MỤC LỤC

CHƯƠNG 1

THỐNG KÊ VÀ CÁC KẾT LUẬN THỐNG KÊ

1.1 Mở đầu

Như chúng ta biết Lý thuyết xác suất được sinh ra từ việc nghiên cứu các quy luật ngâu nhiên ẩn sau các bài toán thực tế. Thông qua việc nghiên cứu các bài toán thực tế chúng ta tìm ra các quy luật ngẫu nhiên. Những quy luật đó chúng ta đã được học trong chương 1.

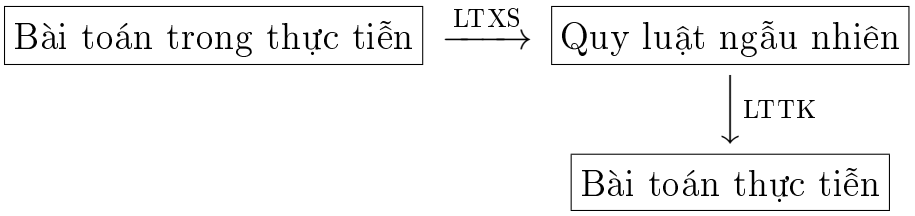
Một câu hỏi tự nhiên được đặt ra là: *Những quy luật ngẫu nhiên đó được nghiên cứu ra để làm gì? Có làm tăng sự hiểu biết của chúng ta về các hiện tượng tự nhiên và xác hội hay không?*

Câu trả lời là: Các quy luật ngẫu nhiên của lý thuyết xác suất sau khi được nghiên cứu ra thì nó được sử dụng để nghiên cứu các bài toán trong thực tế và người ta thường gọi bài toán thống kê. Vậy *Thống kê là gì?* Cho đến nay chúng ta có rất nhiều thuật ngữ thống kê khác nhau trong thực tế chẳng hạn như:

$$\left\{\begin{array}{l} \text{- Thống kê kinh tế} \\ \text{- Thống kê sinh học} \\ \text{- Vật lý thống kê} \\ \ldots\ldots \end{array}\right. \Leftarrow \text{Dùng } \textbf{thống kê toán học} \text{ làm công cụ.}$$

- Vậy *thống kê toán* học là gì? Nó có nhiệm vụ gì? Gồm những nội dung nào?

Thống kê toán học là phần ứng dụng của lý thuyết xác suất



Nhiệm vụ của thống kê Toán học

Thống kê toán học nghiên cứu các phương pháp thu thập, phân tích, xữ lý các số liệu thống kê để đưa ra các quyết định có cơ sở khoa học phục vụ cho việc quản lý xã hội.

Nội dung của thống kê toán học

Lý thuyết thống kê toán học có các nội dung cơ bản sau đây:

- Bài toán về lý thuyết chọn mẫu
- Bài toán ước lượng tham số
- Bài toán kiểm định giả thiết
- Bài toán phân tích tương quan và hồi quy

1.2 Thống kê mô tả

1.2.1 Tổng thể và mẫu ngẫu nhiên

Trong thực tế, nhiều khi ta cần quan tâm đến một số đặc điểm (định tính hoặc định lượng) của các phần tử thuộc về một tập hợp nào đó, chẳng hạn tuổi thọ của một loại sản phẩm nào đó, thu nhập trung bình

củ người dân ở một quốc gia, tỉ lệ sản phẩm đạt tiêu chuẩn, tỉ lệ người dân bỏ phiếu cho một ứng cử viên nào đó, tỉ lệ cá thể nhiễm bệnh trong quần thể, ... Tập hợp các phần tử cần nghiên cứu này được gọi là *đám đông* hay *tổng thể*, ký hiệu là $\mathcal{C}$.

Việc tiến hành thu thập thông tin trên các phần tử của đám đông được gọi là *quan sát*.

Thuộc tính của một đối tượng mà chúng ta quan tâm thường là chúng ta chưa biết hoặc biết chưa đầy đủ và chúng ta coi đó là được coi như một đại lượng ngẫu nhiên, ký hiệu là X và được gọi là *đại lượng ngẫu nhiên gốc đám đông* $\mathcal{C}$. Quá trình đi nghiên cứu đám đông của $\mathcal{C}$ thực chất là quá trình đi tìm quy luật phân phối của đại lượng ngẫu nhiên X, nhiều khi đó là quá trình đi tìm các số đặc trưng của X.

Đặc điểm của đám đông (tổng thể) thường được nghiên cứu dưới hai phương diện:

$\diamond$ Phương diện định lượng: Khi ta cần quan tâm đến các giá trị về lượng của đại lượng ngẫu nhiên X như: trọng lượng, năng suất, tuổi thọ, ... và ta thường quan tâm đến hai đặc trưng:

- Kỳ vọng $\mathbb{E}X = \mu$: đặc trưng giá trị trung bình của đặc điểm định lượng cần quan tâm trên đám đông $\mathcal{C}$.

- Phương sai $\mathbb{D}X = \sigma^2$: đặc trưng cho mức độ biến động giá trị của đặc điểm định lượng cần quan tâm trên đám đông $\mathcal{C}$.

$\diamond$ Phương diện định tính: Khi ta cần quan tâm đến một tính chất A nào đó trên đám đông, các phần tử của đám đông hoặc có tính chất A hoặc không có tính chất A như: chất lượng sản phẩm, sự nẩy mầm của một giống lúa, chất độc hại trong nguồn nước, ... Giá trị mà đại lượng

ngẫu nhiên X có thể nhận được

$$X = \begin{cases} 1 & \text{khi phần tử đó có tính chất } A\,; \\ 0 & \text{khi phần tử đó không có tính chất } A\,, \end{cases}$$

và ta thường quan tâm đến xác suất $\mathbb{E}X = p$.

Chúng ta khó có thể quan sát hết tất cả các phần tử của đám đông vì những lý do như thời gian, chi phí tốn kém, ... Chính vì vậy, người ta chỉ lấy ra một số phần tử đại diện cho đám đông và nghiên cứu trên tập phần tử này, tập hợp các phần tử đại diện cho đám đông đó được gọi là *mẫu*. Phương pháp nghiên cứu trên mẫu đại diện cho đám đông được gọi là *phương pháp mẫu* và cách thức thực hiện quá trình lấy mẫu được gọi là *phương pháp lấy mẫu*.

Khi cần quan tâm đến đặc điểm là đại lượng ngẫu nhiên X của đám đông $\mathcal{C}$, ta chọn ra mẫu có n phần tử, trong đó việc chọn phần tử thứ i là quá trình thực hiện một phép thử rút ngẫu nhiên một phần tử của đám đông $\mathcal{C}$, giá trị ngẫu nhiên này được gán cho đại lượng ngẫu nhiên X_i. Với cách chọn này, các đại lượng ngẫu nhiên X_i độc lập với nhau và có cùng luật phân phối với đại lượng ngẫu nhiên X. Mẫu này được gọi là *mẫu ngẫu nhiên* có kích thước n của đám đông $\mathcal{C}$. Vậy *mẫu ngẫu nhiên là gì?*.

Định nghĩa 1.2.1. Cho $X_1, \cdots X_n$ là dãy các biến ngẫu nhiên độc lập cùng phân phối với biến ngẫu nhiên X. Khi đó véc tơ ngẫu nhiên $(X_1, \cdots X_n)$ được gọi là mẫu ngẫu nhiên cỡ n lấy từ X. Một bộ giá trị $(x_1, ..., x_n)$ của véc tơ ngẫu nhiên $(X_1, \cdots X_n)$ được gọi là một thể hiện của mẫu ngẫu nhiên hay thường gọi là *một mẫu cụ thể*.

Ví dụ 1. Thống kê về số chấm của một con xúc xắc khi gieo 5 lần.

Mẫu ngẫu nhiên: $(X_1,\, X_2, ...,\, X_5)$; mẫu cụ thể: $(2,\, 3,\, 1,\, 6,\, 2)$.

Các phương pháp lấy mẫu

Việc lấy mẫu được coi là tốt nếu như thông tin thu được từ mẫu phản ánh càng gần với đặc điểm của đám đông (tính chất đại diện cao). Chính vì vậy, trong thống kê việc lấy mẫu là một công việc hết sức quan trọng. Người ta thường sử dụng một số phương pháp lấy mẫu như sau:

Lấy mẫu ngẫu nhiên đơn giản

Là phương pháp lấy mẫu thỏa mãn các điều kiện: mỗi lần chỉ được chọn một phần tử từ đám đông, khả năng được chọn của tất cả các phần tử trong đám đông đều như nhau. Có hai cách thức tiến hành chọn, đó là chọn hoàn lại và chọn không hoàn lại, tuy nhiên khi kích thước của đám đông lớn hơn nhiều so với kích thước mẫu thì có thể coi hai phương pháp chọn này là giống nhau.

Phương pháp lấy mẫu ngẫu nhiên đơn giản ở trên có tính chất đại diện cho đám đông cao, tuy nhiên nó khó thực hiện và cần nhiều thời gian cũng như kinh phí. Ta có thể xem phương pháp lấy mẫu này là hoàn toàn ngẫu nhiên hay ngẫu nhiên không có định hướng.

Lấy mẫu ngẫu nhiên có định hướng

◇ Lấy mẫu theo nhóm: là phương pháp chia đám đông thành các nhóm thuần nhất, từ mỗi nhóm này ta lấy ra một mẫu ngẫu nhiên đơn giản với một kích thước tương ứng. Tập hợp tất cả các phần tử thu được từ các mẫu ngẫu nhiên đơn giản đó lập nên mẫu ngẫu nhiên theo nhóm.

◇ Lấy mẫu theo chùm: là phương pháp chia đám đông thành nhiều chùm (đám đông con) sao cho giữa các chùm có sự đồng đều về quy mô, từ các chùm đó ta lấy một mẫu ngẫu nhiên đơn giản. Tập hợp tất cả phần tử thu được từ các mẫu ngẫu nhiên đơn giản của các chùm lập nên mẫu ngẫu nhiên theo chùm.

Phương pháp này dễ quy hoạch, có thể tiết kiệm được thời gian và kinh phí nhưng sai số chọn mẫu cao hơn các phương pháp nói trên.

Ví dụ 2. Chúng ta muốn đi tìm hiểu về tổng thu nhập trong một năm của toàn bộ cán bộ công chức của một tỉnh.

- Chia đám đông này thành các nhóm theo từng cơ cấu ngành nghề: quốc phòng, an ninh, giáo dục, y tế, kinh doanh, ... Trong mỗi cơ cấu ngành nghề có sự thuần nhất về mức lương (nếu có sự sai khác về thu nhận chủ yếu là do thâm niên và chức vụ công tác). Như vậy, phương pháp lấy mẫu bằng việc gom lại các mẫu ngẫu nhiên đơn giản của từng nhóm ngành nghề chính là phương pháp lấy mẫu theo nhóm.

- Chia đám đông này theo các huyện trong tỉnh A. Giữa các huyện, có sự đồng đều về quy mô (đầy đủ các thành phần) và phương pháp lấy mẫu bằng việc gom lại các mẫu ngẫu nhiên đơn giản của từng huyện chính là phương pháp lấy mẫu theo chùm.

1.2.2 Cách biểu diễn mẫu

Bảng tần số và bảng tần suất

Ta thực hiện n lần quan sát trên đám đông $\mathcal{C}$, khi đó ta sẽ thu được mẫu cụ thể gồm k giá trị khác nhau $(x_1, x_2, ..., x_k),\ k \leqslant n$. Giá trị x_i có n_i lần xuất hiện, n_i được gọi là *tần số xuất hiện* của x_i và tỉ số $\dfrac{n_i}{n}$ được gọi là *tần suất xuất hiện* của x_i, ký hiệu là f_i. Ta có biểu diễn kết quả của mẫu bằng bảng tần số và tần suất như sau

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

x_i	x_1	x_2	...	x_k
f_i	f_1	f_2	...	f_k

trong đó

$$n = \sum_{i=1}^{k} n_i; \quad \sum_{i=1}^{k} f_i = 1.$$

Ví dụ 1. Thống kê điểm số kết thúc học phần của một lớp gồm 40 sinh viên, ta có

x_i	4	5	6	7	8
n_i	5	10	12	8	5

x_i	4	5	6	7	8
f_i	5/40	10/40	12/40	8/40	5/40

Trong trường hợp mẫu cụ thể $(x_1, x_2, ..., x_n)$ có nhiều giá trị khác nhau, khi đó ta thực hiện việc ghép lớp. Nguyên tắc ghép lớp được tiến hành như sau

- Số lớp chia k được xác định trên cơ sở $k = \min\{l : 2^l > n\}$.
- Độ dài mỗi lớp: $l = \dfrac{\text{giá trị lớn nhất - giá trị nhỏ nhất}}{k}$.
- Trong 2 lớp liền nhau $x_{i-1} \to x_i$, $x_i \to x_{i+1}$ thì x_i thuộc lớp $x_{i-1} \to x_i$.

Ngoài phương pháp ghép lớp đã trình bày ở trên, còn có một số phương pháp ghép lớp khác, với những mẫu cụ thể rời rạc người ta có thể chia thành các lớp có độ dài khác nhau, các lớp được chia rời nhau. Trong phạm vi giáo trình này, chúng ta không đề cập cụ thể các kiểu ghép lớp này.

Ví dụ 2. Thống kê về chiều cao của 30 sinh viên với chiều cao nằm trong khoảng từ 1m50 đến 1m75.

Nhận thấy $2^5 > 30$ và $2^4 < 30$ nên ta chọn $k = 5$. Bảng tần số, tần suất như sau:

Lớp	Giá trị	Tần số	Tần suất
150-155	152,5	4	4/30
155-160	157,5	7	7/30
160-165	162,5	6	6/30
165-170	167,5	10	10/30
170-175	172,5	3	3/30

1.2.3 Đa giác tần số và tổ chức đồ

Đối với số liệu chưa ghép lớp

- Chấm trên mặt phẳng các điểm $(x_i, n_i), i = 1, 2, \ldots, n$.

