


CĐ 3
MỘT SỐ YẾU TỐ THỐNG KÊ
I Kiến thức cơ bản
1 Các số đặc trưng của mẫu số liệu không ghép nhóm
a) Trung bình, môđ, phương sai và độ lệch chuẩn

- ☑ Giả sử ta có một mẫu số liệu là $x_1, x_2, \dots, x_n$.

Cỡ mẫu của dãy số liệu trên là n .

Số trung bình (hay số trung bình cộng) của mẫu số liệu này, kí hiệu là $\bar{x}$, được tính bởi công thức

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

- ☑ Phương sai của mẫu số liệu này, kí hiệu là S^2 , được tính bởi công thức

$$S^2 = \frac{1}{n} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2] = \frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2) - \bar{x}^2.$$

Độ lệch chuẩn, kí hiệu S , là căn bậc hai của phương sai.

- ☑ Giả sử mẫu số liệu được cho dưới dạng bảng tần số

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

Khi đó, công thức tính số trung bình trở thành

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \dots + n_kx_k}{n},$$

trong đó cỡ mẫu $n = n_1 + n_2 + \dots + n_k$.

Công thức tính phương sai trở thành

$$S^2 = \frac{1}{n} [n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \dots + n_k(x_k - \bar{x})^2] = \frac{1}{n} (n_1x_1^2 + n_2x_2^2 + \dots + n_kx_k^2) - \bar{x}^2.$$

- ⚠ Kí hiệu $f_k = \frac{n_k}{n}$ là tần số tương đối (hay còn gọi là tần suất) của x_k trong mẫu số liệu.

Khi đó, số trung bình còn có thể biểu diễn là

$$\bar{x} = f_1x_1 + f_2x_2 + \dots + f_kx_k.$$

Mốt là giá trị có tần số lớn nhất của mẫu số liệu. Một mẫu số liệu có thể không có mốt, có một mốt hoặc có nhiều mốt.

b) Khoảng biến thiên và khoảng tứ phân vị

Sắp xếp lại mẫu số liệu $x_1, x_2, \dots, x_n$ theo thứ tự không giảm, ta được dãy

$$x_1^* \leq x_2^* \leq \dots \leq x_n^*.$$

Khoảng biến thiên của một mẫu số liệu, kí hiệu R , là hiệu giữa giá trị lớn nhất và giá trị nhỏ nhất của mẫu số liệu đó, tức là

$$R = x_n^* - x_1^*.$$

Số trung vị của mẫu $x_1, x_2, \dots, x_n$ là giá trị ở chính giữa dãy $x_1^*, x_2^*, \dots, x_n^*$. Cụ thể



- Nếu $n = 2k + 1$, $k \in \mathbb{N}$, thì số trung vị của mẫu là x_{k+1}^* .
- Nếu $n = 2k$, $k \in \mathbb{N}$, thì số trung vị của mẫu là $\frac{x_k^* + x_{k+1}^*}{2}$.

Tứ phân vị của một mẫu số liệu gồm 3 giá trị, đó là *tứ phân vị thứ nhất*, *thứ hai* và *thứ ba* (lần lượt kí hiệu là Q_1 , Q_2 , Q_3). Ba giá trị này chia tập hợp dữ liệu đã sắp xếp thành bốn phần đều nhau. Cụ thể:

- Giá trị tứ phân vị thứ hai, Q_2 , chính là số trung vị của mẫu.
- Giá trị tứ phân vị thứ nhất, Q_1 , là số trung vị của nửa số liệu đã sắp xếp bên trái Q_2 (không bao gồm Q_2 nếu n lẻ).
- Giá trị tứ phân vị thứ ba, Q_3 , là số trung vị của nửa số liệu đã sắp xếp bên phải Q_2 (không bao gồm Q_2 nếu n lẻ).

Khoảng tứ phân vị, kí hiệu Δ_Q , là hiệu giữa Q_3 và Q_1 , tức là $\Delta_Q = Q_3 - Q_1$.

Một giá trị của mẫu số liệu được gọi là *giá trị ngoại lệ* hay *giá trị bất thường* nếu nó nhỏ hơn $Q_1 - 1,5\Delta_Q$ hoặc lớn hơn $Q_3 + 1,5\Delta_Q$.

