \geq \frac{s^2(u_{1-\alpha/2})^2}{\varepsilon^2}} \tag{4.65}$$

Như vậy, nếu tìm được ước lượng cho s thì ta sẽ tìm được n từ bất đẳng thức này. Người ta thường lấy một mẫu sơ bộ cỡ mẫu n đủ lớn ($n \geq 30$) để tính s rồi sau đó tìm n từ bất đẳng thức (4.65).

Ví dụ 4.20. Ta muốn xây dựng một ước lượng với độ tin cậy 95% và độ chính xác 2 dặm cho vận tốc trung bình của ô tô trên đường cao tốc. Một mẫu điều tra sơ bộ với cỡ mẫu 50 cho ta $s = 9$ dặm. Hỏi cần lấy mẫu cỡ tối thiểu là bao nhiêu để đạt được độ chính xác và độ tin cậy đã đặt ra.

Lời giải Ví dụ 4.20 Ta có $\varepsilon = 2, s = 9, \gamma = 95\%$ nên $u_{1-\alpha/2} = 1,96$. Từ bất đẳng thức (4.65) ta có

$$n \geq \frac{s^2(u_{1-\alpha/2})^2}{\varepsilon^2} = \frac{9^2(1,96)^2}{2^2} = 77,79.$$

Nghĩa là ta phải lấy cỡ mẫu ít nhất là 78. Tuy nhiên vì ta đã có mẫu sơ bộ với cỡ 50, nên thực ra ta chỉ cần bổ sung thêm 28 quan sát nữa.

Chú ý 4.6. 1. Việc lấy mẫu sơ bộ để tìm s là hợp lý vì ta biết rằng s là ước lượng vững cho σ, nghĩa là s hội tụ đến σ khi $n \to \infty$, do đó khi n lớn thì các giá trị của s khá "ổn định" và "gần" σ.

2. Trong ví dụ trên, phương sai của tổng thể dữ liệu được ước lượng thông qua độ phân tán của mẫu sơ bộ. Ta cũng có thể ước lượng phương sai của tổng thể dữ liệu theo những cách khác.

Ví dụ 4.21. Ta muốn ước lượng số năm đi học của những người trưởng thành trong một vùng với độ tin cậy 95% và độ chính xác không quá 1 năm thì cần lấy mẫu tối thiểu là bao nhiêu?

Lời giải Ví dụ 4.21 Ta có $\varepsilon = 1$, $\gamma = 95\%$ nên $u_{1-\alpha/2} = 1,96$. Để ước lượng phương sai của tổng thể dữ liệu, ta nhận thấy rằng số năm học dao động từ) đến 18. Nếu phân phối của số năm đi học là phân phối chuẩn thì hầu hết các giá trị quan sát sẽ thuộc khoảng $(\mu - 3\sigma; \mu + 3\sigma)$. Khoảng này có độ dài 6σ do đó ta dùng ước lượng $6\sigma = 18$ hay $\sigma = 3$. Từ đó ta tìm được

$$n \geq \frac{3^2(1,96)^2}{1^2} = 35.$$

Ta cũng có thể đoán rằng số năm học không thể quá 24 và do đó độ lệch tiêu chuẩn σ không thể quá 4 để có ước lượng thô về σ từ đó tìm n.

(b) Trường hợp ước lượng cho tỷ lệ

Ta xét một tổng thể mà mỗi cá thể hoặc có tính chất A hoặc không có tính chất A nào đó. Gọi p là tỷ lệ cá thể có tính chất A trong tổng thể. Thông thường p chưa biết. Giả sử trên một mẫu ngẫu nhiên cỡ n có m cá thể có tính chất A. Khi đó tần suất $f_n = \dfrac{m}{n}$ có phân phối xấp xỉ phân phối chuẩn với trung bình là p và phương sai $\dfrac{p(1-p)}{n}$. Do đó

$$\boxed{P\left(|f_n - p| \leq u_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = \gamma} \tag{4.66}$$

ở đây $\alpha = 1 - \gamma$. Do đó nếu n thỏa mãn

$$u_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \varepsilon$$

hay tương đương với

$$\boxed{n \geq \frac{(u_{1-\alpha/2})^2 p(1-p)}{\varepsilon^2}} \tag{4.67}$$

thì

$$P\left(|f_n - p| \leq \varepsilon\right) \geq \gamma.$$

Như vậy ta cần lấy n là số nguyên dương nhỏ nhất thỏa mãn (4.67).