- Nối các điểm $(x_i, 0)$ với các điểm (x_i, n_i), ta được *biểu đồ tần số hình gậy.*

- Nối liên tiếp điểm (x_i, n_i) với các điểm (x_{i+1}, n_{i+1}) ta được *biểu đồ đa giác tần số.*

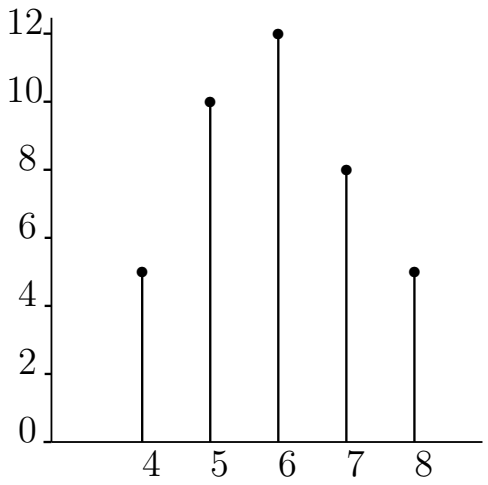
Hoàn toàn tương tự đối với tần suất:

- Chấm trên mặt phẳng các điểm $(x_i, f_i), i = 1, 2, \ldots, n$.

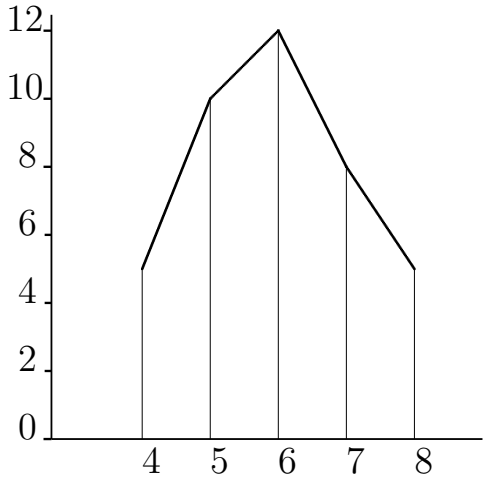
- Nối các điểm $(x_i, 0)$ với các điểm (x_i, f_i), ta được *biểu đồ tần suất hình gậy.*

- Nối liên tiếp điểm (x_i, f_i) với các điểm (x_{i+1}, f_{i+1}) ta được *biểu đồ đa giác tần suất.*

Ví dụ 3. Minh họa số liệu của ví dụ thống kê điểm



Biểu đồ tần số hình gậy



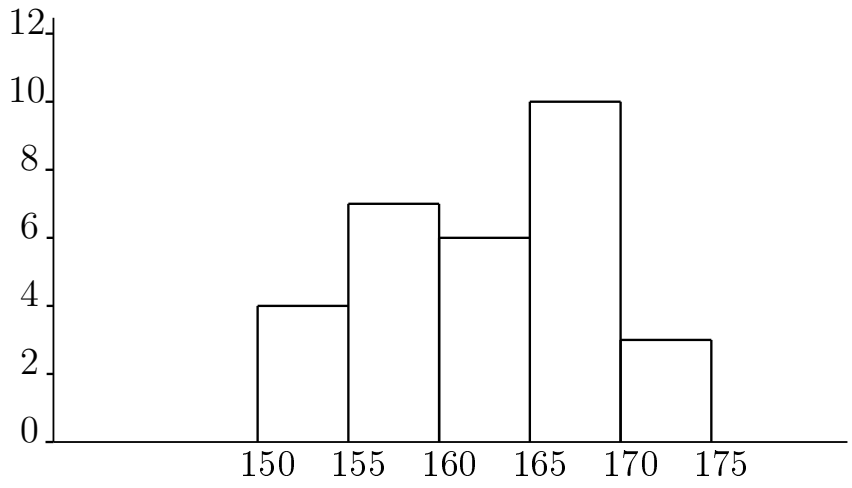
Biểu đồ đa giác tần số

Đối với số liệu đã ghép lớp.

- Trên mỗi lớp ta dựng hình chữ nhật có chiều cao bằng tần số (hay tần suất) tương ứng với lớp đó.

- Tô đậm hoặc kẻ chéo bằng các đường song song các hình chữ nhật này ta thu được *tổ chức đồ tần số* (hay *tổ chức đồ tần suất*).

Ví dụ 4. Minh họa số liệu của Ví dụ 2.



Biểu đồ đa giác tần số

1.2.4 Phân phối mẫu và các đặc trưng của mẫu

Trong nội dung Chương 1 chúng ta đã được làm quen với việc tính các đặc trưng của đại lượng ngẫu nhiên thông qua phân phối xác suất đã biết trước.

Tuy nhiên, trong thực tế thật khó khăn để xác định được tường minh phân phối xác suất của một đại lượng ngẫu nhiên gốc đám đông. Chính vì vậy, trên cơ sở của các thông tin thu thập được từ các mẫu, người ta đem ra một số công thức giúp chúng ta tính được các đặc trưng của mẫu.

Các giá trị này rất quan trọng và có sự tương ứng với những số đặc trưng của đại lượng ngẫu nhiên đã trình bày ở phần trước.

Hàm phân phối mẫu

X là đại lượng ngẫu nhiên gốc đám đông có hàm phân phối xác suất $F(x)$ chưa biết. Khi ta thực hiện n quan sát, gọi hàm $F_n(x) = \dfrac{m_x}{n}$ với

m_x: là số quan sát có giá trị x_i bé hơn x ($i = \overline{1,n}$) là *hàm phân phối mẫu*.

Tính chất của hàm phân phối mẫu $F_n(x)$:

+ $0 \leqslant F_n(x) \leqslant 1$,

+ $F_n(x)$ là hàm đơn điệu tăng,

+ $F_n(x)$ là hàm liên tục bên trái.

Khi kích thước mẫu lớn thì phân phối mẫu $F_n(x)$ càng gần với phân phối xác suất của đại lượng ngẫu nhiên X. Khi n đủ lớn, ta có thể dùng $F_n(x)$ thay thế cho $F(x)$ chưa biết hoặc dựa vào $F_n(x)$ ta có thể sơ lược về dáng điệu của $F(x)$ và đưa ra những dự đoán về dạng của $F(x)$ cũng như tính toán các số đặc trưng có liên quan.

Ví dụ 1.2.1. Bảng tần số từ ví dụ thống kê điểm

x_i	4	5	6	7	8
n_i	5	10	12	8	5

Hàm phân phối mẫu

$$F_n(x) = \begin{cases} 0 & \text{với } x \leqslant 4, \\ \dfrac{5}{40} & \text{với } 4 < x \leqslant 5, \\ \dfrac{15}{40} & \text{với } 5 < x \leqslant 6, \\ \dfrac{27}{40} & \text{với } 6 < x \leqslant 7, \\ \dfrac{35}{40} & \text{với } 7 < x \leqslant 8, \\ 1 & \text{với } x > 8. \end{cases}$$

Trung bình mẫu

Định nghĩa 1.2.2. Giả sử $n(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên có kích thước n của đám đông X, khi đó $\frac{1}{n}\sum_{i=1}^{n} X_i$ được gọi là *trung bình mẫu* và ký hiệu là $\overline{X}$.

Trong thực hành tính toán

Đối với một mẫu cụ thể $(x_1, x_2, ..., x_n)$ thì trung bình mẫu thực nghiệm được xác định như sau $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Trường hợp mẫu cụ thể đã được ghép bộ có bảng tần số

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

thì trung bình mẫu thực nghiệm là $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\, x_i$.

Ví dụ 1.2.2. Bảng tần số từ ví dụ thống kê điểm

x_i	4	5	6	7	8
n_i	5	10	12	8	5

Khi đó $\overline{x} = \frac{1}{40}\sum_{i=1}^{5} n_i\, x_i = \frac{238}{40} = 5{,}95$.

Nhận xét 1. Công thức tính trung bình mẫu ở trên là dạng tổng quát, tuy nhiên do đặc trưng số nên ta thường dùng khi nghiên cứu về một đặc điểm định lượng nào đó của đám đông. Đối với đặc điểm định tính A ta có khái niệm tỉ lệ mẫu

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

trong đó X_i chỉ nhận 2 giá trị là 0 và 1 (bằng 1 nếu quan sát đó có tính chất A, bằng 0 nếu quan sát đó không có tính chất A). Với $m = \sum_{i=1}^{n} X_i$ chính là số quan sát có tính chất A, công thức tính tỉ lệ mẫu là $F = \dfrac{m}{n}$.

Phương sai mẫu và phương sai hiệu chỉnh mẫu

Định nghĩa 1.2.3. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên có kích thước n của đám đông X, khi đó $\dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ được gọi là *phương sai mẫu* và ký hiệu là $\hat{S}^2$.

Ngoài ra, chúng ta thường dùng một đặc trưng mẫu khá quan trọng là *phương sai hiệu chỉnh mẫu*, ký hiệu là S^2, được xác định $S^2 = \dfrac{n}{n-1}\hat{S}^2$.

Mệnh đề 1.2.4. *Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên có kích thước n của đám đông X. Ta có*

$$\hat{S}^2 = \overline{X^2} - (\overline{X})^2 \quad \textit{trong đó} \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Chứng minh.

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

□

Trong thực hành tính toán

Dối với một mẫu cụ thể đã được ghép bộ có bảng tần số

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

thì *phương sai mẫu thực nghiệm* và *phương sai hiệu chỉnh mẫu thực nghiệm* được xác định như sau

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{k} n_i\big(x_i - \overline{x}\big)^2 = \overline{x^2} - \big(\overline{x}\big)^2;$$

$$s^2 = \frac{n}{n-1}\,\hat{s}^2 = \frac{n}{n-1}\big(\overline{x^2} - \big(\overline{x}\big)^2\big).$$

s được gọi là *độ lệch chuẩn mẫu.*

Việc đưa ra các khái niệm trung bình mẫu thực nghiệm (phương sai mẫu thực nghiệm, phương sai hiệu chỉnh mẫu thực nghiệm) chỉ nhằm nhấn mạnh đó là giá trị bằng số cụ thể, được xác định từ thực nghiệm.

Ví dụ 1.2.3. Bảng tần số từ ví dụ thống kê điểm

x_i	4	5	6	7	8
n_i	5	10	12	8	5

x_i	n_i	$n_i\,x_i$	$n_i\,x_i^2$
4	5	20	80
5	10	50	250
6	12	72	432
7	8	56	392
8	5	40	320
Tổng	**40**	**238**	**1474**

Ta có $\overline{x} = \dfrac{238}{40} = 5{,}95;\ \overline{x}^2 = \dfrac{1474}{40} = 36{,}85.$

$\hat{s}^2 = 36{,}85 - 5{,}95^2 = 1{,}4475;\ \ s^2 \approx 1{,}485.$

Nhận xét 2. Đối với mẫu được ghép lớp, việc tính các số đặc trưng của mẫu cũng theo trình tự tiến hành như trên, trong mỗi lớp ta sử dụng giá trị trung điểm $x_i' = \dfrac{x_i + x_{i+1}}{2}$ của lớp.

1.3 Ước lượng tham số

1.3.1 Mở đầu

Giả sử đại lượng ngẫu nhiên X có luật phân phối phụ thuộc vào một tham số hoặc một véctơ tham số θ chưa biết chẳng $X \sim P(\lambda),\ N(\mu, \sigma)$; $B(n,p), \mathcal{E}(\lambda), ...$ nhưng chưa biết tham số λ, μ, σ, p .v.v. Khi đó để xác định hoàn toàn phân phối xác suất của X ta phải xác định được giá trị tham số. Để có được điều đó người ta phải quan sát X xay dựng mẫu ngâu nhiên $(X_1, X_2, ...X_n)$ từ đó tìm cách ước lượng tham số. Đây chính là *bài toán ước lượng tham số.*

Trong thực tế người ta xét 2 loại ước lượng tham số cơ bản đó là: *ước lượng điêm* và *ước lượng khoảng.*

Ước lượng điểm: Xuất phát tư mẫu ngẫu nhiên $(X_1, X_2, ...X_n)$ người ta xây dựng thống kê $\hat{\theta}(X_1, X_2, ...X_n)$ dùng để ước lượng tham số θ theo các nghĩa khác nhau như *Ước lượng không chêch (không có sai số hệ thống), ước lượng vững , ước lương hiệu quả, ước lượng hợp lý cực đai*

Ước lượng khoảng: Xuất phát tư mẫu ngẫu nhiên $(X_1, X_2, ...X_n)$ người ta xây dựng thống kê $\theta_1 := \hat{\theta}_1(X_1, X_2, ...X_n)$ và $\theta_2 := \hat{\theta}_2(X_1, X_2, ...X_n)$ sao cho

$$\mathbb{P}\{\hat{\theta}_1(X_1, X_2, ...X_n) \leq \theta \leq \hat{\theta}_2(X_1, X_2, ...X_n)\} \geq 1 - \alpha$$

trong đó α được gọi là *mức ý nghĩa* và $\beta = 1 - \alpha$ được gọi là *độ tin cậy.*

Khi đó, người ta nói rằng với độ tin cậy α hay mức ý nghĩa α khoảng tin cậy đối với θ là $(\theta_1; \theta_2)$.

1.3.2 Ước lượng điểm

$\diamond$ Ước lượng $\hat{\theta}(X_1, X_2, ..., X_n)$ được gọi là *ước lượng không chệch* của θ, nếu thỏa mãn $\mathbb{E}\hat{\theta} = \theta$.

$\diamond$ Ước lượng $\hat{\theta}(X_1, X_2, ..., X_n)$ được gọi là *ước lượng vững* của θ, nếu với n lớn vô hạn thì $\hat{\theta}$ hội tụ theo xác suất về θ, nghĩa là với mọi $\varepsilon > 0$ tùy ý thì

$$\lim_{n\to\infty} \mathbb{P}[|\hat{\theta} - \theta| < \varepsilon] = 1.$$

$\diamond$ Ước lượng $\hat{\theta}(X_1, X_2, ..., X_n)$ được gọi là *ước lượng hợp lý tối đa* của θ, nếu

$$L(x, \theta) = \prod_{i=1}^{n} p(X_i, \theta)$$

đạt cực đại tại $\hat{\theta}$. $L(x, \theta)$ được gọi là *hàm hợp lý* của X, trong đó $p(x, \theta)$ là hàm mật độ xác suất hoặc là hàm tính xác suất của đại lượng ngẫu nhiên X.

