

### Bài 3. Các số đặc trưng đo mức độ phân tán cho mẫu số liệu không ghép nhóm

#### A. Lý thuyết

##### I. Khoảng biến thiên. Khoảng tứ phân vị

###### 1. Định nghĩa

- Trong một mẫu số liệu, *khoảng biến thiên* là hiệu số giữa giá trị lớn nhất và giá trị nhỏ nhất của mẫu số liệu đó.

Ta có thể tính khoảng biến thiên  $R$  của mẫu số liệu theo công thức sau:  $R = x_{\max} - x_{\min}$ , trong đó  $x_{\max}$  là giá trị lớn nhất,  $x_{\min}$  là giá trị nhỏ nhất của mẫu số liệu đó.

- Giả sử  $Q_1, Q_2, Q_3$  là tứ phân vị của mẫu số liệu. Ta gọi hiệu  $\Delta_Q = Q_3 - Q_1$  là *khoảng tứ phân vị* của mẫu số liệu đó.

**Chú ý:** Khoảng tứ phân vị của mẫu số liệu còn gọi là khoảng trải giữa (tiếng Anh là InterQuartile Range – IQR) của mẫu số liệu đó.

**Ví dụ:** Mẫu số liệu thống kê cân nặng (đơn vị: kg) của 8 học sinh trong một tổ như sau:

45    52    41    37    50    52    66    49

a) Tìm khoảng biến thiên của mẫu số liệu trên.

b) Tìm khoảng tứ phân vị của mẫu số liệu trên.

###### Hướng dẫn giải

a) Trong mẫu có số liệu lớn nhất là 66, số liệu nhỏ nhất là 37.

Khi đó, khoảng biến thiên của mẫu là  $R = x_{\max} - x_{\min} = 66 - 37 = 29$  (kg).

Vậy khoảng biến thiên của mẫu là  $R = 29$  kg.

b) Sắp xếp mẫu theo thứ tự tăng dần, ta được:

37    41    45    49    50    52    52    66

Khi đó, trung vị của mẫu là:  $Q_2 = \frac{49+50}{2} = 49,5$ .

$Q_1$  là trung vị của mẫu 37, 41, 45, 49 nên  $Q_1 = \frac{41+45}{2} = 43$ .

$Q_3$  là trung vị của mẫu 50, 52, 52, 66 nên  $Q_3 = \frac{52+52}{2} = 52$ .

Khi đó, ta có khoảng tứ phân vị là:

$$\Delta_Q = Q_3 - Q_1 = 52 - 43 = 9 \text{ (kg)}.$$

Vậy khoảng tứ phân vị của mẫu là  $\Delta_Q = 9 \text{ kg}$ .

## 2. Ý nghĩa

**a) Ý nghĩa của khoảng biến thiên:** *Khoảng biến thiên của mẫu số liệu phản ánh sự “dao động”, “sự dàn trải” của các số liệu trong mẫu đó. Khoảng biến thiên được sử dụng trong nhiều tình huống thực tiễn, chẳng hạn: tìm ra sự phân tán điểm kiểm tra của một lớp học hay xác định phạm vi giá cả của một dịch vụ ...*

Theo cách nhìn như ở trong vật lí, ở đó biên độ dao động phản ánh khoảng cách từ điểm cân bằng đến điểm xa nhất của dao động, nếu coi số trung bình cộng là “điểm cân bằng” của mẫu số liệu thì khoảng biến thiên của mẫu số liệu có thể xem như hai lần biên độ dao động của các số liệu trong mẫu đó quanh điểm cân bằng.

Trong các đại lượng đo mức độ phân tán của mẫu số liệu, khoảng biến thiên là đại lượng dễ hiểu, dễ tính toán và tương đối tốt đối với các mẫu số liệu nhỏ. Tuy nhiên, do khoảng biến thiên chỉ sử dụng hai giá trị  $x_{\max}$  và  $x_{\min}$  của mẫu số liệu nên đại lượng đó chưa diễn giải đầy đủ sự phân tán của số liệu trong mẫu. Ngoài ra, giá trị

của khoảng biến thiên sẽ bị ảnh hưởng bởi các giá trị bất thường của mẫu số liệu đó. Trong những trường hợp như vậy, khoảng biến thiên của mẫu số liệu không phản ánh chính xác độ dàn trải của mẫu số liệu.

