Chương 4

Thống kê. Ước lượng tham số

TUẦN 11

4.1 Lý thuyết mẫu

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 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

Tập mẫu (sample) là tập con của tổng thể và có tính chất tương tự như tổng thể. Số phần tử của tập mẫu được gọi là kích thước mẫu (cỡ mẫu), ký hiệu là n.

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

Một 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:

1. Chọn mẫu ngẫu nhiên có hoàn lại: Lấy ngẫu nhiên một phần tử từ tổng thể và khảo sát nó. Sau đó trả phần tử đó lại tổng thể trước khi lấy một phần tử khác. Tiếp tục như thế n lần ta thu được một mẫu có hoàn lại gồm n phần tử.

2. Chọn mẫu ngẫu nhiên không hoàn lại: Lấy ngẫu nhiên một phần tử từ tổng thể và khảo sát nó rồi để qua một bên, không trả lại tổng thể. Sau đó lấy ngẫu nhiên một phần tử khác, tiếp tục như thế n lần ta thu được một mẫu không hoàn lại gồm n phần tử.

3. Chọn mẫu phân nhóm: Đầu tiên ta chia tập nền thành các nhóm tương đối thuần nhất, từ mỗi nhóm đó chọn ra một mẫu ngẫu nhiên. Tập hợp tất cả mẫu đó cho ta một mẫu phân nhóm. Phương pháp này dùng khi trong tập nền có những sai khác lớn. Hạn chế là phụ thuộc vào việc chia nhóm.

4. Chọn mẫu có suy luận: Dựa trên ý kiến của chuyên gia về đối tượng nghiên cứu để chọn mẫu.

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	...	x_n
P	$P(X = x_1)$	$P(X = x_2)$	...	$P(X = x_n)$

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

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:

(a) Trung bình tổng thể: $E(X) = \mu$.

(b) Phương sai tổng thể: $V(X) = \sigma^2$.

(c) Độ lệch chuẩn của tổng thể: $\sigma(X) = \sigma$.

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 X có hàm phân phối xác suất $F_X(x)$. Một mẫu ngẫu nhiên cỡ 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 $x_2, \ldots, X_n$ nhận giá trị x_n ta thu được một mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$.

Ví dụ 4.2. 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 $W_x = (6, 3, 1)$.

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: $x_i = \dfrac{1}{2}(a_{i-1} + a_i)$.

Phân phối thực nghiệm

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

Thông thường ta biểu diễn phân phối tần số, tần suất bằng đồ thị. Có hai dạng biểu diễn đồ thị hay dùng là biểu đồ và đa giác tần số (sinh viên tự đọc).

4.1.4 Đại lượng thống kê và các đặc trưng của mẫu ngẫu nhiên

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.

Định nghĩa 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.

Khi có mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_2)$, ta tính được giá trị cụ thể của G, ký hiệu là $g = f(x_1, x_2, \ldots, x_n)$, còn gọi là giá trị quan sát của thống kê.

Nhận xét 4.1. 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 xét các đặc trưng của thống kê này.

Trung bình mẫu ngẫu nhiên

Cho mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$. Trung bình mẫu của mẫu ngẫu nhiên W_X của biến ngẫu nhiên gốc X được định nghĩa và ký hiệu

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

Nếu biến ngẫu nhiên gốc có kỳ vọng $E(X) = \mu$, phương sai $V(X) = \sigma^2$ thì theo Tính chất 2.4(c) và Tính chất 2.5(c) của kỳ vọng và phương sai, 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.

Phương sai mẫu ngẫu nhiên

Phương sai mẫu của mẫu ngẫu nhiên W_X của biến ngẫu nhiên gốc X được ký hiệu và định nghĩa

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

Độ lệch chuẩn mẫu ngẫu nhiên được ký hiệu và xác định bởi

$$\boxed{\hat{S} = \sqrt{\hat{S}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}} \tag{4.4}$$

Sử dụng Tính chất 2.4(c) của kỳ vọng, ta có

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

Để kỳ vọng của phương sai mẫu ngẫu nhiên trùng với phương sai của biến ngẫu nhiên gốc ta cần một sự hiệu chỉnh. Đó là phương sai hiệu chỉnh mẫu ngẫu nhiên.

Phương sai hiệu chỉnh mẫu ngẫu nhiên

Phương sai hiệu chỉnh mẫu của mẫu ngẫu nhiên W_X của biến ngẫu nhiên gốc X được ký hiệu và định nghĩa

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

Độ lệch chuẩn hiệu chỉnh mẫu ngẫu nhiên được ký hiệu và xác định bởi

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

Theo Tính chất 2.4(c) của kỳ vọng ta nhận được

$$E(S^2) = \sigma^2.$$

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

$$\boxed{f = \frac{m}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}} \tag{4.7}$$

Như vậy tần suất mẫu là trung bình mẫu của biến ngẫu nhiên X có phân bố Béc-nu-li tham số p. Ngoài ra theo Tính chất 2.4(c) và Tính chất 2.5(c), ta có

$$\boxed{E(f) = p, \quad V(f) = \frac{p(1-p)}{n}} \tag{4.8}$$

4.1.5 Cách tính giá trị cụ thể của trung bình mẫu và phương sai mẫu

Giả sử ta có mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$ cỡ n.

(a) Mẫu cho dưới dạng liệt kê. (Tần số của các x_i bằng 1) (a1) Trung bình mẫu:

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i} \tag{4.9}$$

(a2) Phương sai mẫu:

$$\boxed{\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \tag{4.10}$$

(a3) Phương sai hiệu chỉnh mẫu:

$$\boxed{s^2 = \frac{n}{n-1}\hat{s}^2} \tag{4.11}$$

(a4) Các độ lệch chuẩn:

$$\boxed{\hat{s} = \sqrt{\hat{s}^2}; \quad s = \sqrt{s^2}} \tag{4.12}$$

Để tính các công thức (4.9)–(4.12), 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
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} x_i^2$

(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:

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i} \tag{4.13}$$

(b2) Phương sai mẫu:

$$\boxed{\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{k} n_i x_i\right)^2} \tag{4.14}$$

(b3) Phương sai hiệu chỉnh mẫu:

$$\boxed{s^2 = \frac{n}{n-1}\hat{s}^2} \tag{4.15}$$

(b4) Các độ lệch chuẩn:

$$\boxed{\hat{s} = \sqrt{\hat{s}^2}; \quad s = \sqrt{s^2}} \tag{4.16}$$

Để tính các công thức (4.13)–(4.16), 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:

$$\boxed{\overline{x} = x_0 + h\overline{u} = x_0 + \frac{h}{n}\sum_{i=1}^k n_i u_i} \tag{4.17}$$

(c2) Phương sai mẫu:

$$\boxed{\hat{s}^2 = h^2\Big[\frac{1}{n}\sum_{i=1}^k n_i u_i^2 - \Big(\frac{1}{n}\sum_{i=1}^k n_i u_i\Big)^2\Big] = h^2 \hat{s}_u^2} \tag{4.18}$$

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 (4.17)–(4.18), 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** → **1** → **4(VAR)** → **2**
- Độ lệch tiêu chuẩn mẫu hiệu chỉnh (s): **SHIFT** → **1** → **4** → **4**

Ví dụ 4.3. 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
Cách 1: Gọi X là số khuyết tật ở mỗi đơn vị. Lập bảng tính toán

x_i	n_i	n_ix_i	$n_ix_i^2$
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
$\sum$	$n = 150$	499	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 $\overline{x} = 3,3267$; $s = 1,6648$.

4.1.6 Phân phối xác suất của các thống kê trung bình mẫu, phương sai mẫu, tần suất mẫu ngẫu nhiên

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 $W_X=(X_1,X_2,\ldots,X_n)$ cỡ 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.

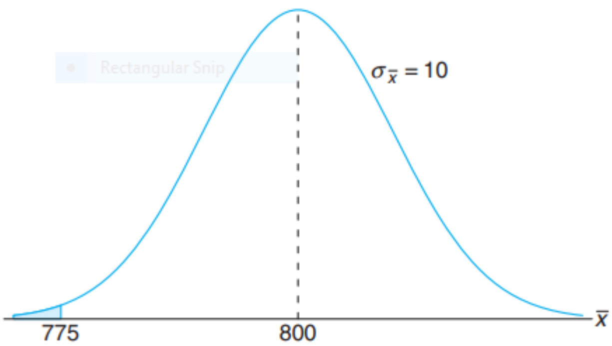
Phân phối của thống kê trung bình mẫu

Thống kê trung bình mẫu $\overline{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$ có phân phối chuẩn $\mathcal{N}\Big(\mu;\dfrac{\sigma^2}{n}\Big)$ và do đó thống kê $U=\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}$ có phân phối chuẩn tắc (xem Định lý giới hạn trung tâm)

$$\boxed{\overline{X}\sim\mathcal{N}\Big(\mu,\frac{\sigma^2}{n}\Big),\quad U=\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\sim\mathcal{N}(0,1)}\tag{4.19}$$

Ví dụ 4.4. Một công ty điện sản xuất bóng đèn có tuổi thọ là biến ngẫu nhiên phân phối xấp xỉ chuẩn, với tuổi thọ trung bình là 800 giờ và độ lệch chuẩn là 40 giờ. Tìm xác suất để một mẫu ngẫu nhiên gồm 16 bóng đèn sẽ có tuổi thọ trung bình dưới 775 giờ.

Lời giải Ví dụ 4.4 Gọi X là tuổi thọ của bóng đèn. $X\sim\mathcal{N}(800,40^2)$. Khi đó, tuổi thọ trung bình của mẫu ngẫu nhiên $\overline{X}$ có phân phối xấp xỉ chuẩn với $\mu_{\overline{X}}=800$ và $\sigma_{\overline{X}}=40/\sqrt{16}=10$. Xác suất cần tính là diện tích của vùng bóng mờ trong Hình 4.1.



Hình 4.1: Minh họa của Ví dụ 4.4

Vì $\overline{X}\sim\mathcal{N}(800,10^2)$, nên

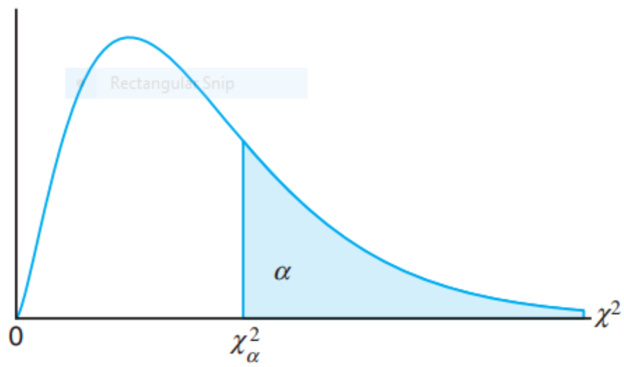
$$P(\overline{X}<775)=0,5+\phi\Big(\frac{775-800}{10}\Big)=0,5+\phi(-2.5)=0,5-0,49379=0.00621,$$

trong đó $\phi(-2,5)=-0,49379$ tra từ bảng giá trị hàm số Láp-la-xơ (Phụ lục 2).

Phân phối của thống kê phương sai mẫu

Thống kê $\chi^2 = \dfrac{n\hat{S}^2}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2}$ có phân phối khi bình phương với $n-1$ bậc tự do

$$\boxed{\frac{n\hat{S}^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}} \tag{4.20}$$



Hình 4.2: Phân phối khi bình phương

(sinh viên tự đọc phân phối này).