$\diamond$ Ước lượng $\hat{\theta}(X_1, X_2, ..., X_n)$ được gọi là *ước lượng hiệu quả* của θ, nếu như nó là ước lượng không chệch và có phương sai bé nhất trong tất cả các ước lượng không chệch của θ.

Nếu hàm mật độ xác suất của đại lượng ngẫu nhiên X thỏa mãn thêm một số điều kiện nhất định thì ta có bất đẳng thức Cramer-Rao

$$D(\theta^*) \geqslant \frac{1}{n\mathbb{E}\Big(\dfrac{\partial \ln p(X,\, \theta)}{\partial \theta}\Big)^2} \;;\;\; \forall \theta^* \;:\;\; \mathbb{E}(\theta^*) = \theta.$$

do đó, ước lượng không chệch $\hat{\theta}$ là ước lượng hiệu quả của θ khi

$$V(\hat{\theta}) = \frac{1}{n\mathbb{E}\Big(\dfrac{\partial \ln p(X,\,\theta)}{\partial\theta}\Big)^2}.$$

Từ bất đẳng thức Cramer-Rao, ta thấy một điều lý thú đó là: đã là ước lượng thì phải chấp nhận sai số, bất đẳng thức cho ta cận dưới của sai số.

Ví dụ 1.3.1. Giả sử X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, có kỳ vọng μ và phương sai hữu hạn, khi đó trung bình mẫu $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ chính là ước lượng không chệch, lước lượng vững, ước lượng hiệu quả, ước lượng hợp lý cực đại của μ.

Ví dụ 1.3.2. Giả sử X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, có phương sai $\mathbb{D}X = \sigma^2$ cần ước lượng, khi đó phương sai hiệu chỉnh mẫu $S^2 = \frac{1}{n}\sum(X_i - \bar{X})$ chính là ước lượng không chệch của σ^2. Như vậy S^2 là ước lượng không chệch của σ^2. Mặt khác $\hat{S}^2 = \dfrac{n-1}{n}S^2$ nên $\hat{S}^2$ không phải là ước lượng không chệch của σ^2. Tuy nhiên người ta chứng minh được rằng cả S^2 và $\hat{S}^2$ đều là ước lượng vững của σ^2.

Ví dụ 1.3.3. Giả sử X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, ta cần quan tâm đến một tính chất A có xác suất $p = \mathbb{P}(A) = \mathbb{E}X$ cần ước lượng, khi đó tỉ lệ mẫu F chính là ước lượng không chệch của xác suất p.

Khẳng định trên là hiển nhiên vì thực chất tỉ lệ mẫu cũng là trung bình mẫu khi đặc điểm định tính được số hóa dưới dạng

$$X_i = \begin{cases} 1 & \text{khi phần tử đó có tính chất } A\,; \\ 0 & \text{khi phần tử đó không có tính chất } A\,, \end{cases}$$

và $\mathbb{E}F = \mathbb{E}\overline{X} = \mathbb{E}X = p$.

Ngoài ra người ta còn chứng minh được F cũng chính là ước lượng vững của xác suất p.

1.3.3 Ước lượng khoảng

Trong nội dung của phần trước, chúng ta đã đề cập đến ước lượng điểm của tham số. Do θ là tham số chưa biết nên ước lượng điểm chỉ cho ta một cách nhìn hết sức tương đối và có phần chưa thỏa đáng. Sau đây chúng ta sẽ suy nghĩ đến một cách tiếp cận khác để tìm ra miền giá trị của θ.

1.3.4 Khái niệm về khoảng tin cậy

Cho X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, có tham số θ cần ước lượng. Căn cứ vào mẫu ngẫu nhiên từ n quan sát độc lập $(X_1, X_2, ..., X_n)$, ta cần đưa ra khoảng (θ_1, θ_2) chứa được hầu hết các giá trị θ với xác suất lớn, nghĩa là

$$\mathbb{P}(\theta_1 < \theta < \theta_2) = 1 - \alpha\,.$$

Một số khái niệm

$\diamond$ (θ_1, θ_2): được gọi là *khoảng tin cậy* của ước lượng.

$\diamond$ $\theta_1 - \theta_2 = 2\varepsilon$: được gọi là *độ dài khoảng tin cậy* của ước lượng.

$\diamond$ ε: được gọi là *độ chính xác* của ước lượng.

$\diamond$ $1 - \alpha$: được gọi là *độ tin cậy* của của ước lượng.

$\diamond$ Bài toán đi tìm khoảng tin cậy cho tham số θ với độ tin cậy $1 - \alpha$ được gọi là *bài toán ước lượng khoảng tin cậy.*

1.3.5 Khoảng tin cậy cho giá trị trung bình

Cho X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, có trung bình $\mathbb{E}X = \mu$ cần ước lượng và phương sai $\mathbb{D}X = \sigma^2$ (đã biết trước hoặc chưa biết), từ mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ ta xác định được $\overline{X}$.

a. Ước lượng hai phía

Vấn đề đặt ra ở đây là với độ tin cậy $1-\alpha$ cho trước, tìm khoảng ước lượng $(\overline{X}-\varepsilon, \overline{X}+\varepsilon)$ của μ để

$$\mathbb{P}[\overline{X}-\varepsilon < \mu < \overline{X}+\varepsilon] \geq 1-\alpha.$$

Ta chia bài toán thành 3 trường hợp để giải quyết:

Trường hợp 1. Phương sai σ^2 đã biết.

Khi đó $\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n} \simeq \mathcal{N}(0,1)$, đặt $t_{\alpha/2} = \varphi^{-1}(1-\frac{\alpha}{2})$, trong đó φ là hàm phân phối chuẩn $\mathcal{N}(0,1)$ và $t_{\alpha/2}$ là mức phân vị $\alpha/2$ cho phân phối chuẩn. Ta có

$$\begin{aligned}\mathbb{P}\big[-t_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma}\sqrt{n} < t_{\alpha/2}\big] &= \varphi(t_{\alpha/2}) - \varphi(-t_{\alpha/2}) \\ &= \varphi(t_{\alpha/2}) - (1-\varphi(t_{\alpha/2})) = 1-\alpha,\end{aligned}$$

hay

$$\mathbb{P}\Big[\overline{X} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big] = 1-\alpha.$$

Quy tắc thực hành

◇ Xác định mức phân vị $t_{\alpha/2}$

Tính giá trị $1-\frac{\alpha}{2}$, tra bảng hàm phân phối $\mathcal{N}(0,1)$ (xem bảng 4 phần phụ lục), tra từ giữa ra hai biên.

◇ Xác định khoảng ước lượng $(\overline{x}-\varepsilon, \overline{x}+\varepsilon)$ với độ chính xác của ước

lượng

$$\varepsilon = t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Chú ý 1.3.1. *Nếu như kích thước mẫu $n < 30$ cần bổ sung thêm điều kiện X tuân theo luật phân phối chuẩn, khi đó* $\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$.

Ví dụ 1.3.4. Tìm khoảng ước lượng cho giá trị trung bình với độ tin cậy 95% từ mẫu của một đám đông tuân theo luật phân phối chuẩn, $\sigma^2 = 16$. Biết mẫu đó có kích thước 16 và trung bình mẫu là 15.

Giải. $\sigma^2 = 16,\ n = 15;\ \overline{x} = 15;\ \alpha$ =0,05 tra bảng hàm phân phối chuẩn ứng với $1-\alpha/2$ =0,975 được $t_{\alpha/2}$ =1,96. Độ chính xác của ước lượng

$$\varepsilon = t_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1{,}96\frac{4}{\sqrt{16}} = 1{,}96.$$

Khoảng ước lượng cho giá trị trung bình:

$$(15 - 1{,}96 < \mu < 15 + 1{,}96) \quad \text{hay} \quad (13{,}04 < \mu < 16{,}96).$$

Trường hợp 2. Phương sai σ^2 chưa biết và $n \geqslant 30$.

Khi đó $\dfrac{\overline{X}-\mu}{S}\sqrt{n} \simeq \mathcal{N}(0,1)$, việc thiết lập tương tự như ở trường hợp 1, ta được

$$\mathbb{P}\Big[\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\Big] = 1-\alpha\,.$$

Như vậy, với một mẫu cụ thể, ta sẽ xác định được độ chính xác của ước lượng $\varepsilon = t_{\alpha/2}\dfrac{s}{\sqrt{n}}$ và khoảng ước lượng

$$(\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}).$$

Ví dụ 1.3.5. Để ước lượng khối lượng trung bình mỗi bao xi măng của nhà máy, người ta kiểm tra ngẫu nhiên 49 bao thu được khối lượng trung bình là 49,7kg và độ lệch chuẩn mẫu 0,5kg. Với độ tin cậy là 94%, hãy ước lượng khoảng khối lượng trung bình của một bao xi măng.

Giải. α =0,06, $t_{\alpha/2}$ =1,88. Độ chính xác của ước lượng

$$\varepsilon = t_{\alpha/2}\frac{s}{\sqrt{n}} = 1{,}88\frac{0{,}5}{\sqrt{49}} = 0{,}13.$$

Khoảng ước lượng cho giá trị trung bình: $(49{,}57 < \mu < 49{,}83)$.

Trường hợp 3. Phương sai σ^2 chưa biết và $n < 30$.

Nếu $X \sim \mathcal{N}(\mu, \sigma^2)$ thì $\dfrac{\overline{X} - \mu}{S}\sqrt{n} \sim t(n-1)$. Mức phân vị $\alpha/2$ cho phân phối Student với $n-1$ bậc tự do ký hiệu là $t_{(n-1,\alpha/2)}$ là giá trị thỏa mãn $\mathbb{P}(\dfrac{\overline{X} - \mu}{S}\sqrt{n} > t_{(n-1,\alpha/2)}) = \alpha/2$. Khi đó

$$\begin{aligned}
&\mathbb{P}\big[-t_{(n-1,\alpha/2)} < \frac{\overline{X} - \mu}{S}\sqrt{n} < t_{(n-1,\alpha/2)}\big] \\
&= \mathbb{P}\big[t_{(n-1,1-\alpha/2)} < \frac{\overline{X} - \mu}{S}\sqrt{n} < t_{(n-1,\alpha/2)}\big] \\
&= 1 - \alpha/2 - \alpha/2 = 1 - \alpha.
\end{aligned}$$

Quy tắc thực hành

◇ Xác định mức phân vị $t_{(n-1,\alpha/2)}$.

Tra bảng phân phối Student (xem bảng 5 phần phụ lục), $t_{(n-1,\alpha/2)}$ là giá trị trong bảng ứng với giá trị hàng là $n-1$ và cột là $\alpha/2$.

◇ Xác định khoảng ước lượng $(\overline{x} - \varepsilon, \overline{x} + \varepsilon)$ với độ chính xác của ước lượng

$$\varepsilon = t_{(n-1,\alpha/2)}\frac{s}{\sqrt{n}}.$$

Ví dụ 1.3.6. Độ chịu lực của mỗi tấm bê tông tuân theo luật phân phối chuẩn. Đo độ chịu lực của 20 tấm bê tông cùng loại người ta thu được trung bình mẫu độ chịu lực 220kg/cm^2 và độ lệch chuẩn mẫu 32,4kg/cm^2. Với độ tin cậy 90%, tìm khoảng ước lượng trung bình độ chịu lực của mỗi tấm bê tông.

Giải. Tra bảng hàm phân phối Student ta được $t_{(19;0,05)}$ =1,729. Độ chính xác của ước lượng

$$\varepsilon = t_{(n-1,\alpha/2)}\frac{s}{\sqrt{n}} \approx 12{,}5.$$

Khoảng ước lượng cho giá trị trung bình: $(187{,}5 < \mu < 212{,}5)$.

Các dạng toán phát sinh

Xuất phát từ các công thức tương ứng với từng trường hợp

$$\varepsilon = t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ; \quad \varepsilon = t_{\alpha/2}\frac{s}{\sqrt{n}}\ ; \quad \varepsilon = t_{(n-1,\,\alpha/2)}\frac{s}{\sqrt{n}}\,.$$

◇ Cho $1-\alpha$ và n. Tìm độ chính xác của ước lượng ε.

◇ Cho $1-\alpha$ và ε. Tìm kích thước mẫu n.

◇ Cho ε và n. Tìm độ tin cậy của ước lượng $1-\alpha$.

Một số trong số các vấn đề này sẽ được đề cập ở phần sau.

b. Ước lượng một phía

Vấn đề đặt ra ở đây là với độ tin cậy $1-\alpha$ cho trước, tìm khoảng ước lượng một phía:

◇ Khoảng ước lượng bên trái $(-\infty,\, \overline{X}+\varepsilon)$: $\mathbb{P}[-\infty < \mu < \overline{X}+\varepsilon] = 1-\alpha$.

◇ Khoảng ước lượng bên phải $(\overline{X}-\varepsilon,\, +\infty)$: $\mathbb{P}[\overline{X}-\varepsilon < \mu < +\infty] = 1-\alpha$.

Nhận xét 3. Khoảng tin cậy bên trái cho ta biết giá trị tối đa, khoảng tin cậy bên phải cho ta biết giá trị tối thiểu của μ với độ tin cậy $1-\alpha$.

Ta cũng chia thành 3 trường hợp, điểm khác biệt là thay thế $\alpha/2$ bởi α.

Trường hợp 1. Phương sai σ^2 đã biết.

Đặt $t_\alpha = \varphi^{-1}(1-\alpha)$, ta có

$$\mathbb{P}\big[-t_\alpha < \frac{\overline{X}-\mu}{\sigma}\sqrt{n} < +\infty\big] = 1-\varphi(-t_\alpha) = 1-\alpha,$$
$$\mathbb{P}\big[-\infty < \frac{\overline{X}-\mu}{\sigma}\sqrt{n} < t_\alpha\big] = \varphi(t_\alpha) = 1-\alpha,$$

hay $\mathbb{P}\Big[-\infty < \mu < \overline{X} + t_\alpha\frac{\sigma}{\sqrt{n}}\Big] = \mathbb{P}\big[\overline{X} - t_\alpha\frac{\sigma}{\sqrt{n}} < \mu < +\infty\big] = 1-\alpha.$

Như vậy, với một mẫu cụ thể, khoảng ước lượng bên trái và bên phải lần lượt là $(-\infty, \overline{x}+\varepsilon)$, $(\overline{x}-\varepsilon, +\infty)$ trong đó $\varepsilon = t_\alpha\dfrac{\sigma}{\sqrt{n}}$.