Tuy nhiên, vì giá trị của p chưa biết, nên vế phải của (4.67) chưa xác định. Có hai cách để khắc phục tình trạng này.

1. Cách thứ nhất là lấy một mẫu sơ bộ kích thước k để thu được tần suất f_k và lấy f_k làm ước lượng ban đầu cho p. Khi đó bất đẳng thức (4.67) trở thành

$$\boxed{n \geq \frac{(u_{1-\alpha/2})^2 f_k(1-f_k)}{\varepsilon^2}} \tag{4.68}$$

với điều kiện

$$\boxed{kf_k > 5 \quad \text{và} \quad k(1-f_k) > 5} \tag{4.69}$$

Ta sẽ lấy n là số nguyên dương nhỏ nhất thỏa mãn (4.68) và (4.69).

2. Cách thứ hai, sử dụng bất đẳng thức Cauchy $p(1-p) \leq \frac{1}{4}$ ta nhận được

$$u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq \frac{u_{1-\alpha/2}}{2\sqrt{n}}.$$

Nếu ta chọn n thỏa mãn điều kiện

$$\boxed{\frac{u_{1-\alpha/2}}{2\sqrt{n}} \quad \text{hay} \quad n \geq \frac{(u_{1-\alpha/2})^2}{4\varepsilon^2}} \tag{4.70}$$

thì n thỏa mãn điều kiện (4.67). Vậy ta sẽ lấy n là số nguyên dương nhỏ nhất thỏa mãn (4.70).

Ví dụ 4.22. Một kỹ sư muốn ước lượng tỷ lệ sản phẩm loại A của một nhà máy với độ tin cậy 90% và sai số không quá 0,02. Hỏi kỹ sư đó phải lấy một mẫu cỡ n bằng bao nhiêu?

Lời giải Ví dụ 4.22

1. Nếu làm theo cách thứ nhất, trước hết người kỹ sư lấy một mẫu cỡ $n = 1000$ sản phẩm kiểm tra thấy có 640 sản phẩm loại A. Khoảng tin cậy của tỷ lệ sản phẩm loại A của nhà máy dựa trên mẫu điều tra này là

$$\begin{aligned} & f - u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \; ; \; f + u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \\ &= (0,64 - 1,645\sqrt{\frac{0,64(1-0,64)}{1000}} \;\; 0,64 + 1,645\sqrt{\frac{0,64(1-0,64)}{1000}}) \\ &= (0,64-,25 \; ; \; 0,64+0,25). \end{aligned}$$

Sai số của ước lượng là 0,25 lớn hơn 0,02. Vậy cần lấy mẫu lớn hơn nữa. Cỡ mẫu n phải thỏa mãn (4.68), tức là

$$n \geq \frac{(1,645)^2(0,64)(0,36)}{(0,02)^2} = 1558,67.$$

Vậy $n = 1559$.

2. Nếu sử dụng cách thứ hai thì phải chọn n thỏa mãn (4.70), tức là

$$n \geq \frac{(1,645)^2}{(4)(0,02)^2} = 1691,266.$$

Do đó $n = 1692$.

Ví dụ 4.23. Một cuộc thăm dò dư luận tại thành phố A được tiến hành để hỏi xem có nên thực hiện một dự án quan trọng với kinh phí lớn trong thành phố hay không. Tỷ lệ người ủng hộ việc thực hiện dự án trên mẫu thăm dò cho ta ước lượng tỷ lệ người ủng hộ việc thực hiện dự án trên toàn bộ số dân của thành phố. Nếu ước lượng tỷ lệ này với độ tin cậy 95% và sai số không quá 0,03 thì phải lấy cỡ mẫu bao nhiêu?

