Bài 4. Các số đặc trưng đo mức độ phân tán của mẫu số liệu

A. Lý thuyết

1. Khoảng biến thiên và khoảng tứ phân vị

1.1. Khoảng biến thiên và khoảng tứ phân vị

Sắp xếp mẫu số liệu theo thứ tự không giảm, ta được:

$$x_1 \le x_2 \le \ldots \le x_n.$$

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

$$R = x_n - x_1.$$

• **Khoảng tứ phân vị**, kí hiệu là Δ_Q, là hiệu giữa Q_3 và Q_1, tức là:

$$\Delta_Q = Q_3 - Q_1.$$

Ví dụ: Hãy tính khoảng biến thiên và khoảng tứ phân vị của mẫu số liệu:

10; 3; 5; 7; 20; 1; 4; 9.

Hướng dẫn giải

Sắp xếp mẫu số liệu theo thứ tự không giảm, ta được: 1; 3; 4; 5; 7; 9; 10; 20.

- Khoảng biến thiên của mẫu số liệu là R = 20 – 1 = 19.

- Cỡ mẫu là n = 8, là số chẵn nên giá trị tứ phân vị thứ hai là $Q_2 = 6$.

- Tứ phân vị thứ nhất là trung vị của mẫu: 10; 3; 5; 7. Do đó $Q_1 = 4$.

- Tứ phân vị thứ 3 là trung vị của mẫu: 7; 9; 10; 20. Do đó $Q_3 = 9,5$.

- Khoảng tứ phân vị của mẫu là: $\Delta_Q = 9,5 - 4 = 5,5$.

1.2. Ý nghĩa của khoảng biến thiên và khoảng tứ phân vị:

Khoảng biến thiên đặc trưng cho độ phân tán của toàn bộ mẫu số liệu.

Khoảng tứ phân vị đặc trưng cho độ phân tán của một nửa các số liệu, có giá trị thuộc đoạn từ Q_1 đến Q_3 trong mẫu.

Khoảng tứ phân vị không bị ảnh hưởng bởi các giá trị rất lớn hoặc rất bé trong mẫu.

Ví dụ: Dưới đây là bảng số liệu thống kê của Biểu đồ nhiệt độ trung bình các tháng trong năm 2019 của hai tỉnh Lai Châu và Lâm Đồng (được đề cập đến ở hoạt động khởi động của bài học).

Tháng	1	2	3	4	5	6	7	8	9	10	11	12
Lai Châu	14,8	18,8	20,3	23,5	24,7	24,2	23,6	24,6	22,7	21,0	18,6	14,2
Lâm Đồng	16,3	17,4	18,7	19,8	20,2	20,3	19,5	19,3	18,6	18,5	17,5	16,0

a) Hãy tìm khoảng biến thiên và khoảng tứ phân vị của nhiệt độ trung bình mỗi tháng của tỉnh Lai Châu và Lâm Đồng.

b) Hãy cho biết trong một năm, nhiệt độ ở địa phương nào ít thay đổi hơn.

Hướng dẫn giải

a)

* Tỉnh Lai Châu:

Sắp xếp các số liệu theo thứ tự không giảm, ta được:

14,2; 14,8; 18,6; 18,8; 20,3; 21,0; 22,7; 23,5; 23,6; 24,2; 24,6; 24,7.

+ Khoảng biến thiên của mẫu số liệu là: R = 24,7 – 14,2 = 10,5.

+ Cỡ mẫu là n = 12 là số chẵn nên giá trị tứ phân vị thứ hai là:

$Q_2 = \frac{1}{2}(21,0 + 22,7) = 21,85$.

+ Tứ phân vị thứ nhất là trung vị của mẫu: 14,2; 14,8; 18,6; 18,8; 20,3; 21,0.

Do đó $Q_1 = \frac{1}{2}(18,6 + 18,8) = 18,7$.