Trường hợp 2. Phương sai σ^2 chưa biết và $n \geqslant 30$.

Lý luận hoàn toàn tương tự, khoảng ước lượng bên trái và bên phải lần lượt là $(-\infty, \overline{x}+\varepsilon)$, $(\overline{x}-\varepsilon, +\infty)$ trong đó $\varepsilon = t_\alpha\dfrac{s}{\sqrt{n}}$.

Trường hợp 3. Phương sai σ^2 chưa biết và $n < 30$.

Khoảng ước lượng bên trái và bên phải lần lượt là $(-\infty, \overline{x}+\varepsilon)$, $(\overline{x}-\varepsilon, +\infty)$ trong đó $\varepsilon = t_{(n-1,\alpha)}\dfrac{s}{\sqrt{n}}$.

Ước lượng khoảng cho giá trị trung bình ứng với 3 trường hợp trên được mô tả qua bảng tổng hợp sau

Loại ước lượng	ε	Độ chính xác của ước lượng: ε		
		TH1	TH2	TH3
Hai phía	$(\overline{x}-\varepsilon, \overline{x}+\varepsilon)$	$t_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$	$t_{\alpha/2}\dfrac{s}{\sqrt{n}}$	$t_{(n-1,\alpha/2)}\dfrac{s}{\sqrt{n}}$
Bên trái	$(-\infty, \overline{x}+\varepsilon)$	$t_{\alpha}\dfrac{\sigma}{\sqrt{n}}$	$t_{\alpha}\dfrac{s}{\sqrt{n}}$	$t_{(n-1,\alpha)}\dfrac{s}{\sqrt{n}}$
Bên phải	$(\overline{x}-\varepsilon, +\infty)$			

Ví dụ 1.3.7. Để đánh giá về mức doanh thu hàng tháng tại các đại lý nhỏ trên một địa bàn, người ta lấy mẫu gồm 36 đại lý. Kết quả thu được như sau: doanh thu trung bình là 155,3 triệu đồng và độ lệch chuẩn mẫu là 16 triệu đồng. Với độ tin cậy 99%, ước lượng doanh thu trung bình tối đa và tối thiểu của mỗi đại lý.

Giải. $1-\alpha = 0{,}99;\ t_\alpha = 2{,}33$. Độ chính xác của ước lượng

$$\varepsilon = t_\alpha \frac{s}{\sqrt{n}} = 2{,}33\frac{16}{\sqrt{36}} \approx 6{,}21.$$

Doanh thu tối thiểu: $\overline{x}-\varepsilon = 149{,}09$;

Doanh thu tối đa: $\overline{x}+\varepsilon = 161{,}51$.

1.3.6 Khoảng tin cậy cho tỉ lệ

a. Ước lượng hai phía

Đám đông X có tỉ lệ p cần ước lượng, từ mẫu ngẫu nhiên chúng ta xác định được tỉ lệ F, vấn đề đặt ra ở đây là với độ tin cậy $1-\alpha$ cho trước, tìm khoảng ước lượng $(F-\varepsilon,\ F+\varepsilon)$ của p để

$$\mathbb{P}[F-\varepsilon < p < F+\varepsilon] = 1-\alpha.$$

Khi n đủ lớn $\dfrac{F-p}{\sqrt{F(1-F)}}\sqrt{n} \simeq \mathcal{N}(0,\,1)$, đặt $t_{\alpha/2} = \varphi^{-1}(1-\frac{\alpha}{2})$, ta có

$$\mathbb{P}\big[-t_{\alpha/2} < \frac{F-p}{\sqrt{F(1-F)}}\sqrt{n} < t_{\alpha/2}\big] = \varphi(t_{\alpha/2}) - \varphi(-t_{\alpha/2})$$
$$= \varphi(t_{\alpha/2}) - (1-\varphi(t_{\alpha/2})) = 1-\alpha,$$

hay $\mathbb{P}\Big[F - t_{\alpha/2}\sqrt{\frac{F(1-F)}{n}} < p < F + t_{\alpha/2}\sqrt{\frac{F(1-F)}{n}}\ \Big] = 1-\alpha.$

Quy tắc thực hành: khi $nf \geqslant 10$ và $n(1-f) \geqslant 10$.

$\diamond$ Xác định mức phân vị $t_{\alpha/2}$.

$\diamond$ Xác định khoảng ước lượng $(f-\varepsilon, f+\varepsilon)$ với độ chính xác của ước lượng

$$\varepsilon = t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}.$$

Ví dụ 1.3.8. Để ước lượng tỉ lệ phế phẩm của một kho hàng. Người ta kiểm tra 100 sản phẩm, phát hiện có 20 sản phẩm là phế phẩm. Với độ tin cậy 95%, hãy ước lượng khoảng tỉ lệ phế phẩm của kho hàng.

Giải. $t_{\alpha/2}$ =1,96; f =0,2; n =100. Độ chính xác của ước lượng

$$\varepsilon = t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} = 0{,}0784.$$

Khoảng ước lượng cho tỉ lệ phế phẩm: $(0{,}1216 < p < 0{,}2784)$.

b. Ước lượng một phía

Với các bước thiết lập tương tự ta thu được khoảng ước lượng của p bên trái là $p < f+\varepsilon$ và bên phải là $p > f-\varepsilon$, trong đó $\varepsilon = t_\alpha\sqrt{\frac{f(1-f)}{n}}$.

Ví dụ 1.3.9. Cho giả thiết như Ví dụ 5. Hãy ước lượng tỉ lệ phế phẩm tối đa và tối thiểu.

Giải. t_α =1,64; f =0,2; n =100. Độ chính xác của ước lượng

$$\varepsilon = t_\alpha \sqrt{\frac{f(1-f)}{n}} = 0{,}0656.$$

Tỉ lệ sản phẩm tối thiểu: $f - \varepsilon$ =0,1344;

Tỉ lệ sản phẩm tối đa: $f + \varepsilon$ =0,2656.

Ví dụ 1.3.10. Một lô hàng nhập cảng gồm 5.000 thiết bị điện tử đã qua sử dụng. Cơ quan quản lý kiểm tra ngẫu nhiên 100 thiết bị từ lô hàng thì có 82 thiết bị có thể tiếp tục sử dụng được. Với độ tin cậy 90%, lô hàng có tối thiểu bao nhiêu thiết bị có thể tiếp tục sử dụng được?

Giải. $t_\alpha = 1,28$; $f = 0,82$; $n = 100$; $N = 5.000$. Độ chính xác xủa ước lượng

$$\varepsilon = t_\alpha \sqrt{\frac{f(1-f)}{n}} = 0{,}0492.$$

Tỉ lệ sản phẩm tối thiểu: $f - \varepsilon$ =0,7708.

Vậy, số thiết bị tối thiểu có thể tiếp tục sử dụng được: $N(f - \varepsilon) =$ 4864.

Các dạng toán phát sinh

Xuất phát từ công thức

$$\varepsilon = t_{\alpha/2} \sqrt{\frac{f(1-f)}{n}}\,.$$

◇ Cho $1 - \alpha$ và n. Tìm độ chính xác của ước lượng ε.

◇ Cho $1 - \alpha$ và ε. Tìm kích thước mẫu n.

◇ Cho ε và n. Tìm độ tin cậy của ước lượng $1 - \alpha$.

1.3.7 Độ chính xác của ước lượng

Trong các nội dung trước chúng ta đã giải quyết bài toán xây dựng ước lượng khoảng cho trung bình và ước lượng khoảng cho tỉ lệ, nghĩa là từ mẫu cụ thể, độ tin cậy $1-\alpha$ ta sẽ xác định được khoảng ước lượng cho tham số θ là (θ_1, θ_2) trong đó độ chính xác của ước lượng $\varepsilon = \frac{\theta_2-\theta_1}{2}$.

Trong các trường hợp đã trình bày thì ε phụ thuộc vào kích thước mẫu n. Bây giờ ta đặt ra bài toán ngược: với độ tin cậy $1-\alpha$ đã biết, cho độ chính xác của ước lượng ε, tìm kích thước mẫu n cần thiết để nhận được ước lượng với độ chính xác đã cho. Chúng ta sẽ giải quyết bài toán này đối với trường hợp 1 của bài toán ước lượng khoảng trung bình. Các trường hợp còn lại là hoàn toàn tương tự (giành cho bạn đọc).

Trong trường hợp này, khoảng ước lượng là $(\overline{x}-\varepsilon, \overline{x}+\varepsilon)$ và công thức xác định độ chính xác của ước lượng $\varepsilon = t_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$. Kích thước mẫu điều tra cần thiết nếu độ chính xác của ước lượng ε_0 là

$$n = \left[\frac{t^2_{\alpha/2}\,\sigma^2}{\varepsilon_0^2}\right] + 1,$$

trong đó ký hiệu $[x]$ là phần nguyên của x, chẳng hạn $[20,36] = 20$.

Ví dụ 1.3.11. Với giả thiết như ở Ví dụ 1: $\sigma^2 = 16$; $1-\alpha$ =0,95. Muốn có ước lượng có độ chính xác là 1 thì phải điều tra mẫu có kích thước bao nhiêu?

Giải. Như vậy $\varepsilon_0 = 1$, khi đó

$$n = \left[\frac{t^2_{\alpha/2}\,\sigma^2}{\varepsilon_0^2}\right] + 1 = 62.$$

Ngoài ra, chúng ta còn giải quyết được bài toán ngược dạng tìm độ tin cậy của ước lượng khi biết độ chính xác của ước lượng và kích thước mẫu. Vấn đề này được đề cập trong ví dụ sau đây:

Ví dụ 1.3.12. Một mẫu thống kê có kích thước $n = 36$, có trung bình mẫu là 100 và độ lệch chuẩn mẫu là 5. Tìm độ tin cậy của ước lượng nếu khoảng ước lượng là (99; 101).

Giải. Tính mức phân vị: $t_{\frac{\alpha}{2}} = \dfrac{\varepsilon\sqrt{n}}{s} = 2$. Độ tin cậy của ước lượng

$$1 - \alpha = 2\varphi(t_{\alpha/2}) = 0{,}955.$$

1.4 Kiểm định giả thiết

1.4.1 Đặt vấn đề

Trong thực tế cuộc sống chúng ta thường gặp 2 quan điểm trái ngược nhau về một vấn đề nào đó. Chẳng hạn, các nhà sản xuất cho rằng có 95% sản phẩm của công ty đảm bảo tiêu chuẩn, trong khi đó các nhà quản lý thị trường lại cho rằng không phải như vậy thực tế thấp hơn nhiều; trước cuộc bầu cử tổng thống đảng phái A cho rằng có 65% cử tri ủng hộ UWCV của đảng phái họ, trong khi đó đảng đối lập lại cho rằng thực tế thấp hơn nhiều.

Vấn đề đặt ra là, thông qua số liệu thống kê hãy chỉ ra chấp nhận ý kiến nào trong 2 ý kiến đó với một mức ý nghĩa α cho trước.

Tổng quát: Chúng ta thường có bài toán

$$\begin{cases} H: & \text{(Giả thiết)} \quad \text{có tính chất A} \\ K: & \text{(Giả thiết)} \quad \text{không có tính chất A} \end{cases}$$

Từ số liệu thống kê hãy đưa ra kết luận cho bài toán trên.

Trong kiểm định giả thiết thường gặp 2 loại sai lầm:

- Sai lầm loại 1: Bác bỏ H trong khi H đúng

- Sai lầm loại 2: Chấp nhận H trong khi H sai.

Mục đích của các nhà thống kê là làm giảm cả 2 loại sai lầm này. Tuy vậy điều này không thể vì giảm sai lầm này thì khả năng mắc sai lầm loại kia tăng lên.

Trong thực tế thống kê người ta thấy mỗi loại sai lầm sẽ gây ra một tác hại khác nhau. Tuy vậy người ta thấy cần phải giảm sai lầm loại 1 với một xác suất xãy ra bé. Chẳng hạn như trong xã hội hiện đại người ta cho rằng "*Kết án người vô tội nguy hiểm hơn rất nhiều so với việc tha bổng một người có tội*". .. Do đó, Neyman- Pearson đã cho rằng chúng ta chỉ xét những bài toán thống kê với

$$\mathbb{P}(\text{Sai lầm loại 1}) = \mathbb{P}(\text{Bác bỏ H}|\ \text{H đúng}) \leq \alpha$$

trong đó α là một số bé và gọi là mức ý nghĩa. Thông thường $\alpha \leq 10\%$.

1.5 Kiểm định giả thiết về giá trị trung bình và về tỉ lệ

1.5.1 Kiểm định giả thiết về giá trị trung bình

Đây là một dạng bài toán kiểm định số đặc trưng $\mathbb{E}X = \mu$ của biến ngẫu nhiên gốc đám đông X (so sánh giá trị kỳ vọng của đại lượng ngẫu nhiên X với giá trị μ_0 cho trước). Có 2 dạng bài toán kiểm định giả thiết về giá trị trung bình.

a. Kiểm định hai phía

Vấn đề đặt ra ở đây là với mức ý nghĩa α và một giá trị μ_0 cho trước, đánh giá về cặp giả thiết thống kê

$$H:\ \mu = \mu_0\ ; \quad K:\ \mu \neq \mu_0.$$

Trường hợp 1. Phương sai σ^2 đã biết.

Khoảng ước lượng của μ với độ tin cậy $1-\alpha$

$$(\overline{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

Chấp nhận giả thiết H khi $\mu_0 \in [\overline{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ;\ \overline{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$ hay

$$\overline{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu_0 \leqslant \overline{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

tương đương với $\left|\dfrac{\overline{x}-\mu_0}{\sigma}\sqrt{n}\right| \leqslant t_{\alpha/2}$.