2 Các số đặc trưng của mẫu số liệu ghép nhóm

a) Số liệu ghép nhóm

Mẫu số liệu ghép nhóm thường được trình bày dưới dạng bảng tần số ghép nhóm có dạng như dưới đây.

Giá trị chính giữa mỗi nhóm được dùng làm giá trị đại diện cho nhóm đó.

Nhóm	$[u_1; u_2)$	$[u_2; u_3)$	$\dots$	$[u_k; u_{k+1}]$
Giá trị đại diện	c_1	c_2	$\dots$	c_k
Tần số	n_1	n_2	$\dots$	n_k

Với $c_j = \frac{u_j + u_{j+1}}{2}$.

b) Trung bình, mốt, phương sai và độ lệch chuẩn

Số trung bình của mẫu số liệu ghép nhóm, kí hiệu $\bar{x}$, được tính như sau

$$\bar{x} = \frac{n_1 c_1 + n_2 c_2 + \dots + n_k c_k}{n},$$

với cỡ mẫu $n = n_1 + n_2 + \dots + n_k$.

Phương sai của mẫu số liệu ghép nhóm, kí hiệu là S^2 , được tính bởi công thức

$$S^2 = \frac{1}{n} [n_1 (c_1 - \bar{x})^2 + n_2 (c_2 - \bar{x})^2 + \dots + n_k (c_k - \bar{x})^2] = \frac{1}{n} (n_1 c_1^2 + n_2 c_2^2 + \dots + n_k c_k^2) - \bar{x}^2.$$

Độ lệch chuẩn của mẫu số liệu ghép nhóm là $S = \sqrt{S^2}$.

Nhóm chứa mốt của mẫu số liệu ghép nhóm là nhóm có tần số lớn nhất.

Mốt của mẫu số liệu ghép nhóm được dùng đại diện cho giá trị có tần số xuất hiện cao nhất.

Giả sử nhóm chứa mốt là $[u_k; u_{k+1})$, khi đó mốt của mẫu số liệu ghép nhóm, kí hiệu M_0 , được xác định bởi công thức

$$M_0 = u_m + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} \cdot (u_{m+1} - u_m)$$



! Nếu không có nhóm kề trước của nhóm chứa một thì $n_{m-1} = 0$. Ngược lại, nếu không có nhóm kề sau của nhóm chứa một thì $n_{m+1} = 0$.
 Một mẫu số liệu có thể có nhiều nhóm chứa một và có nhiều một.

c) Khoảng biến thiên và tứ phân vị

Khoảng biến thiên của mẫu số liệu ghép nhóm, kí hiệu R , là hiệu số giữa đầu mút phải của nhóm cuối cùng và đầu mút trái của nhóm đầu tiên có chứa dữ liệu của mẫu số liệu.

Nếu n_1 và n_k cùng khác 0 thì $R = u_{k+1} - u_1$.

Tứ phân vị thứ i , kí hiệu Q_i , với $i = 1, 2, 3$ của mẫu số liệu ghép nhóm được xác định như sau

$$Q_i = u_m + \frac{i \cdot \frac{n}{4} - C}{n_m} \cdot (u_{m+1} - u_m),$$

trong đó

- ✓ $n = n_1 + n_2 + \dots + n_k$ là cỡ mẫu;
- ✓ $[u_m; u_{m+1})$ là nhóm chứa tứ phân vị thứ i ;
- ✓ n_m là tần số của nhóm chứa tứ phân vị thứ i ;
- ✓ $C = n_1 + n_2 + \dots + n_{m-1}$.

Tứ phân vị thứ hai, Q_2 , chính là *trung vị* M_e của mẫu số liệu ghép nhóm.

Khoảng tứ phân vị của mẫu số liệu ghép nhóm là $\Delta_Q = Q_3 - Q_1$.

Một giá trị của mẫu số liệu được gọi là *giá trị ngoại lệ* hay *giá trị bất thường* nếu nó nhỏ hơn $Q_1 - 1,5\Delta_Q$ hoặc lớn hơn $Q_3 + 1,5\Delta_Q$.

! Nếu tứ phân vị thứ k là $\frac{x_m + x_{m+1}}{2}$, trong đó x_m và x_{m+1} thuộc hai nhóm liên tiếp, chẳng hạn $x_m \in [u_{j-1}; u_j)$ và $x_{m+1} \in [u_j; u_{j+1})$ thì $Q_k = u_j$.