+ Tứ phân vị thứ ba là trung vị của mẫu: 22,7; 23,5; 23,6; 24,2; 24,6; 24,7.

Do đó $Q_3 = \frac{1}{2}(23,6 + 24,2) = 23,9$.

+ Khoảng tứ phân vị của mẫu là: $\Delta_Q = 23,9 – 18,7 = 5,2$.

* Tỉnh Lâm Đồng:

Sắp xếp các số liệu theo thứ tự không giảm, ta được:

16,0; 16,3; 17,4; 17,5; 18,5; 18,6; 18,7; 19,3; 19,5; 19,8; 20,2; 20,3.

+ Khoảng biến thiên của mẫu số liệu là: $R' = 20,3 – 16,0 = 4,3$.

+ Cỡ mẫu là n = 12 là số chẵn nên giá trị tứ phân vị thứ hai là:

$Q'_2 = \frac{1}{2}(18,6 + 18,7) = 18,65$.

+ Tứ phân vị thứ nhất là trung vị của mẫu: 16,0; 16,3; 17,4; 17,5; 18,5; 18,6.

Do đó $Q'_1 = \frac{1}{2}(17,4 + 17,5) = 17,45$.

+ Tứ phân vị thứ ba là trung vị của mẫu: 18,7; 19,3; 19,5; 19,8; 20,2; 20,3.

Do đó $Q'_3 = \frac{1}{2}(19,5 + 19,8) = 19,65$.

+ Khoảng tứ phân vị của mẫu là: $\Delta'_Q = 19,65 - 17,45 = 2,2$.

b) Xét về cả khoảng biến thiên và khoảng tứ phân vị của nhiệt độ trung bình mỗi tháng của cả hai tỉnh, ta thấy: $10,5 > 4,3$ hay $R > R'$ và $5,2 > 2,2$ hay $\Delta_Q > \Delta'_Q$.

Điều đó có nghĩa là trong một năm, nhiệt độ ở Lâm Đồng ít thay đổi hơn.

1.3. Giá trị ngoại lệ

Khoảng tứ phân vị được dùng để xác định các *giá trị ngoại lệ* trong mẫu, đó là các giá trị quá nhỏ hay quá lớn so với đa số các giá trị của mẫu. Cụ thể, phần tử x trong mẫu là giá trị ngoại lệ nếu $x > Q_3 + 1,5\Delta_Q$ hoặc $x < Q_1 - 1,5\Delta_Q$.

Sự xuất hiện của các giá trị ngoại lệ làm cho số trung bình và phạm vi của mẫu thay đổi lớn. Do đó, khi mẫu có giá trị ngoại lệ, người ta thường sử dụng trung vị và khoảng tứ phân vị để đo mức độ tập trung và mức độ phân tán của đa số các phần tử trong mẫu số liệu.

Ví dụ: Trong ví dụ ở phần 1.1, ta có:

$Q_1 - 1,5\Delta_Q = 4 - 1,5 \cdot 5,5 = -4,25$

$Q_3 + 1,5\Delta_Q = 9,5 + 1,5 \cdot 5,5 = 17,75$

Do đó, mẫu có một giá trị ngoại lệ là 20.

2. Phương sai và độ lệch chuẩn

2.1. Công thức tính phương sai và độ lệch chuẩn

* Giả sử ta có một mẫu số liệu là $x_1, x_2, \ldots, x_n$.

• **Phương sai** của mẫu số liệu này, kí hiệu là S^2, được tính bởi công thức:

$$S^2 = \frac{1}{n}\left[\left(x_1 - \overline{x}\right)^2 + \left(x_2 - \overline{x}\right)^2 + \ldots + \left(x_n - \overline{x}\right)^2\right],$$

trong đó $\overline{x}$ là số trung bình của mẫu số liệu.

• Căn bậc hai của phương sai được gọi là **độ lệch chuẩn**, kí hiệu là S.