Quy tắc thực hành

$\diamond$ Từ mẫu cụ thể xác định giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}.$$

$\diamond$ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Chú ý

- Nếu như kích thước mẫu $n < 30$ thì ta cần bổ sung thêm điều kiện X tuân theo luật phân phối chuẩn.

- Như vậy miền bác bỏ $W_\alpha = (-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, +\infty)$, điều này là hợp lý. Giả sử $H:\ \mu = \mu_0$ đúng thì $T = \dfrac{\overline{X}-\mu_0}{\sigma}\sqrt{n} \simeq \mathcal{N}(0,1)$, khi đó

$$\begin{aligned}\mathbb{P}[T \in W_\alpha | H \text{ đúng}] &= \mathbb{P}\Big[\Big|\frac{\overline{X}-\mu_0}{\sigma}\sqrt{n}\Big| > t_{\alpha/2}\Big] \\ &= 1 - \big(\varphi(t_{\alpha/2}) - \varphi(-t_{\alpha/2})\big) = \alpha.\end{aligned}$$

Nghĩa là xác suất phạm sai lầm loại 1 được ấn định bởi một giá trị tương đối nhỏ α nào đó, việc chứng minh xác suất phạm sai lầm loại 2 cực tiểu bạn đọc tham khảo tài liệu [5].

Ví dụ 1.5.1. Một máy tiện tự động cho ra những trục máy có đường kính là 120mm và độ lệch chuẩn cho phép là 3mm. Kiểm tra ngẫu nhiên 50 trục máy, kết quả thu được đường kính trung bình là 119,2mm. Với mức ý nghĩa là 10%, máy tiện trên có hoạt động bình thường không?

Giải. Máy tiện được gọi là hoạt động bình thường khi nó sản xuất những trục máy với sai số không vượt quá mức cho phép. Cặp giả thiết thống kê

$$H:\ \mu = \mu_0 = 120\ ; \quad K:\ \mu \neq \mu_0.$$

$\mu_0 = 120;\ \sigma = 3;\ \alpha = 0{,}1;\ t_{\alpha/2} = 1{,}64;\ n = 50;\ \overline{x} = 119{,}2$. Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{119{,}2 - 120}{3}\sqrt{50} \approx -1{,}89,$$

Vì $|t_{tn}| > t_{\alpha/2}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận khẳng định cho rằng máy tiện trên hoạt động không bình thường.

Trường hợp 2. Phương sai σ^2 chưa biết và $n \geqslant 30$.

Tương tự như ở trường hợp 1, đặt $t_{tn} = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$, khi đó

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Trường hợp 3. Phương sai σ^2 chưa biết và $n < 30$.

Giả sử X tuân theo luật phân phối chuẩn $\mathcal{N}(0,1)$. Đặt $t_{tn} = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$, khi đó

$|t_{tn}| \leqslant t_{(n-1,\alpha/2)}$: chấp nhận H.

$|t_{tn}| > t_{(n-1,\alpha/2)}$: bác bỏ H, chấp nhận K.

Ví dụ 1.5.2. Thể tích sơn chứa trong mỗi thùng sơn nước nhãn hiệu A là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn với trung bình

18 lít. Kiểm tra ngẫu nhiên 25 thùng thu được kết quả: thể tích trung bình là 17,92 lít và độ lệch chuẩn mẫu là 0,24 lít. Với mức ý nghĩa 5%, thể tích sơn trong các thùng sơn có đúng tiêu chuẩn không?

Giải. Cặp giả thiết thống kê

$$H: \ \mu = \mu_0 = 18 \ ; \quad K: \ \mu \neq \mu_0.$$

$\mu_0 = 18; \ s = 0{,}24; \ \alpha = 0,05; \ t_{(n-1,\alpha/2)} = 2{,}11; \ n = 25; \ \overline{x} = 17{,}92.$

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{17{,}92 - 18}{0{,}24}\sqrt{25} \approx -1{,}67,$$

Vì $|t_{tn}| \leqslant t_{\alpha/2}$ nên ta chấp nhận H. Nghĩa là chấp nhận khẳng định cho rằng thể tích sơn trong các thùng sơn đúng tiêu chuẩn.

b. Kiểm định một phía

Trong thực tế xuất hiện một số dạng toán về kiểm định như:

- Sau chiến dịch quảng cáo, doanh số bán ra một loại hàng có tăng lên hay không? (kiểm định lớn hơn)

- Kiểm tra xem khối lượng đóng gói các bao gạo của một kho có nhỏ hơn giá trị in trên bao bì hay không? (kiểm định nhỏ hơn)

Các dạng bài toán này được gọi là *bài toán kiểm định một phía.*

◇ Kiểm định lớn hơn: $H: \ \mu = \mu_0 \ ; \quad K: \ \mu > \mu_0.$

◇ Kiểm định nhỏ hơn: $H: \ \mu = \mu_0 \ ; \quad K: \ \mu < \mu_0.$

Giải quyết bài toán kiểm định một phía được phân chia các trường hợp giống như trong bài toán kiểm định hai phía. Tiêu chuẩn kiểm định ứng với 3 trường hợp của bài toán kiểm định giá trị trung bình được mô tả qua bảng tổng hợp sau đây

Trường hợp	t_{tn}	Điều kiện chấp nhận $H:\ \mu = \mu_0$		
		$K:\ \mu = \mu_0$	$K:\ \mu > \mu_0$	$K:\ \mu < \mu_0$
TH1	$\dfrac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$	$\lvert t_{tn}\rvert \leqslant t_{\alpha/2}$	$t_{tn} \leqslant t_{\alpha}$	$t_{tn} \geqslant -t_{\alpha}$
TH2	$\dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$	$\lvert t_{tn}\rvert \leqslant t_{\alpha/2}$	$t_{tn} \leqslant t_{\alpha}$	$t_{tn} \geqslant -t_{\alpha}$
TH3	$\dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$	$\lvert t_{tn}\rvert \leqslant t_{(n-1,\alpha/2)}$	$t_{tn} \leqslant t_{(n-1,\alpha)}$	$t_{tn} \geqslant -t_{(n-1,\alpha)}$

Ví dụ 1.5.3. Một nhà máy cung cấp nước sạch cho rằng khối lượng trung bình của một loại chất độc hại trong một lít nước của nhà máy là 14mg. Người ta nghi ngờ số liệu này thấp hơn thực tế. Kiểm tra ngẫu nhiên với 64 mẫu nước thu được kết quả: $\bar{x} = 14{,}2$ và $s = 0{,}24$. Hãy cho kết luận về nghi ngờ nói trên với mức ý nghĩa 8%.

Giải. Cặp giả thiết thống kê: $H:\ \mu = \mu_0 = 120$; $K:\ \mu > \mu_0$.

$\mu_0 = 14;\ s = 0{,}24;\ \alpha = 0{,}08;\ t_\alpha = 1{,}4;\ n = 64;\ \overline{x} = 14{,}2.$

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{14{,}2 - 14}{0{,}24}\sqrt{64} \approx 6{,}67,$$

Vì $t_{tn} > t_\alpha$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận nghi ngờ trên.

1.5.2 Kiểm định giả thiết về tỉ lệ

Đây là dạng bài so sánh giá trị tỉ lệ p của đám đông X với giá trị p_0 cho trước. Có hai dạng bài toán kiểm định giả thiết về tỉ lệ.

a. Kiểm định hai phía

Vấn đề đặt ra ở đây là với mức ý nghĩa α và một giá trị p_0 cho trước, đánh giá về cặp giả thiết thống kê

$$H:\ p = p_0\ ;\quad K:\ p \neq p_0.$$

Với n đủ lớn và $H:\ p = p_0$ đúng thì $T = \dfrac{F - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \simeq \mathcal{N}(0,1)$, khi đó

$$\mathbb{P}[T \in W_\alpha | H \text{ đúng}] = \mathbb{P}\Big[\Big|\frac{F - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}\Big| > t_{\alpha/2}\Big] = \alpha,$$

trong đó miền bác bỏ $W_\alpha = (-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, +\infty)$.

Quy tắc thực hành: Khi $np_0 \geqslant 5;\ n(1-p_0) \geqslant 5$.

$\diamond$ Xác định giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}.$$

$\diamond$ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Ví dụ 1.5.4. Một hãng sản xuất đĩa cứng công bố rằng: có 10% đĩa cứng của hãng phải bảo hành trong thời gian 2 năm đầu sử dụng. Người ta điều tra ngẫu nhiên 200 khách hàng đã sử dụng đĩa cứng của hãng thì có 29 đĩa cứng phải bảo hành trong thời gian 2 năm đầu sử dụng. Với mức ý nghĩa 5%, tỉ lệ trong công bố trên có đúng với thực tế không?

Giải. Cặp giả thiết thống kê: $H:\ p = p_0 = 0{,}1\ ;\quad K:\ p \neq p_0$.

$n = 200;\ f = 0{,}145;\ p_0 = 0{,}1;\ \alpha = 0,05;\ t_{\alpha/2} = 1{,}96$.

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{0,145 - 0,1}{\sqrt{0,1 \times 0,9}}\sqrt{200} \approx 2{,}12.$$

Vì $|t_{tn}| > t_{\alpha/2}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận khẳng định cho rằng tỉ lệ trong công bố trên không đúng với thực tế.

b. Kiểm định một phía

Tương tự như bài toán kiểm định về giá trị trung bình, bài toán kiểm định tỉ lệ cũng có hai dạng kiểm định một phía như sau:

◇ Kiểm định lớn hơn: $H: \ p = p_0$; $\quad K: \ p > p_0$.

◇ Kiểm định nhỏ hơn: $H: \ p = p_0$; $\quad K: \ p < p_0$.

Bảng dưới đây sẽ trình bày các tiêu chuẩn kiểm định của bài toán kiểm định tỉ lệ

t_{tn}	Điều kiện chấp nhận $H: \ p = p_0$		
	$K: \ p = p_0$	$K: \ p > p_0$	$K: \ p < p_0$
$\dfrac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$\lvert t_{tn}\rvert \leqslant t_{\alpha/2}$	$t_{tn} \leqslant t_{\alpha}$	$t_{tn} \geqslant -t_{\alpha}$

Ví dụ 1.5.5. Một trung tâm đào tạo nghề báo cáo rằng tỷ lệ người học tại trung tâm kiếm được việc làm ngay sau khi tốt nghiệp là 70%. Một mẫu ngẫu nhiên gồm 200 người đã tốt nghiệp ở trung tâm cho thấy có 130 người kiếm được việc làm ngay sau khi tốt nghiệp. Với mức ý nghĩa 5%, kiểm định xem phải chăng tỉ lệ trong báo cáo của trung tâm là cao hơn thực tế.

Giải. Cặp giả thiết thống kê: $H: \ p = p_0 = 0{,}7$; $\quad K: \ p < p_0$.

$n = 200$; $f = 0{,}65$; $p_0 = 0{,}7$; $\alpha = 0,05$; $t_{\alpha} = 1{,}64$.

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{0,65 - 0,7}{\sqrt{0,7 \times 0,3}}\sqrt{200} \approx -1{,}54.$$

Vì $t_{tn} > -t_{\alpha}$ nên ta chấp nhận H. Nghĩa là chấp nhận ý kiến cho rằng tỉ lệ trong báo cáo của trung tâm là đúng thực tế.

1.5.3 Boài toán so sánh

Giả sử chúng ta có hai đám đông $\mathcal{C}_1$ và $\mathcal{C}_2$ có chung một đặc điểm cần nghiên cứu nào đó; hai đại lượng ngẫu nhiên gốc đám đông tương ứng lần lượt là X_1 và X_2. Trong mục này chúng ta đề cập đến dạng bài toán so sánh hai giá trị đặc trưng của hai đại lượng ngẫu nhiên này.

So sánh hai giá trị trung bình

Hai đám đông $\mathcal{C}_1$ và $\mathcal{C}_2$ có hai giá trị trung bình là $\mathbb{E}X_1 = \mu_1$ và $\mathbb{E}X_2 = \mu_2$ cần so sánh. Vấn đề đặt ra ở đây là với mức ý nghĩa α cho trước, đánh giá về cặp giả thiết thống kê

$$H: \ \mu_1 = \mu_2 \ ; \quad K: \ \mu_1 \neq \mu_2.$$

Giả sử $\mathbb{D}X_1 = \sigma_1^2$, $\mathbb{D}X_2 = \sigma_2^2$. Từ hai mẫu cụ thể $(x_1, x_2, ..., x_{n_1})$ của đám đông $\mathcal{C}_1$ và $(y_1, y_2, ..., y_{n_2})$ của đám đông $\mathcal{C}_2$ chúng ta xác định được trung bình mẫu và phương sai hiệu chỉnh mẫu lần lượt là $\overline{x}_1, \overline{x}_2, s_1^2, s_2^2$.

Quy tắc thực hành

Trường hợp 1. σ_1^2, σ_2^2 đã biết.

◇ Xác định giá trị kiểm định từ thực nghiệm

$$t_{tn} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}.$$

◇ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Chú ý. Nếu như kích thước mẫu $n_1 < 30$ hoặc $n_2 < 30$ thì ta cần bổ sung thêm điều kiện X_1, X_2 tuân theo luật phân phối chuẩn.

Ví dụ 1.5.6. Người ta muốn so sánh tuổi thọ của hai loại thiết bị điện tử (trong điều kiện hoạt động liên tục) được sản xuất bởi hai công nghệ khác nhau. Biết rằng độ lệch chuẩn tuổi thọ của thiết bị được sản xuất từ công nghệ thứ nhất và công nghệ thứ hai tương ứng là 120 giờ và 125 giờ. Thử nghiệm 50 thiết bị cho mỗi công nghệ trên thu được tuổi thọ trung bình của chúng tương ứng là 264 giờ và 245 giờ. Với mức ý nghĩa 5%, tuổi thọ của hai loại thiết bị điện tử được sản xuất từ hai công nghệ trên có khác nhau không?

Giải. Cặp giả thiết thống kê: $H: \mu_1 = \mu_2$; $K: \mu_1 \neq \mu_2$.