Lời giải Ví dụ 4.23 Nếu sử dụng cách thứ hai, ta phải chọn n thỏa mãn (4.70) hay

$$n \geq \frac{(1,96)^2}{(4)(0,03)^2} = 1067,111.$$

Do đó $n = 1068$. Điều thú vị nhất trong tính toán trên là cỡ mẫu tối thiểu n chỉ phụ thuộc vào độ chính xác và độ tin cậy mà ta mong muốn, chứ không phụ thuộc vào tổng số dân trên thành phố. Nếu hỏi ý kiến 1068 người, thì ta đạt được độ tin cậy là 95% và độ chính xác 0,03 mà không phụ thuộc vào tổng số dân trên thành phố là 5 triệu, là 50 triệu hay 500 triệu!

Chú ý 4.7.

1. Cái khó khăn ở đây là không phải thu thập ý kiến của càng nhiều người càng tốt. Vấn đề lớn là phải đảm bảo đây là mẫu ngẫu nhiên. Chẳng hạn nếu thực hiện ý kiến qua mạng:

   - Gửi email đến n địa chỉ ngẫu nhiên. Giả sử trường hợp tốt nhất, cả n người đều trả lời. Vấn đề là không phải ai cũng dùng email.
   - Ngay cả trường hợp tất cả mọi người đều dùng email, thì không phải ai cũng trả lời. Quyết định trả lời và câu trả lời cũng liên quan đến nhau. Nếu xe hơi của bạn chạy tốt, ít khi bạn trả lời những câu hỏi liên quan đến chất lượng của hãng; nhưng nếu có trục trặc thì khả năng này rất cao. Nếu ta thấy 30% khách trên mạng than thở về chất lượng của xe cũng không nói lên là 30% số người mua xe gặp vấn đề.

2. Nếu p khá gần 0,5 thì sự khác nhau giữa số n tìm được theo hai cách sẽ không nhiều lắm. Nhưng nếu n khá gần 0 hoặc 1 thì sự khác biệt sẽ rất lớn. Do đó, nếu cảm thấy tỷ lệ p rất bé hoặc rất lớn thì nên dùng cách thứ nhất.

Hướng dẫn sử dụng phần mềm thống kê R

R là phần mềm thống kê miễn phí, được phát triển tại phòng thí nghiệm AT&T bởi Rick Becker, John Chambers và các cộng sự. Phiên bản đầu tiên của R được viết vào năm 1976. Tham khảo tại [2].

[2]Nguyễn Văn Tuấn (2015). Phân tích dữ liệu với R. NXB tổng hợp thành phố Hồ Chí Minh.

1. Để tìm khoảng tin cậy cho giá trị trung bình trong trường hợp dữ liệu có phân phối chuẩn và phương sai chưa biết đơn giản nhất là ta dùng hàm *t.test()*.

2. Để tìm khoảng tin cậy cho giá trị trung bình trong trường hợp dữ liệu có phân phối chuẩn và phương sai đã biết ta dùng hàm *z.test()*.

3. Để tìm khoảng tin cậy cho tỷ lệ ta dùng hàm *prop.test()*.

Đối với các khoảng tin cậy một phía ta đặt giá trị "less", "greater" cho tham số *alternative* (giá trị mặc định của tham số này là "two.sided").

Bài tập Chương 4

Ước lượng khoảng cho kỳ vọng

Bài tập 4.1. Xác suất để một sinh viên Đại học Bách khoa Hà Nội thi trượt môn Giải tích 2 là p. Một mẫu lớn n sinh viên được lựa chọn ngẫu nhiên và ký hiệu X là số sinh viên đã trượt môn Giải tích 2 trong mẫu.

(a) Giải thích tại sao có thể sử dụng $\frac{X}{n}$ để ước lượng cho p?

(b) Trình bày cách tính xấp xỉ xác suất sự sai khác giữa $\frac{X}{n}$ và p nhỏ hơn $0,01$? Áp dụng cho $n = 500$ và $p = 0,2$.

Bài tập 4.2. Tuổi thọ của một loại bóng đèn do một dây chuyền công nghệ sản xuất ra có độ lệch chuẩn là 305 giờ. Người ta lấy ngẫu nhiên ra 45 bóng đèn loại này thấy tuổi thọ trung bình là 2150 giờ. Với độ tin cậy 95% hãy ước lượng tuổi thọ trung bình của loại bóng đèn nói trên.