**b) Ý nghĩa của khoảng tứ phân vị:** Khoảng tứ phân vị là đại lượng cho biết mức độ phân tán của 50% số liệu chính giữa của mẫu số liệu đã sắp xếp và có thể giúp xác định các giá trị bất thường của mẫu số liệu đó. Khoảng tứ phân vị thường được sử dụng thay cho khoảng biến thiên vì nó loại trừ hầu hết giá trị bất thường của mẫu số liệu.

## II. Phương sai

### 1. Định nghĩa

- Mỗi hiệu số giữa số liệu và số trung bình cộng gọi là *độ lệch* của số liệu đó đối với số trung bình cộng.

- Cho mẫu số liệu thống kê có  $n$  giá trị  $x_1, x_2, \dots, x_n$  và số trung bình cộng là  $\bar{x}$ .

Ta gọi số 
$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$
 là *phương sai* của mẫu số liệu trên.

### Nhận xét:

- Khi có các số liệu bằng nhau, ta có thể tính phương sai theo công thức sau:

+ Đối với bảng tần số:

Giá trị	$x_1$	$x_2$	...	$x_k$
Tần số	$n_1$	$n_2$	...	$n_k$

Phương sai của mẫu số liệu thống kê trong bảng phân bố tần số là:

$$s^2 = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \dots + n_k(x_k - \bar{x})^2}{n},$$

trong đó  $n = n_1 + n_2 + \dots + n_k$ ;  $\bar{x}$  là số trung bình cộng của các số liệu đã cho.

+ Đối với bảng phân bố tần số tương đối:

Giá trị	$x_1$	$x_2$	...	$x_k$
Tần số tương đối	$f_1$	$f_2$	...	$f_k$

Phương sai của mẫu số liệu thống kê trong bảng phân bố tần số tương đối là:

$$s^2 = f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \dots + f_k(x_k - \bar{x})^2,$$

trong đó  $\bar{x}$  là số trung bình cộng của số liệu đã cho.

- Trong thực tế, người ta còn dùng công thức sau để tính phương sai của mẫu số liệu:

$$\hat{s}^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1},$$

trong đó:  $x_i$  là giá trị của quan sát thứ  $i$ ;  $\bar{x}$  là giá trị trung bình và  $n$  là số quan sát trong mẫu số liệu đó.

**Ví dụ:** Hai lớp 10A và 10B của một trường THPT đồng thời làm bài thi môn Toán theo cùng một đề thi. Kết quả được ghi lại trong bảng tần số sau:

Điểm thi của lớp 10A:

Điểm thi	5	6	7	8	9	10
----------	---	---	---	---	---	----

Số học sinh	3	7	12	14	3	1
-------------	---	---	----	----	---	---

Điểm thi của lớp 10B:

Điểm thi	6	7	8	9
Số học sinh	8	18	10	4

a) Tính phương sai của từng mẫu số liệu ở hai bảng trên.

b) Xét xem kết quả bài thi của lớp nào đồng đều hơn.

### Hướng dẫn giải

Ta có điểm thi trung bình của lớp 10A là:

$$\bar{x}_{10A} = \frac{3.5 + 7.6 + 12.7 + 14.8 + 3.9 + 1.10}{3 + 7 + 12 + 14 + 3 + 1} = 7,25$$

Điểm thi trung bình của lớp 10A là:

$$\bar{x}_{10B} = \frac{8.6 + 18.7 + 10.8 + 4.9}{8 + 18 + 10 + 4} = 7,25$$

Ta có phương sai của mẫu số liệu lớp 10A là:

$$s_{10A}^2 = \frac{3(5 - 7,25)^2 + 7(6 - 7,25)^2 + \dots + 1.(10 - 7,25)^2}{3 + 7 + 12 + 14 + 3 + 1} = 1,2875$$

$$\Rightarrow s_{10A}^2 = 1,2875$$

Ta có phương sai của mẫu số liệu lớp 10B là:

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

$$\Rightarrow s_{10B}^2 = 0,7875.$$

Ta thấy điểm thi trung bình của lớp 10A và 10B bằng nhau đều bằng 7,25. Nhưng phương sai bảng điểm của lớp 10A lại lớn hơn phương sai bảng điểm lớp 10B nên kết quả làm bài thi của lớp 10B đồng đều hơn lớp 10A.