Phân phối của thống kê $T = \dfrac{\overline{X}-\mu}{S}\sqrt{n}$ **hoặc** $T = \dfrac{\overline{X}-\mu}{\hat{S}}\sqrt{n-1}$

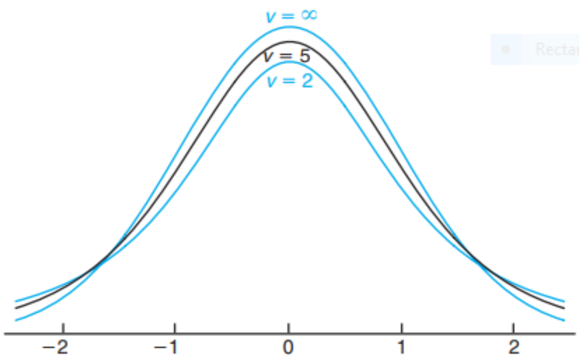
Thống kê $T = \dfrac{\overline{X}-\mu}{S}\sqrt{n} = \dfrac{\overline{X}-\mu}{\hat{S}}\sqrt{n-1}$ có phân phối Student với $n-1$ bậc tự do.

$$\boxed{T = \frac{\overline{X}-\mu}{S}\sqrt{n} = \frac{\overline{X}-\mu}{\hat{S}}\sqrt{n-1} \sim \mathcal{T}^{(n-1)}} \tag{4.21}$$

Nhận xét 4.2. **(a)** Phân phối Student (của thống kê $T = T = \dfrac{\overline{X}-\mu}{S}\sqrt{n}$) có cùng dạng và tính đối xứng như phân phối chuẩn (của thống kê $U = \dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}$) nhưng nó phản ánh tính biến đổi của phân phối sâu sắc hơn (do thực tế là giá trị T phụ thuộc vào sự biến động của hai đại lượng $\overline{X}$ và S^2, trong khi U chỉ phụ thuộc vào những thay đổi của $\overline{X}$ từ mẫu này sang mẫu khác).

(b) Phân phối chuẩn không thể dùng để xấp xỉ phân phối khi mẫu có kích thước nhỏ. Trong trường hợp này ta dùng phân phối Student.

(c) Khi bậc tự do n tăng lên ($n \geq 30$) thì phân phối Student tiến nhanh về phân phối chuẩn. Do đó khi $n \geq 30$ ta có thể dùng phân phối chuẩn thay thế cho phân phối Student.



Hình 4.3: Phân phối Student với số bậc tự do $\nu = 2, 5$ và ∞

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 $\mathcal{N}(0,1)$.

Nếu $T \sim \mathcal{T}(n)$ thì $P(T < t_{\alpha}^{(n)}) = \alpha$. Giá trị $t_{\alpha}^{(n)}$ được tra từ bảng phân phối Student (Phụ lục 4). Chẳng hạn với $n = 10$, $\alpha = 0,5$ thì $t_{1-\alpha/2}^{(n)} = t_{1-0,025}^{(10)} = t_{0,975}^{(10)} = 2,228$.

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

$$\boxed{U = \frac{f-p}{\sqrt{p(1-p)}}\sqrt{n} \sim \mathcal{N}(0,1)} \tag{4.22}$$

4.2 Ước điểm cho kỳ vọng, phương sai và tỷ lệ

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 Ước lượng điểm

Khái niệm ước lượng điểm

Cho biến ngẫu nhiên gốc X có thể đã biết hoặc chưa biết quy luật phân phối xác suất dạng tổng quát, nhưng chưa biết tham số θ nào đó. Hãy ước lượng θ bằng phương pháp mẫu. Vì θ là một hằng số nên có thể dùng một số nào đó để ước lượng θ. Ước lượng như vậy gọi là ước lượng điểm.

Phương pháp hàm ước lượng

- Giả sử cần ước lượng tham số θ của biến ngẫu nhiên X. Từ X ta lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ cỡ n. Chọn 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.

- 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 $G = f(X_1, X_2, \ldots, X_n)$ là hàm ước lượng của θ.

4.2.2 Các 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.

Ước lượng không chệch (unbiased estimator)

Thống kê G được gọi là ước lượng không chệch của θ nếu

$$\boxed{E(G) = \theta \quad \text{với mọi} \quad \theta} \tag{4.23}$$

Nếu $E(G) \neq \theta$ thì G là ước lượng chệch của θ.

Điều kiện (4.23) 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.

Ước lượng hiệu quả (efficient estimator)

Thống kê G được gọi là ước lượng hiệu quả (hay ước lượng phương sai bé nhất) của θ nếu G là ước lượng không chệch của θ và phương sai của G nhỏ hơn bất kỳ phương sai của một hàm ước lượng không chệch nào khác.

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: *Cho mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ cỡ 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 (nếu là biến ngẫu nhiên liên tục) hay bảng phân phối xác suất (nếu là biến ngẫu nhiên rời rạc) thỏa mãn một số điều kiện nhất* *định (thường được thỏa mãn trong thực tế, ít ra là các phân phối xác suất đã xét trong Chương 2) và G là ước lượng không chệch bất kỳ của θ thì*

$$\boxed{V(G) \geq \frac{1}{nE\left(\frac{\partial(\ln f(X,\theta))}{\partial\theta}\right)^2}} \tag{4.24}$$

Ước lượng vững (consistent estimator)

Thống kê G được gọi là ước lượng vững của tham số θ nếu G hội tụ theo xác suất đến θ khi $n \to +\infty$.

4.2.3 Ước lượng điểm cho kỳ vọng, phương sai và xác suất

- Chọn hàm $G = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ nếu ước lượng kỳ vọng $E(X) = \mu$. Kỳ vọng mẫu ngẫu nhiên $\overline{X}$ là ước lượng không chệch, hiệu quả và vững của kỳ vọng $E(X) = \mu$ của biến ngẫu nhiên gốc của tổng thể.
- Chọn $G = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = S^2$ nếu ước lượng phương sai $V(X) = \sigma^2$. Phương sai hiệu chỉnh mẫu ngẫu nhiên S^2 là ước lượng không chệch, hiệu quả và vững của phương sai $V(X) = \sigma^2$ của biến ngẫu nhiên gốc của tổng thể.
- Chọn $G = \frac{m}{n} = f$ nếu ước lượng cho xác suất p. Tần suất mẫu ngẫu nhiên f là ước lượng không chệch, hiệu quả và vững của xác suất p của tổng thể.

Ví dụ 4.5. 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 4.5 Ước lượng điểm cần tìm là $f = \frac{960}{1600} = 0,6 = 60\%$.

4.2.4 Một số phương pháp tìm ước lượng điểm

(a) Phương pháp hợp lý cực đại (maximum-likelihood estimation)

(b) Phương pháp mô men (moment estimation)

(c) Phương pháp Bayes, phương pháp minimax, phương pháp bootstrap ...

(Sinh viên tự đọc).

TUẦN 12

4.3 Phương pháp ước lượng bằng khoảng tin cậy

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.

4.3.1 Khoảng tin cậy của kỳ vọng của biến ngẫu nhiên phân phối chuẩn

Bài toán 4.1. Giả sử biến ngẫu nhiên X tuân theo luật phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$ với kỳ vọng $E(X) = \mu$ chưa biết. Hãy ước lượng $E(X)$.

Các bước tiến hành: Từ tổng thể, ta lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ cỡ n và xét các trường hợp sau.

Trường hợp đã biết phương sai $V(X) = \sigma^2$

Bước 1 Chọn thống kê

$$\boxed{U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}} \tag{4.25}$$

Theo Mục 4.1.6, thống kê U có phân phối chuẩn tắc $\mathcal{N}(0;1)$.

Bước 2 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,

$$\begin{aligned} 1 - \alpha = P(-u_{1-\alpha_1} < U < u_{1-\alpha_2}) &= P\Big(-u_{1-\alpha_1} < \frac{\overline{X} - \mu}{\sigma}\sqrt{n} < u_{1-\alpha_2}\Big) \\ &= P\Big(\overline{X} - u_{1-\alpha_2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\alpha_1}\frac{\sigma}{\sqrt{n}}\Big). \end{aligned}$$

Bước 3 Lập 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:

$$\boxed{\Big(\overline{x} - u_{1-\alpha_2}\frac{\sigma}{\sqrt{n}} \quad ; \quad \overline{x} + u_{1-\alpha_1}\frac{\sigma}{\sqrt{n}}\Big)} \tag{4.26}$$

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 đối xứng $(\alpha_1 = \alpha_2 = \alpha/2)$

$$\boxed{\Big(\overline{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad ; \quad \overline{x} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\Big)} \tag{4.27}$$

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

$$\boxed{\Phi(u_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}} \tag{4.28}$$

Sai số của ước lượng: $\varepsilon = u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}$ được gọi là sai số (độ chính xác) của ước lượng. Với phương sai σ^2 đã biết không đổi và độ tin cậy γ không đổi thì giá trị $u_{1-\frac{\alpha}{2}}$ không đổi, do đó sai số của ước lượng chỉ phụ thuộc vào kích thước mẫu n. Khi n càng lớn thì ε càng bé, do đó khoảng ước lượng càng chính xác.

Tìm kích thước mẫu: Nếu muốn ước lượng kỳ vọng với độ chính xác ε_0 và độ tin cậy γ cho trước, kích thước mẫu cần thiết là số tự nhiên n nhỏ nhất thỏa mãn:

$$\boxed{n \geq \frac{\sigma^2 u^2_{1-\frac{\alpha}{2}}}{\varepsilon_0^2}} \tag{4.29}$$

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

$$\boxed{\left(-\infty \quad ; \quad \overline{x} + u_{1-\alpha}\frac{\sigma}{\sqrt{n}} \right)} \tag{4.30}$$

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

$$\boxed{\Phi(u_{1-\alpha}) = 1 - \alpha}. \tag{4.31}$$

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

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

Ví dụ 4.6. Trọng lượng của một loại sản phẩm là biến ngẫu nhiên tuân theo luật phân phối chuẩn với độ lệch tiêu chuẩn là 1 gam. Cân thử 25 sản phẩm loại này ta thu được kết quả sau:

Trọng lượng (gam)	18	19	20	21
Số sản phẩm	3	5	15	2

(a) 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.

(b) Không cần tính toán, nếu độ tin cậy 99% thì khoảng ước lượng trung bình sẽ rộng hơn, hẹp hơn hay bằng như trong ý (a)?

(c) Nếu muốn độ chính xác của ước lượng tăng lên gấp đôi, độ tin cậy không đổi thì cần nghiên cứu mẫu có kích thước là bao nhiêu?

Lời giải Ví dụ 4.6

(a) Gọi X là trọng lượng sản phẩm, $X \sim \mathcal{N}(\mu, \sigma^2)$ với $\sigma = 1$. Trọng lượng trung bình của sản phẩm là $E(X) = \mu$ chưa biết cần ước lượng.

Bước 1: Chọn thống kê $U = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n}$. Thống kê $U \sim \mathcal{N}(0;1)$.

Bước 2: Áp dụng khoảng tin cậy đối xứng $\left(\overline{x} - u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}} \, ; \, \overline{x} + u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}} \right)$.

Với $\alpha = 0,05$, $\Phi(u_{1-\frac{\alpha}{2}}) = 1 - \frac{0,05}{2} = 0,975$, tra bảng giá trị hàm phân phối chuẩn tắc (Phụ lục 3) nhận được $u_{1-\frac{\alpha}{2}} = 1,96$.

Bước 3: Từ số liệu đã cho ta có $n = 25, \sigma = 1$ và tính được $\overline{x} = 19,64$, suy ra khoảng tin cậy đối xứng của $E(X) = \mu$ là $\left(19,64 - 1,96 \times \dfrac{1}{\sqrt{25}}\ \ 19,64 + 1,96 \times \dfrac{1}{\sqrt{25}}\right)$ hay $(19,248\,;\,20,032)$.