Bài tập 4.3. Một kỹ sư cho biết trọng lượng tạp chất trong một sản phẩm tuân theo luật phân phối chuẩn với độ lệch chuẩn bằng 3,8gam. Một mẫu ngẫu nhiên gồm 9 sản phẩm được tiến hành kiểm tra và thấy lượng tạp chất như sau (đơn vị tính là gam):

18,2 13,7 15,9 17,4 21,8 16,6 12,3 18,8 16,2

(a) Tìm khoảng tin cậy cho trọng lượng trung bình tạp chất của sản phẩm với độ tin cậy 99%.

(b) Không cần tính toán, nếu độ tin cậy 95% thì khoảng ước lượng trung bình sẽ rộng hơn, hẹp hơn hay bằng như trong ý (a)?

Bài tập 4.4. Giả sử chiều dài của một chi tiết sản phẩm là biến ngẫu nhiên tuân theo luật phân phối chuẩn với độ lệch chuẩn là 0,2m. Người ta sản xuất thử nghiệm 35 sản phẩm loại này và tính được chiều dài trung bình là 25m. Với độ tin cậy 95% hãy ước lượng khoảng cho chiều dài trung bình của chi tiết sản phẩm đang được thử nghiệm.

Bài tập 4.5. Để xác định trọng lượng trung bình của các bao gạo được đóng gói bằng máy tự động, người ta chọn ngẫu nhiên ra 20 bao gạo và thấy trung bình mẫu là 49,2kg và độ lệch chuẩn mẫu hiệu chỉnh là 1,8kg. Biết rằng trọng lượng các bao gạo xấp xỉ phân phối chuẩn. Hãy tìm khoảng tin cậy cho trọng lượng trung bình của một bao gạo với độ tin cậy 99%.

Bài tập 4.6. Thời gian đợi phục vụ tại một cửa hàng ăn nhanh là biến ngẫu nhiên xấp xỉ phân phối chuẩn. Người ta khảo sát 16 người thì thấy thời gian đợi trung bình là 4 phút và độ lệch chuẩn mẫu hiệu chỉnh là 1,8 phút. Với độ tin cậy 99% hãy tìm khoảng tin cậy cho thời gian chờ đợi trung bình của một khách hàng tại cửa hàng ăn nhanh này.

Bài tập 4.7. Một mẫu ngẫu nhiên gồm 16 thùng hàng được chọn ra từ tất cả các thùng hàng được sản xuất bởi nhà máy trong một tháng. Trọng lượng của 16 thùng hàng lần lượt như sau (đơn vị tính là kg):

18,6	18,4	19,2	19,8	19,4	19,5	18,9	19,4
19,7	20,1	20,2	20,1	18,6	18,4	19,2	19,8

Tìm khoảng tin cậy cho trọng lượng trung bình tổng thể của tất cả các thùng hàng của nhà máy với độ tin cậy 95%, biết rằng trọng lượng thùng hàng được chọn ngẫu nhiên là biến ngẫu nhiên tuân theo luật phân phối chuẩn.

Bài tập 4.8. Để định mức thời gian gia công một chi tiết máy, người ta theo dõi ngẫu nhiên quá trình gia công 35 chi tiết máy và thu được số liệu:

Thời gian (phút)	16-17	17-18	18-19	19-20	20-21	21-22
Số chi tiết máy	3	4	10	9	5	4

Giả sử thời gian gia công chi tiết máy là biến ngẫu nhiên tuân theo luật phân phối chuẩn. Với độ tin cậy 95% hãy ước lượng khoảng tin cậy cho thời gian gia công trung bình một chi tiết máy nói trên.

Bài tập 4.9. Đo áp lực X (tính bằng kg/cm^2) của 18 thùng chứa ta được bảng kết quả sau:

Áp lực (kg/cm^2)	19,6	19,5	19,9	20,0	19,8	20,5	21,0	18,5	19,7
Số thùng	1	2	2	4	2	3	2	1	1

Với độ tin cậy 99% hãy tìm khoảng ước lượng đối xứng của áp lực trung bình của các thùng trên. Biết rằng áp lực là biến ngẫu nhiên tuân theo luật phân phối chuẩn.

