BÀI GIẢNG XÁC SUẤT VÀ THỐNG KÊ TOÁN HỌC

THỐNG KÊ TOÁN HỌC

I. MỞ ĐẦU VỀ THỐNG KÊ

1. Tổng thể nghiên cứu và mẫu

• **Tổng thể hoặc tập nền** là **tập hợp các phần tử (hay cá thể) cần nghiên cứu tính chất định tính hoặc định lượng nào đó.** Số phần tử của tổng thể được gọi là **cỡ của tổng thể.**

Ví dụ:

a) Điều tra ngẫu nhiên về thu nhập của nữ giới trong độ tuổi lao động của một nước.

Dấu hiệu nghiên cứu: Thu nhập cá nhân. Phần tử cần nghiên cứu: Các nữ giới trong độ tuổi lao động.

b) Điều tra ngẫu nhiên về nhu cầu sử dụng nước của một hộ gia đình trong khu vực $\boldsymbol{X}$.

Dấu hiệu nghiên cứu: nhu cầu sử dụng nước. Phần tử cần nghiên cứu: Các hộ gia đình trong khu vực $\boldsymbol{X}$.

c) Kiểm tra ngẫu nhiên về chất lượng các sản phẩm của một nhà máy.

Dấu hiệu nghiên cứu: chất lượng của sản phẩm. Phần tử cần nghiên cứu: Các sản phẩm của nhà máy.

• **Biến nghiên cứu:** Do các phần tử của tổng thể có thể có một hoặc cả hai loại dấu hiệu nghiên cứu định tính hoặc định lượng, nên ta có hai loại biến nghiên cứu như sau:

+) **Biến định lượng là các số đo của phần tử.**

Ví dụ: Chiều cao, cận nặng, thu nhập, ...

+) **Biến định tính là tính chất nào đó của đối tượng nghiên cứu.**

Ví dụ: Giới tính, chất lượng, dân tộc,...

• **Mã hóa các biến nghiên cứu:** Ta có các cách mã hóa biến như sau:

+) *Mã hóa biến định lượng:* Ta lấy giá trị của biến định lượng làm mã của biến.

+) *Mã hóa biến định tính:* Ta gán tính chất định tính của biến ứng với các số nguyên.

Do vậy, **các biến nghiên cứu đều được chuyển thành các biến ngẫu nhiên. Việc quan sát trên mỗi phần tử của tổng thể chính là đang quan sát một hoặc một số biến ngẫu nhiên nào đó.**

Ví dụ:

a) Điều tra ngẫu nhiên về giới tính của người dân trong một nước. Ta có hai giới tính là: nữ và nam.

Ta mã hóa thành biến ngẫu nhiên $\boldsymbol{X}$ như sau: nếu kết quả chọn được một người có giới tính "Nam" thì gán $\boldsymbol{X}$ nhận giá trị là $\mathbf{1}$; nếu kết quả chọn được một người có giới tính "Nữ" thì gán $\boldsymbol{X}$ nhận giá trị là $\mathbf{0}$.

b) Điều tra ngẫu nhiên về thu nhập của một hộ gia đình trong khu vực $\boldsymbol{T}$. Ta xem xét các mức sau: nghèo, trung bình và giàu.

Ta mã hóa thành biến ngẫu nhiên $\boldsymbol{Y}$ như sau: nếu kết quả chọn được một hộ nghèo thì gán $\boldsymbol{Y}$ nhận giá trị là $\mathbf{-1}$; nếu kết quả chọn được một hộ trung bình thì gán $\boldsymbol{Y}$ nhận giá trị là $\mathbf{0}$; nếu kết quả chọn được một hộ giàu thì gán $\boldsymbol{Y}$ nhận giá trị là $\mathbf{1}$.

• **Mẫu:** Trong quá trình nghiên cứu tổng thể, ta có thể không thể nghiên cứu cặn kẽ từng phần tử của tổng thể, do đó ta phải hạn chế trên một nhóm nhỏ hơn được rút ra từ tổng thể, nhóm nhỏ này được gọi là **mẫu**, và từ đó rút ra kết luận cho tổng thể, do vậy ta mong muốn mẫu đại diện tốt nhất cho tổng thể.

Nói chung, để có được một mẫu đại diện tốt nhất cho tổng thể người ta thường phải tiến hành xây dựng mẫu theo một quy trình chọn ngẫu nhiên các phần tử của mẫu. Một mẫu như vậy được gọi là **mẫu ngẫu nhiên.**

Các phương pháp lấy mẫu ngẫu nhiên: Có rất nhiều phương pháp chọn mẫu để thỏa mãn tính đại diện tốt nhất cho tổng thể và phù hợp mục tiêu nghiên cứu. Tuy nhiên, ta chỉ nghiên cứu một số phương pháp chủ yếu sau:

+) **Lấy mẫu ngẫu nhiên đơn giản:** Mỗi cá thể của tổng thể được chọn một cách độc lập với xác suất như nhau.

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

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

Ví dụ: Kiểm tra chất lượng sản phẩm đã đóng thùng của nhà máy cơ khí.

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

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

Ví dụ: Một trường đại học có 4 hệ đào tạo: Chính quy, liên thông, văn bằng hai và sau đại học. Để khảo sát về chất lượng và độ hài lòng của người học, ta có thể tiến hành như sau: Coi mỗi hệ đào tạo là một tầng; sau đó thực hiện lấy mẫu ngẫu nhiên ở mỗi tầng.

2. Mẫu ngẫu nhiên

• **Giả sử tổng thể cần nghiên cứu về biến ngẫu nhiên X. Ta quan sát n lần độc lập về X. Gọi X_i là việc quan sát lần i của X. Khi đó, bộ n quan sát $W = (X_1, X_2, \cdots, X_n)$ xác định một mẫu ngẫu nhiên cỡ n cảm sinh từ X.**

+) **Cho một mẫu ngẫu nhiên $W = (X_1, X_2, \cdots, X_n)$ có cỡ n cảm sinh từ biến ngẫu nhiên X.**

- **Ta thấy rằng thực chất $W = (X_1, X_2, \cdots, X_n)$ là n biến ngẫu nhiên độc lập, và có cùng phân phối xác suất với X.**

- **Nếu ta gọi x_i là kết quả quan sát lần i. Khi đó ta gọi bộ giá trị $(x_1, x_2, \cdots, x_n)$ là một mẫu dữ liệu mà $W = (X_1, X_2, \cdots, X_n)$ nhận.**

Mẫu số liệu có thể được biểu diễn dưới dạng điểm hoặc dạng khoảng.

Ví dụ:

Nghiên cứu thời gian hoạt động ($\boldsymbol{X}$) của các bóng đèn do một công ty sản xuất. Ta lấy ngẫu nhiên $\boldsymbol{n}$ bóng đèn để nghiên cứu.

Khi đó, ta thu được một mẫu ngẫu nhiên $\boldsymbol{(X_1, X_2, ..., X_n)}$ cỡ $\boldsymbol{n}$, trong đó $\boldsymbol{X_i}$ là quan sát thứ $\boldsymbol{i}$ về thời gian hoạt động của bóng đèn; với $\boldsymbol{1 \leq i \leq n.}$

Các giá trị mẫu có dạng $\boldsymbol{(x_1, x_2, ..., x_n),}$ trong đó $\boldsymbol{x_i}$ là giá trị quan sát thứ $\boldsymbol{i}$ về thời gian hoạt động của bóng đèn; với $\boldsymbol{1 \leq i \leq n.}$

Mô tả mẫu dữ liệu của mẫu ngẫu nhiên: **Một mẫu dữ liệu của mẫu ngẫu nhiên cỡ n thường được trình bày số liệu như sau:**

+) **Cách 1:** Liệt kê tất cả các giá trị của mẫu dữ liệu và ghi thành dãy $x_1, x_2, \cdots, x_n$.