Bước 4: Kết luận, với độ tin cậy 95%, trọng lượng trung bình của loại sản phẩm nói trên từ 19,248 gam đến 20,032 gam.

(b) Nếu độ tin cậy $1-\alpha$ tăng từ 95% lên 99% thì khoảng ước lượng sẽ rộng hơn khoảng ước lượng xét trong ý (a), do giá trị của $u_{1-\alpha/2}$ tăng từ $1,96$ lên $2,58$.

(c) Theo ý (a), độ chính xác của ước lượng là $\varepsilon = u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}} = 0,392$. Để độ chính xác tăng lên gấp đôi, tức là $\varepsilon_0 = \frac{0,392}{2} = 0,196$. Theo (4.29) ta cần mẫu có kích thước nhỏ nhất là

$$n = \left[\frac{\sigma^2 u^2_{1-\frac{\alpha}{2}}}{\varepsilon_0^2}\right] = \left[\frac{1^2 \times (1,96)^2}{(0,196)^2}\right] \simeq 100.$$

Chú ý 4.2. **(a)** Chú ý rằng không thể viết $P(19,248 < X < 20,032) = 0,95$ 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 (19,248; 20,032) nên $(19,248 < \mu < 20,032)$ không phải là sự kiện ngẫu nhiên.

(b) Ta có thể xác định $u_{1-\frac{\alpha}{2}} = 1,96$ ở ý Ví dụ 4.6(a) từ bảng giá trị hàm Láp-la-xơ (Phụ lục 2) từ hệ thức $\phi(u_{1-\alpha}) = \dfrac{1-\alpha}{2}$.

(c) Từ (4.29) ta nhận thấy khi kích thước mẫu tăng và độ tin cậy giữ nguyên thì ε giảm hay ước lượng chính xác hơn; nếu tăng độ tin cậy và giữ nguyên kích thước mẫu, do giá trị phân vị chuẩn tăng nên sai số của ước lượng ε tăng.

Trường hợp chưa biết phương sai, cỡ mẫu $n < 30$

Do σ chưa biết nên ta thay thế bằng S và chọn thống kê

$$\boxed{T = \frac{\overline{X} - \mu}{S}\sqrt{n}} \tag{4.33}$$

Như đã biết (Mục 4.1.6) thống kê T có phân phối Student với $n-1$ bậc tự do. Ta có các kết luận sau đây.

(a) Khoảng tin cậy đối xứng

$$\boxed{\left(\overline{x} - t^{(n-1)}_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\quad ;\quad \overline{x} + t^{(n-1)}_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)} \tag{4.34}$$

Sai số của ước lượng là $\varepsilon = t_{1-\frac{\alpha}{2}}^{(n-1)} \dfrac{s}{\sqrt{n}}$. Kích thước mẫu được suy từ sai số hay độ chính xác của ước lượng, là số tự nhiên n nhỏ nhất thỏa mãn:

$$\boxed{n \geq \frac{\left(t_{1-\frac{\alpha}{2}}^{(n-1)}\right)^2 \times s^2}{\varepsilon^2}} \tag{4.35}$$

(b) Khoảng tin cậy trái

$$\boxed{\left(-\infty \quad ; \quad \overline{x} + t_{1-\alpha}^{(n-1)} \frac{s}{\sqrt{n}}\right)} \tag{4.36}$$

(c) Khoảng tin cậy phải

$$\boxed{\left(\overline{x} - t_{1-\alpha}^{(n-1)} \frac{s}{\sqrt{n}} \quad ; \quad +\infty\right)} \tag{4.37}$$

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

Ví dụ 4.7. Theo dõi mức xăng hao phí (X) cho một loại ô tô đi từ A đến B thu được bảng số liệu sau:

Mức xăng hao phí (lít)	19-19,5	19,5-20,0	20,0-20,5	20,5-21,0
Số lần đi	2	10	8	5

Với độ tin cậy $1-\alpha = 95\%$ hãy tính mức xăng hao phí trung bình tối thiểu khi đi từ A đến B biết X tuân theo luật phân phối chuẩn.

Lời giải Ví dụ 4.7 Gọi X là lượng xăng hao phí của loại ô tô trên đoạn đường AB, $X \sim \mathcal{N}(\mu, \sigma^2)$ với phương sai σ^2 chưa biết. Mức xăng hao phí trung bình là $E(X) = \mu$ chưa biết, cần ước lượng.

Bước 1: Vì phương sai chưa biết và $n = 25 < 30$, chọn thống kê $T = \dfrac{\overline{X} - \mu}{S}\sqrt{n}$. Thống kê T có phân phối Student với $n-1$ bậc tự do.

Bước 2: Sử dụng khoảng tin cậy phải cho $E(X) = \mu$:

$$\left(\overline{x} - t_{1-\alpha}^{(n-1)} \frac{s}{\sqrt{n}} \quad ; \quad +\infty\right)$$

trong đó $t_{1-\alpha}^{(n-1)} = t_{0,95}^{(24)} = 1,711$ được xác định từ bảng phân phối Student (Phụ lục 4).

Bước 3: Từ số liệu của đầu bài, tính được $n = 25$, $\overline{x} = 20,07$, $s = 0,45$. Suy ra khoảng tin cậy phải của μ là $\left(20,07 - 1,711 \times \dfrac{0,45}{\sqrt{25}} < \mu < +\infty\right)$ hay $(19,92 < \mu < +\infty)$.

Bước 4: Kết luận mức xăng hao phí trung bình tối thiểu khi đi từ A đến B là 19,92 lít với độ tin cậy 95%.

Trường hợp chưa biết phương sai, cỡ mẫu $n \geq 30$

Khi $n \geq 30$ thống kê T trong (4.33) sẽ có phân phối tiệm cận chuẩn tắc $\mathcal{N}(0,1)$. Hay thống kê

$$\boxed{U = \frac{\overline{X} - \mu}{S}\sqrt{n} \sim \mathcal{N}(0,1)} \tag{4.38}$$

Do đó,

(a) Khoảng tin cậy đối xứng

$$\boxed{\left(\overline{x} - u_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \quad ; \quad \overline{x} + u_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)} \tag{4.39}$$

Sai số của ước lượng là $\varepsilon = u_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$. Kích thước mẫu được suy từ sai số hay độ chính xác của ước lượng, là số tự nhiên n nhỏ nhất thỏa mãn:

$$\boxed{n \geq \frac{\left(u_{1-\frac{\alpha}{2}}\right)^2 \times s^2}{\varepsilon^2}} \tag{4.40}$$

(b) Khoảng tin cậy trái

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

(c) Khoảng tin cậy phải

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

Ví dụ 4.8. 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 4.8 Gọi X là trọng lượng loại trái cây A, $X \sim \mathcal{N}(\mu, \sigma^2)$ với phương sai σ^2 chưa biết. Trọng lượng trung bình của loại trái cây A là $E(X) = \mu$ chưa biết, cần ước lượng.

Bước 1: Chọn thống kê $U = \frac{\overline{X} - \mu}{S}\sqrt{n}$. Vì $n = 100 > 30$ nên thống kê $U \sim \mathcal{N}(0,1)$.

Bước 2: 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).

Bước 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 $(45,573\ ;\ 46,546)$.

Bước 4: Kết luận, 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 Ước lượng khoảng cho tỷ lệ

Bài toán 4.2. Xác suất xảy ra sự kiện A là p. Do không biết p nên người ta thực hiện n phép thử độc lập, cùng điều kiện, trong đó có m phép thử xảy ra A. Khi đó tần suất xuất hiện A là $f = m/n$ là ước lượng điểm không chệch cho p. Với độ tin cậy $\gamma = 1 - \alpha$ hãy ước lượng khoảng cho p.

Phương pháp tiến hành

Bước 1 Chọn thống kê $Z = \dfrac{f - p}{\sqrt{p(1-p)}}\sqrt{n}$. Theo Mục 4.1.6, Z có phân phối chuẩn tắc $\mathcal{N}(0;1)$.

Trong trường hợp n khá lớn ta có thể dùng f để thay thế cho p. Khi đó,

$$\boxed{Z = \frac{f - p}{\sqrt{f(1-f)}}\sqrt{n} \sim \mathcal{N}(0;1)} \tag{4.43}$$

Bước 2: Khi có mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$, ta tính được giá trị cụ thể của f và suy ra khoảng ước lượng cho p với độ tin cậy $\gamma = 1 - \alpha$ là:

$$\boxed{\left(f - u_{1-\alpha_2}\sqrt{\frac{f(1-f)}{n}}\quad ;\quad f + u_{1-\alpha_1}\sqrt{\frac{f(1-f)}{n}}\right)} \tag{4.44}$$

với $\alpha = \alpha_1 + \alpha_2$.

Các trường hợp ước lượng hay dùng:

(a) Khoảng tin cậy đối xứng

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

Độ chính xác của ước lượng $\varepsilon = u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}}$. Với độ tin cậy $\gamma = 1 - \alpha$ và độ chính xác ε_0 cho trước thì kích thước mẫu cần thiết là số tự nhiên n nhỏ nhất thỏa mãn:

$$\boxed{n \ge \frac{\left(u_{1-\frac{\alpha}{2}}\right)^2 \times f \times (1-f)}{\varepsilon_0^2}} \tag{4.46}$$

(b) Khoảng tin cậy trái

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

(c) Khoảng tin cậy phải

$$\boxed{\left(f - u_{1-\alpha}\sqrt{\frac{f(1-f)}{n}} \quad ; \quad +\infty\right)} \tag{4.48}$$

Chú ý 4.3. **(a)** Do tỷ lệ chỉ nhận giá trị từ 0 đến 1 nên ta có thể thay giá trị $-\infty$ bằng 0 và $+\infty$ bằng 1 trong khoảng tin cậy trái (phải).

(b) Các khoảng tin cậy trên được xây dựng khi kích thước mẫu n đủ lớn thỏa mãn $nf \geq 5$ và $n(1-f) \geq 5$.

Ví dụ 4.9. 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.

Lời giải Ví dụ 4.9 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 $n(1-f) = 100 \times 0,4 = 40 > 5$.

Bước 1: Chọn thống kê $Z = \dfrac{f-p}{\sqrt{f(1-f)}}\sqrt{n}$. Thống kê $Z \sim \mathcal{N}(0,1)$.

Bước 2: 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 (Phụ lục 3).

Bước 3: Với $n = 100$, $m = 60$, $f = \dfrac{m}{n} = 0,6$, suy ra khoảng tin cậy đối xứng của p là

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

Bước 4: Kết luận, 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%.

Chương 5

Kiểm định giả thuyết thống kê

TUẦN 13

Một dạng khác của quy nạp thống kê là kiểm định giả thuyết thống kê. Đây là một phương pháp quan trọng cho phép giải quyết nhiều bài toán trong thực tế. Nội dung của kiểm định giả thuyết thống kê là dựa vào mẫu cụ thể và các quy tắc hay thủ tục quyết định dẫn đến bác bỏ hay chấp nhận giả thuyết của tổng thể.

5.1 Các khái niệm

Thông thường ta nghiên cứu biến ngẫu nhiên trong trường hợp thông tin không đầy đủ, thể hiện ở nhiều mặt. Cụ thể là:

1. Chưa biết chính xác tham số θ, hoặc quy luật phân phối xác xuất của biến ngẫu nhiên X, nhưng có cơ sở nào đó để nêu lên giả thuyết, chẳng hạn $\theta = \theta_0$ (θ_0 đã biết), hoặc X tuân theo quy luật phân phối chuẩn.