## 2. Ý nghĩa

Phương sai là số đặc trưng đo mức độ phân tán của mẫu số liệu. Mẫu số liệu nào có phương sai nhỏ hơn thì mức độ phân tán (so với số trung bình cộng) của các số liệu trong mẫu đó sẽ thấp hơn.

## III. Độ lệch chuẩn

### 1. Định nghĩa

Căn bậc hai (số học) của phương sai gọi là *độ lệch chuẩn* của mẫu số liệu thống kê.

**Nhận xét:** Vì độ lệch chuẩn có cùng đơn vị đo với số liệu thống kê nên khi cần chú ý đến đơn vị đo thì ta sử dụng độ lệch chuẩn mà không sử dụng phương sai.

**Ví dụ:** Cho mẫu số liệu:

23    22    20    12    35

Tính độ lệch chuẩn của mẫu trên.

### Hướng dẫn giải

Mẫu trên có 5 số liệu.

Số trung bình của mẫu trên là:  $\bar{x} = \frac{23 + 22 + 20 + 12 + 35}{5} = 22,4$ .

Phương sai của mẫu số liệu đó là:

$$s^2 = \frac{(23 - 22,4)^2 + (22 - 22,4)^2 + (20 - 22,4)^2 + (12 - 22,4)^2 + (35 - 22,4)^2}{5} = 54,64$$

$$\Rightarrow s = \sqrt{s^2} = \sqrt{54,64} \approx 7,39.$$

Vậy độ lệch chuẩn của mẫu là 7,39.

## 2. Ý nghĩa

Cũng như phương sai, khi hai mẫu số liệu thống kê có cùng đơn vị đo và có số trung bình cộng bằng nhau (hoặc xấp xỉ nhau), mẫu số liệu nào có độ lệch chuẩn nhỏ hơn thì mức độ phân tán (so với số trung bình cộng) của các số liệu trong mẫu đó sẽ thấp hơn. Độ lệch chuẩn là số đặc trưng đo mức độ phân tán của mẫu số liệu thống kê có cùng đơn vị đo.

#### IV. Tính hợp lí của số liệu thống kê

Ta có thể sử dụng các số đặc trưng đo mức độ phân tán cho mẫu số liệu không ghép nhóm để chỉ ra được những số liệu bất thường của mẫu số liệu đó. Ta thường sử dụng khoảng tứ phân vị để xác định số liệu bất thường của mẫu số liệu. Cụ thể như sau:

Giả sử  $Q_1, Q_2, Q_3$  là tứ phân vị của mẫu số liệu và hiệu  $\Delta_Q = Q_3 - Q_1$  là khoảng tứ phân vị của mẫu số liệu đó. Một giá trị trong mẫu số liệu được coi là một giá trị bất thường nếu nó nhỏ hơn  $Q_1 - \frac{3}{2}\Delta_Q$  hoặc lớn hơn  $Q_3 + \frac{3}{2}\Delta_Q$ . Như vậy, khoảng tứ phân vị cho ta cách nhận biết giá trị bất thường của mẫu số liệu.

**Chú ý:** Ta cũng có thể xác định số liệu bất thường của mẫu số liệu bằng số trung bình cộng và độ lệch chuẩn. Cụ thể như sau:

Giả sử  $\bar{x}$ ,  $s$  lần lượt là số trung bình cộng và độ lệch chuẩn của mẫu số liệu. Một giá trị trong mẫu số liệu cũng được coi là một giá trị bất thường nếu nó nhỏ hơn  $\bar{x} - 3s$  hoặc lớn hơn  $\bar{x} + 3s$ . Như vậy, số trung bình cộng và độ lệch chuẩn cho ta cách nhận ra giá trị bất thường của mẫu số liệu.

**Ví dụ:** Hãy tìm các giá trị bất thường của mẫu số liệu sau:

12    4    10    -5    6    7    9    30

#### Hướng dẫn giải

Mẫu được sắp xếp theo thứ tự tăng dần là:

-5    4    6    7    9    10    12    30

Mẫu có 8 số liệu

Trung vị của mẫu là:  $\frac{7+9}{2} = 8$ . Suy ra  $Q_2 = 8$ .

Trung vị nửa dưới  $-5, 4, 6, 7$  là  $\frac{4+6}{2} = 5$ . Suy ra  $Q_1 = 5$ .

Trung vị nửa trên  $9, 10, 12, 30$  là  $\frac{10+12}{2} = 11$ . Suy ra  $Q_3 = 11$ .