+) **Cách 2:** Nếu các giá trị của mẫu gồm k giá trị có thể có là $x_1, x_2, \cdots, x_k$ với số lần xuất hiện (hay tần số) trong mẫu lần lượt là $n_1, n_2, \cdots, n_k$ thì ta mô tả mẫu dữ liệu dưới dạng **bảng tần số** như sau:

Giá trị x_i	x_1	x_2	$\cdots$	x_k
Tần số n_i	n_1	n_2	$\cdots$	n_k

trong đó, các giá trị $x_1, x_2, \cdots, x_k$ đôi một khác nhau; n_i là số lần xuất hiện x_i (hay tần số xuất hiện x_i) trong mẫu dữ liệu; và $n_1 + n_2 + \cdots + n_k = n$.

- Nếu trong bảng tần số ở trên, ta thay $\boldsymbol{n_i}$ bởi $\boldsymbol{f_i = \dfrac{n_i}{n}}$ (tần suất xuất hiện $\boldsymbol{x_i}$ trong mẫu) thì ta mô tả mẫu dữ liệu dưới dạng **bảng tần suất** như sau:

Giá trị $\boldsymbol{x_i}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{\cdots}$	$\boldsymbol{x_k}$
Tần suất $\boldsymbol{f_i}$	$\boldsymbol{f_1}$	$\boldsymbol{f_2}$	$\boldsymbol{\cdots}$	$\boldsymbol{f_k}$

trong đó, $\boldsymbol{f_i = \frac{n_i}{n}}$, được gọi là tần suất xuất hiện $\boldsymbol{x_i}$ trong mẫu dữ liệu và $\boldsymbol{f_1 + f_2 + \ldots + f_k = 1}$.

- **Hàm phân phối (phân bố) xác suất thực nghiệm của mẫu** được cho bởi công thức

$$\boldsymbol{F_n(x) = \sum_{x_j < x} f_j, \forall x \in \mathbb{R};}$$

ở đây $\boldsymbol{f_j}$ là tần suất xuất hiện $\boldsymbol{x_j}$ trong mẫu.

Chú ý: Khi $\boldsymbol{n}$ đủ lớn, $\boldsymbol{F_n(x)}$ xấp xỉ hàm phân phối xác suất $\boldsymbol{F_X(x)}$ của $\boldsymbol{X}$.

+) **Cách 3:** Ta có thể gộp các giá trị thành các nhóm để có thể nhận ra sự phân bố một cách dễ dàng hơn. Các nhóm thường có dạng giá trị trong một khoảng hoặc một đoạn nào đó.

Chẳng hạn, các giá trị của mẫu dữ liệu được xem xét nằm trong k khoảng $[a_0, a_1), [a_1, a_2), ..., [a_{k-1}, a_k)$. Ta ký hiệu n_i là số các giá trị của mẫu dữ liệu rơi vào khoảng $[a_{i-1}, a_i)$, với $i = 1, 2, ..., k$. Khi đó ta mô tả mẫu dữ liệu dưới dạng bảng ghép lớp như sau:

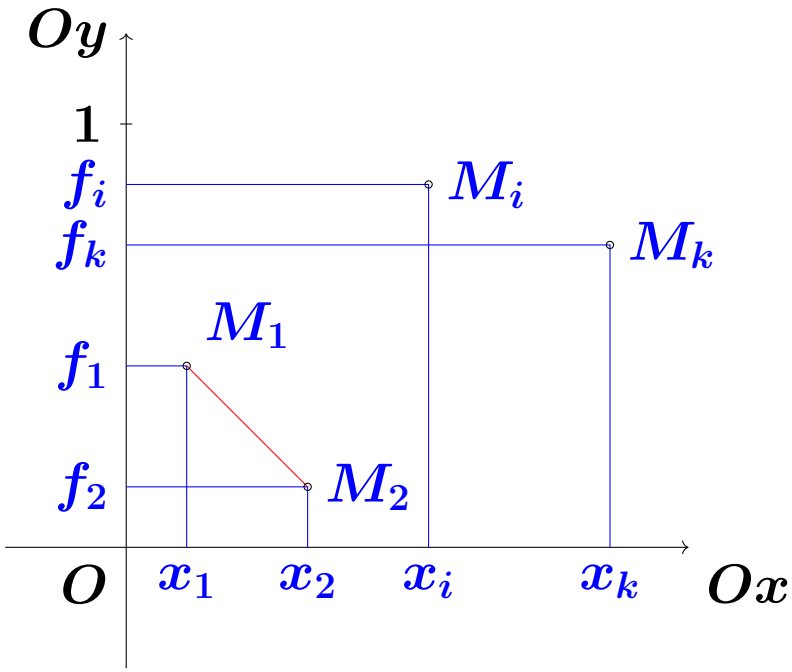
Giá trị thuộc khoảng $[a_{i-1}, a_i)$	$[a_0, a_1)$	$[a_1, a_2)$	$\cdots$	$[a_{k-1}, a_k)$
Tần số n_i	n_1	n_2	$\cdots$	n_k

trong đó, $n_1 + n_2 + \cdots + n_k = n$.

Lưu ý: Đôi khi chúng ta viết tắt $[a_i, a_{i+1})$ thành $(a_i - a_{i+1})$.

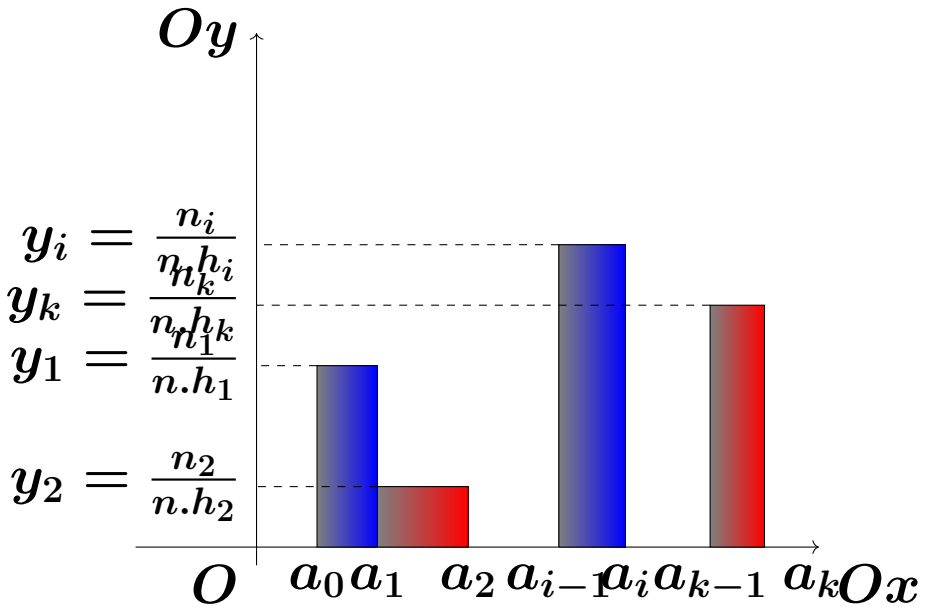
Biểu diễn trực quan giá trị của mẫu ngẫu nhiên

+) **Đa giác tần số, tần suất:** Khi mẫu dữ liệu cho ở dạng bảng tần suất, trong hệ tọa độ (Oxy), ta chấm các điểm $M_i(x_i, f_i)$ với $i = 1, \cdots, k$. Kẻ các đoạn thẳng M_iM_{i+1} với $1 \leq i \leq k$ thì ta thu được **đa giác tần suất.**



Trong biểu diễn trên, thay tần suất f_i bởi tần số n_i, thì ta thu được **đa giác tần số.**

+) **Tổ chức đồ:** Khi mẫu dữ liệu được cho ở dạng bảng ghép lớp, ta có biểu diễn như sau: Trong mặt phẳng (Oxy), trên trục Ox, biểu diễn các khoảng $[a_{i-1}, a_i)$, trên trục Oy biểu diễn các giá trị $y_i = \frac{n_i}{n.h_i}$, với $h_i = a_i - a_{i-1}$ là độ dài khoảng $[a_{i-1}, a_i)$. Trên mỗi đoạn $[a_{i-1}, a_i]$ trong trục Ox, dựng hình chữ nhật có một cạnh là đoạn $[a_{i-1}, a_i]$ và cạnh còn lại có độ dài là y_i. Hình thu được tạo bởi các hình chữ nhật này được gọi là **tổ chức đồ.**



3. Đại lượng thống kê và một số đặc trưng đặc biệt của mẫu ngẫu nhiên

• Cho một mẫu ngẫu nhiên $W = (X_1, X_2, \cdots, X_n)$ cảm sinh từ biến ngẫu nhiên gốc X.