2. Khi nghiên cứu hai hay nhiều biến ngẫu nhiên, một trong những vấn đề cần quan tâm nhất là: các biến ngẫu nhiên này độc lập với nhau hay có sự phụ thuộc tương quan? Hơn nữa, các tham số của chúng có bằng nhau hay không? Những câu hỏi này thường chưa được trả lời khẳng định mà mới chỉ nêu lên như một giả thuyết.

5.1.1 Giả thuyết thống kê

Giả thuyết thống kê là giả thuyết về biến ngẫu nhiên gốc của tổng thể, bao gồm: dạng phân phối xác suất, các đặc trưng tham số của biến ngẫu nhiên gốc hoặc giả thuyết về sự độc lập của các biến ngẫu nhiên gốc.

Giả thuyết thống kê. Kiểm định giả thuyết thống kê

1. Bất kỳ giả thuyết nào nói về tham số, dạng quy luật phân phối xác suất hay tính độc lập của các biến ngẫu nhiên, đều được gọi là giả thuyết thống kê.

2. Việc tìm ra kết luận về tính thừa nhận được hay không thừa nhận được của giả thuyết gọi là kiểm định giả thuyết thống kê.

Trong khuôn khổ của chương trình, ta chỉ đề cập đến giả thuyết về tham số của biến ngẫu nhiên.

Giả thuyết cơ bản. Giả thuyết đối

1. Giả sử cần nghiên cứu tham số θ của biến ngẫu nhiên X và có cơ sở nào đó để nêu lên giả thuyết $\theta = \theta_0$. Giả thuyết này ký hiệu là H_0, còn gọi là giả thuyết cần kiểm định hay giả thuyết cơ bản hay giả thuyết không (null hypothesis).

2. Mệnh đề đối lập với giả thuyết H_0 ký hiệu là H_1, còn gọi là đối thuyết (alternative hypothesis). Dạng tổng quát nhất của H_1 là $\theta \neq \theta_0$. Trong nhiều trường hợp giả thuyết đối được phát biểu cụ thể là $H_1 : \theta > \theta_0$ hoặc $H_1 : \theta < \theta_0$.

Như vậy, giả thuyết cơ bản hay giả thuyết đối thường được phát biểu thành cặp:

Giả thuyết H_0	$\theta = \theta_0$	$\theta = \theta_0$	$\theta = \theta_0$
Đối thuyết H_1	$\theta \neq \theta_0$	$\theta > \theta_0$	$\theta < \theta_0$

Nhiệm vụ của lý thuyết kiểm định giả thuyết thống kê là kiểm tra bằng thực nghiệm, thông qua mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$, tính đúng sai của giả thuyết H_0.

5.1.2 Tiêu chuẩn kiểm định. Mức ý nghĩa. Miền bác bỏ

Quy tắc kiểm định dựa trên hai nguyên lý sau:

1. Nguyên lý xác suất nhỏ: "Nếu một sự kiện có xác rất nhỏ thì trong một phép thử sự kiện đó coi như không xảy ra".

2. Phương pháp phản chứng: "Để bác bỏ A ta giả sử A đúng; nếu A đúng dẫn đến một điều vô lý thì bác bỏ A".

Dựa vào hai nguyên lý này ta đưa ra phương pháp chung để kiểm định một giả thuyết thống kê như sau.
Cơ sở lập luận: Giả sử giả thuyết H_0 đúng. Trên cơ sở đó xây dựng một sự kiện A nào đó, sao cho xác suất xảy ra A bằng α bé đến mức có thể sử dụng nguyên lý xác suất nhỏ, tức là có thể coi A không xảy ra trong phép thử về sự kiện này. Thực hiện một phép thử đối với sự kiện A:

1. Nếu A xảy ra thì bác bỏ giả thuyết H_0;

2. Nếu A không xảy ra thì chưa có cơ sở để bác bỏ H_0.

Các bước tiến hành:

Bước 1 Từ biến ngẫu nhiên X, lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ cỡ n và chọn thống kê

$$G(X, \theta) = f(X_1, X_2, \ldots, X_n, \theta) \tag{5.1}$$

sao cho nếu H_0 đúng thì quy luật phân phối xác suất của G hoàn toàn xác định. Thống kê G gọi là tiêu chuẩn kiểm định.

Bước 2 Tìm miền W_α sao cho $P(G \in W_\alpha) = \alpha$ (với giả thuyết H_0 đúng), tức là

$$P(G \in W_\alpha | H_0) = \alpha. \tag{5.2}$$

Vì α nhỏ, nên theo nguyên lý xác suất nhỏ có thể coi G không nhận giá trị trong miền W_α đối với một phép thử.

Bước 3 Thực hiện một phép thử đối với mẫu ngẫu nhiên W_X ta thu được mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_n)$ và tính được giá trị cụ thể của tiêu chuẩn kiểm định G trong (5.1), gọi là giá trị quan sát, ký hiệu là g hay g_{qs}.

Bước 4 Xét xem giá trị quan sát g có thuộc miền W_α hay không để kết luận.

(a) Nếu $g \in W_\alpha$ thì bác bỏ H_0 thừa nhận H_1.

(b) Nếu $g \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Xác suất α gọi là mức ý nghĩa của tiêu chuẩn kiểm định (thông thường yêu cầu $\alpha \le 0,05$). Miền W_α gọi là miền bác bỏ giả thuyết H_0 với mức ý nghĩa α nếu $P(G \in W_\alpha | H_0 = \alpha)$.

Chú ý 5.1. Cùng mức ý nghĩa α đối với một tiêu chuẩn kiểm định G có thể có vô số miền bác bỏ giả thuyết H_0.

5.1.3 Sai lầm loại I. Sai lầm loại II

Sai lầm loại I: Bác bỏ giả thuyết H_0 trong khi H_0 đúng. Xác suất mắc sai lầm này chính bằng α: $P(G \in W_\alpha | H_0) = \alpha$.

Sai lầm loại 1 phát sinh do kích thước mẫu quá nhỏ, do phương pháp lấy mẫu . . .

Sai lầm loại 2: Thừa nhận H_0 trong khi H_0 sai, hay giá trị quan sát g không thuộc miền bác bỏ W_α trong khi H_1 đúng. Xác suất mắc sai lầm loại II là

$$\beta = P(G \notin W_\alpha | H_1) = 1 - P(G \in W_\alpha | H_1). \tag{5.3}$$

Suy ra xác suất bác bỏ giả thuyết H_0 nếu nó sai là $P(G \in W_\alpha | H_1) = 1 - \beta$. Xác suất này gọi là hiệu lực của kiểm định, nó chính là xác suất "không mắc sai lầm loại II".

Các tình huống có thể xảy ra trong kiểm định giả thuyết thống kê được tóm tắt trong bảng dưới đây.

Quyết định \ Thực tế	H_0 đúng	H_0 sai
Bác bỏ H_0	Sai lầm loại I Xác suất bằng α	Quyết định đúng Xác suất bằng $1 - \beta$
Không bác bỏ H_0	Quyết định đúng Xác suất bằng $1 - \alpha$	Sai lầm loại II Xác suất bằng β

Bảng 5.1: Các tình huống có thể xảy ra trong kiểm định giả thuyết thống kê

Mục tiêu là phải cực tiểu cả hai sai lầm. Tuy nhiên, điều đó là khó thực hiện. Người ta tìm cách cố định sai lầm loại I và cực tiểu sai lầm loại II.

Lựa chọn miền bác bỏ để xác suất mắc sai lầm loại 2 là bé nhất: Khi kiểm định giả thuyết thống kê, nếu mức ý nghĩa α đã chọn, cỡ mẫu n đã xác định, vấn đề còn lại là trong vô số miền bác bỏ, ta chọn miền W_α sao cho xác suất mắc sai lầm loại II là nhỏ nhất hay hiệu lực của kiểm định lớn nhất.

Định lý Neymann–Pearson chỉ ra rằng nhiều bài toán quan trọng trong thực tiễn có thể tìm được miền bác bỏ W_α thỏa mãn yêu cầu trên, nghĩa là

$$\boxed{P(G \in W_\alpha | H_0) = \alpha \quad \text{và} \quad P(G \in W_\alpha | H_1) = 1 - \beta \to \max} \tag{5.4}$$

Trong thực hành, quy tắc được xây dựng dưới đây có miền bác bỏ thỏa mãn tính chất trên.

5.1.4 Thủ tục kiểm định giả thuyết thống kê

Qua nội dung trình bày ở trên ta có thể xây dựng một thủ tục kiểm định giả thuyết thống kê bao gồm:

1. Phát biểu giả thuyết H_0 và đối thuyết H_1.

2. Từ tổng thể nghiên cứu lập mẫu ngẫu nhiên kích thước n. Chọn tiêu chuẩn kiểm định G và xác định quy luật phân phối xác suất của G với điều kiện giả thuyết H_0 đúng.

3. Với mức ý nghĩa α, xác định miền bác bỏ giả thuyết H_0 (ký hiệu là W_α) tốt nhất tùy thuộc vào đối thuyết H_1.

4. Từ mẫu cụ thể tính giá trị quan sát g_{qs} của tiêu chuẩn kiểm định.

5. So sánh giá trị quan sát g_{qs} của tiêu chuẩn kiểm định với miền bác bỏ W_α và kết luận.

5.2 Kiểm định giả thuyết về kỳ vọng của biến ngẫu nhiên có phân phối chuẩn

Bài toán 5.1. Giả sử biến ngẫu nhiên gốc X trong tổng thể có phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$, trong đó $E(X) = \mu$ chưa biết nhưng có cơ sở để nêu lên giả thuyết $H_0 : \mu = \mu_0$ với μ_0 là tham số đã biết. Hãy kiểm định giả thuyết này với các thuyết đối $H_1 : \mu \neq \mu_0$ hoặc $\mu > \mu_0$ hoặc $\mu < \mu_0$.

Tiêu chuẩn kiểm định và miền bác bỏ giả thuyết H_0 phụ thuộc các trường hợp sau.

5.2.1 Trường hợp đã biết phương sai

Giả sử phương sai σ^2 của biến ngẫu nhiên gốc X trong tổng thể có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$ đã biết. Từ tổng thể rút ra một mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ kích thước n.

Bước 1 Chọn tiêu chuẩn kiểm định:

$$\boxed{U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}} \tag{5.5}$$

Nếu giả thuyết H_0 đúng thì

$$\boxed{U = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}} \tag{5.6}$$

Theo (4.19) thống kê U có phân phối chuẩn tắc $\mathcal{N}(0; 1)$.

Bước 2 Xây dựng miền bác bỏ W_α phụ thuộc vào thuyết đối H_1.

(a) $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ **(bài toán kiểm định hai phía).** Với mức ý nghĩa α cho trước, giả thuyết H_0 bị bác bỏ nếu $P\Big\{|U| > u_{1-\alpha/2}\Big|(\mu = \mu_0)\Big\} = \alpha$, trong đó $u_{1-\alpha/2}$ được xác định từ hệ thức $\Phi(u_{1-\alpha/2}) = 1 - \alpha/2$. Do đó, miền bác bỏ giả thuyết H_0 là

$$W_\alpha = (-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty).$$

(b) $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ **(bài toán kiểm định một phía).** Với mức ý nghĩa α cho trước, ta tìm giá trị $u_{1-\alpha}$ sao cho $P\Big\{U > u_{1-\alpha}\Big|(\mu = \mu_0)\Big\} = \alpha$ từ bảng giá trị hàm phân phối chuẩn tắc (Phụ lục 3) và xác định được miền bác bỏ giả thuyết H_0 là

$$W_\alpha = (u_{1-\alpha}; +\infty).$$

(c) $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$ **(bài toán kiểm định một phía).** Với mức ý nghĩa α cho trước, ta tìm giá trị $u_{1-\alpha}$ sao cho $P\Big\{U < -u_{1-\alpha}\Big|(\mu = \mu_0)\Big\} = \alpha$ và xác định được miền bác bỏ giả thuyết H_0 là

$$W_\alpha = (-\infty; -u_{1-\alpha}).$$

Tóm lại, miền bác bỏ giả thuyết H_0 được xác định như sau:

H_0	H_1	Miền bác bỏ W_α
$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty)$
$\mu = \mu_0$	$\mu > \mu_0$	$(u_{1-\alpha}; +\infty)$
$\mu = \mu_0$	$\mu < \mu_0$	$(-\infty; -u_{1-\alpha})$

trong đó $u_{1-\alpha/2}$ và $u_{1-\alpha}$ được xác định từ bảng giá trị hàm phân phối chuẩn tắc $\Phi(x)$ (Phụ lục 3).