Khoảng tứ phân vị là  $\Delta_Q = Q_3 - Q_1 = 11 - 5 = 6$ .

Ta có:  $Q_1 - \frac{3}{2}\Delta_Q = 5 - \frac{3}{2}.6 = -4$ ;  $Q_3 + \frac{3}{2}\Delta_Q = 11 + \frac{3}{2}.6 = 20$ .

Ta thấy  $-5 < -4$  và  $30 > 20$  nên các giá trị  $-5$  và  $30$  là các giá trị bất thường của mẫu. Vậy mẫu có hai giá trị bất thường là  $-5$  và  $30$ .

## B. Bài tập tự luyện

### B.1 Bài tập tự luận

**Bài 1.** Cho mẫu số liệu thống kê chiều cao của 13 cây xoan đào (đơn vị: mét) như sau:

2,4   3   2,5   4   0,9   5   4,3   4   3,7   3,9   4,4   7,3   1,7

- a) Tìm khoảng biến thiên của mẫu số liệu trên.
- b) Tìm khoảng tứ phân vị của mẫu số liệu.
- c) Tìm các giá trị bất thường của mẫu.

### Hướng dẫn giải

a) Sắp xếp mẫu theo thứ tự tăng dần:

0,9   1,7   2,4   2,5   3   3,7   3,9   4   4   4,3   4,4   5   7,3

Mẫu có 13 số liệu, số liệu nhỏ nhất là 0,9; số liệu lớn nhất là 8.

Khi đó khoảng biến thiên của mẫu là  $R = 8 - 0,9 = 7,1$  (m).

Vậy khoảng biến thiên của mẫu là 7,1 mét.

b) Trung vị của mẫu là 3,9; tức là  $Q_2 = 3,9$ .



Trung vị nửa dưới của mẫu 0,9; 1,7; 2,4; 2,5; 3; 3,7 là  $\frac{2,4+2,5}{2} = 2,45$ . Suy ra  $Q_1 = 2,45$  (m).

Trung vị nửa dưới của mẫu 4; 4; 4,3; 4,4; 5; 8 là  $\frac{4,3+4,4}{2} = 4,35$ . Suy ra  $Q_3 = 4,35$ (m).

Khi đó khoảng tứ phân vị của mẫu là:  $\Delta_Q = Q_3 - Q_1 = 4,35 - 2,45 = 1,9$  (m).

Vậy khoảng tứ phân vị là 1,9 m.

c) Ta có:  $Q_1 - \frac{3}{2}\Delta_Q = 2,45 - \frac{3}{2}.1,9 = -0,4$ ;  $Q_3 + \frac{3}{2}\Delta_Q = 4,35 + \frac{3}{2}.1,9 = 7,2$ .

Ta thấy  $7,3 > 7,2$  nên giá trị 7,3 là các giá trị bất thường của mẫu.

Vậy mẫu có giá trị bất thường là 7,3.

**Bài 2.** Thời gian 5 lần chạy cự li 100 m (đơn vị: giây) của hai bạn Long và Nam được ghi lại trong bảng sau:

<b>Long</b>	20	23	19	21	22
<b>Nam</b>	18	21	20	22	22

a) Tính thời gian chạy trung bình của mỗi bạn Long và Nam. Bạn nào có kết quả chạy tốt hơn.

b) Tính phương sai, độ lệch chuẩn của mẫu số liệu thống kê kết quả 5 lần chạy của mỗi bạn. Từ đó cho biết bạn nào có kết quả chạy ổn định hơn.

### Hướng dẫn giải

a) Thời gian chạy trung bình của Long là:  $\overline{x_L} = \frac{20+23+19+21+22}{5} = 21$  (s).

Thời gian chạy trung bình của Nam là:  $\overline{x_N} = \frac{18+21+20+22+22}{5} = 20,6$  (s).

Do  $\overline{x_L} > \overline{x_N}$  ( $21 > 20,6$ ) nên kết quả chạy của Nam tốt hơn của Long.