Bài tập 4.10. Một bài báo trong Nuclear Engineering International (tháng 2 năm 1988, trang 33) mô tả một số đặc điểm của các thanh nhiên liệu được sử dụng trong một lò phản ứng hạt nhân của một công ty điện lực ở Na Uy. Người ta đo tỷ lệ làm giàu của 12 thanh và có được dữ liệu sau:

2,94 3,00 2,90 2,90 2,75 2,95 2,75 3,00 2,95 2,82 2,81 3,05

Giả sử tỷ lệ làm giàu của các thanh nhiên liệu tuân theo luật phân phối chuẩn. Hãy ước lượng khoảng cho tỷ lệ làm giàu trung bình của các thanh nhiên liệu với độ tin cậy 95%.

Bài tập 4.11. Trọng lượng những viên gạch trong một quá trình sản xuất gạch được giả sử là tuân theo luật phân phối chuẩn. Một mẫu ngẫu nhiên gồm 25 viên gạch vừa sản xuất ra trong ngày có trọng lượng trung bình 2,45 kg và độ lệch chuẩn mẫu hiệu chỉnh là 0,15 kg.

(a) Tìm khoảng tin cậy của trọng lượng trung bình của tất cả các viên gạch trong ngày với độ tin cậy 99%.

(b) Không cần tính toán, với độ tin cậy 95% thì khoảng tin cậy trung bình sẽ rộng hơn, hẹp hơn hay bằng với kết quả ý (a)?

(c) Một mẫu ngẫu nhiên gồm 20 viên gạch sẽ được chọn ra trong ngày mai. Không cần tính toán, với độ tin cậy 99% thì khoảng tin cậy cho trọng lượng trung bình của tất cả các viên gạch sản xuất ra trong ngày mai sẽ rộng hơn, hẹp hơn hay bằng như trong ý (a)?

(d) Sự thật rằng, độ lệch chuẩn mẫu của các viên gạch sản xuất trong ngày mai là 0,10kg. Không cần tính toán, với độ tin cậy 99% thì khoảng tin cậy cho trọng lượng trung bình của tất cả các viên gạch sản xuất ra trong ngày mai sẽ rộng hơn, hẹp hơn hay bằng như trong ý (a)?

Bài tập 4.12. Một trường đại học lớn đang quan tâm về lượng thời gian sinh viên tự nghiên cứu mỗi tuần. Người ta tiến hành khảo sát một mẫu ngẫu nhiên gồm 16 sinh viên, dữ liệu cho thấy thời gian nghiên cứu trung bình của một sinh viên là 15,26 giờ/tuần và độ lệch chuẩn hiệu chỉnh là 6,43 giờ. Giả sử thời gian nghiên cứu của sinh viên của trường đại học trên là tuân theo luật phân phối chuẩn.

(a) Tìm khoảng tin cậy cho lượng thời gian tự nghiên cứu trung bình mỗi tuần cho tất cả sinh viên trường đại học này với độ tin cậy 95%.

(b) Không cần tính toán, khoảng tin cậy của trung bình tổng thể khi ước lượng sẽ rộng hơn hay hẹp hơn với ba điều kiện sau:

b1. Mẫu gồm 30 sinh viên được chọn ra, với tất cả các điều kiện khác giống như ý (a)?

b2. Độ lệch chuẩn mẫu là 4,15 giờ, tất cả các điều kiện khác giống như ý (a)?

b3. Độ tin cậy 99%, tất cả các điều kiện khác giống như ý (a)?

Bài tập 4.13. Một kỹ sư nghiên cứu về cường độ nén của bê tông đang được thử nghiệm. Anh ta tiến hành kiểm tra 12 mẫu vật và có được các dữ liệu sau đây:

2216 2234 2225 2301 2278 2255 2249 2204 2286 2263 2275 2295

Giả sử cường độ nén của bê tông đang thử nghiệm tuân theo luật phân phối chuẩn.