Bước 3 Lập mẫu cụ thể $W_x = (x_1, x_2, .., x_n)$, tính giá trị quan sát của tiêu chuẩn kiểm định:

$$\boxed{u_{qs} = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}} \tag{5.7}$$

Bước 4 Xét xem u_{qs} có thuộc W_α hay không để kết luận.

(a) Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

(b) Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Ví dụ 5.1. Một hãng bảo hiểm thông báo rằng số tiền trung bình hãng chi trả cho khách hàng bị tai nạn ô tô là 8500 USD. Để kiểm tra lại, người ta kiểm tra ngẫu nhiên hồ sơ chi trả của 25 khách hàng thì thấy số tiền trung bình chi trả là 8900 USD. Giả sử số tiền chi trả tuân theo luật phân phối chuẩn với độ lệch chuẩn là 2600 USD. Hãy kiểm định lại thông báo của hãng bảo hiểm trên với mức ý nghĩa 5%.

Lời giải Ví dụ 5.1 Gọi X là số tiền hãng bảo hiểm chi trả cho khách hàng. $X \sim \mathcal{N}(\mu, \sigma^2)$ với $\sigma = 2600$. Số tiền trung bình hãng chi trả cho khách hàng là $E(X) = \mu$ chưa biết. Đây là bài toán kiểm định giả thuyết về kỳ vọng của biến ngẫu nhiên phân phối chuẩn trường hợp đã biết phương sai.

Bước 1: Đặt giả thuyết $H_0 : \mu = \mu_0$, đối thuyết $H_1 : \mu \neq \mu_0$ với $\mu_0 = 8500$.

Bước 2: Chọn tiêu chuẩn kiểm định $U = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$ nếu giả thuyết H_0 đúng. $U \sim \mathcal{N}(0, 1)$.

Bước 3: Với $\alpha = 0,05$, $u_{1-\alpha/2} = u_{0,975} = 1,96$, tra từ bảng giá trị hàm phân phối chuẩn tắc (Phụ lục 3). Miền bác bỏ giả thuyết H_0 là

$$W_\alpha = (-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty) = (-\infty; -1,96) \cup (1,96; +\infty).$$

Bước 4: Từ số liệu của đầu bài ta có $n = 25$, $\mu_0 = 8500$, $\overline{x} = 8900$, $\sigma = 2600$ suy ra giá trị quan sát

$$u_{qs} = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = \frac{8900 - 8500}{2600}\sqrt{25} \simeq 0,77.$$

Bước 5: Vì $u_{qs} = 0,77 \notin W_\alpha$ nên chưa có cơ sở để bác bỏ giả thuyết H_0. Tức là chưa có cơ sở để bác bỏ thông báo của hãng bảo hiểm với mức ý nghĩa 5%.

Ví dụ 5.2. Nếu máy móc hoạt động bình thường thì trọng lượng sản phẩm là biến ngẫu nhiên có phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$ với trọng lượng trung bình $\mu_0 = 100$ gam, độ lệch tiêu chuẩn $\sigma = 2$ gam. Qua một thời gian sản xuất người ta nghi ngờ trọng lượng sản phẩm có xu hướng tăng lên, cân thử 100 sản phẩm thì trọng lượng trung bình của chúng là 100,4 gam. Với mức ý nghĩa $\alpha = 5\%$ hãy kết luận về điều nghi ngờ trên.

Lời giải Ví dụ 5.2 Gọi X là trọng lượng sản phẩm thì $X \sim \mathcal{N}(\mu, \sigma^2)$ với $\sigma = 2$. Đây là bài toán kiểm định giả thuyết về kỳ vọng của biến ngẫu nhiên phân phối chuẩn trường hợp đã biết phương sai.

Bước 1: Đặt giả thuyết $H_0 : \mu = \mu_0$, đối thuyết $H_1 : \mu > \mu_0$ với $\mu_0 = 100$.

Bước 2: Chọn tiêu chuẩn kiểm định $U = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$ nếu giả thuyết H_0 đúng. $U \sim \mathcal{N}(0, 1)$.

Bước 3: Với $\alpha = 0,05$, $u_{1-\alpha} = u_{0,95} = 1,65$, được tra từ bảng giá trị hàm phân phối chuẩn tắc (Phụ lục 3). Miền bác bỏ giả thuyết H_0 là $W_\alpha = (u_{1-\alpha}; +\infty) = (1,65; +\infty)$.

Bước 4: Từ số liệu đầu bài với $n = 100$, $\mu_0 = 100$, $\sigma = 2$, $\overline{x} = 100,4$ suy ra giá trị quan sát

$$u_{qs} = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = \frac{100,4 - 100}{2}\sqrt{100} = 2.$$

Bước 5: Vì $u_{qs} = 2 \in W_\alpha$ nên bác bỏ giả thuyết H_0. Tức là điều nghi ngờ nói trên là có cơ sở với mức ý nghĩa 5%.

5.2.2 Trường hợp chưa biết phương sai, kích thước mẫu $n < 30$

Bước 1 Chọn tiêu chuẩn kiểm định:

$$\boxed{T = \frac{\overline{X} - \mu}{S}\sqrt{n}} \tag{5.8}$$

Nếu giả thuyết H_0 đúng thì

$$\boxed{T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}} \tag{5.9}$$

Theo (4.21), T có phân phối Student với $n - 1$ bậc tự do.

Bước 2 Miền bác bỏ giả thuyết H_0 được xây dựng phụ thuộc vào thuyết đối H_1 như sau:

H_0	H_1	Miền bác bỏ W_α
$\mu = \mu_0$	$\mu \neq \mu_0$	$\big(-\infty; -t^{(n-1)}_{1-\alpha/2}\big) \cup \big(t^{(n-1)}_{1-\alpha/2}; +\infty\big)$
$\mu = \mu_0$	$\mu > \mu_0$	$\big(t^{(n-1)}_{1-\alpha}; +\infty\big)$
$\mu = \mu_0$	$\mu < \mu_0$	$\big(-\infty; -t^{(n-1)}_{1-\alpha}\big)$

trong đó $t^{(n-1)}_{1-\alpha/2}$ và $t^{(n-1)}_{1-\alpha}$ được xác định từ bảng phân phối Student (Phụ lục 4).

Bước 3 Lập mẫu cụ thể $W_x = (x_1, x_2, .., x_n)$, tính giá trị quan sát của tiêu chuẩn kiểm định:

$$\boxed{t_{qs} = \frac{\overline{x} - \mu_0}{s}\sqrt{n}} \tag{5.10}$$

Bước 4 Xét xem t_{qs} có thuộc W_α hay không để kết luận.

(a) Nếu $t_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

(b) Nếu $t_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Ví dụ 5.3. Một công ty sản xuất hạt giống tuyên bố rằng một loại giống mới của họ có năng suất trung bình là 21,5 tạ/ha. Gieo thử hạt giống mới này tại 16 vườn thí nghiệm và thu được kết quả:

$$19,2; 18,7; 22,4; 20,3; 16,8; 25,1; 17,0; 15,8; 21,0; 18,6; 23,7; 24,1; 23,4; 19,8; 21,7; 18,9.$$

Dựa vào kết quả này hãy xác nhận xem quảng cáo của công ty có đúng không với mức ý nghĩa $\alpha = 0,05$. Biết rằng năng suất giống cây trồng là một biến ngẫu nhiên tuân theo luật phân phối chuẩn.

Lời giải Ví dụ 5.3 Gọi X là năng suất giống cây trồng. $X \sim \mathcal{N}(\mu, \sigma^2)$. Đây là bài toán kiểm định giả thuyết về kỳ vọng của biến ngẫu nhiên phân phối chuẩn trường hợp chưa biết phương sai, mẫu cỡ $n = 16 < 30$.

Bước 1: Đặt giả thuyết $H_0 : \mu = \mu_0$, đối thuyết $H_1 : \mu \neq \mu_0$ với $\mu_0 = 21,5$.

Bước 2: Chọn tiêu chuẩn kiểm định: $T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$ nếu giả thuyết H_0 đúng. $T \sim \mathcal{T}^{(n-1)}$.

Bước 3: Với $\alpha = 0,05$ tra bảng phân phối Student được $t^{(n-1)}_{1-\alpha/2} = t^{(15)}_{0,975} = 2,131$. Miền bác bỏ giả thuyết H_0 là

$$W_\alpha = \big(-\infty; -t^{(n-1)}_{1-\alpha/2}\big) \cup \big(t^{(n-1)}_{1-\alpha/2}; +\infty\big) = (-\infty; -2,131) \cup (2,131; +\infty).$$

Bước 4: Từ số liệu đầu bài tính được $n = 16$, $\overline{x} = 20,406$, $s = 3,038$ với $\mu_0 = 21,5$ suy ra giá trị quan sát

$$t_{qs} = \frac{\overline{x} - \mu_0}{s}\sqrt{n} = \frac{20,406 - 21,5}{3,038}\sqrt{16} = -1,44.$$

Bước 5: Vì $t_{qs} = -1,44 \notin W_\alpha$ nên chưa có cơ sở để bác bỏ giả thuyết H_0, nghĩa là với số liệu này có thể chấp nhận lời quảng cáo của công ty với mức ý nghĩa 5%.

5.2.3 Trường hợp chưa biết phương sai, cỡ mẫu $n \geq 30$

Chú ý 5.2. Như đã biết phân phối Student xấp xỉ phân phối chuẩn khi n khá lớn. Trong thực tế khi $n \geq 30$ coi T có phân phối chuẩn.

Bước 1 Chọn tiêu chuẩn kiểm định:

$$\boxed{U = \frac{\overline{X} - \mu}{S}\sqrt{n}} \tag{5.11}$$

Nếu giả thuyết H_0 đúng thì

$$\boxed{U = \frac{\overline{X} - \mu_0}{S}\sqrt{n}} \tag{5.12}$$

Như đã biết $U \sim \mathcal{N}(0;1)$.

Bước 2 Xây dựng miền bác bỏ giả thuyết H_0 phụ thuộc vào thuyết đối H_1:

H_0	H_1	Miền bác bỏ W_α
$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty)$
$\mu = \mu_0$	$\mu > \mu_0$	$(u_{1-\alpha}; +\infty)$
$\mu = \mu_0$	$\mu < \mu_0$	$(-\infty; -u_{1-\alpha})$

trong đó $u_{1-\alpha/2}$ và $u_{1-\alpha}$ được xác định từ bảng giá trị hàm phân phối chuẩn tắc $\Phi(x)$ (Phụ lục 3).

Bước 3 Lập mẫu cụ thể $W_x = (x_1, \ldots, x_n)$, tính giá trị quan sát của tiêu chuẩn kiểm định:

$$\boxed{u_{qs} = \frac{\overline{x} - \mu_0}{s}\sqrt{n}} \tag{5.13}$$

Bước 4 Xét xem u_{qs} có thuộc W_α hay không để kết luận.