Vậy thời gian chạy trung bình của Long là 21 giây, của Nam là 20,6 giây và kết quả chạy của Nam tốt hơn của Long.

b) Ta có phương sai của mẫu số liệu thống kê kết quả 5 lần chạy của Long là:

$$s_L^2 = \frac{(20-21)^2 + (23-21)^2 + (19-21)^2 + (21-21)^2 + (22-21)^2}{5} = 2$$

$$\Rightarrow s_L = \sqrt{s_L^2} = \sqrt{2} \approx 1,41.$$

Phương sai của mẫu số liệu thống kê kết quả 5 lần chạy của Nam là:

$$s_N^2 = \frac{(18-20,6)^2 + (21-20,6)^2 + (20-20,6)^2 + (22-20,6)^2 + (22-20,6)^2}{5} = 2,24.$$

$$\Rightarrow s_N = \sqrt{s_N^2} = \sqrt{2,24} \approx 1,5.$$

Ta thấy  $s_L^2 < s_N^2$  ( $2 < 2,24$ ) nên bạn Long có kết quả chạy ổn định hơn bạn Nam.

Vậy phương sai của mẫu số liệu thống kê kết quả 5 lần chạy của Long và Nam lần lượt là 2 và 2,24. Độ lệch chuẩn của mẫu số liệu thống kê kết quả 5 lần chạy của Long và Nam lần lượt là 1,41 và 1,5. Bạn Long có kết quả chạy ổn định hơn bạn Nam.

## B.2 Bài tập trắc nghiệm

**Câu 1.** Năng suất lúa hè thu (tạ/ha) năm 1998 của 31 tỉnh thành ở Việt Nam được thống kê trong bảng sau:

Năng suất lúa (tạ/ha)	25	30	35	40	45
Tần số	4	7	9	6	5

Hãy tính khoảng biến thiên của mẫu số liệu trên.

A. 40;

B. 20;

C. 61;

D. 1.

**Hướng dẫn giải**

**Đáp án đúng là: B**

Theo bảng số liệu trên ta có: Giá trị lớn nhất của số liệu là 45; giá trị nhỏ nhất của số liệu là 25.

Khoảng biến thiên :  $R = 45 - 25 = 20$ .

**Câu 2.** Tiến hành đo huyết áp của 8 người ta thu được kết quả sau:

77 105 117 84 96 72 105 124

Hãy tìm khoảng tứ phân vị của mẫu số liệu trên.

A. 10;

B. 20;

C. 10,5;

D. 30,5.

**Hướng dẫn giải**

**Đáp án đúng là: D**

Mẫu số liệu trên được sắp xếp theo thứ tự tăng dần như sau:

72 77 84 96 105 105 117 124

Trung vị của mẫu số liệu trên là:  $\frac{96+105}{2} = 100,5 \Rightarrow Q_2 = 100,5$ .

Nửa dãy phía dưới số 100,5 (nghĩa là những số nhỏ hơn 100,5) gồm: 72 77 84 96

có trung vị là  $\frac{77+84}{2} = 80,5 \Rightarrow Q_1 = 80,5$ .

Nửa dãy phía trên số 100,5 (nghĩa là những số lớn hơn 100,5) gồm: 105 105 117

124 có trung vị là  $\frac{105+117}{2} = 111 \Rightarrow Q_3 = 111$ .

Do đó, tứ phân vị của mẫu số liệu:  $Q_1 = 80,5$ ;  $Q_2 = 100,5$ ;  $Q_3 = 111$ .

Vậy khoảng tứ phân vị của mẫu số liệu là:  $\Delta_Q = Q_3 - Q_1 = 111 - 80,5 = 30,5$ .

**Câu 3.** Thời gian chạy 50 m của 20 học sinh được ghi lại trong bảng sau đây:

Thời gian (giây)	8,3	8,4	8,5	8,7	8,8
Tần số	2	3	9	5	1

Hãy tìm độ lệch chuẩn của mẫu số liệu đã cho.

A. 0,14;

B. 0,0191;

C. 8,53;

D. 8,5.

**Hướng dẫn giải**

**Đáp án đúng là: A**

Ta có:  $\bar{x} = \frac{2.8,3 + 3.8,4 + 9.8,5 + 5.8,7 + 1.8,8}{20} = 8,53$ .

Phương sai là:

$$s^2 = \frac{2.(8,3 - 8,53)^2 + 3.(8,4 - 8,53)^2 + 9.(8,5 - 8,53)^2 + 5.(8,7 - 8,53)^2 + (8,8 - 8,53)^2}{20}$$

$$= 0,0191$$

$$\Rightarrow \text{Độ lệch chuẩn: } s = \sqrt{0,0191} \approx 0,14.$$