- Hàm $T = T(X_1, X_2, \cdots, X_n)$ phụ thuộc các giá trị của mẫu được gọi là một đại lượng thống kê (hay gọi tắt là thống kê).

+) Ta thấy rằng T là một biến ngẫu nhiên có luật phân phối và các đặc trưng.

+) Khi mẫu $(X_1, X_2, \cdots, X_n)$ nhận giá trị là $(x_1, x_2, \cdots, x_n)$ thì T nhận giá trị là $T_{qs} = T(x_1, x_2, \cdots, x_n)$, và T_{qs} được gọi là giá trị quan sát của T tại giá trị $(x_1, x_2, \cdots, x_n)$.

Một số đặc trưng đặc biệt của mẫu ngẫu nhiên

a) Kỳ vọng mẫu (hay trung bình mẫu)

Cho một mẫu ngẫu nhiên $\boldsymbol{W = (X_1, X_2, \cdots, X_n)}$ cảm sinh từ biến ngẫu nhiên gốc $\boldsymbol{X}$. Khi đó thống kê

$$\overline{\boldsymbol{X}} = \frac{\boldsymbol{X_1 + X_2 + \cdots + X_n}}{\boldsymbol{n}}$$

được gọi là *kỳ vọng mẫu.*

- Ta có $\boldsymbol{E\overline{X} = EX = \mu}$, $\boldsymbol{D\overline{X} = \frac{DX}{n} = \frac{\sigma^2}{n}}$.

b) **Phương sai mẫu**

Cho một mẫu ngẫu nhiên $\boldsymbol{W = (X_1, X_2, \cdots, X_n)}$ cảm sinh từ biến ngẫu nhiên gốc $\boldsymbol{X}$. Khi đó, thống kê

$$\boldsymbol{\hat{S}^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n}}$$

được gọi là *phương sai mẫu.*

- Ta có: $\boldsymbol{\hat{S}^2 = \frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n} - (\overline{X})^2}$, và $\boldsymbol{E(\hat{S}^2) = \frac{n-1}{n} DX = \frac{n-1}{n}\sigma^2}$.

c) **Phương sai mẫu hiệu chỉnh**

Cho một mẫu ngẫu nhiên $\boldsymbol{W = (X_1, X_2, \cdots, X_n)}$ cảm sinh từ biến ngẫu nhiên gốc $\boldsymbol{X}$. **Phương sai mẫu hiệu chỉnh** được xác định như sau

$$\boldsymbol{S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}},$$

và $\boldsymbol{S = \sqrt{S^2}}$ được gọi là *độ lệch chuẩn mẫu hiệu chỉnh.* **Thông thường, người ta lấy S là độ lệch chuẩn của mẫu**

- Ta có:

$$\boldsymbol{S^2 = \frac{X_1^2 + X_2^2 + \cdots + X_n^2 - n(\overline{X})^2}{n-1}},$$

$$\boldsymbol{S^2 = \frac{n}{n-1}\hat{S}^2},$$

và $\boldsymbol{E(S^2) = DX = \sigma^2}$.

c) Trong trường hợp $\boldsymbol{X}$ đã biết kỳ vọng $\boldsymbol{EX = \mu}$, **phương sai mẫu được chọn là**

$$\boldsymbol{(S^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}$$

e) **Tần suất mẫu:** Xét tổng thể gồm các phần tử mang dấu hiệu nghiên cứu $\boldsymbol{A}$ và không mang dấu hiệu $\boldsymbol{A}$. **Gọi $\boldsymbol{p}$ là xác suất chọn một phần tử trong tổng thể mà phần tử mang dấu hiệu $\boldsymbol{A}$.**

- Mã hóa biến định tính của tổng thể bởi biến ngẫu nhiên $\boldsymbol{Z}$ như sau: **$\boldsymbol{Z = 1}$ nếu phần tử được chọn ngẫu nhiên có dấu hiệu $\boldsymbol{A}$ và $\boldsymbol{Z = 0}$ nếu phần tử được chọn ngẫu nhiên không có dấu hiệu $\boldsymbol{A}$.**

+) Xét mẫu ngẫu nhiên $\boldsymbol{W = (X_1, X_2, \cdots, X_n)}$ được cảm sinh từ $\boldsymbol{Z}$, thực chất mẫu này dùng để quan sát số lần xuất hiện phần tử mang dấu hiệu $\boldsymbol{A}$ trong mẫu. Ta có $\boldsymbol{X_i}$ có phân bố Bernoulli tham số $\boldsymbol{p}$. Gọi $\boldsymbol{X}$ là số các phần tử có dấu hiệu $\boldsymbol{A}$ có trong mẫu. Ta có $\boldsymbol{X = X_1 + \cdots + X_n \sim B(n, p)}$.

- **Tần suất mẫu là thống kê được cho bởi công thức $\boldsymbol{f = \dfrac{X}{n}}$.**

- Ta có $\boldsymbol{Ef = p, Df = \dfrac{p(1-p)}{n}}$.

Cách tính $\overline{x}$, $\hat{s}^2$, s^2, f

+) Nếu mẫu dữ liệu cho dưới dạng: $\{x_1, x_2, ..., x_n\}$ thì

$\overline{x} = \frac{x_1+x_2+...+x_n}{n}$

và $\hat{s}^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+...+(x_n-\overline{x})^2}{n}$

và $s^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+...+(x_n-\overline{x})^2}{n-1}$

+) Nếu mẫu dữ liệu cho ở dạng bảng sau:

x_i	x_1	x_2	$\cdots$	x_k
n_i	n_1	n_2	$\cdots$	n_k

, thì $n = n_1 + n_2 + \cdots + n_k$, $\overline{x} = \frac{n_1.x_1 + n_2.x_2 + \cdots + n_k.x_k}{n}$,

và $\hat{s}^2 = \frac{n_1.(x_1 - \overline{x})^2 + n_2.(x_2 - \overline{x})^2 + \cdots + n_k.(x_k - \overline{x})^2}{n}$,

và $s^2 = \frac{n_1.(x_1 - \overline{x})^2 + n_2.(x_2 - \overline{x})^2 + \cdots + n_k.(x_k - \overline{x})^2}{n-1}$.

+) Nếu mẫu dữ liệu cho dạng bảng sau

$[a_{i-1}, a_i)$	$[a_0, a_1)$	$[a_1, a_2)$	$\cdots$	$[a_{k-1}, a_k)$
n_i	n_1	n_2	$\cdots$	n_k

thì ta thường thay khoảng $[a_{i-1}, a_i)$ bằng $x_i = \frac{a_{i-1}+a_i}{2}$, với $i = 1, 2, ..., k$.