(a) Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

(b) Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Ví dụ 5.4. Một công ty có một hệ thống máy tính có thể xử lý 1200 hóa đơn trong một giờ. Công ty mới nhập một hệ thống máy tính mới. Hệ thống này khi chạy kiểm tra trong 40 giờ cho thấy số hóa đơn được xử lý trung bình trong một giờ là 1260 với độ lệch chuẩn hiệu chỉnh 215. Với mức ý nghĩa 5% hãy nhận định xem hệ thống mới có tốt hơn hệ thống cũ hay không?

Lời giải Ví dụ 5.4 Gọi X là số hóa đơn mà hệ thống máy tính mới xử lý được trong vòng một giờ. Ta thấy $E(X) = \mu$ là số hóa đơn trung bình mà hệ thống máy tính mới xử lý được trong một giờ chưa biết. Đây là bài toán kiểm định giả thuyết về kỳ vọng của biến ngẫu nhiên phân phối chuẩn trường hợp chưa biết phương sai mẫu cỡ $n = 40 > 30$.

Bước 1: Kiểm tra giả thuyết $H_0 : \mu = \mu_0$, đối thuyết $H_1 : \mu > \mu_0$ với $\mu_0 = 1200$.

Bước 2: Chọn tiêu chuẩn kiểm định: $U = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$ nếu H_0 đúng. $U \sim \mathcal{N}(0, 1)$.

Bước 3: Với $\alpha = 0,05$ tra bảng giá trị hàm phân phối chuẩn tắc được $u_{1-\alpha} = u_{0,95} = 1,65$. Miền bác bỏ giả thuyết H_0 là $W_\alpha = (u_{1-\alpha}; +\infty) = (1,65; +\infty)$.

Bước 4: Từ số liệu đầu bài ta có $\mu_0 = 1200$, $n = 40$, $\overline{x} = 1250$, $s = 215$ suy ra giá trị quan sát

$$u_{qs} = \frac{\overline{x} - \mu_0}{s}\sqrt{n} = \frac{1260 - 1200}{215}\sqrt{40} = 1,76.$$

Bước 5: Vì $u_{qs} = 1,76 \in W_\alpha$ nên bác bỏ giả thuyết H_0, nghĩa là với số liệu này có thể coi hệ thống máy mới tốt hơn hệ thống máy cũ với mức ý nghĩa 5%.

Nhận xét 5.1. Nếu tổng thể của biến ngẫu nhiên X không tuân theo quy luật phân phối chuẩn thì ta có thể tiến hành chọn mẫu có kích thước lớn $n \geq 30$, khi đó ta có thể tiến hành kiểm định tương tự như tiến hành kiểm định đối với biến ngẫu nhiên có phân phối chuẩn. Do đó, trong nhiều trường hợp người ta có thể bỏ qua giả thiết chuẩn của biến ngẫu nhiên gốc X (khi mẫu kích thước lớn).

Do đó,

(a) Nếu kích thước mẫu $n < 30$ thì ta phải có điều kiện $X \sim \mathcal{N}(\mu, \sigma^2)$.

(b) Nếu $n \geq 30$ ta có thể bỏ qua giả thiết chuẩn của biến ngẫu nhiên gốc X.

TUẦN 14

5.3 Kiểm định giả thuyết về tỷ lệ hay xác suất

5.3.1 Bài toán

Bài toán 5.2. Giả sử ta quan tâm đến một đặc trưng A nào đó mà mỗi cá thể của tổng thể có thể có tính chất này hoặc không. Gọi p là tần suất có đặc trưng A của tổng thể (p cũng là xác suất cá thể có đặc trưng A của tổng thể). Dấu hiệu nghiên cứu này là một biến ngẫu nhiên X tuân theo luật phân phối Béc-nu-li với kỳ vọng bằng p. Nếu p chưa biết, nhưng có cơ sở để nêu lên giả thuyết

$$H_0 : p = p_0 \quad \text{với} \quad p_0 \text{ là tỷ lệ đã biết.}$$

Hãy kiểm định giả thuyết này với thuyết đối

$$H_1 : p \neq p_0 \quad \text{hoặc} \quad p > p_0 \quad \text{hoặc} \quad p < p_0.$$

Do không biết p nên người ta thực hiện n phép thử độc lập, cùng điều kiện, trong đó có m phép thử xảy ra A. Tần suất mẫu $f = m/n$ là ước lượng điểm không chệch cho p. Ta có f có phân phối xấp xỉ phân phối chuẩn với kỳ vọng $E(f) = p$ và phương sai $V(f) = \dfrac{p(1-p)}{n}$. Từ đó bài toán kiểm định giả thuyết về tỷ lệ không có khác biệt căn bản so với bài toán kiểm định giả thuyết về kỳ vọng.

5.3.2 Các bước tiến hành

Bước 1 Với giả thuyết H_0 đúng xét thống kê

$$\boxed{U = \frac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}} \tag{5.14}$$

Theo (4.22) khi n đủ lớn thống kê (5.14) xấp xỉ phân phối chuẩn tắc $\mathcal{N}(0;1)$. Trong thực tế khi $np_0 \geq 5$ và $n(1-p_0) \geq 5$ thì có thể xem thống kê U trong (5.14) tuân theo luật phân phối chuẩn tắc $\mathcal{N}(0;1)$.

Bước 2 Xây dựng miền bác bỏ giả thuyết H_0 phụ thuộc vào thuyết đối H_1 như sau:

H_0	H_1	Miền bác bỏ W_α
$p = p_0$	$p \neq p_0$	$(-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty)$
$p = p_0$	$p > p_0$	$(u_{1-\alpha}; +\infty)$
$p = p_0$	$p < p_0$	$(-\infty; -u_{1-\alpha})$

trong đó $u_{1-\alpha/2}$ và $u_{1-\alpha}$ được xác định từ bảng giá trị hàm phân phối chuẩn tắc $\Phi(x)$ (Phụ lục 3).

Bước 3 Lập mẫu cụ thể, tính giá trị quan sát của tiêu chuẩn kiểm định:

$$\boxed{u_{qs} = \frac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}, \quad f = \frac{m}{n}} \tag{5.15}$$

Bước 4 Xét xem u_{qs} có thuộc W_α hay không để kết luận.

(a) Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

(b) Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Ví dụ 5.5. Một công ty A sản xuất bánh kẹo tuyên bố rằng $\frac{1}{2}$ số trẻ em thích ăn bánh kẹo của công ty. Trong một mẫu gồm 100 trẻ em được hỏi, có 47 em tỏ ra thích ăn bánh của công ty. Với mức ý nghĩa 5%, số liệu trên có chứng tỏ là tuyên bố của công ty là đúng hay không?

Lời giải Ví dụ 5.5 Gọi p là tỷ lệ trẻ em thích bánh của công ty. Đây là bài toán kiểm định giả thuyết về tỷ lệ của tổng thể.

Bước 1: Đặt giả thuyết $H_0 : p = p_0$, đối thuyết $H_1 : p \neq p_0$ với $p_0 = 0,5$.

Bước 2: Chọn tiêu chuẩn kiểm định $U = \dfrac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$ nếu giả thuyết H_0 đúng.

Vì $np_0 = n(1-p_0) = 100 \times 0,5 = 50$ khá lớn nên $U \sim \mathcal{N}(0,1)$.

Bước 3: Với $\alpha = 0,05$ tra bảng giá trị hàm phân phối chuẩn tắc được $u_{1-\alpha/2} = u_{0,975} = 1,96$.

Miền bác bỏ giả thuyết H_0 là

$$W_\alpha = (-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty) = (-\infty; -1,96) \cup (1,96; +\infty).$$

Bước 4: Từ số liệu đã cho ta có $n = 100$, $m = 47$ tính được $f = \dfrac{m}{n} = \dfrac{47}{100} = 0,47$, với $p_0 = 0,5$ suy ra giá trị quan sát

$$u_{qs} = \frac{f - p_0}{\sqrt{p_0 \times (1-p_0)}}\sqrt{n} = \frac{0,47 - 0,5}{\sqrt{0,5 \times 0,5}}\sqrt{100} = -0,6.$$

Bước 5: Vì $u_{qs} = -0,6 \notin W_\alpha$ nên chưa có cơ sở để bác bỏ giả thuyết H_0 hay tuyên bố của công ty là có cơ sở với mức ý nghĩa 5%.

TÓM TẮT về tiêu chuẩn kiểm định và miền bác bỏ giả thuyết H_0:

Trường hợp	Tiêu chuẩn kiểm định nếu H_0 đúng	H_0	H_1	Miền bác bỏ W_α
σ^2 đã biết	$U = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$	$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty)$
			$\mu > \mu_0$	$(u_{1-\alpha}; +\infty)$
			$\mu < \mu_0$	$(-\infty; -u_{1-\alpha})$
σ^2 chưa biết $n < 30$	$T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$	$\mu = \mu_0$	$\mu \neq \mu_0$	$\big(-\infty; -t^{(n-1)}_{1-\alpha/2}\big) \cup \big(t^{(n-1)}_{1-\alpha/2}; +\infty\big)$
			$\mu > \mu_0$	$\big(t^{(n-1)}_{1-\alpha}; +\infty\big)$
			$\mu < \mu_0$	$\big(-\infty; -t^{(n-1)}_{1-\alpha}\big)$
σ^2 chưa biết $n \geq 30$	$U = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$	$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty)$
			$\mu > \mu_0$	$(u_{1-\alpha}; +\infty)$
			$\mu < \mu_0$	$(-\infty; -u_{1-\alpha})$
$np_0 \geq 5$ $n(1-p_0) \geq 5$	$U = \dfrac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$p = p_0$	$p \neq p_0$	$(-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty)$
			$p > p_0$	$(u_{1-\alpha}; +\infty)$
			$p < p_0$	$(-\infty; -u_{1-\alpha})$

5.4 So sánh hai kỳ vọng của hai biến ngẫu nhiên phân phối chuẩn

Bài toán 5.3. Cho hai biến ngẫu nhiên phân phối chuẩn $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Bài toán đặt ra là cần so sánh giá trị kỳ vọng μ_1 với μ_2:

Giả thuyết H_0	$\mu_1 = \mu_2$	$\mu_1 = \mu_2$	$\mu_1 = \mu_2$
Đối thuyết H_1	$\mu_1 \neq \mu_2$	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$

5.4.1 Trường hợp phương sai σ_1^2, σ_2^2 đã biết

Từ tổng thể rút ra hai mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_{n_1})$, $W_Y = (Y_1, Y_2, \ldots, Y_{n_2})$.

Bước 1 Chọn tiêu chuẩn kiểm định

$$\boxed{U = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}} \tag{5.16}$$

Nếu giả thuyết H_0 đúng thì $\mu_1 - \mu_2 = 0$ và

$$\boxed{U = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}} \tag{5.17}$$

Như đã biết $U \sim \mathcal{N}(0;1)$.

Bước 2 Miền bác bỏ giả thuyết H_0 được xác định phụ thuộc vào thuyết đối H_1 như sau:

H_0	H_1	Miền bác bỏ W_α
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(-\infty; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)$
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$(u_{1-\alpha}; +\infty)$
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$(-\infty; -u_{1-\alpha})$

trong đó $u_{1-\alpha/2}$ và $u_{1-\alpha}$ được xác định từ bảng giá trị hàm phân phối chuẩn tắc $\Phi(x)$ (Phụ lục 3).