Chú ý: Có thể biến đổi công thức tính phương sai ở trên thành:

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

Trong thống kê, người ta cũng quan tâm đến phương sai hiệu chỉnh, kí hiệu là $\hat{s}^2$, được tính bởi công thức:

$$\hat{s}^2 = \frac{1}{n-1}\left[\left(x_1 - \overline{x}\right)^2 + \left(x_2 - \overline{x}\right)^2 + ... + \left(x_n - \overline{x}\right)^2\right].$$

* Giả sử mẫu số liệu được cho dưới dạng bảng tần số:

Giá trị	x_1	x_2	…	x_k
Tần số	n_1	n_2	…	n_k

Khi đó, công thức tính phương sai trở thành:

$$S^2 = \frac{1}{n}\left[n_1\left(x_1 - \overline{x}\right)^2 + n_2\left(x_2 - \overline{x}\right)^2 + ... + n_k\left(x_k - \overline{x}\right)^2\right],$$

trong đó $n = n_1 + n_2 + \ldots + n_k$.

Có thể biến đổi công thức tính phương sai trên thành

$$S^2 = \frac{1}{n}\left(n_1x_1^2 + n_2x_2^2 + ... + n_kx_k^2\right) - \overline{x}.$$

Ví dụ: Tính phương sai và độ lệch chuẩn của mẫu số liệu sau:

8; 10; 9; 7; 6; 10; 6; 7; 8; 9.

Hướng dẫn giải

Cỡ mẫu n = 10.

Số trung bình: (8 + 10 + 9 + 7 + 6 + 10 + 6 + 7 + 8 + 9) : 10 = 8.

Phương sai mẫu số liệu là:

$S^2 = \frac{1}{10}(8^2 + 10^2 + 9^2 + 7^2 + 6^2 + 10^2 + 6^2 + 7^2 + 8^2 + 9^2) - 8^2 = 2.$

Độ lệch chuẩn mẫu số liệu là $S = \sqrt{S^2} = \sqrt{2} \approx 1,41.$

Ví dụ: Điều tra số con của mỗi hộ gia đình trong tổ dân cư xóm 2, kết quả được ghi lại ở bảng sau:

Số con	0	1	2	3	4
Số hộ gia đình	4	4	8	3	1

Tính phương sai và độ lệch chuẩn của mẫu số liệu.

Hướng dẫn giải

Tổng số hộ gia đình là: n = 4 + 4 + 8 + 3 + 1 = 20 (hộ gia đình).

Số trung bình của mẫu số liệu trên là

$$\bar{x} = \frac{1}{20}(4 . 0 + 4 . 1 + 8 . 2 + 3 . 3 + 1 . 4) = 1,65$$

Phương sai của mẫu số liệu trên là:

$$S^2 = \frac{1}{20}(4 \cdot 0^2 + 4 \cdot 1^2 + 8 \cdot 2^2 + 3 \cdot 3^2 + 1 \cdot 4^2) - 1{,}65^2 = 1{,}2275$$

Độ lệch chuẩn của mẫu số liệu trên là:

$$S = \sqrt{S^2} = \sqrt{1{,}2275} \approx 1{,}11.$$

2.2. Ý nghĩa của phương sai và độ lệch chuẩn

Phương sai là trung bình cộng của các bình phương độ lệch từ mỗi giá trị của mẫu số liệu đến số trung bình.

Phương sai và độ lệch chuẩn được dùng để đo mức độ phân tán của các số liệu trong mẫu quanh số trung bình. Phương sai và độ lệch chuẩn càng lớn thì các giá trị của mẫu càng cách xa nhau (có độ phân tán lớn).

Ví dụ: Bảng dưới đây thống kê tổng số giờ nắng trong năm 2019 theo từng tháng được đo bởi hai trạm quan sát khí tượng đặt ở Tuyên Quang và Cà Mau.