Khi đó, $n = n_1 + n_2 + \cdots + n_k$, $\overline{x} = \dfrac{n_1.x_1 + n_2.x_2 + \cdots + n_k.x_k}{n}$,

và $\hat{s}^2 = \dfrac{n_1.(x_1 - \overline{x})^2 + n_2.(x_2 - \overline{x})^2 + \cdots + n_k.(x_k - \overline{x})^2}{n}$,

và $s^2 = \dfrac{n_1.(x_1 - \overline{x})^2 + n_2.(x_2 - \overline{x})^2 + \cdots + n_k.(x_k - \overline{x})^2}{n-1}$.

+) Cách tính giá trị của tần suất mẫu:

Cho mẫu quan sát cỡ n, có m phần tử mang dấu hiệu A. Khi đó $f = \dfrac{m}{n}$,

và f được gọi là tỷ lệ phần tử mang dấu hiệu A trong mẫu.

Quy trình bấm máy tính cầm tay FX570-VN plus

Bước 1: Khai báo có bảng dữ liệu có tần số SHIFT → MODE → PHÍM TRỎ DƯỚI (▼) → STAT → ON

Bước 2: Nhập dữ liệu MODE → STAT → VAR → AC

Bước 3: Tính các giá trị đặc trưng SHIFT → STAT → VAR.

Chú ý: n là số các dữ liệu nhập vào (cỡ mẫu), $\overline{x}$ là giá trị trung bình của mẫu, σX là $\hat{s}$ và SX là s (độ lệch chuẩn mẫu hiệu chỉnh).

Ví dụ: Để đưa ra một nhận định nào đó về trọng lượng (đơn vị kg) của một trẻ mới sinh, người ta tiến hành cân ngẫu nhiên 150 trẻ và thu được bảng dữ liệu sau:

Trọng lượng	1,8	2,0	2,5	2,8	3,0	3,2	3,5	3,7	3,9	4,0
Số trẻ được khảo sát	8	15	20	25	28	20	12	12	5	5

Tính trọng lượng trung bình của mẫu. Tính phương sai và phương sai hiệu chỉnh của mẫu. Biết rằng một trẻ có trọng lượng trong đoạn $\mathbf{[2,7;3,8]}$ được gọi là đạt chuẩn, hãy tính tỉ lệ trẻ có trọng lượng đạt chuẩn của mẫu.

Giải: Gọi $\boldsymbol{X}$ là trọng lượng của một trẻ. Mẫu ngẫu nhiên cảm sinh bởi $\boldsymbol{X}$ có cỡ $\boldsymbol{n = 150.}$

Trọng lượng trung bình:

$$\overline{\boldsymbol{x}} = \frac{\boldsymbol{x_1 + x_2 + \cdots + x_n}}{\boldsymbol{n}}.$$

Suy ra $\overline{\boldsymbol{x}} = \boldsymbol{2,922.}$

Phương sai của mẫu:

$$\hat{\boldsymbol{s}}^{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \left((\boldsymbol{x_1} - \overline{\boldsymbol{x}})^{\boldsymbol{2}} + \cdots + (\boldsymbol{x_n} - \overline{\boldsymbol{x}})^{\boldsymbol{2}} \right).$$

Suy ra $\hat{\boldsymbol{s}}^{\boldsymbol{2}} \simeq \boldsymbol{0,335583.}$

Phương sai hiệu chỉnh của mẫu: $(\boldsymbol{s})^{\boldsymbol{2}} = \dfrac{\boldsymbol{1}}{\boldsymbol{n-1}} \left((\boldsymbol{x_1} - \overline{\boldsymbol{x}})^{\boldsymbol{2}} + \cdots + (\boldsymbol{x_n} - \overline{\boldsymbol{x}})^{\boldsymbol{2}} \right).$

Suy ra $(\boldsymbol{s})^{\boldsymbol{2}} \simeq \boldsymbol{0,337835.}$

Tỉ lệ trẻ có trọng lượng đạt chuẩn trong mẫu là $\boldsymbol{f} = \dfrac{\boldsymbol{97}}{\boldsymbol{150}} \simeq \boldsymbol{64,667\%.}$

4. Phân bố xác suất của một số thống kê mẫu

4.1. Trường hợp X có phân phối chuẩn

Giả sử $\boldsymbol{X \sim N(\mu; \sigma^2)}$ và $\boldsymbol{W = (X_1, X_2, \ldots, X_n)}$ là mẫu ngẫu nhiên cảm sinh từ $\boldsymbol{X}$. Khi đó

a) $\boldsymbol{\overline{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)}$ và $\boldsymbol{U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0; 1).}$

b) $\boldsymbol{\chi^2 = \frac{n(S^*)^2}{\sigma^2} \sim \chi^2(n)}$ và $\boldsymbol{T = \frac{(\overline{X} - \mu)\sqrt{n}}{S} \sim T(n-1).}$

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

4.2. Phân bố xác suất của tần suất

Giả sử $\boldsymbol{X}$ có phân phối xác suất Bernoulli tham số $\boldsymbol{p}$ (tức $\boldsymbol{X}$ là biến ngẫu nhiên quan sát số lần xuất hiện phần tử mang dấu hiệu $\boldsymbol{A}$ trong phép chọn một phần tử ngẫu nhiên của tổng thể và biết xác suất chọn 1 phần tử mà nó mang dấu hiệu $\boldsymbol{A}$ là $\boldsymbol{p}$).

Lấy mẫu ngẫu nhiên $\boldsymbol{W = (X_1, \ldots, X_n)}$ cảm sinh về $\boldsymbol{X}$. Khi đó tần suất mẫu $\boldsymbol{f = \dfrac{X_1 + X_2 + \ldots + X_n}{n}}$ có $\boldsymbol{Ef = p, Df = \dfrac{p(1-p)}{n}}$

Khi $\boldsymbol{n}$ đủ lớn thỏa mãn $\boldsymbol{np > 5}$ và $\boldsymbol{n(1-p) > 5}$ hoặc $\boldsymbol{np(1-p) > 20}$, ta có thể coi

$$\boldsymbol{U = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim N(0;1).}$$

II. BÀI TOÁN ƯỚC LƯỢNG THAM SỐ

1. Ước lượng điểm

Giả $\boldsymbol{a}$ là tham số liên quan biến ngẫu nhiên $\boldsymbol{X}$**.** Ta cần xấp xỉ giá trị cho $\boldsymbol{a}$**.** Bài toán xấp xỉ $\boldsymbol{a}$ bởi một giá trị nào đó được xác định dựa trên mẫu quan sát về $\boldsymbol{X}$ được gọi là **bài toán ước điểm cho tham số** $\boldsymbol{a}$**.**

A. Phương pháp hàm ước lượng.

Từ mẫu quan sát $\boldsymbol{x_1, x_2, \cdots, x_n}$, của mẫu ngẫu nhiên $\boldsymbol{(X_1, X_2, \cdots, X_n)}$ cảm sinh từ $\boldsymbol{X}$, ta tìm được số $\boldsymbol{\hat{a} = T(x_1, x_2, \cdots, x_n)}$ để xấp xỉ tham số $\boldsymbol{a}$. Số $\boldsymbol{\hat{a}}$ là giá trị của hàm thống kê $\boldsymbol{T(X_1, X_2, \cdots, X_n)}$ tại mẫu quan sát đã cho.

Khi đó, **hàm thống kê $\boldsymbol{T = T(X_1, X_2, \cdots, X_n)}$ được gọi là hàm ước lượng hay ước lượng điểm của tham số $\boldsymbol{a}$.**

Các tiêu chuẩn ước lượng.

+) **Ước lượng điểm $T = T(X_1, X_2, \cdots, X_n)$ được gọi là ước lượng không chệch của tham số a** nếu $ET = a$.

Ước lượng điểm $T = T(X_1, X_2, \cdots, X_n)$ được gọi là ước lượng chệch với độ lệch C của tham số a nếu $ET = a + C$.