Bước 3 Từ mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_{n_1})$, $W_y = (y_1, y_2, \ldots, y_{n_2})$, ta tính được giá trị quan sát của tiêu chuẩn kiểm định:

$$\boxed{u_{qs} = \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}} \tag{5.18}$$

Bước 4 Xét xem u_{qs} có thuộc W_α hay không để kết luận:

(a) Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

(b) Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

5.4.2 Trường hợp phương sai σ_1^2, σ_2^2 chưa biết, cỡ mẫu $n_1 < 30, n_2 < 30$

Giả sử $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Bước 1 Chọn tiêu chuẩn kiểm định:

$$\boxed{T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}}} \tag{5.19}$$

Nếu giả thuyết H_0 đúng thì

$$\boxed{T = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}}} \tag{5.20}$$

Nếu giả thiết hai biến ngẫu nhiên gốc có phương sai giống nhau thì thống kê T trong (5.20) có phân phối Student với $n_1 + n_2 - 2$ bậc tự do, $T \sim \mathcal{T}^{(n_1+n_2-2)}$.

Bước 2 Miền bác bỏ giả thuyết H_0 được xác định phụ thuộc vào thuyết đối H_1 như sau:

H_0	H_1	Miền bác bỏ W_α
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\big(-\infty; -t^{(n_1+n_2-2)}_{1-\frac{\alpha}{2}}\big) \cup \big(t^{(n_1+n_2-2)}_{1-\frac{\alpha}{2}}; +\infty\big)$
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$\big(t^{(n_1+n_2-2)}_{1-\alpha}; +\infty\big)$
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$\big(-\infty; -t^{(n_1+n_2-2)}_{1-\alpha}\big)$

trong đó $t^{(n_1+n_2-2)}_{1-\frac{\alpha}{2}}$ và $t^{(n_1+n_2-2)}_{1-\alpha}$ được xác định từ bảng phân phối Student (Phụ lục 4).

Bước 3 Từ mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_{n_1})$, $W_y = (y_1, y_2, \ldots, y_{n_2})$, ta tính được giá trị quan sát của tiêu chuẩn kiểm định:

$$\boxed{t_{qs} = \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}}} \tag{5.21}$$

Bước 4 Xét xem u_{qs} có thuộc W_α hay không để kết luận:

(a) Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

(a) Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Chú ý 5.3. Trường hợp mẫu cặp (x_i, y_i), $i = 1, 2, \ldots, n$ ta có thể thiết lập hiệu $z_i = x_i - y_i$ và đưa về bài toán kiểm định giả thuyết $H_0 : E(Z) = 0$, $(Z = X - Y)$, đối thuyết $H_1 : E(Z) \neq 0$ hoặc $H_1 : E(Z) > 0$ hoặc $H_1 : E(Z) < 0$.

Ví dụ 5.6. Để so sánh hai chế độ bón phân cho một loại cây trồng A, trên 8 mảnh ruộng người ta chia mỗi mảnh thành hai nửa: nửa thứ nhất áp dụng phương pháp bón phân I, nửa thứ hai theo phương pháp bón phân II (với các chế độ chăm sóc khác nhau). Sau khi thu hoạch ta được số liệu về năng suất loại cây trồng A như sau:

Mảnh	1	2	3	4	5	6	7	8
Năng suất nửa thứ nhất	5	20	16	22	24	14	18	20
Năng suất nửa thứ hai	15	22	14	25	29	16	20	24

Đánh giá xem hai chế độ bón phân có giống nhau không với mức ý nghĩa 1%. Biết rằng năng suất loại cây trồng A (sau hai phương pháp bón phân) có phân phối chuẩn và có cùng phương sai.

Lời giải Ví dụ 5.6

Cách 1: Gọi X, Y lần lượt là năng suất loại cây trồng A ở nửa thứ nhất, thứ hai (sử dụng hai phương pháp bón phân tương ứng). Ta có $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Đây là bài toán so sánh hai kỳ vọng của hai biến ngẫu nhiên có phân phối chuẩn trường hợp chưa biết phương sai, mẫu cỡ $n_1 = n_2 = 8 < 30$ và $\sigma_1^2 = \sigma_2^2$.

Bước 1: Đặt giả thuyết $H_0 : \mu_1 = \mu_2$, đối thuyết $H_1 : \mu_1 \neq \mu_2$.

Bước 2: Chọn tiêu chuẩn kiểm định $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$ nếu H_0 đúng.

Vì X và Y có cùng phương sai, nên $T \sim \mathcal{T}^{(n_1+n_2-2)}$.

Bước 3: Với $\alpha = 0,05$ tra bảng phân phối Student được $t_{1-\frac{\alpha}{2}}^{(n_1+n_2-2)} = t_{0,995}^{(14)} = 2,977$. Miền bác bỏ giả thuyết H_0 là

$$W_\alpha = \left(-\infty; -t_{1-\frac{\alpha}{2}}^{(n_1+n_2-2)}\right) \cup \left(t_{1-\frac{\alpha}{2}}^{(n_1+n_2-2)}; +\infty\right) = (-\infty; -2,977) \cup (2,977; +\infty).$$

Bước 4: Từ số liệu đã cho tính được $n_1 = n_2 = 8$, $\overline{x} = 17,375$, $s_1^2 = 35,125$, $\overline{y} = 20,625$, $s_2^2 = 28,5535$ suy ra giá trị quan sát

$$t_{qs} = \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} = \frac{-3,25}{2,6391} = -1,2314.$$

Bước 5: Vì $t_{qs} = -1,2314 \notin W_\alpha$ nên chưa có cơ sở để bác bỏ giả thuyết H_0, hay có thể xem hai phương pháp bón phân cho kết quả như nhau với mức ý nghĩa 1%.

Cách 2: Đặt $Z = X - Y$, thiết lập hiệu $z_i = x_i - y_i$, $i = 1, \ldots, 8$ với

x_i	5	20	16	22	24	14	18	20
y_i	15	22	14	25	29	16	20	24
z_i	-10	-2	2	-3	-4	-2	-2	-2

Ta đưa về bài toán kiểm định giả thuyết $H_0 : \mu_Z = 0$, đối thuyết $H_1 : \mu_Z \neq 0$, $\mu_Z = E(Z)$.

5.4.3 Trường hợp phương sai σ_1^2, σ_2^2 chưa biết, cỡ mẫu $n_1 \geq 30, n_2 \geq 30$

Bước 1 Chọn tiêu chuẩn kiểm định:

$$\boxed{U = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}} \tag{5.22}$$

Nếu giả thuyết H_0 đúng thì $\mu_1 - \mu_2 = 0$ và

$$\boxed{U = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}} \tag{5.23}$$

Như đã biết $U \sim \mathcal{N}(0; 1)$.

Bước 2 Miền bác bỏ giả thuyết H_0 được xác định cho ba trường hợp như sau:

H_0	H_1	Miền bác bỏ W_α
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$(-\infty; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)$
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$(u_{1-\alpha}; +\infty)$
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$(-\infty; -u_{1-\alpha})$

trong đó $u_{1-\alpha/2}$ và $u_{1-\alpha}$ được xác định từ bảng giá trị hàm phân phối chuẩn tắc $\Phi(x)$ (Phụ lục 3).

Bước 3 Từ mẫu cụ thể $W_x = (x_1, x_2, \ldots, x_{n_1})$, $W_y = (y_1, y_2, \ldots, y_{n_2})$, ta tính được giá trị quan sát của tiêu chuẩn kiểm định:

$$\boxed{u_{qs} = \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}} \tag{5.24}$$

Bước 4 Xét xem u_{qs} có thuộc W_α hay không để kết luận.

(a) Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

(b) Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Ví dụ 5.7. Hai máy tự động dùng để cắt những thanh kim loại do cùng một kỹ thuật viên phụ trách và căn chỉnh. Từ mỗi máy lấy ra 31 thanh kim loại để kiểm tra và thu được kết quả sau:

Máy 1: Trung bình mẫu là 12 cm, độ lệch chuẩn hiệu chỉnh là 1,2 cm.

Máy 2: Trung bình mẫu là 12,3 cm, độ lệch hiệu chỉnh là 1,4 cm.

Với mức ý nghĩa $\alpha = 0,01$ có thể cho rằng chiều dài của các thanh kim loại do máy 2 sản xuất khác chiều dài do máy 1 sản xuất hay không. Biết chiều dài thanh kim loại do các máy sản xuất có phân phối chuẩn.

Lời giải Ví dụ 5.7 Gọi X, Y lần lượt là chiều dài các thanh kim loại do máy 1, 2 sản xuất. Khi đó $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Đây là bài toán so sánh hai kỳ vọng của hai biến ngẫu nhiên có phân phối chuẩn trường hợp chưa biết phương sai, mẫu cỡ $n_1 = n_2 = 31 > 30$.

Bước 1: Đặt giả thuyết $H_0 : \mu_1 = \mu_2$, đối thuyết $H_1 : \mu_1 \neq \mu_2$.

Bước 2: Chọn tiêu chuẩn kiểm định $U = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ nếu giả thuyết H_0 đúng. $U \sim \mathcal{N}(0, 1)$.

Bước 3: Với $\alpha = 0,01$ tra bảng giá trị hàm phân phối chuẩn tắc được $u_{1-\frac{\alpha}{2}} = u_{0,995} = 2,58$. Miền bác bỏ giả thuyết H_0 là

$$W_\alpha = (-\infty; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty) = (-\infty; -2,58) \cup (2,58; +\infty).$$

Bước 4: Từ số liệu đã cho ta có $n_1 = n_2 = 31$, $\overline{x} = 12$, $s_1 = 1,2$, $\overline{y} = 12,3$, $s_2 = 1,4$, suy ra giá trị quan sát

$$u_{qs} = \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{12 - 12,3}{\sqrt{\dfrac{1,44}{31} + \dfrac{1,94}{31}}} = -0,9085.$$

Bước 5: Vì $u_{qs} = -0,9085 \notin W_\alpha$ nên chưa có cơ sở để bác bỏ giả thuyết H_0, hay có thể xem chiều dài các thanh kim loại do hai nhà máy sản xuất là như nhau với mức ý nghĩa 1%.

Chú ý 5.4. (a) Nếu cỡ mẫu n_1, n_2 nhỏ thì ta phải thêm giả thuyết biến ngẫu nhiên gốc tuân theo phân phối chuẩn; nếu n_1 và n_2 khá lớn ta có thể bỏ giả thiết chuẩn của đầu bài.

(b) Hai đối thuyết $\mu_1 > \mu_2$ và $\mu_1 < \mu_2$ dễ dàng chuyển đổi cho nhau bằng cách thay đổi thứ tự của hai mẫu.

5.5 So sánh hai tỷ lệ

5.5.1 Bài toán

Bài toán 5.4. Giả sử p_1, p_2 tương ứng là tỷ lệ các phần tử mang dấu hiệu A nào đó của tổng thể thứ nhất và tổng thể thứ hai. Mẫu của tổng thể thứ nhất: Thực hiện n_1 phép thử độc lập cùng điều kiện, có m_1 phép thử xảy ra sự kiện A. Mẫu của tổng thể thứ hai: Thực hiện n_2 phép thử độc lập cùng điều kiện, có m_2 phép thử xảy ra sự kiện A. Hãy so sánh p_1 với p_2.
Cặp giả thuyết đặt ra là:

Giả thuyết H_0	$p_1 = p_2$	$p_1 = p_2$	$p_1 = p_2$
Đối thuyết H_1	$p_1 \neq p_2$	$p_1 > p_2$	$p_1 < p_2$

5.5.2 Các bước tiến hành

Đặt

$$\overline{f} = \frac{m_1 + m_2}{n_1 + n_2}. \tag{5.25}$$

Bước 1 Chọn tiêu chuẩn kiểm định:

$$\boxed{U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\overline{f}(1 - \overline{f})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}} \tag{5.26}$$

Nếu giả thuyết H_0 đúng thì $p_1 = p_2$ và

$$\boxed{U = \frac{f_1 - f_2}{\sqrt{\overline{f}(1 - \overline{f})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}} \tag{5.27}$$

Ta có $U \sim \mathcal{N}(0;1)$.