Tháng	1	2	3	4	5	6	7	8	9	10	11	12
Tuyên Quang	25	89	72	117	106	177	156	203	227	146	117	145
Cà Mau	180	223	257	245	191	111	141	134	130	122	157	173

(Nguồn: Tổng cục Thống kê)

a) Hãy tính phương sai và độ lệch chuẩn của dữ liệu từng tỉnh.

b) Nêu nhận xét về sự thay đổi tổng số giờ nắng theo từng tháng ở mỗi tỉnh.

Hướng dẫn giải

a)

* Tỉnh Tuyên Quang:

+ Số trung bình:

$$\overline{x_1} = \frac{25+89+72+117+106+177+156+203+227+146+117+145}{12} \approx 131,67.$$

+ Phương sai mẫu số liệu ở tỉnh Tuyên Quang là:

$$S_1^2 = \frac{1}{12}\left(25^2+89^2+72^2+117^2+106^2+177^2+156^2+203^2+227^2+146^2+117^2+145^2\right)$$

$$-\left(131,67\right)^2 \approx 2920,34.$$

+ Độ lệch chuẩn mẫu số liệu ở tỉnh Tuyên Quang là:

$$S_1 = \sqrt{S_1^2} = \sqrt{2920,34} \approx 54,04.$$

* Tỉnh Cà Mau:

+ Số trung bình:

$$\overline{x_2} = \frac{180+223+257+245+191+111+141+134+130+122+157+173}{12} = 172.$$

+ Phương sai mẫu số liệu ở tỉnh Cà Mau là:

$$S_2^2 = \frac{1}{12}(180^2 + 223^2 + 257^2 + 245^2 + 191^2 + 111^2 + 141^2 + 134^2 + 130^2 + 122^2 + 157^2$$

$$+ 173^2) - 172^2 = 2183.$$

+ Độ lệch chuẩn mẫu số liệu ở tỉnh Cà Mau là:

$$S_2 = \sqrt{S_2^2} = \sqrt{2183} \approx 46,72.$$

b) Phương sai mẫu và độ lệch chuẩn mẫu số liệu ở tỉnh Tuyên Quang cao hơn tỉnh Cà Mau nên tổng số giờ nắng trong năm 2019 theo từng tháng ở tỉnh Tuyên Quang

có độ phân tán cao hơn ở tỉnh Cà Mau. Do đó, sự thay đổi tổng số giờ nắng theo từng tháng ở tỉnh Cà Mau ổn định (có ít sự thay đổi) hơn so với tỉnh Tuyên Quang.

B. Bài tập tự luyện

Bài 1. Hãy tìm độ lệch chuẩn, khoảng biến thiên, khoảng tứ phân vị và các giá trị ngoại lệ (nếu có) của mẫu số liệu sau: 6; 8; 3; 4; 5; 6; 7; 2; 4.

Hướng dẫn giải

Số trung bình: $\bar{x} = \frac{6+8+3+4+5+6+7+2+4}{9} = 5$.

Phương sai mẫu số liệu là:

$S^2 = \frac{1}{9}(6^2 + 8^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 + 2^2 + 4^2) - 5^2 = \frac{10}{3}$.

Độ lệch chuẩn mẫu số liệu là: $S = \sqrt{S^2} = \sqrt{\frac{10}{3}} = \frac{\sqrt{30}}{3}$.

Sắp xếp các số liệu theo thứ tự không giảm, ta được:

2; 3; 4; 4; 5; 6; 6; 7; 8.

Khoảng biến thiên của mẫu là: R = 8 – 2 = 6.

Vì cỡ mẫu là 9 là số lẻ nên tứ phân vị thứ hai là $Q_2 = 5$.

Tứ phân vị thứ nhất là trung vị của mẫu: 2; 3; 4; 4. Do đó $Q_1 = 3,5$.

Tứ phân vị thứ ba là trung vị của mẫu: 6; 6; 7; 8. Do đó $Q_3 = 6,5$.