+) **Ước lượng hiệu quả: Trong các ước lượng không chệch $T = T(X_1, ..., X_n)$ của tham số a, ước lượng không chệch có phương sai nhỏ nhất được gọi là ước lượng hiệu quả của a**

Chú ý: Nếu X có hàm mật độ xác suất $f(x, a)$ thì

$$DT \geq \frac{1}{nE\left(\frac{\partial \ln f(x,a)}{\partial a}\right)^2},$$

với T là ước lượng không chệch bất kỳ của a.

+) Ước lượng vững: Ước lượng điểm $T = T(X_1, ..., X_n)$ của tham số a được gọi là ước lượng vững của a nếu $\lim_{n\to\infty} P(|T - a| > \epsilon) = 0$, với mọi $\epsilon > 0$.

B. Một số kết quả đặc biệt.

a) **Trung bình mẫu $\overline{X}$ là ước lượng không chệch, hiệu quả, vững cho kỳ vọng $\mu = EX$ của X.**

b) **Phương sai mẫu S^2 và $(S^*)^2$ (trong trường hợp $\mu = EX$ đã biết) là ước lượng không chệch, vững của phương sai $\sigma^2 = DX$ của X.**

c) **Tần suất mẫu f là ước ượng không chệch, hiệu quả, vững cho tham số p (là tỉ lệ phần tử có dấu hiệu A trong tổng thể hoặc xác suất chọn 1 phần tử mà nó mang dấu hiệu A).**

Ví dụ: Để đưa ra một nhận định nào đó về trọng lượng (đơn vị kg) của một trẻ mới sinh, người ta tiến hành cân ngẫu nhiên 150 trẻ và thu được bảng dữ liệu sau:

Trọng lượng	1,8	2,0	2,5	2,8	3,0	3,2	3,5	3,7	3,9	4,0
Số trẻ được khảo sát	8	15	20	25	28	20	12	12	5	5

Biết rằng một trẻ có trọng lượng trong đoạn $[\mathbf{2,7;3,8}]$ được gọi là đạt chuẩn. Ước lượng không chệch cho trọng lượng trung bình và phương sai về trọng lượng của mỗi trẻ mới sinh. Ước lượng không chệch cho tỉ lệ trẻ mới sinh có trọng lượng đạt chuẩn.

Giải: Gọi X là trọng lượng của một trẻ. Mẫu ngẫu nhiên cảm sinh bởi X có cỡ $n = 150$. Gọi EX, DX lần lượt là kỳ vọng và phương sai của X. Gọi p là tỉ lệ trẻ mới sinh có trọng lượng đạt chuẩn. Từ mẫu đã cho, ta xác định được:

Trọng lượng trung bình: $\overline{x} = 2,922$.

Phương sai hiệu chỉnh của mẫu: $(s)^2 \simeq 0,337835$.

Tỉ lệ trẻ có trọng lượng đạt chuẩn trong mẫu là $f = \dfrac{97}{150} \simeq 64,667\%$.

$+$) Trung bình mẫu $\overline{X}$ là ước lượng không chệch cho EX. Ta có $EX \simeq \overline{x} = 2,922$.

$+$) Phương sai hiệu chỉnh $(S)^2$ là ước lượng không chệch cho DX. Ta có $DX \simeq (s)^2 \simeq 0,337835$.

+) Tần suất mẫu f là ước lượng chệch cho p. Ta có $p \simeq f \simeq 0,65$.

2. Ước lượng khoảng tin cậy

Giả a là tham số liên quan biến ngẫu nhiên X. Chúng ta cần ước lượng giá trị cho a.

Bài toán: Từ mẫu quan sát về X, hãy tìm hai giá trị $a_1 < a_2$ sao cho khoảng (a_1, a_2) được kỳ vọng chứa a, được gọi là bài toán ước lượng khoảng tin cậy cho tham số a.

• **Định nghĩa: Xét mẫu ngẫu nhiên $(X_1, X_2, \cdots, X_n)$ cảm sinh từ X.**

Khoảng (F, G) (với F, G là hai thống kê) được gọi là khoảng ước lượng của tham số a với độ tin cậy $1 - \alpha$ nếu tham số a thuộc khoảng trên với xác suất $1 - \alpha$ hay

$$P(F < a < G) = 1 - \alpha.$$

Chú ý:

a) Xác suất để a không nằm trong $(F;G)$ là α.

b) Giá trị $1-\alpha$ được gọi là độ tin cậy và hiệu $G-F$ được gọi là độ dài khoảng tin cậy.

c) Khoảng tin cậy được gọi là đối xứng nếu

$$P(a \leq F) = P(a \geq G),$$

và khi đó $\dfrac{G-F}{2}$ được gọi là độ chính xác (hay sai số) của ước lượng khoảng tin cậy đối xứng.

d) Với mẫu quan sát ban đầu, ta gọi f, g lần lượt là giá trị quan sát của F, G tại mẫu này. Khi đó, ta thu được khoảng tin cậy cụ thể là $a \in (f;g)$.

Nguyên tắc chung:

1) Tìm một thống kê $G = G(X_1, X_2, \ldots, X_n, a)$ mà sao cho phân phối xác suất của G hoàn toàn xác định và không chứa tham số a.

2) Với độ tin cậy $1 - \alpha = \gamma$ cho trước, ta cố định giá trị $\alpha_1 \geq 0$ và đặt $\alpha_2 = \alpha - \alpha_1 \geq 0$. Sau đó, tìm g_{α_1} và $g_{1-\alpha_2}$ sao cho

$$P(G < g_{\alpha_1}) = \alpha_1 \text{ và } P(G > g_{1-\alpha_2}) = \alpha_2.$$

Do đó, ta thu được $P(g_{\alpha_1} < G(X_1, X_2, \ldots, X_n, a) < g_{1-\alpha_2}) = 1 - \alpha$.

3) Biến đổi bất đẳng thức của $G(X_1, X_2, \ldots, X_n, a)$, ta nhận được $a_1 < a < a_2$ và

$$P(a_1 < a < a_2) = 1 - \alpha.$$

4) Tại mẫu dữ liệu đã cho, tính được giá trị của hai thống kế a_1 và a_2 lần

lượt là $\hat{a}_1, \hat{a}_2$. Khoảng tin cậy là $a \in (\hat{a}_1,\ \hat{a}_2)$.

A. Ước lượng khoảng tin cậy đối xứng cho kỳ vọng của BNN X

Bài toán: Cho X là biến ngẫu nhiên. Đặt $\mu = EX, DX = \sigma^2$. Từ mẫu ngẫu nhiên cỡ n và độ tin cậy $1-\alpha = \gamma$, hãy xác định khoảng tin cậy đối xứng cho $\mu = EX$.

i) Trường hợp 1: phương sai $\sigma^2 = \sigma_0^2$ đã biết và X có phân phối chuẩn.

Khi đó, thống kê: $G = \dfrac{\overline{X} - \mu}{\sigma_0}.\sqrt{n} \sim N(0,1)$.

Do đó, khoảng tin cậy đối xứng cho μ với độ tin cậy $1-\alpha$ có dạng:

$$\mu = EX \in \left(\overline{x} - U_{\frac{\alpha}{2}} \cdot \frac{\sigma_0}{\sqrt{n}};\ \overline{x} + U_{\frac{\alpha}{2}} \cdot \frac{\sigma_0}{\sqrt{n}}\right),$$

với n là cỡ mẫu; $\overline{x}$ là giá trị của trung bình mẫu; $\sigma_0 = \sqrt{DX}$ là độ lệch chuẩn của X; và $U_{\frac{\alpha}{2}}$ là giá trị tới hạn chuẩn mức $\dfrac{\alpha}{2}$, nghĩa là $\Phi(U_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

Chú ý: Sai số (hay độ chính xác) của khoảng tin cậy đối xứng trong trường hợp này là $\epsilon = U_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$. Độ dài của khoảng là 2ϵ.