Bước 2 Miền bác bỏ giả thuyết H_0 được xác định phụ thuộc vào thuyết đối H_1 như sau:

H_0	H_1	Miền bác bỏ W_α
$p_1 = p_2$	$p_1 \neq p_2$	$(-\infty; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)$
$p_1 = p_2$	$p_1 > p_2$	$(u_{1-\alpha}; +\infty)$
$p_1 = p_2$	$p_1 < p_2$	$(-\infty; -u_{1-\alpha})$

trong đó $u_{1-\alpha/2}$ và $u_{1-\alpha}$ được xác định từ bảng giá trị hàm phân phối chuẩn tắc $\Phi(x)$ (Phụ lục 3).

Bước 3 Từ mẫu thu thập, ta tính được giá trị quan sát của tiêu chuẩn kiểm định:

$$\boxed{u_{qs} = \frac{f_1 - f_2}{\sqrt{\overline{f}(1-\overline{f})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}} \tag{5.28}$$

với $f_1 = \dfrac{m_1}{n_1}$, $f_2 = \dfrac{m_2}{n_2}$, $\overline{f} = \dfrac{m_1 + m_2}{n_1 + n_2} = \dfrac{n_1.f_1 + n_2.f_2}{n_1 + n_2}$.

Bước 4 Xét xem u_{qs} có thuộc W_α hay không để kết luận.

(a) Nếu $u_{qs} \in W_\alpha$ thì bác bỏ giả thuyết H_0.

(b) Nếu $u_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

Ví dụ 5.8. Từ kho đồ hộp thứ nhất lấy ngẫu nhiên 1000 hộp để kiểm tra thấy có 20 hộp bị hỏng. Từ kho đồ hộp thứ hai lấy ngẫu nhiên 900 hộp kiểm tra thấy 30 hộp bị hỏng. Hỏi chất lượng bảo quản của 2 kho có thực sự giống nhau hay không với mức ý nghĩa 5%.

Lời giải Ví dụ 5.8 Gọi p_1, p_2 lần lượt là tỷ lệ hộp hỏng ở kho đồ hộp thứ nhất và thứ hai tương ứng. Đây là bài toán so sánh hai tỷ lệ.

Bước 1: Đặt giả thuyết $H_0 : p_1 = p_2$, đối thuyết $H_1 : p_1 \neq p_2$.

Bước 2: Chọn tiêu chuẩn kiểm định $U = \dfrac{f_1 - f_2}{\sqrt{\overline{f}(1-\overline{f})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ nếu giả thuyết H_0 đúng. Ta thấy $U \sim \mathcal{N}(0,1)$.

Bước 3: Với $\alpha = 0,05$ tra bảng giá trị hàm phân phối chuẩn tắc được $u_{1-\alpha/2} = 1,96$. Miền bác bỏ giả thuyết H_0 là:

$$W_\alpha = (-\infty; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty) = (-\infty; -1,96) \cup (1,96; +\infty).$$

Bước 4: Theo đầu bài $n_1 = 1000$, $n_2 = 900$, $m_1 = 20$, $m_2 = 30$, $f_1 = \dfrac{2}{100}$, $f_2 = \dfrac{3}{90}$, $\overline{f} = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \dfrac{20+30}{1900} = \dfrac{5}{190}$, suy ra

$$u_{qs} = \frac{f_1 - f_2}{\sqrt{\overline{f}(1-\overline{f})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}} = -1,8129.$$

Bước 5: Kết luận: Vì $u_{qs} = -1,8129 \notin W_\alpha$ nên chưa có cơ sở để bác bỏ giả thuyết H_0, nghĩa là có thể xem chất lượng bảo quản của hai kho hàng là như nhau với mức ý nghĩa 5%.

Ví dụ 5.9. Một bệnh viện điều trị loại bệnh A theo hai phương pháp. Sau một thời gian thấy kết quả như sau:

Trong 102 bệnh nhân điều trị phương pháp I có 82 bệnh nhân khỏi bệnh.

Trong 98 bệnh nhân điều trị phương pháp II có 69 bệnh nhân khỏi bệnh.

Hỏi có phải phương pháp I điều trị tốt hơn phương pháp II hai hay không với mức ý nghĩa 5%.

Lời giải Ví dụ 5.9 Gọi p_1, p_2 lần lượt là tỷ lệ bệnh nhân khỏi bệnh khi điều trị bằng phương pháp I và II tương ứng. Đây là bài toán so sánh hai tỷ lệ.

Bước 1: Đặt giả thuyết $H_0 : p_1 = p_2$, đối thuyết $H_1 : p_1 > p_2$.

Bước 2: Chọn tiêu chuẩn kiểm định $U = \dfrac{f_1 - f_2}{\sqrt{\overline{f}(1-\overline{f})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}}$ nếu giả thuyết H_0 đúng. Ta thấy $U \sim \mathcal{N}(0,1)$.

Bước 3: Với $\alpha = 0,05$ tra bảng giá trị hàm phân phối chuẩn tắc được $u_{1-\alpha} = 1,65$. Miền bác bỏ giả thuyết H_0 là $W_\alpha = (u_{1-\alpha}; +\infty) = (1,65; +\infty)$.

Bước 4: Theo đầu bài $n_1 = 102$, $n_2 = 98$, $m_1 = 82$, $m_2 = 69$, $f_1 = \dfrac{82}{102}$, $f_2 = \dfrac{69}{98}$, $\overline{f} = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \dfrac{82+69}{102+98} = \dfrac{151}{200}$, suy ra

$$u_{qs} = \frac{f_1 - f_2}{\sqrt{\overline{f}(1-\overline{f})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}} = 1,641.$$

Bước 5: Vì $u_{qs} = 1,641 \notin W_\alpha$ nên chưa có cơ sở để bác bỏ giả thuyết H_0, nghĩa là chưa thể xem phương pháp I điều trị tốt hơn phương pháp II với mức ý nghĩa 5%.

Ví dụ 5.10 (Đề thi cuối kỳ 20191)**.** Để điều tra doanh thu của các gia đình kinh doanh loại mặt hàng A tại địa phương B, người ta khảo sát 100 gia đình kinh doanh loại mặt hàng này trong một tháng của năm 2019 thu được bảng số liệu

Doanh thu (triệu VNĐ)	25	30	35	40	45	50	55	60	65
Số gia đình	4	9	17	25	20	10	8	4	3

(a) Với độ tin cậy 95% hãy ước lượng doanh thu trung bình/tháng của các gia đình kinh doanh loại mặt hàng A tại địa phương B.

(b) Một tài liệu thống kê cho biết doanh thu trung bình/tháng của các gia đình kinh doanh loại mặt hàng A tại địa phương B là 40 triệu VNĐ. Hãy cho kết luận về tài liệu nói trên với mức ý nghĩa 5%.

(c) Điều tra doanh thu của 200 gia đình kinh doanh loại mặt hàng A ở địa phương C người ta tính được doanh thu trung bình/tháng là 43 triệu VNĐ và độ lệch tiêu chuẩn mẫu hiệu chỉnh là 8,912 triệu VNĐ. Doanh thu trung bình loại mặt hàng A ở địa phương C và B (với số liệu ở Câu 4) có như nhau hay không? Hãy kết luận với mức ý nghĩa 1%.

Lời giải Ví dụ 5.10

(a) Gọi X(triệu VNĐ/tháng) là biến ngẫu nhiên chỉ doanh thu của các gia đình kinh doanh mặt hàng A của địa phương B. Ký hiệu $E(X) = \mu_X$. Đây là bài toán ước lượng khoảng của kỳ vọng trường hợp chưa biết phương sai, mẫu cỡ $n = 100 > 30$.

Chọn thống kê $U = \dfrac{\overline{X} - \mu_X}{S_X}\sqrt{n}$. Thống kê $U \sim \mathcal{N}(0,1)$.

Áp dụng khoảng tin cậy đối xứng $\left(\overline{x} - u_{1-\frac{\alpha}{2}}\dfrac{s_X}{\sqrt{n}}\ ;\ \overline{x} + u_{1-\frac{\alpha}{2}}\dfrac{s_X}{\sqrt{n}}\right)$.

Với $\gamma = 95\%$, $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$. Từ bảng số liệu tính được $n = 100$, $\overline{x} = 42,4$, $s_{\mathcal{X}} = 9,2245$, suy ra khoảng tin cậy cần tìm $(40,592\ ;\ 44,208)$.

Vậy doanh thu trung bình của các gia đình kinh doanh loại mặt hàng A tài địa phương B là từ 40,592 triệu đồng/tháng đến 44,208 triệu đồng/tháng với độ tin cậy 95%.

(b) Đây là bài toán kiểm định giả thuyết của kỳ vọng trường hợp chưa biết phương sai, mẫu cỡ $n = 100 > 30$.

Kiểm định cặp giả thuyết $H_0 : \mu_X = \mu_0$, $H_1 : \mu_X \neq \mu_0$, $\mu_0 = 40$.

Chọn thống kê $U = \dfrac{\overline{X} - \mu_0}{S_X}\sqrt{n}$. Thống kê $U \sim \mathcal{N}(0,1)$.

Với $\alpha = 5\%$, miền bác bỏ H_0 là $W_\alpha = (-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2;+\infty}) = (-\infty; -1,96) \cup (1,96; +\infty)$.

Từ bảng số liệu tính được $n = 100$, $\overline{x} = 42,4$, $s_{\mathcal{X}} = 9,2245$, suy ra giá trị quan sát $u_{qs} = \dfrac{\overline{x} - \mu_0}{s_X}\sqrt{n} = \dfrac{42,4 - 40,0}{9,2245} \times 10 \simeq 2,6018$.

Vì $u_{qs} = 2,6018 \in W_\alpha$ nên bác bỏ H_0, chấp nhận H_1, nghĩa là tài liệu thống kê chưa có cơ sở với mức ý nghĩa 5%.

(c) Gọi Y(triệu đồng/tháng) là biến ngẫu nhiên chỉ doanh thu của các gia đình ở địa phương C. Ký hiệu $E(Y) = \mu_Y$. Đây là bài toán so sánh hai giá trị trung bình của hai tổng thể trường hợp chưa biết phương sai, mẫu cỡ $n = 100 > 30$ và $m = 200 > 30$.

Kiểm định cặp giả thuyết $H_0 : \mu_X = \mu_Y$, $H_1 : \mu_X \neq \mu_Y$.

Chọn thống kê $U = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$. Thống kê $U \sim \mathcal{N}(0,1)$ khi H_0 đúng.

Với $\alpha = 5\%$, miền bác bỏ H_0 là $W_\alpha = (-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2;+\infty}) = (-\infty; -2,575) \cup (2,575; +\infty)$.

Từ bảng số liệu tính được $n = 100$, $\overline{x} = 42,4$, $s_{\mathcal{X}} = 9,2245$, $m = 200$, $\overline{y} = 43$, $s_{\mathcal{Y}} = 8,912$ suy ra giá trị quan sát $u_{qs} = \dfrac{\overline{x} - \overline{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \simeq -0,5371$.

Vì $u_{qs} = -0,5371 \notin W_\alpha$ nên chấp nhận H_0, nghĩa là doanh thu trung bình mặt hàng A ở địa phương C và B là như nhau mức ý nghĩa 5%.