$\sigma_1 = 120;\ \sigma_2 = 125;\ n_1 = n_2 = 50;\ \overline{x}_1 = 264;\ \overline{x}_2 = 245;$

$\alpha = 0{,}05;\ t_{\alpha/2} = 1{,}96.$

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{264 - 245}{\sqrt{\dfrac{120^2}{50} + \dfrac{125^2}{50}}} \approx 0{,}78.$$

Vì $|t_{tn}| \leqslant t_{\alpha/2}$ nên ta chấp nhận H. Nghĩa là chấp nhận khẳng định rằng tuổi thọ của hai loại thiết bị điện tử được sản xuất từ hai công nghệ trên là giống nhau.

Trường hợp 2. σ_1^2, σ_2^2 chưa biết và $n_1 \geqslant 30, n_2 \geqslant 30$.

$\diamond$ Xác định giá trị kiểm định từ thực nghiệm

$$t_{tn} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}.$$

◇ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Trường hợp 3. X_1, X_2 có phân phối chuẩn, $\sigma_1^2 = \sigma_2^2$ chưa biết và $n_1 < 30, n_2 < 30$.

◇ Xác định giá trị kiểm định từ thực nghiệm

$$t_{tn} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s^2\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}}, \quad \text{trong đó} \quad s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\,.$$

◇ So sánh giá trị của t_{tn} với mức phân vị $t_{(n_1+n_2-2,\alpha/2)}$, nếu

$|t_{tn}| \leqslant t_{(n_1+n_2-2,\alpha/2)}$: chấp nhận H.

$|t_{tn}| > t_{(n_1+n_2-2,\alpha/2)}$: bác bỏ H, chấp nhận K.

Ví dụ 1.5.7. Hai máy tự động dùng cắt những thanh kim loại với cùng một yêu cầu. Từ máy thứ nhất lấy ra 12 sản phẩm thu được chiều dài trung bình là 55cm và độ lệch chuẩn mẫu là 0,3cm, từ máy thứ 2 lấy ra 18 sản phẩm có các kết quả tương ứng là : 55,2cm và 0,2cm. Với mức ý nghĩa là 0,1, đánh giá về nhận định: hai máy đó sản xuất ra các thiết bị cùng kích cỡ. Giả sử rằng kích cỡ sản phẩm từ 2 máy có phân phối chuẩn và có cùng phương sai.

Giải. Cặp giả thiết thống kê: $H:\ \mu_1 = \mu_2\ ; \quad K:\ \mu_1 \neq \mu_2$.

$s_1 = 0{,}3\text{cm};\ s_2 = 0{,}2;\ n_1 = 12;\ n_2 = 18;\ \overline{x_1} = 55\text{cm};\ \overline{x_2} = 55{,}2;$

$\alpha = 0{,}1;\ t_{(28;0{,}05)} = 1,701.$

Giá trị kiểm định thực nghiệm

$$s^2 = \frac{11 \times 0,3^2 + 17 \times 0,2^2}{28} \approx 0,06;$$
$$t_{tn} = \frac{55 - 55,2}{\sqrt{0,06\left(\frac{1}{12} + \frac{1}{18}\right)}} \approx -2,2.$$

Vì $|t_{tn}| > t_{(28;0,05)}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận nhận định cho rằng hai máy đó sản xuất ra các thiết bị không cùng kích cỡ.

Dối với bài toán so sánh 2 giá trị trung bình, có hai dạng bài toán kiểm định một phía như sau:

◇ Kiểm định lớn hơn: $H: \mu_1 = \mu_2$; $K: \mu_1 > \mu_2$.

◇ Kiểm định nhỏ hơn: $H: \mu_1 = \mu_2$; $K: \mu_1 < \mu_2$.

Giải quyết bài toán kiểm định một phía được phân chia các trường hợp giống như trong bài toán kiểm định hai phía. Tiêu chuẩn kiểm định ứng với 3 trường hợp được mô tả qua bảng tổng hợp sau:

Trường hợp	t_{tn}	Điều kiện chấp nhận $H:\ \mu_1 = \mu_2$	
		$K:\ \mu_1 > \mu_2$	$K:\ \mu_1 < \mu_2$
TH1	$\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$t_{tn} \leqslant t_\alpha$	$t_{tn} \geqslant -t_\alpha$
TH2	$\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	$t_{tn} \leqslant t_\alpha$	$t_{tn} \geqslant -t_\alpha$
TH3	$\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{s^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	$t_{tn} \leqslant t_{(n_1+n_2-2,\alpha)}$	$t_{tn} \geqslant -t_{(n_1+n_2-2,\alpha)}$

Ví dụ 1.5.8. Với giả thiết như ở Ví dụ 2: $s_1 = 0{,}3$cm; $s_2 = 0{,}2$; $n_1 = 12$; $n_2 = 18$; $\overline{x_1} = 55$cm; $\overline{x_2} = 55{,}2$. Đánh giá nhận định: máy thứ hai sản xuất ra thiết bị có kích cỡ lớn hơn máy thứ nhất.

Giải. Cặp giả thiết thống kê: $H:\ \mu_1 = \mu_2$; $K:\ \mu_1 < \mu_2$.

$\alpha = 0{,}1$; $t_{(28;0,1)}$ =1,313. Giá trị kiểm định thực nghiệm

$$s^2 \approx 0,06;\ t_{tn} \approx -2,2.$$

Vì $t_{tn} < -t_{(28;0,1)}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận nhận định cho rằng máy thứ hai sản xuất ra thiết bị có kích cỡ lớn hơn máy thứ nhất.

So sánh hai tỉ lệ

Hai đám đông $\mathcal{C}_1$ và $\mathcal{C}_2$ có hai tỉ lệ p_1 và p_2 cần so sánh. Vấn đề đặt ra ở đây là với mức ý nghĩa α cho trước, đánh giá về cặp giả thiết thống

kê

$$H:\ p_1 = p_2\ ;\quad K:\ p_1 \neq p_2.$$

Từ mẫu cụ thể kích thước n_1 của đám đông $\mathcal{C}_1$ ta xác định được k_1 phần tử có đặc điểm cần nghiên cứu, do đó tỉ lệ mẫu là $f_1 = k_1/n_1$; tương tự đối với mẫu kích thước n_2 của đám đông $\mathcal{C}_2$ ta xác định được k_2 và $f_2 = k_2/n_2$.

Quy tắc thực hành: Khi n_1, n_2 đủ lớn.

◇ Xác định giá trị kiểm định từ thực nghiệm

$$t_{tn} = \frac{f_1 - f_2}{\sqrt{f(1-f)\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}},\quad \text{trong đó}\quad f = \frac{k_1 + k_2}{n_1 + n_2}.$$

◇ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Nhận xét 4. Khi kích thước mẫu điều tra càng lớn thì kết quả kiểm định càng chính xác, ở mức độ tương đối khái niệm n_1, n_2 đủ lớn ở đây được hiểu là thỏa mãn hai điều kiện: $(n_1 + n_2)f \geqslant 10,\ (n_1 + n_2)(1 - f) \geqslant 10$.

Ví dụ 1.5.9. Người ta kiểm tra ngẫu nhiên 400 sản phẩm từ dây chuyền thứ nhất thì có 24 phế phẩm, kiểm tra 800 sản phẩm từ dây chuyền thứ hai thấy có 42 phế phẩm. Với mức ý nghĩa $\alpha = 0{,}05$, tỉ lệ phế phẩm của 2 dây chuyền trên có như nhau hay không ?

Giải. Cặp giả thiết thống kê: $H:\ p_1 = p_2$; $K:\ p_1 \neq p_2$.

$t_{\alpha/2} = 1,96;\ f_1 = 0{,}06;\ f_2 = 0{,}0525;\ f = 0{,}055.$

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{0,06 - 0,0525}{\sqrt{0,055 \times 0,945\big(1/400 + 1/800\big)}} \approx 0,537.$$

Vì $|t_{tn}| < t_{\alpha/2}$ nên ta chấp nhận H. Nghĩa là chấp nhận khẳng định cho rằng tỉ lệ phế phẩm của 2 dây chuyền trên là như nhau.

Tương tự như bài toán kiểm định về giá trị trung bình, bài toán kiểm định tỉ lệ cũng có hai dạng kiểm định một phía như sau:

◇ Kiểm định lớn hơn: $H: \ p_1 = p_2 \ ; \quad K: \ p_1 > p_2$.

◇ Kiểm định nhỏ hơn: $H: \ p_1 = p_2 \ ; \quad K: \ p_1 < p_2$.

Bảng dưới đây sẽ trình bày các tiêu chuẩn kiểm định của bài toán kiểm định tỉ lệ:

t_{tn}	Điều kiện chấp nhận $H: \ p_1 = p_2$		
	$K: \ p_1 \neq p_2$	$K: \ p_1 > p_2$	$K: \ p_1 < p_2$
$\dfrac{f_1 - f_2}{\sqrt{f(1-f)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	$\lvert t_{tn}\rvert \leqslant t_{\alpha/2}$	$t_{tn} \leqslant t_{\alpha}$	$t_{tn} \geqslant -t_{\alpha}$

Ví dụ 1.5.10. Dùng thuốc A cho 200 bệnh nhân thì có 160 người khỏi bệnh. Dùng thuốc B cho 300 bệnh nhân thì có 210 người khỏi bệnh. Với mức ý nghĩa $\alpha = 0{,}04$, hiệu quả của thuốc A có cao hơn thuốc B hay không?

Giải. Cặp giả thiết thống kê: $H: \ p_1 = p_2$; $K: \ p_1 > p_2$.

$t_{\alpha} = 1{,}75$; $f_1 = 0{,}8$; $f_2 = 0{,}7$; $f = 0{,}74$.

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{(0,8 - 0,7)}{\sqrt{0,74 \times 0,26\big(1/200 + 1/300\big)}} \approx 2,497.$$

Vì $t_{tn} > t_{\alpha}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận khẳng định cho rằng hiệu quả của thuốc A cao hơn thuốc B.

1.6 Hồi quy và tương quan

1.6.1 Mở đầu

Trên cùng một đám đông $\mathcal{C}$ có hai đặc điểm định lượng cần nghiên cứu, hai đại lượng ngẫu nhiên gốc đám đông tương ứng lần lượt là X và Y. Bài toán đặt ra ở đây là tìm hiểu mức độ phụ thuộc giữa hai đại lượng ngẫu nhiên và tìm biểu thức biểu diễn sự liên hệ giữa chúng.

Đây là một vấn đề hoàn toàn thực tế, sự phụ thuộc của hai đại lượng ngẫu nhiên X và Y có thể phân thành ba loại:

$\diamond$ Sự phụ thuộc hàm số: tồn tại hàm φ để $Y = \varphi(X)$.

$\diamond$ Sự phụ thuộc thống kê: khi X thay đổi thì phân phối xác suất của Y cũng thay đổi.

$\diamond$ Sự phụ thuộc tương quan: X thay đổi thì kỳ vọng có điều kiện $\mathbb{E}(Y|X)$ cũng thay đổi, nghĩa là $\mathbb{E}(Y|X) = \varphi(X) \neq$ hằng số.

Nếu $\varphi(X) = AX + B$ thì ta nói X và Y có *tương quan tuyến tính*, trong trường hợp ngược lại thì ta nói X và Y có *tương quan phi tuyến.*

Phụ thuộc tương quan là trường hợp riêng của phụ thuộc thống kê, nghĩa là nếu phụ thuộc tương quan thì có sự phụ thuộc về phân phối xác suất. Khi phân tích độ phụ thuộc tương quan giữa hai đại lượng ngẫu nhiên X và Y thì ta không cần xét đến trường hợp nó độc lập với nhau.

1.6.2 Hệ số tương quan mẫu

Chúng ta đã được làm quen với khái niệm hệ số tương quan giữa hai đại lượng ngẫu nhiên X và Y

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\mathbb{D}X\,\mathbb{D}Y}} = \frac{\mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y}{\sqrt{\mathbb{D}X\,\mathbb{D}Y}}\,.$$

Đó là số đo mức độ phụ thuộc tuyến tính giữa hai đại lượng ngẫu nhiên X và Y, nhưng nếu chưa biết được phân phối xác suất thì hệ số tương quan lý thuyết $\rho(X, Y)$ chưa xác định được. Do đó ta tìm cách ước lượng $\rho(X, Y)$ bởi một giá trị thu được từ mẫu quan sát, giá trị đó được gọi là *hệ số tương quan mẫu*.

Giả sử ta có n cặp quan sát $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ của (X, Y), *hệ số tương quan mẫu* được tính theo công thức

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Do vậy

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \Big(\sum_{i=1}^{n} y_i\Big)^2}} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\hat{s}_X\,\hat{s}_Y},$$

trong đó $\overline{xy} = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i$.

Tương tự như hệ số tương quan, hệ số tương quan mẫu cũng có tính chất $|r| \leqslant 1$. Biểu diễn các cặp (x_i, y_i) của mẫu lên một mặt phẳng tọa độ tạo thành đám mây điểm. Hình ảnh của đám mây điểm thể hiện mối quan hệ giữa X và Y. Nếu đám mây điểm có xu hướng tập trung quanh một đường thẳng nào đó (có hệ số góc khác 0) thì $|r|$ càng gần 1 và ta có thể kết luận X, Y có quan hệ gần với quan hệ tuyến tính (tương quan tuyến tính), còn nếu nó phân tán thành hình tròn hay hình vuông thì $|r|$ gần bằng 0.

Ví dụ 1.6.1. Bảng số liệu sau đây là kết quả thu thập từ một công ty về doanh thu (X) và số tiền dành cho quảng cáo (Y) của một số tháng như sau:

X (tỉ đồng)	5	7	8	11	9
Y (triệu đồng)	45	60	75	90	80

Hãy xác định hệ số tương quan mẫu.