VD 1. Người ta đo ngẫu nhiên chiều cao của 100 cây giống $\boldsymbol{T}$ được hai năm tuổi và thu được trung bình chiều cao là $\mathbf{1,69}\,\boldsymbol{m}$. Biết rằng chiều cao của cây giống $\boldsymbol{T}$ được hai năm tuổi có phân phối chuẩn, hãy xác định:

a) Ước lượng không chệch cho trung bình chiều cao của cây giống $\boldsymbol{T}$ được hai năm tuổi.

b) Với độ tin cậy $\mathbf{95\%}$, ước lượng khoảng tin cậy đối xứng cho trung bình chiều cao của cây giống $\boldsymbol{T}$ được hai năm tuổi biết rằng độ lệch chuẩn của chiều cao là $\mathbf{0,2}$. Sau đó, xác định sai số của khoảng ước lượng tin cậy này.

Hướng dẫn:

Gọi X là chiều cao của cây giống T được hai năm tuổi, khi đó $X \sim N(\mu, \sigma^2)$ với $\mu = EX, \sigma^2 = DX$. Mẫu quan sát có cỡ mẫu $n = 100$ và $\overline{x} = 1,69$. Ta ước lượng cho chiều cao trung bình μ.

a) $\mu = EX \simeq 1,69\, m$.

b) Độ tin cậy $1-\alpha=0,95$ suy ra $\alpha=0,05$. Đây là bài toán ước lượng khoảng đối xứng cho μ khi biết $\sigma=0,2$.

Khoảng tin cậy đối xứng: $\left(\overline{x}-U_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}};\overline{x}+U_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right)=(1,6508,\ 1,7292)$.

ở đây $\sigma=0,2, n=100$, $\overline{x}=1,69$, và $U_{\frac{\alpha}{2}}=U_{0,025}$. Tra bảng phân vị chuẩn (tra phân vị chuẩn t ứng với mức $0,975$), ta được $U_{0,025}=1,96$.

- Sai số của khoảng tin cậy là $\epsilon=U_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}=1,96\cdot\frac{0,2}{\sqrt{100}}=....$

ii) Trường hợp 2: phương sai chưa biết, X có phân phối chuẩn và cỡ mẫu $n \leq 30$.

Khi đó, thống kê $G = \dfrac{\overline{X} - \mu}{S}.\sqrt{n} \sim T(n-1)$, ở đây $T(k)$ là ký hiệu cho phân phối Student với k bậc tự do, S là độ lệch chuẩn mẫu hiệu chỉnh.

Do đó, khoảng tin cậy đối xứng cho μ với độ tin cậy $1-\alpha$ có dạng:

$$\mu = EX \in \left(\overline{x} - T_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}};\ \overline{x} + T_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}}\right),$$

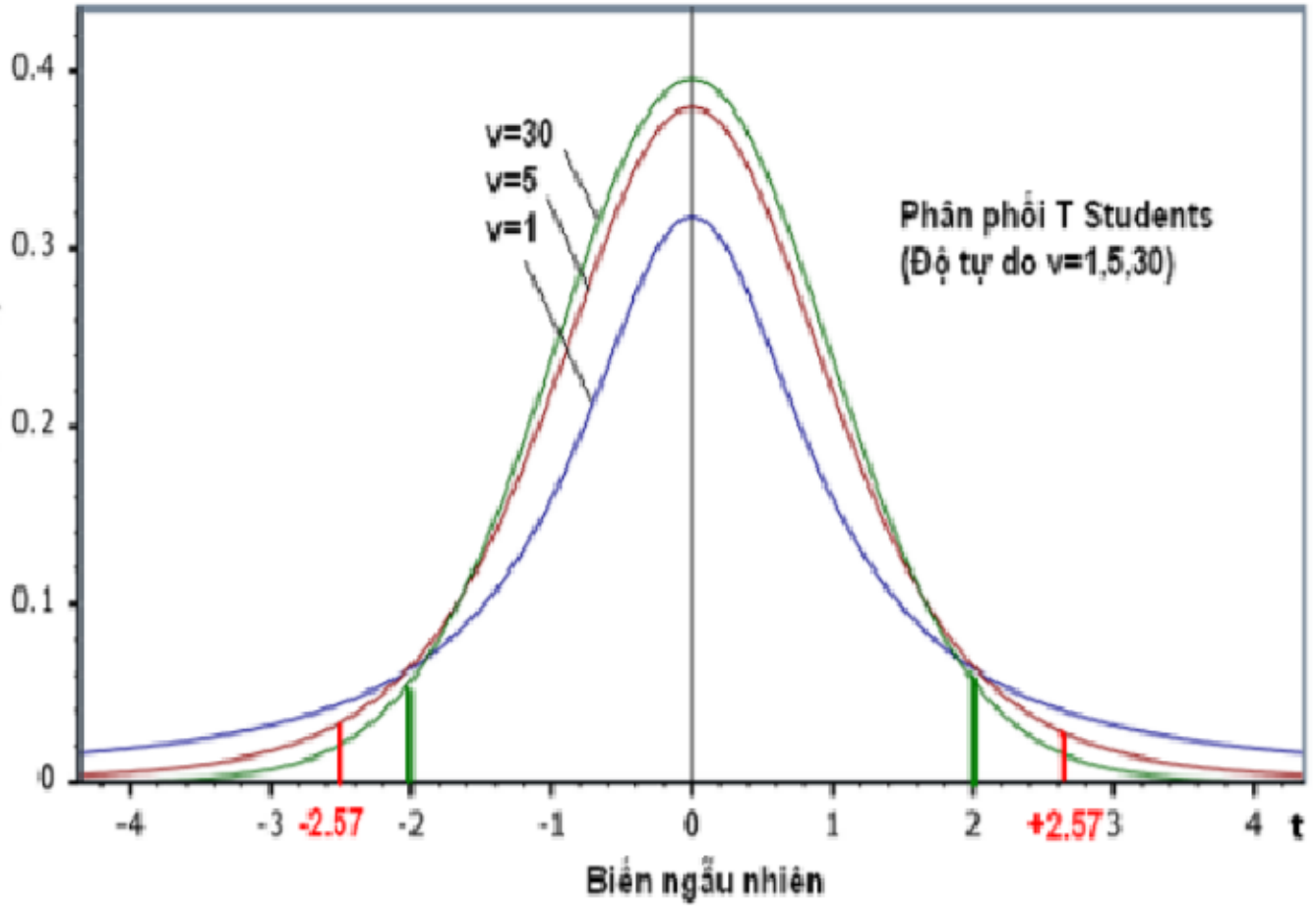
với n là cỡ mẫu; $\overline{x}$ là giá trị của trung bình mẫu; s là độ lệch mẫu hiệu chỉnh; và $T_{\frac{\alpha}{2}}(n-1)$ là giá trị tới hạn Student mức $\dfrac{\alpha}{2}$ với $n-1$ bậc tự do.