(a) Hãy ước lượng khoảng với độ tin cậy 95% cho cường độ nén trung bình của bê tông đang được thử nghiệm.

(b) Hãy ước lượng khoảng tin cậy phải cho cường độ nén trung bình của bê tông đang được thử nghiệm với độ tin cậy 99%.

Bài tập 4.14. Người ta chọn ngẫu nhiên ra 49 sinh viên của một trường đại học và thấy chiều cao trung bình mẫu là 163cm và độ lệch chuẩn mẫu hiệu chỉnh là 12cm. Hãy tìm khoảng ước lượng với độ tin cậy 99% cho chiều cao trung bình của sinh viên của trường đó.

Bài tập 4.15. Một trường đại học tiến hành một nghiên cứu xem trung bình một sinh viên tiêu hết bao nhiêu tiền điện thoại trong một tháng. Họ điều tra 60 sinh viên và cho thấy số tiền trung bình mẫu là 95 nghìn và độ lệch chuẩn mẫu hiệu chỉnh là 36 nghìn. Hãy ước lượng khoảng với độ tin cậy 95% cho số tiền điện thoại trung bình trong một tháng của mỗi sinh viên.

Bài tập 4.16. Người ta điều tra 35 người nghiện thuốc lá được chọn ngẫu nhiên từ số lượng người nghiện hút thuốc lá của một thành phố thấy số điếu thuốc hút trong 5 ngày của họ là:

31	37	48	40	59	97	98	87	80	68	64	45
48	62	74	76	79	85	83	81	93	82	85	79
34	57	95	49	59	63	48	79	50	55	63	

Hãy tìm khoảng ước lượng cho số điếu thuốc hút trung bình trong 5 ngày của những người nghiện thuốc lá của thành phố đó với độ tin cậy 99%.

Bài tập 4.17. Để nghiên cứu về thời gian xem ti vi của một thanh niên từ 18 đến 35 tuổi trong vòng một tuần, người ta tiến hành khảo sát trên 40 người và cho ta bảng số liệu sau:

39	02	43	35	15	54	23	21	25	07	24	33	17	
23	24	43	11	15	17	15	19	06	43	35	25	37	
15	14	08	11	29	12	13	25	15	28	24	06	16	7

Hãy tìm khoảng ước lượng cho thời gian xem ti vi trung bình của thanh niên trong độ tuổi trên trong vòng một tuần với độ tin cậy 99%.

Bài tập 4.18. Để điều tra tiền điện phải trả trong một tháng của một hộ dân cư ở phường A, người ta kiểm tra ngẫu nhiên 200 hộ gia đình ở phường này và được kết quả sau:

Số tiền (nghìn đồng)	[80,180)	[180,280)	[280,380)	[380,480)	[480,580)	[580,680)	[680,780]
Số hộ gia đình	14	25	43	46	39	23	10

Ước lượng khoảng cho số tiền trung bình một hộ dân phải trả ở phường đó với độ tin cậy 95%.

Bài tập 4.19. Để ước lượng số lượng xăng hao phí trên một tuyến đường của một hãng xe khách, người ta tiến hành chạy thử nghiệm 55 lần liên tiếp trên tuyến đường này và có được số liệu:

Lượng xăng hao phí	10,5-11	11-11,5	11,5-12	12-12,5	12,5-13	13-13,5
Tần số	5	12	15	13	6	4

Hãy ước lượng lượng xăng hao phí trung bình cho một xe với độ tin cậy 95%.

Bài tập 4.20. Để xác định giá trung bình đối với một loại hàng hóa trên thị trường, người ta điều tra ngẫu nhiên tại 100 cửa hàng thu được số liệu sau:

Giá (nghìn đồng)	83	85	87	89	91	93	95	97	99	101
Số cửa hàng	5	8	13	14	30	11	8	6	4	1

Với độ tin cậy 95% hãy ước lượng giá trung bình của loại hàng đó tại thời điểm đang xét. Biết rằng giá hàng hóa là biến ngẫu nhiên tuân theo luật phân phối chuẩn.