Giải. Bảng tính

x_i	y_i	x_iy_i	x_i^2	y_i^2
5	45	225	25	2025
7	60	420	49	3600
8	75	600	64	5625
11	90	990	121	8100
9	80	720	81	6400
40	**350**	**2955**	**340**	**25750**

Hệ số tương quan mẫu

$$r = \frac{5 \cdot 2955 - 40 \cdot 350}{\sqrt{5 \cdot 340 - 40^2}\ \sqrt{5 \cdot 25750 - 350^2}} \approx 0{,}98.$$

Nhận xét 5. Trường hợp số liệu thu thập có kích thước lớn, dạng bảng có tần số chúng ta cũng lập bảng tính trung gian như trên sau đó sử dụng công thức: $r = \dfrac{\overline{xy} - \overline{x}\ \overline{y}}{\hat{s}_X\ \hat{s}_Y}$

1.6.3 Phương trình hồi quy thực nghiệm

Phương trình hồi quy

Mệnh đề 1.6.1. *Trong tất cả các hàm $h(X)$ dùng để ước lượng Y thì $\varphi(X) = \mathbb{E}(Y|X)$ là hàm có sai số bình phương trung bình nhỏ nhất.*

Nghĩa là

$$\mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)^2 \leqslant \mathbb{E}\big(Y - h(X)\big)^2.$$

Chứng minh.

$$\begin{aligned}\mathbb{E}\big(Y - h(X)\big) &= \mathbb{E}\big(Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - h(X)\big)^2 \\ &= \mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)^2 + \mathbb{E}\big(\mathbb{E}(Y|X) - h(X)\big)^2 + \\ &\quad 2\mathbb{E}\Big[\big(Y - \mathbb{E}(Y|X)\big)\big(\mathbb{E}(Y|X) - h(X)\big)\Big].\end{aligned}$$

Với mọi hàm $k(X)$ ta luôn có

$$\begin{aligned}\mathbb{E}\big(k(X)\,E(Y|X)\big) &= \int \Big[k(x)\int y\,p(y|x)\,dy\Big]p_X(x)\,dx \\ &= \int\int k(x)\,y\,p(y|x)p_X(x)\,dx\,dy \\ &= \int\int k(x)\,y\,p(x,\,y)\,dx\,dy = \mathbb{E}\big(k(X)\,Y\big).\end{aligned}$$

Đặt $k(X) = \mathbb{E}(Y|X) - h(X)$, suy ra

$$\mathbb{E}\Big[\big(Y - \mathbb{E}(Y|X)\big)\big(\mathbb{E}(Y|X) - h(X)\big)\Big] = \mathbb{E}\Big[\big(Y - \mathbb{E}(Y|X)\big)\,k(X)\Big]$$
$$= \mathbb{E}\big[k(X)\,Y\big] - \mathbb{E}\big[k(X)\,\mathbb{E}(Y|X)\big] = 0.$$

Do đó

$$\begin{aligned}\mathbb{E}\big(Y - h(X)\big) &= \mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)^2 + \mathbb{E}\big(\mathbb{E}(Y|X) - h(X)\big)^2 \\ &\geqslant \mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)^2.\end{aligned}$$

□

Như vậy $\mathbb{E}(Y|X)$ là hàm ước lượng Y có sai số bình phương trung bình nhỏ nhất. Phương trình $\varphi(X) = \mathbb{E}(Y|X)$ được gọi là *phương trình hồi quy* của Y theo X.

1.6.4 Hệ số hồi quy tuyến tính thực nghiệm

Giả sử X là đại lượng ngẫu nhiên độc lập còn Y là đại lượng ngẫu nhiên phụ thuộc và giữa chúng có tương quan tuyến tính

$$\mathbb{E}(Y|X) = AX + B, \quad A \neq 0,$$

trong đó A, B chưa biết và được gọi là *hệ số hồi quy lý thuyết.*

Bài toán. Căn cứ vào n cặp quan sát (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) của (X, Y), ta cần đi tìm một phương trình $y = ax+b$ ước lượng cho phương trình hồi quy tuyến tính lý thuyết $\mathbb{E}(Y|X) = AX + B$.

Phương trình $y = ax + b$ được gọi là *phương trình hồi quy tuyến tính thực nghiệm*; a và b được gọi là *hệ số hồi quy tuyến tính thực nghiệm* của Y theo X. Chúng ta sử dụng phương pháp bình phương bé nhất để xác định giá trị của a và b.

Như vậy, giữa giá trị thực nghiệm và giá trị xác định từ phương trình hồi quy tuyến tính thực nghiệm tại x_i có sai số $|y_i-(ax_i+b)|$. Tiêu chuẩn để xác định phương trình hồi quy tuyến tính thực nghiệm $y = ax + b$ là đảm bảo được yêu cầu

$$F(a, b) = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2 \Rightarrow \min.$$

Tìm cực tiểu của $F(a, b)$ dẫn đến hệ phương trình

$$\begin{cases} \dfrac{\partial F(a,\, b)}{\partial a} = -2\displaystyle\sum_{i=1}^{n}(y_i - ax_i - b)\, x_i = 0; \\ \dfrac{\partial F(a,\, b)}{\partial b} = -2\displaystyle\sum_{i=1}^{n}(y_i - ax_i - b) = 0, \end{cases}$$

tương đương với hệ

$$\begin{cases} \Big(\sum\limits_{i=1}^{n} x_i^2\Big)a + \Big(\sum\limits_{i=1}^{n} x_i\Big)b = \sum\limits_{i=1}^{n} x_i\, y_i; \\ \Big(\sum\limits_{i=1}^{n} x_i\Big)a + nb = \sum\limits_{i=1}^{n} y_i. \end{cases}$$

Giải hệ phương trình bậc nhất đối với a và b, ta được

$$\begin{cases} a = \dfrac{n\sum\limits_{i=1}^{n} x_i\, y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \Big(\sum\limits_{i=1}^{n} x_i\Big)^2}; \\ b = \dfrac{\sum\limits_{i=1}^{n} y_i - a\sum\limits_{i=1}^{n} x_i}{n}. \end{cases}$$

Ngoài ra, hệ số hồi quy tuyến tính thực nghiệm còn có thể xác định nhờ công thức tương đương

$$\begin{cases} a = \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{\hat{s}_X^2}; \\ b = \overline{y} - a\,\overline{x}\,. \end{cases}$$

Ví dụ 1.6.2. Với giả thiết như ở Ví dụ 1.6.1:

$$n = 5;\ \sum x_i = 40;\ \sum y_i = 350;\ \sum x_i^2 = 340;\ \sum x_i y_i = 2955.$$

a. Tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x.

b. Nếu doanh thu của một tháng nào đó là 10 tỉ đồng, hãy dự đoán chi phí quảng cáo của công ty tháng đó là bao nhiêu.

Giải. a. Hệ số hồi quy tuyến tính thực nghiệm

$$a = \frac{5 \times 2955 - 40 \times 350}{5 \times 340 - (40)^2} = 7,75; \quad b = \frac{350 - 7,75 \times 40}{5} = 8.$$

Phương trình hồi quy tuyến tính thực nghiệm: $y = 7,75\,x + 8$.

b. $x = 10$ suy ra $y = 85{,}5$. Vậy chi phí quảng cáo của tháng đó khoảng 85,5 triệu đồng.

BÀI TẬP

1. Cho ví dụ về đám đông, một số đặc điểm có thể nghiên cứu và các phương pháp thực hiện việc lấy mẫu trên đám đông đó.

2. Phân biệt sự khác nhau giữa mẫu ngẫu nhiên và mẫu cụ thể, cho ví dụ minh họa.

3. Phân biệt sự khác nhau giữa đặc điểm định lượng và đặc điểm định tính. Cho ví dụ về hai đặc điểm cùng nghiên cứu trên một đám đông.

4. Khi đo độ dài của 36 chi tiết được lấy ngẫu nhiên từ một loại sản phẩm, người ta thu được bảng số liệu sau đây:

   15 14 16 14 15 12 13 16 13 12 15 13 16 13 15

   13 16 13 16 13 15 12 15 15 14 14 15 15 16 15

   a. Lập bảng tần số và bảng tần suất.

   b. Vẽ biểu đồ đa giác tần số và tần suất.

   c. Tìm hàm phân phối mẫu.

5. Dưới đây là số liệu được lấy ngẫu nhiên về thời gian đợi của các khách hành (tính bằng giây) tại quầy thanh toán tiền ở một siêu thị đối với 48 khách hàng:

3	24	34	5	14	22	3	19	13	32	19	4	24	30	48	24
14	16	3	4	5	14	19	41	43	16	48	4	58	13	10	60
12	14	14	22	3	16	14	4	34	32	4	19	12	24	13	26

a. Lập bảng tần số ghép lớp và bảng tần suất ghép lớp.

b. Vẽ bảng tổ chức đồ tần số và tần suất.

c. Tính trung bình mẫu, phương sai mẫu và phương sai hiệu chỉnh mẫu.

6. Mẫu điều tra có kích thước 35 đối với hai đặc điểm X và Y của một loại sản phẩm được kết quả cho bởi bảng số liệu dưới đây:

$X\backslash Y$	64	65	66
6-10	3	8	3
10-14	0	5	2
14-16	6	1	0
16-20	0	3	4

a. Lập bảng tần số, tần suất của Y.

b. Những sản phẩm được gọi là đạt chất lượng nếu $X \leqslant 16$ và $Y > 64$. Tính tỉ lệ sản phẩm đạt chất lượng.

c. Lập bảng tần số và tính trung bình mẫu của chỉ tiêu Y đối với các sản phẩm có $X > 10$.

7. Cơ quan quản lý thị trường lấy số liệu về giá thành bán lẻ của một loại sản phẩm tại 40 đại lý (đơn vị: ngàn), người ta thu được bảng tần số như sau:

x_i	19	20	21	22
n_i	8	16	6	10

a. Tìm hàm phân phối mẫu.

b. Tính trung bình mẫu và độ lệch chuẩn mẫu.

8. Tìm hàm phân phối mẫu, trung bình mẫu, phương sai hiệu chỉnh mẫu đối với hai mẫu cụ thể sau:

a.

x_i	19,2	19,8	20,1	20,3	20,7
n_i	6	2	4	2	6

b.

x_i	460	480	490	505
n_i	5	6	10	4

9. Điều tra ngẫu nhiên ý kiến của 2500 số khách hàng thường xuyên đi xe taxi về chất lượng phục vụ của 3 hãng taxi thu được kết quả sau đây:

Chất lượng phục vụ	Hãng taxi		
	A	B	C
Rất tốt	140	110	205
Khá	230	150	350
Bình thường	350	225	520
Kém	80	15	125

Hãy tính đặc trưng mẫu cho từng hãng taxi và nêu đánh giá sơ bộ từ số liệu điều tra trên.

10. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên kích thước n của đám đông X có $\mathbb{E}X = \mu$. Chứng minh rằng

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \quad \text{và} \quad \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

đều là các ước lượng không chệch của phương sai $\mathbb{D}X$.

11. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên kích thước n từ phân phối với hàm mật độ là:

$$p(x,\theta) = \begin{cases} \dfrac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{với } x > 0,\ \theta > 0, \\ 0 & \text{với } x \leqslant 0. \end{cases}$$

Tìm ước lượng hiệu quả của θ.

12. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên kích thước n từ phân phối Poisson với tham số $\mathbb{E}X = \mathbb{D}X = \lambda > 0$. Tìm ước lượng hợp lý tối đa của λ.

13. Để xác định độ chính xác của một chiếc cân, người ta tiến hành cân một quả tạ. Kết quả thu được sau 7 lần cân như sau:

159,8 159,7 160,2 159,6 160,4 159,5 160,6 (kg)

(a) Tìm ước lượng không chệch của khối lượng quả cân.

(b) Tìm ước lượng không chệch của phương sai số đo trong hai trường hợp:

- Biết khối lượng quả cân là 160 kg.
- Chưa biết khối lượng của quả cân.

14. Cơ quan quản lý thị trường lấy số liệu về giá thành bán lẻ của một loại sản phẩm tại 40 đại lý, người ta thu được bảng tần số như sau: (đơn vị: ngàn đồng)

x_i	39	40	41	42
n_i	8	16	4	12

(a) Tính trung bình mẫu $\overline{x}$ và phương sai mẫu hiệu chỉnh s^2.

(b) Với độ tin cậy 95%, ước lượng khoảng giá thành bán lẻ trung bình mỗi sản phẩm.

15. Một dây chuyền sản xuất những thanh kim loại có chiều dài tuân theo luật phân phối chuẩn. Người ta chọn ngẫu nhiên ra một số thanh và đo chiều dài (đơn vị: cm) của chúng, thu được dãy số liệu sau:

149; 151; 148; 152; 151; 152; 149; 148; 149; 151; 152; 149; 151; 149; 152.

(a) Tính trung bình mẫu $\overline{x}$ và phương sai mẫu hiệu chỉnh s^2.

(b) Với độ tin cậy 90%, ước lượng khoảng độ dài trung bình của mỗi thanh kim loại.

16. Một dây chuyền tự động đóng gói một loại bao gạo có khối lượng tuân theo luật phân phối chuẩn với độ lệch chuẩn là 0,5. Người ta cân kiểm tra 20 bao gạo, thu được bảng tần số như sau: (đơn vị: kg)

x_i	49,3	49,5	49,9	50,2
n_i	6	2	4	8

(a) Tính trung bình mẫu $\overline{x}$ và phương sai mẫu hiệu chỉnh s^2.

(b) Với độ tin cậy 98%, ước lượng khoảng khối lượng trung bình của mỗi bao gạo.

17. Nhà sản xuất muốn ước lượng khối lượng sắt trong mỗi cuộn được sản xuất từ một dây chuyền công nghệ quốc gia. Theo tiêu chuẩn của công nghệ, độ lệch chuẩn là 8 kg. Điều tra một mẫu 50 cuộn được khối lượng sắt trung bình là 97kg.

(a) Với độ tin cậy là 99%, ước lượng khối lượng sắt trung bình của một cuộn.

(b) Với độ tin cậy là 99%, ước lượng khối lượng sắt trung bình tối thiểu của một cuộn.

(c) Nếu nhà sản xuất muốn ước lượng khối lượng sắt trung bình của mỗi cuộn đảm bảo độ chính xác là 2 kg thì cần điều tra thêm bao nhiêu cuộn nữa.