Khoảng tứ phân vị của mẫu là: $\Delta_Q = 6,5 – 3,5 = 3$.

Ta có: $Q_3 + 1,5\Delta_Q = 6,5 + 1,5 \cdot 3 = 11$ và $Q_1 – 1,5\Delta_Q = 3,5 – 1,5 \cdot 3 = – 1$.

Do đó mẫu số liệu không có giá trị ngoại lệ.

Bài 2. Hai lớp 10A, 10B của một trường Trung học phổ thông đồng thời làm bài thi môn Toán theo cùng một đề thi. Kết quả thi được trình bày ở hai bảng phân bố tần số sau đây:

Điểm thi Toán của lớp 10A

Điểm thi	5	6	7	8	9	10	Cộng
Tần số	3	7	12	14	3	1	40

Điểm thi Toán của lớp 10B

Điểm thi	6	7	8	9	Cộng
Tần số	8	18	10	4	40

a) Tính các số trung bình cộng, phương sai, độ lệch chuẩn của các mẫu số liệu đã cho.

b) Xét xem kết quả làm bài thi môn Toán ở lớp nào đồng đều hơn?

Hướng dẫn giải

a)

* Lớp 10A:

Số trung bình: $\overline{x_A} = \frac{1}{40}(3 . 5 + 7 . 6 + 12 . 7 + 14 . 8 + 3 . 9 + 1 . 10) = 7,25.$

Phương sai mẫu số liệu:

$S_A^2 = \frac{1}{40}(3 . 5^2 + 7 . 6^2 + 12 . 7^2 + 14 . 8^2 + 3 . 9^2 + 1 . 10^2) - 7,25^2 = 1,2875.$

Độ lệch chuẩn: $S_A = \sqrt{S_A^2} = \sqrt{1,2875} \approx 1,135.$

* Lớp 10B:

Số trung bình: $\overline{x_B} = \frac{1}{40}(8 \,.\, 6 + 18 \,.\, 7 + 10 \,.\, 8 + 4 \,.\, 9) = 7{,}25.$

Phương sai mẫu số liệu:

$S_B^2 = \frac{1}{40}(8 \,.\, 6^2 + 18 \,.\, 7^2 + 10 \,.\, 8^2 + 4 \,.\, 9^2) - 7{,}25^2 = 0{,}7875.$

Độ lệch chuẩn: $S_B = \sqrt{S_B^2} = \sqrt{0{,}7875} \approx 0{,}887$.

b) Vì 0,887 < 1,135 nên $S_B < S_A$ hay độ lệch chuẩn của mẫu số liệu lớp 10B nhỏ hơn lớp 10A.

Vậy kết quả làm bài thi của học sinh lớp 10B đồng đều hơn.

Bài 3. Kết quả điều tra mức lương hằng tháng của một số công nhân của hai nhà máy A và B được cho ở bảng sau (đơn vị: triệu đồng):

Công nhân nhà máy A	4	5	5	47	5	6	4	4	
Công nhân nhà máy B	2	9	9	8	10	9	9	11	9

a) Hãy tìm số trung bình, mốt, tứ phân vị và độ lệch chuẩn của hai mẫu số liệu lấy từ nhà máy A và nhà máy B.

b) Hãy tìm các giá trị ngoại lệ trong mỗi mẫu số liệu trên. Công nhân nhà máy nào có mức lương cao hơn? Tại sao?

Hướng dẫn giải

a)

* Nhà máy A:

+ Số trung bình mức lương hàng tháng: $\overline{x_A} = \frac{4+5+5+47+5+6+4+4}{8} = 10$.

+ Giá trị 4 và 5 có tần số lớn nhất nên mốt của mẫu số liệu ở nhà máy A là 4 và 5.

+ Sắp xếp các số liệu theo thứ tự không giảm, ta được:

4; 4; 4; 5; 5; 5; 6; 47.