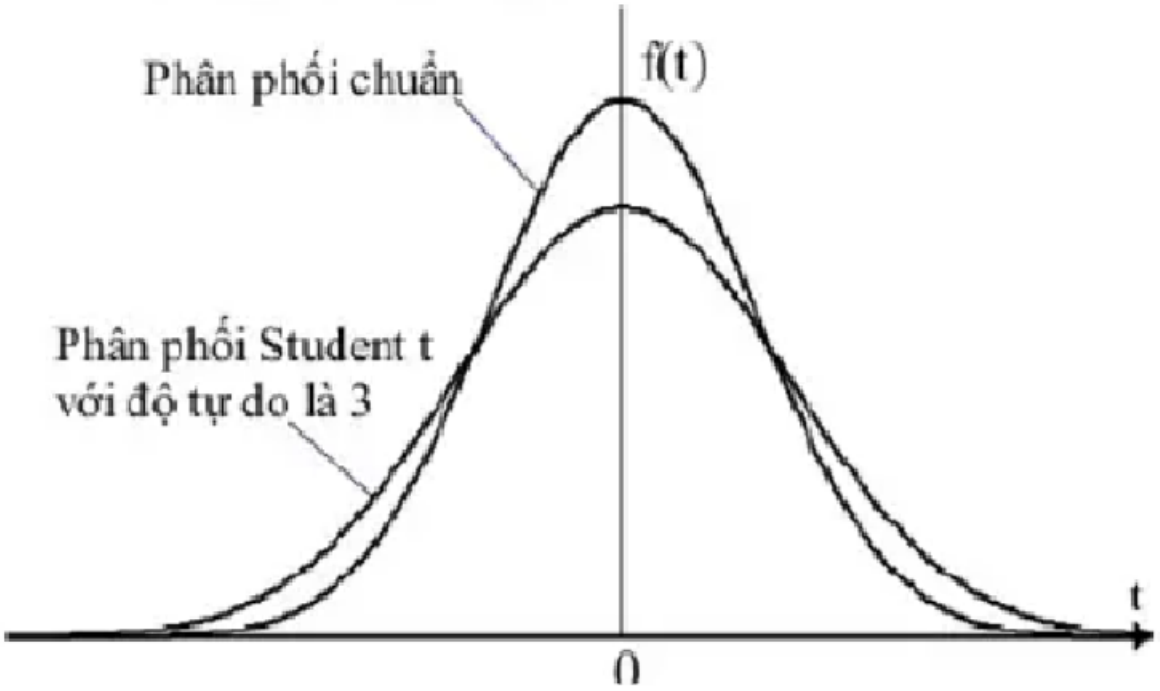
Chú ý: Sai số (hay độ chính xác) của khoảng tin cậy đối xứng trong

trường hợp này là $\epsilon = T_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}}$. Độ dài của khoảng là 2ϵ.

Một vài kết quả về phân phối Student

- Phân phối Student với n bậc tự do được ký hiệu là $T(n)$. Ta viết $X \sim T(n)$, tức là BNN liên tục X có phân phối Student với n bậc tự do.





- Ta ký hiệu T_α^n hoặc $T_\alpha(n)$ hoặc $T_{n,\alpha}$ là giá trị tới hạn Student mức α và n bậc tự do, nghĩa là $P(X > T_\alpha^n) = \alpha$, với $X \sim T(n)$.

- Khi $n \geq 30$, ta xấp xỉ phân phối Student $T(n)$ theo phân phối chuẩn tắc $N(0,1)$.

VD 2. Người ta tiến hành cân trọng lượng (đơn vị kg) ngẫu nhiên của 25 gà và thu được kết quả dạng khoảng $[a, b)$ cho bởi bảng dưới đây:

Trọng lượng	1,1-1,3	1,3-1,5	1,5-1,7	1,7-1,9	1,9-2,1	2,1-2,3	2,3-2,5
Số lượng	2	4	5	7	3	2	2

Biết rằng trọng lượng của gà có phân phối chuẩn, hãy xác định: khoảng tin cậy đối xứng trọng lượng trung bình của gà với độ tin cậy **99%.**

Hướng dẫn:

Gọi X là trọng lượng của gà, khi đó $X \sim N(\mu, \sigma^2)$ với $\mu = EX, \sigma^2 = DX$. Mẫu quan sát có cỡ $n = 25$, $\overline{x} = 1,752$ và độ lêch chuẩn hiệu chỉnh $s \simeq 0,3331$. Ta ước lượng cho trọng lượng trung bình μ.

Ước lượng khoảng tin cậy đối xứng cho μ với độ tin cậy 99% và chưa biết độ lệch chuẩn σ và cỡ mẫu $n = 25 < 30$. Độ tin cậy $1 - \alpha = 0,99$ suy ra $\alpha = 0,01$.

Khoảng tin cậy đối xứng:

$\left(\overline{x} - T_{(\frac{\alpha}{2})}(n-1).\frac{s}{\sqrt{n}}; \overline{x} + T_{(\frac{\alpha}{2})}(n-1).\frac{s}{\sqrt{n}}\right) = (1,56566,\ 1,93834)$, ở đây $T_{(\frac{\alpha}{2})}(n-1) = T_{0,005}(24) = 2,797$ (Tra bảng Student).

iii) Trường hợp 3: phương sai chưa biết, X có phân phối chuẩn hoặc không có phân phối chuẩn, và cỡ mẫu $n > 30$.

Khi đó, thống kê $G = \dfrac{\overline{X} - \mu}{S}.\sqrt{n} \sim T(n-1)$. Do $n - 1 \geq 30$ nên ta coi $G = \dfrac{\overline{X} - \mu}{S}.\sqrt{n} \sim N(0,1)$.

Do đó, khoảng tin cậy đối xứng cho μ với độ tin cậy $1 - \alpha$ có dạng:

$$\mu = EX \in \left(\overline{x} - U_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}};\ \overline{x} + U_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right),$$

với n là cỡ mẫu; $\overline{x}$ là giá trị của trung bình mẫu; s là độ lệch mẫu hiệu chỉnh; và $U_{\frac{\alpha}{2}}$ là giá trị tới hạn chuẩn mức $\dfrac{\alpha}{2}$, nghĩa là $\Phi(U_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

Chú ý: Sai số (hay độ chính xác) của khoảng tin cậy đối xứng trong trường hợp này là $\epsilon = U_{\frac{\alpha}{2}} \cdot \dfrac{s}{\sqrt{n}}$. Độ dài của khoảng là 2ϵ.

VD 3. Người ta đo ngẫu nhiên chiều cao của 100 nam thanh niên tại khu vực K và thu được trung bình chiều cao là $\mathbf{1,69}\,\boldsymbol{m}$ và độ lệch hiệu chỉnh là $\mathbf{0,4}$. Biết rằng chiều cao của nam thanh niên khu vực K có phân phối chuẩn, hãy xác định khoảng tin cậy đối xứng cho trung bình chiều cao của các nam thanh niên khu vực K với độ tin cậy $\mathbf{95\%}$. Sau đó tìm sai số và độ dài cho khoảng tin cậy đối xứng này.

Hướng dẫn:

Gọi X là chiều cao của nam thanh niên khu vực K, khi đó $X \sim N(\mu, \sigma^2)$ với $\mu = EX, \sigma^2 = DX$. Mẫu quan sát có cỡ mẫu $n = 100$ và $\overline{x} = 1,69$, độ lệch chuẩn hiệu chỉnh là $s = 0,4$. Ta ước lượng cho chiều cao trung bình μ.

Độ tin cậy $1 - \alpha = 0,95$ suy ra $\alpha = 0,05$. Đây là bài toán ước lượng khoảng đối xứng cho μ khi chưa biết σ và cỡ mẫu $n > 30$.

Khoảng tin cậy đối xứng: $\left(\overline{x} - U_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}}; \overline{x} + U_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}}\right) = (1,6116,\ 1,7684)$, ở đây $U_{\frac{\alpha}{2}} = U_{0,025} = 1,96; n = 100, s = 0,4$ và $\overline{x} = 1,69$.

- Sai số của khoảng tin cậy đối xứng là: $\epsilon = U_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}} = 0,0784$ và độ dài là $2\epsilon = ...$

B. Ước lượng khoảng tin cậy đối xứng cho tỉ lệ hay xác suất

Bài toán: Xét tổng thể có các phần tử mang dấu hiệu A và không có dấu hiệu A. Gọi p là tỉ lệ phần tử mang dấu hiệu A trong tổng thể. Từ mẫu quan sát cỡ n và độ tin cậy $1-\alpha$, ước lượng khoảng tin cậy đối xứng cho p.

- Ta xét thống kê $G = \dfrac{f-p}{\sqrt{f(1-f)}}.\sqrt{n}$, với f là tần suất mẫu. Gọi f_0 là giá trị của f tại mẫu quan sát. Khi n khá lớn thỏa mãn $nf_0 > 10$ và $n(1-f_0) > 10$, ta có $G \sim N(0,1)$.