18. Một công ty có 500 đại lý, để đánh giá về mức doanh thu, người ta lấy mẫu gồm 36 đại lý. Kết quả thu được như sau: doanh thu trung bình là 84,5 triệu đồng và độ lệch chuẩn mẫu là 3 triệu đồng. Với độ tin cậy 99%, hãy ước lượng doanh thu tối thiểu và tối đa của công ty.

19. Người ta đo chiều sâu của biển bằng một loại thiết bị điện tử, kết quả đo tuân theo luật phân phối chuẩn có phương sai $400m^2$. Với độ tin cậy là 95%, cần phải đo ít nhất bao nhiêu lần để kết quả có sai số không vượt quá $15m$.

20. Một mẫu thống kê có kích thước $n = 64$, tuân theo luật phân phối chuẩn với trung bình mẫu là 200, độ lệch chuẩn mẫu là 3. Tìm độ tin cậy của ước lượng nếu khoảng ước lượng là $(199, 201)$.

21. Để đánh giá hiệu quả của một loại thuốc, người ta đem sử dụng cho 1000 bệnh nhân thì có 820 người khỏi bệnh. Với độ tin cậy 96 %,

(a) Hãy ước lượng khoảng cho tỉ lệ chữa khỏi bệnh của loại thuốc trên.

(b) Hãy ước lượng tỉ lệ chữa bệnh tối đa và tối thiểu của loại thuốc trên.

22. Tỉ lệ chính phẩm của một nhà máy là 90%. Với độ tin cậy 95%, muốn ước lượng tỉ lệ chính phẩm của nhà máy với độ dài khoảng tin cậy không quá 0,02 thì phải kiểm tra ít nhất bao nhiêu sản phẩm?

23. Một kho hàng tồn gồm 10.000 chiếc bút bi. Lấy mẫu gồm 100 chiếc bút từ kho hàng ra kiểm tra thì có 75 chiếc đạt chất lượng. Với độ tin cậy 95%, hãy ước lượng khoảng tỉ lệ số bút không đạt chất lượng và suy ra khoảng tin cậy số bút không đạt chất lượng của kho.

24. Tại một bang có 4 triệu người tham gia bầu cử, người ta phỏng vấn ngẫu nhiên 1000 cử tri thì có 720 cử tri ủng hộ một ứng cử viên A. Với độ tin cậy là 95%, có ít nhất bao nhiêu cử tri của bang đó đã ủng hộ ứng cử viên A?

25. Để đánh giá trữ lượng cá trong một hồ nuôi, người ta bắt 1000 con cá và đánh dấu chúng, sau đó thả lại hồ. Lần thứ hai người ta bắt 200 con thì thấy có 30 con được đánh dấu. Với độ tin cậy là 95%,

    (a) Hãy ước lượng trữ lượng cá trong hồ.

    (b) Nếu muốn sai số của ước lượng giảm đi một nửa thì cần phải bắt bao nhiêu con cá.

26. Quy định của một thiết bị phải có chiều dài là 300cm và độ lệch chuẩn là 3cm. Từ một lô hàng người ta lấy ra 40 chiếc, kết quả thu được độ dài trung bình là 301,2cm. Với mức ý nghĩa 5%, lô hàng trên có đạt tiêu chuẩn hay không?

27. Trong điều kiện chăn nuôi bình thường, lượng sữa thu được trung bình hàng ngày của một loại giống bò sữa là 19,4 (đơn vị: kg/ngày). Lấy mẫu 49 con bò sữa ở một trang trại thu được lượng sữa trung bình của một con trong một ngày là 18,9 và độ lệch chuẩn mẫu là 3,24. Với mức ý nghĩa $\alpha = 0{,}08$, lượng sữa thu được hàng ngày từ bò sữa của trang trại có đúng chuẩn không?

28. Khối lượng chuẩn của một bao gạo được đóng gói bằng dây chuyền tự động là đại lượng ngẫu nhiên có phân phối chuẩn với khối lượng mỗi bao là 50 kg. Sau một thời gian hoạt động người ta nghi ngờ khối lượng đó có xu hướng giảm sút. Cân 28 bao gạo thu được khối lượng trung bình mỗi bao là 49,8 kg và độ lệch chuẩn mẫu là 0,6 kg. Với mức ý nghĩa 1%, hãy kết luận về nghi ngờ nói trên.

29. Thời gian trước đây, số tiền gửi tiết kiệm trung bình của mỗi khách hàng vào ngân hàng A là 1000 USD. Sau đợt tăng lãi suất tiết kiệm, kiểm tra ngẫu nhiên 36 khách hàng thu được kết quả: số tiền gửi trung bình là 1060 USD và độ lệch chuẩn mẫu là 100 USD. Với mức ý nghĩa 4%, việc tăng lãi suất có làm tăng lượng tiền gửi tiết kiệm của mỗi khách hàng không?

30. Một kênh truyền thông tuyên bố rằng 30% khán giả truyền hình yêu thích các chương trình phát sóng của họ. Thăm dò ý kiến ngẫu nhiên qua mạng đối với 800 người xem truyền hình thì có 192 người yêu thích các chương trình của kênh truyền thông đó. Với mức ý nghĩa 0,08, tỉ lệ trong tuyên bố trên có đúng với thực tế không?

31. Tỉ lệ phế phẩm của một nhà máy trước đây là 10%. Sau khi cải tiến kỹ thuật, kiểm tra 400 sản phẩm thì thấy có 38 phế phẩm. Với mức

ý nghĩa là 1%, kiểm tra xem việc cải tiến kỹ thuật có mang lại hiệu quả không?

32. Tỉ lệ người chữa khỏi một loại bệnh bằng loại thuốc cũ là 80%. Người ta thay thế bằng loại thuốc mới để chữa bệnh cho 1000 người thì có 820 người khỏi bệnh. Với mức ý nghĩa 1%, có thể kết luận thuốc mới tốt hơn thuốc cũ không?

33. Hai giống vịt được nuôi sau 4 tháng. Lấy mẫu $n_1 = 50$ ở giống vịt thứ nhất, được $\overline{x}_1 = 1.9kg$ và $s_1^2 = 1$. Lấy mẫu $n_2 = 80$ ở giống vịt thứ hai, được $\overline{x}_2 = 2kg$ và $s_2^2 = 0.8$. Với mức ý nghĩa $\alpha = 10\%$, hai giống vịt này có trọng lượng trung bình bằng nhau không?

34. Chọn ngẫu nhiên 20 đại lý có áp dụng khuyến mãi thu được số lượng bán trung bình mỗi ngày là 140 sản phẩm và độ lệch chuẩn mẫu là 12; còn tại 20 đại lý không có khuyến mãi được 2 số liệu tương ứng là 120 và 10. Giả sử lượng hàng bán được có phân phối chuẩn, có cùng phương sai. Với mức ý nghĩa 5%, hình thức khuyến mãi có làm tăng số lượng hàng bán không?

35. Một công ty bán hàng muốn kiểm tra hiệu quả từ việc thay đổi kiểu đóng gói. Chọn 2 mẫu: mẫu 1 là 35 đại lý bán hàng theo loại gói cũ và mẫu 2 là 35 đại lý bán hàng theo loại gói mới để thống kê về số gói hàng bán ra sau một tháng, thu được 2 giá trị đặc trưng cho 2 mẫu tương ứng như sau: loại gói cũ: $\overline{x_1} = 560$ gói, với $s_1 = 20$; loại gói mới: $\overline{x_2} = 580$ gói, với $s_2 = 30$. Với mức ý nghĩa 1%, hãy đánh giá việc thay đổi kiểu đóng gói có đem lại hiệu quả hay không?

36. Để so sánh tỉ lệ nảy mầm của hai giống cây trong điều kiện độ ẩm thấp. Người ta đem gieo 200 hạt giống loại I thì có 150 hạt nảy mầm,

gieo 300 hạt giống loại II thấy có 210 hạt nảy mầm. Với mức ý nghĩa $\alpha = 0{,}05$, tỉ lệ nảy mầm trong điều kiện độ ẩm thấp của 2 giống cây trên có như nhau không?

37. Lấy số liệu thực tế từ các hộ gia đình vay vốn của ngân hàng nông nghiệp đối với 2 huyện. Huyện A: có 2000 hộ vay thì có 1692 hộ sử dụng tiền vay có hiệu quả; huyện B: có 1000 hộ vay thì có 810 hộ sử dụng tiền vay có hiệu quả. Với mức ý nghĩa 5%, tỉ lệ hộ sử dụng tiền vay có hiệu quả của huyện A có cao hơn ở huyện B không?

38. Để đánh giá về chất lượng sản phẩm của nhà máy do 2 dây chuyền sản xuất. Người ta kiểm tra ngẫu nhiên 200 sản phẩm từ dây chuyền thứ nhất thì có 15 phế phẩm, kiểm tra 300 sản phẩm từ dây chuyền thứ hai thấy có 21 phế phẩm. Từ số liệu thu được có thể đánh giá sơ bộ dây chuyền nào làm việc tốt hơn. Với mức ý nghĩa $\alpha = 0{,}08$, kiểm định đánh giá sơ bộ đó.

39. Có hai phương pháp gieo một loại hạt giống: theo phương pháp A, gieo 125 hạt thấy có 90 hạt nảy mầm; theo phương pháp B, gieo 100 hạt thấy có 85 hạt nảy mầm. Từ số liệu thu được có thể đánh giá sơ bộ phương pháp gieo nào tốt hơn. Với mức ý nghĩa $\alpha = 0{,}05$, kiểm định đánh giá sơ bộ đó.

40. Tại một nhà máy làm việc theo chế độ 3 ca: buổi sáng, buổi chiều và buổi tối, chọn ngẫu nhiên một số sản phẩm để kiểm tra chất lượng, thu được bảng số liệu sau

Chất lượng	Ca		
	Sáng	Chiều	Tối
Chính phẩm	84	64	70
Phế phẩm	2	8	2

Với mức ý nghĩa $\alpha = 0{,}05$, có thể kết luận chất lượng sản phẩm phụ thuộc vào ca làm việc không?

41. Tại một nhà máy có 4 phân xưởng: I, II, III, IV; cùng sản xuất ra một loại sản phẩm với 3 tiêu chí đánh giá chất lượng: Loại A (tốt), loại B (đạt), loại C (chưa đạt). Kiểm tra 1000 sản phẩm khi nhập tổng kho, thu được bảng số liệu sau

Chất lượng / Xưởng	Loại A	Loại B	Loại C
I	105	90	25
II	135	102	13
III	124	100	6
IV	146	138	16

Với mức ý nghĩa $\alpha = 0{,}01$, có thể kết luận chất lượng sản phẩm phụ thuộc vào phân xưởng sản xuất không?

42. Bảng số liệu sau đây là kết quả thống kê về tổng giá trị hàng nông sản (X) và tổng đầu tư xây dựng đường giao thông (Y) của một huyện trong 6 năm như sau: (đơn vị: tỉ đồng)

X	60	45	75	90	80	70
Y	7	5	8	11	9	10

(a) Hãy xác định hệ số tương quan mẫu.

(b) Tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x.

(c) Nếu tiền đầu tư xây dựng đường giao thông của một năm nào đó là 8,6 tỉ đồng, hãy dự đoán tổng giá trị hàng nông sản năm đó là bao nhiêu ?

43. Bảng số liệu sau đây là kết quả thu được của một công ty về số tiền dành cho các hoạt động chăm sóc khách hàng (X) và doanh thu (Y) trong 6 tháng như sau:

X	8	9	7	10	9	11
Y	600	700	500	900	800	1100

(đơn vị: triệu đồng).

(a) Hãy xác định hệ số tương quan mẫu.

(b) Nếu chi phí dành cho các hoạt động chăm sóc khách hàng của một tháng nào đó là 10,5 triệu đồng, hãy dự đoán doanh thu của công ty tháng đó là bao nhiêu ?

44. Thống kê ghi lại dân số của một tỉnh qua 8 năm từ năm 1985 đến 1992 được bảng số sau

Năm	1985	1986	1987	1988	1989	1990	1991	1992
Dân số (10000)	50	51	51	53	54	56	59	60

Để thuận tiện trong tính toán ta đặt $x =$ "năm" $- 1985$ và $y =$ "dân số" $- 50$ (đơn vị 10000 người). Hãy tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x.

45. Tính hệ số tương quan mẫu và phương trình hồi quy tuyến tính thực nghiệm của y theo x dựa vào bảng tần số sau:

x_i	17	14	12	15	12	20
y_i	31	33	25	29	27	40
n_i	2	4	10	3	5	6

46. Bảng số liệu sau đây chỉ ra sự phụ thuộc của năng suất thu hoạch Y theo lượng phân bón X của một loại hoa màu trên 100 thửa ruộng.

Y	X			
	20	25	30	35
400	12	5	1	1
420	6	18	3	2
450	2		10	9
490		1	10	20

Tính hệ số tương quan mẫu và phương trình hồi quy tuyến tính thực nghiệm của năng suất thu hoạch theo lượng phân bón.

TÀI LIỆU THAM KHẢO

1. Nguyễn Quang Báu, *Lý thuyết xác suất và thống kê toán học*, Đại học quốc gia Hà Nội, 2004.

2. Lê Sĩ Dồng, *Xác suất thống kê và ứng dụng*, NXB Giáo dục, 2004.

3. Đặng Hấn, *Xác suất thống kê*, NXB Thống kê, 1996.

4. Đào Hữu Hồ, *Xác suất thống kê*, NXB Đại học quốc gia Hà Nội, 2006.

5. Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như, *Thống kê toán học*, NXB Đại học quốc gia Hà Nội, 2004.

6. Nguyễn Văn Quảng, *Giáo trình xác suất*, NXB Đại học quốc gia Hà Nội, 2007.

7. Đặng Hùng Thắng, *Mở đầu lý thuyết xác suất*, NXB Giáo dục, 2000.

8. Nguyễn Duy Tiến - Vũ Việt Yên, *Lý thuyết xác suất*, NXB Giáo dục, 2000.

9. Y.S. Chow and H. Teicher; *Probabylity Theory: Independence, Interchangeability, martingales,* Springer-Verlag, Berlin and New York, 1988.