Vì cỡ mẫu là 8 là số chẵn nên tứ phân vị thứ hai là $Q_{2A} = 5$.

Tứ phân vị thứ nhất là trung vị của mẫu: 4; 4; 4; 5. Do đó $Q_{1A} = 4$.

Tứ phân vị thứ ba là trung vị của mẫu: 5; 5; 6; 47. Do đó $Q_{3A} = 5,5$.

+ Phương sai mẫu:

$S_A^2 = \frac{1}{8}(4^2 + 5^2 + 5^2 + 47^2 + 5^2 + 6^2 + 4^2 + 4^2) - 10^2 = 196$.

+ Độ lệch chuẩn: $S_A = \sqrt{S_A^2} = \sqrt{196} = 14$.

* Nhà máy B:

+ Số trung bình mức lương hàng tháng: $\overline{x_B} = \frac{2+9+9+8+10+9+9+11+9}{9} \approx 8,4$.

+ Giá trị 9 có tần số lớn nhất nên mốt của mẫu số liệu ở nhà máy B là 9.

+ Sắp xếp các số liệu theo thứ tự không giảm, ta được:

2; 8; 9; 9; 9; 9; 9; 10; 11.

Vì cỡ mẫu là 9 là số lẻ nên tứ phân vị thứ hai là $Q_{2B} = 9$.

Tứ phân vị thứ nhất là trung vị của mẫu: 2; 8; 9; 9. Do đó $Q_{1B} = 8,5$.

Tứ phân vị thứ ba là trung vị của mẫu: 9; 9; 10; 11. Do đó $Q_{3B} = 9,5$.

+ Phương sai mẫu:

$S_B^2 = \frac{1}{9}(2^2 + 8^2 + 9^2 + 9^2 + 9^2 + 9^2 + 9^2 + 10^2 + 11^2) - 8{,}4^2 = 6{,}55.$

+ Độ lệch chuẩn: $S_B = \sqrt{S_B^2} = \sqrt{6{,}55} \approx 2{,}6.$

b)

+ Khoảng tứ phân vị của mẫu số liệu ở nhà máy A là: $\Delta_{QA} = 5{,}5 - 4 = 1{,}5.$

Ta có: $Q_{3A} + 1{,}5\Delta_{QA} = 5{,}5 + 1{,}5 \,.\, 1{,}5 = 7{,}75$ và $Q_{1A} - 1{,}5\Delta_{QA} = 4 - 1{,}5 \,.\, 1{,}5 = 1{,}75.$

Do đó giá trị ngoại lệ trong mẫu số liệu ở nhà máy A là 47.

+ Khoảng tứ phân vị của mẫu số liệu ở nhà máy B là: $\Delta_{QB} = 9{,}5 - 8{,}5 = 1.$

Ta có: $Q_{3B} + 1{,}5\Delta_{QB} = 9{,}5 + 1{,}5 \,.\, 1 = 11$ và $Q_{1B} - 1{,}5\Delta_{QB} = 8{,}5 - 1{,}5 \,.\, 1 = 7.$

Do đó giá trị ngoại lệ trong mẫu số liệu ở nhà máy B là 2.

+ Quan sát các số liệu tính được ở câu a), ta thấy

- Số trung bình mức lương hàng tháng của công nhân ở nhà máy A cao hơn nhà máy B.

- Phương sai mẫu và độ lệch chuẩn mẫu số liệu ở nhà máy A cao hơn nhà máy B nên mức lương hằng tháng của công nhân nhà máy A có độ phân tán cao hơn nhà máy B, do đó mức lương của công nhân nhà máy B ổn định hơn nhà máy A.

- Mức lương xuất hiện nhiều nhất trong mẫu A là 4 và 5 triệu đồng, nhà máy B là 9 triệu đồng.

Do đó, ta có thể khẳng định công nhân nhà máy A có mức lương cao hơn (đều và ổn định hơn).