• Khoảng tin cậy đối xứng cho p.

$$p \in \left(f - U_{\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}},\ f + U_{\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}} \right),$$

với n là cỡ mẫu; f là tần suất mẫu; và $U_{\frac{\alpha}{2}}$ là giá trị tới hạn chuẩn mức

$\frac{\alpha}{2}$, nghĩa là $\Phi(U_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

Chú ý: Sai số (hay độ chính xác) của khoảng tin cậy đối xứng là $\epsilon = U_{\frac{\alpha}{2}} \cdot \sqrt{\dfrac{f(1-f)}{n}}$. **Độ dài của khoảng là** 2ϵ.

VD 4. Một khảo sát ngẫu nhiên 100 sinh viên nam về chiều cao (đơn vị mét) và thu được kết quả dạng khoảng $[a, b)$ cho bởi bảng dưới đây:

Chiều cao	1,53-1,59	1,59-1,65	1,65-1,71	1,71-1,77	1,77-1,83	1,83-1,89	1,89-1,91
Số lượng	9	15	41	15	8	6	6

Một sinh viên nam được gọi là có chiều cao đạt tiêu chuẩn nếu chiều cao nằm giữa $\mathbf{1,65}$ và $\mathbf{1,77}$. Hãy

a) Ước lượng không chệch cho tỉ lệ sinh viên có chiều cao đạt tiêu chuẩn.

b) Với độ tin cậy $\mathbf{99\%}$, ước lượng khoảng tin cậy đối xứng cho tỉ lệ sinh viên có chiều cao đạt tiêu chuẩn.

Hướng dẫn:

Gọi p là tỉ lệ sinh viên có chiều cao đạt tiêu chuẩn.

Mẫu quan sát có cỡ $n = 100$ và tỉ lệ sinh viên có chiều cao đạt tiêu chuẩn của mẫu là $f = \dfrac{56}{100} = 0,56$. Ta ước lượng cho tỉ lệ p.

a) $p \simeq 0,56$.

b) Độ tin cậy $1 - \alpha = 0,99$ suy ra $\alpha = 0,01$.

Khoảng tin cậy đối xứng:

$$\left(f - U_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{f(1-f)}{n}},\ f + U_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{f(1-f)}{n}} \right) = (0,431932,\ 0,688068),$$

ở đây $U_{(\frac{\alpha}{2})} = U_{0,005}$. Tra bảng phân vị chuẩn (tra phân vị chuẩn t mức $0,995$), ta được $U_{0,005} = 2,58$.

C. Ước lượng cỡ mẫu tối thiểu

Xét bài toán ước lượng khoảng tin cậy đối xứng cho:

+) Kỳ vọng $\mu = EX$ với X là một biến ngẫu nhiên nào đó hoặc

+) Xác suất p (xác suất chọn một phần tử của tổng thể mà phần tử này mang dấu hiệu A).

Cho trước mẫu quan sát cỡ m, độ tin cậy $1 - \alpha$.

Ta có bài toán: Cần quan sát một mẫu có cỡ tối thiểu n bao nhiêu biết rằng dựa vào độ tin cậy $1 - \alpha$ và mẫu này, ta xây dựng được khoảng tin cậy đối xứng có sai số hoặc độ dài của khoảng không vượt quá ε_0.

Phương pháp thực hiện như sau:

Bước 1: Giả sử mẫu cần quan sát có cỡ n. Xác định độ tin cậy $1-\alpha$.

Bước 2: Xác định sai số hoặc độ dài của khoảng tin cậy ứng với mẫu có cỡ n :

a) Với bài toán liên quan kỳ vọng $\mu = EX$

- Trường hợp 1: X biết DX, sai số là $\varepsilon = U_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, với $\sigma = \sqrt{DX}$. Lưu ý: Độ dài là 2ϵ.

- Trường hợp 2: X chưa biết DX và mẫu quan sát đã cho ở giả thiết có cỡ $m > 30$, sai số là $\epsilon = U_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$, với s là độ lệch mẫu hiệu chỉnh của mẫu cỡ m. Lưu ý: Độ dài là 2ϵ.

b) Với bài toán liên quan tỉ lệ p : Sai số là $\epsilon = U_{\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}}$, với f là tần suất mẫu cỡ m. Lưu ý: Độ dài là 2ϵ.

Bước 3: Tìm số tự nhiên n nhỏ nhất sao cho $\epsilon \leq \varepsilon_0$ (nếu cho điều kiện của sai số) hoặc $2\epsilon \leq \varepsilon_0$ (nếu cho điều kiện của độ dài).

VD 5. Chiều cao của một giống cây trồng có phân phối chuẩn với độ lệch chuẩn là $\mathbf{0,2}$. Với độ tin cậy $\mathbf{95\%}$, cần quan sát tối thiểu bao nhiêu cây để sai số của khoảng ước lượng đối xứng cho chiều cao trung bình không quá $\mathbf{0,001}$.

Giải: Gọi X là chiều cao của giống cây trồng. Khi đó $X \sim N(\mu, \sigma^2)$ với $\mu = EX, \sigma^2 = DX$. Theo giả thiết, ta có $\sigma = 0,2$ đã biết.

Độ tin cậy $1 - \alpha = 0,95$ suy ra $\alpha = 0,05$.

Giả sử mẫu cần quan sát có cỡ là n.

Sai số của khoảng ước lượng cho μ ứng với mẫu cỡ n là $\epsilon = U_{\frac{\alpha}{2}}.\dfrac{\sigma}{\sqrt{n}}$.

Ta tính được $U_{\frac{\alpha}{2}} = 1,96$ và ta có

$$\epsilon \leq 0,001 \Leftrightarrow 1,96.\frac{0,2}{\sqrt{n}} \leq 0,001 \Leftrightarrow n \geq 153664.$$

Chọn $n = 153664$. Khi đó cần quan sát tối thiểu 153664 cây để khoảng ước lượng có độ chính xác thỏa mãn giả thiết.

VD 6. Khảo sát thu nhập của các hộ gia đình ở một khu dân cư X với quy mô dân số rất lớn, thu được trong 300 hộ khảo sát có 200 hộ thu nhập trên 25 triệu/tháng. Với độ tin cậy **95%**, cần khảo sát tối thiểu bao nhiêu hộ để độ dài khoảng tin cậy đối xứng cho tỉ lệ hộ thu nhập trên 25 triệu/tháng không quá $\mathbf{0,004}$.

Giải: Gọi p là tỉ lệ hộ thu nhập trên 25 triệu/tháng của khu dân cư.

Độ tin cậy $1-\alpha = 0,95$ suy ra $\alpha = 0,05$. Cỡ mẫu đã cho $m = 300$, tỷ lệ hộ thu nhập trên 25 triệu/tháng trong mẫu quan sát là $f = \frac{200}{300} = \frac{2}{3}$. Ta thấy $mf = 300.\frac{2}{3} = 200 > 10$ và $m(1-f) = 100 > 10$.

Gọi n là cỡ mẫu tối thiểu cần quan sát.

Sai số của khoảng tin cậy đối xứng ứng với mẫu cỡ n là

$$\epsilon = U_{\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}},$$

với f là tần suất mẫu đã cho trong giả thiết.

Suy ra, độ dài khoảng tin cậy đối xứng của p là

$$2\epsilon = 2 \cdot U_{\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}}.$$

Ta có $U_{\frac{\alpha}{2}} = 1,96$.

Xét

$$2\epsilon \leq 0,004 \Leftrightarrow 2U_{\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}} \leq 0,004 \Leftrightarrow n \geq 213422,2.$$

Chọn $n = 213423$. Cần quan sát tối thiểu là 213423 hộ.