Ước lượng khoảng cho tỷ lệ hay xác suất

Bài tập 4.21. Để ước lượng cho tỷ lệ những cây bạch đàn có chiều cao đạt chuẩn phục vụ cho việc khai thác ở một nông trường lâm nghiệp, người ta tiến hành đo ngẫu nhiên chiều cao của 135 cây và thấy có 36 cây cao từ 7,5m trở lên. Hãy ước lượng khoảng cho tỷ lệ các cây bạch đàn có chiều cao trên 7,5m với độ tin cậy 95%.

Bài tập 4.22. Để ước lượng số cá có trong hồ người ta bắt từ hồ lên 100 con đánh dấu rồi thả lại vào hồ. Sau đó người ta bắt lên 300 con thì thấy có 32 con bị đánh dấu. Hãy ước lượng khoảng cho số cá có trong hồ với độ tin cậy 99%.

Bài tập 4.23. Để điều tra thị phần xe máy, người ta chọn ngẫu nhiên ra 450 người mua xe máy trong một tháng ở các địa bàn ở một thành phố thì có 275 người mua xe Honda. Tìm khoảng tin cậy cho tỷ lệ người mua xe Honda với độ tin cậy 95%.

Bài tập 4.24. Kiểm tra ngẫu nhiên 400 sản phẩm do một hệ thống máy mới sản xuất thì thấy có 387 chính phẩm. Hãy ước lượng tỷ lệ chính phẩm tối thiểu của hệ thống máy mới với độ tin cậy 95%.

Bài tập 4.25. Thử nghiệm 560 bóng đèn điện tử do một nhà máy sản xuất thì thấy 10 bóng có lỗi kỹ thuật. Hãy tìm ước lượng cho tỷ lệ bóng có lỗi kỹ thuật tối đa với độ tin cậy 95%.

Bài tập 4.26. Mở thử 200 hộp của kho đồ hộp thấy có 10 hộp bị biến chất. Với độ tin cậy 95% hãy ước lượng tỷ lệ hộp bị biến chất tối đa của kho.

Bài tập 4.27. Chọn ngẫu nhiên ra 1000 trường hợp điều trị bệnh ung thư phổi, các bác sĩ thống kê thấy có 823 bệnh nhân bị chết trong vòng 10 năm.

(a) Ước lượng khoảng cho tỷ lệ tử vong của bệnh nhân điều trị bệnh ung thư phổi với độ tin cậy 99%.

(b) Cần phải lấy số lượng mẫu là bao nhiêu để với độ tin cậy 95% các sai số khi dự đoán tỷ lệ bệnh nhân điều trị ung thư phổi tử vong 10 năm là ít hơn 0,03?

Bài tập 4.28. Cần phải lập một mẫu ngẫu nhiên với kích thước là bao nhiêu để tỷ lệ phế phẩm của mẫu là 0,2 và độ dài khoảng tin cậy đối xứng là 0,05 và độ tin cậy của ước lượng là 95%.

Bài tập 4.29. Làm cách nào để ước lượng số thú hiếm trong một khu rừng với độ tin cậy 95%.

Bài tập 4.30. Nghiên cứu về năng suất của loại hoa màu A, người ta kiểm tra năng suất của 64 điểm trồng loại hoa màu này thu được bảng số liệu

Năng suất (tạ/ha)	40–45	45–50	50–55	55-60	60–65	65–70
Số điểm	2	5	15	30	8	4

(a) Hãy ước lượng năng suất trung bình của loại hoa màu A với độ tin cậy 95%; Nếu muốn sai số của ước lượng giảm đi 2 lần thì cần kiểm tra bao nhiêu điểm để đảm bảo yêu cầu nêu trên?

(b) Biết rằng trên toàn miền Bắc có 10.000 điểm trồng loại hoa màu A. Hãy cho biết có khoảng bao nhiêu điểm đạt năng suất trên 60 tạ/ha? Hãy kết luận với độ tin cậy 99%.

(c) Hãy cho biết tỷ lệ những điểm có năng suất trên 60 tạ/ha của loại hoa màu A tối thiểu là bao nhiêu? Hãy kết luận với độ tin cậy 95%?