

CHƯƠNG III

CÁC SỐ ĐẶC TRƯNG ĐO MỨC ĐỘ PHÂN TÁN CỦA MẪU SỐ LIỆU GHÉP NHÓM

Ở lớp 11, chúng ta đã được giới thiệu về các số đặc trưng đo xu thế trung tâm của mẫu số liệu ghép nhóm. Trong chương này, chúng ta tiếp tục tìm hiểu về các số đặc trưng đo mức độ phân tán của mẫu số liệu dạng này.



"Khả năng làm việc với dữ liệu để xử lý dữ liệu, trích xuất dữ liệu, trực quan hoá dữ liệu, truyền đạt dữ liệu sẽ là kỹ năng cực kì quan trọng trong những thập kỷ tới".

*Hal Varian – Trưởng phòng kinh tế của Google
(Theo forbes.com)*

Bài 9

KHOẢNG BIẾN THIÊN VÀ KHOẢNG TỬ PHÂN VỊ

THUẬT NGỮ

- Khoảng biến thiên
- Khoảng tứ phân vị

KIẾN THỨC, KỸ NĂNG

- Tính khoảng biến thiên, khoảng tứ phân vị của mẫu số liệu ghép nhóm.
- Hiểu ý nghĩa, vai trò của khoảng biến thiên, khoảng tứ phân vị trong việc đo mức độ phân tán.

Thống kê số ngày trong tháng Sáu năm 2021 và năm 2022 theo nhiệt độ cao nhất trong ngày tại Hà Nội, người ta thu được bảng sau:

Nhiệt độ (°C)	[28; 30)	[30; 32)	[32; 34)	[34; 36)	[36; 38)	[38; 40)
Số ngày trong tháng 6/2021	0	2	8	5	6	9
Số ngày trong tháng 6/2022	2	3	4	11	8	2

(Theo accuweather.com)

Hỏi tháng Sáu năm nào ở Hà Nội nhiệt độ cao nhất trong ngày biến đổi nhiều hơn?

Để biết tháng Sáu năm nào ở Hà Nội nhiệt độ cao nhất trong ngày biến đổi nhiều hơn, ta cần tính và so sánh các số đặc trưng đo mức độ phân tán của hai mẫu số liệu ghép nhóm trên. Chúng ta cùng tìm hiểu vấn đề này!

1. KHOẢNG BIẾN THIÊN

» **Hỏi.** Trong tình huống mở đầu, gọi $x_1, x_2, \dots, x_{30}$ là nhiệt độ cao nhất trong ngày của 30 ngày tháng Sáu năm 2021 (mẫu số liệu gốc).

- Có thể tính chính xác khoảng biến thiên cho mẫu số liệu gốc hay không?
- Giá trị lớn nhất, giá trị nhỏ nhất x_i có thể nhận là gì?
- Hãy đưa ra một giá trị xấp xỉ cho khoảng biến thiên của mẫu số liệu gốc.

Cho mẫu số liệu ghép nhóm:

Nhóm	$[a_1; a_2)$	...	$[a_i; a_{i+1})$	...	$[a_k; a_{k+1})$
Tần số	m_1	...	m_i	...	m_k

Bảng 3.1. Mẫu số liệu ghép nhóm

trong đó các tần số $m_1 > 0, m_k > 0$ và $n = m_1 + \dots + m_k$ là cỡ mẫu.

Khoảng biến thiên của mẫu số liệu ghép nhóm trên là $R = a_{k+1} - a_1$.

 Chỉ ra rằng khoảng biến thiên của mẫu số liệu ghép nhóm trong Bảng 3.1 lớn hơn khoảng biến thiên của mẫu số liệu gốc.

Ý nghĩa. Khoảng biến thiên của mẫu số liệu ghép nhóm xấp xỉ cho khoảng biến thiên của mẫu số liệu gốc. Khoảng biến thiên được dùng để đo mức độ phân tán của mẫu số liệu ghép nhóm. Khoảng biến thiên càng lớn thì mẫu số liệu càng phân tán.

» **Ví dụ 1.** Thống kê thời gian sử dụng mạng xã hội trong ngày của các bạn Tổ 1, Tổ 2 lớp 12A, được kết quả như bảng sau:

Thời gian sử dụng (phút)	$[0; 10)$	$[10; 30)$	$[30; 60)$	$[60; 90)$
Số học sinh Tổ 1	2	4	3	1
Số học sinh Tổ 2	5	1	3	0

Tìm khoảng biến thiên cho thời gian sử dụng mạng xã hội của học sinh mỗi tổ và giải thích ý nghĩa.

Giải

Gọi R_1, R_2 tương ứng là khoảng biến thiên của mẫu số liệu ghép nhóm về thời gian sử dụng mạng xã hội trong ngày của các bạn Tổ 1 và Tổ 2.

Ta có: $R_1 = 90 - 0 = 90$ và $R_2 = 60 - 0 = 60$.

Do $R_1 > R_2$ nên ta có thể kết luận rằng thời gian sử dụng mạng xã hội trong ngày của các bạn Tổ 1 phân tán hơn thời gian sử dụng mạng xã hội của các bạn Tổ 2.

» **Luyện tập 1.** Thời gian hoàn thành bài kiểm tra môn Toán của các bạn trong lớp 12C được cho trong bảng sau:

Thời gian (phút)	[25; 30)	[30; 35)	[35; 40)	[40; 45)
Số học sinh	8	16	4	2

- Tính khoảng biến thiên R cho mẫu số liệu ghép nhóm trên.
- Nếu biết học sinh hoàn thành bài kiểm tra sớm nhất mất 27 phút và muộn nhất mất 43 phút thì khoảng biến thiên của mẫu số liệu gốc là bao nhiêu?

2. KHOẢNG TỨ PHÂN VỊ

» **HĐ2.** Trong tình huống mở đầu, gọi $y_1, y_2, \dots, y_{30}$ là nhiệt độ cao nhất trong ngày của 30 ngày tháng Sáu năm 2022 (mẫu số liệu gốc).

- Có thể tính chính xác khoảng tứ phân vị của mẫu số liệu gốc hay không?
- Tìm tứ phân vị thứ nhất Q_1 và tứ phân vị thứ ba Q_3 cho mẫu số liệu ghép nhóm.
- Hãy đưa ra một giá trị xấp xỉ cho khoảng tứ phân vị của mẫu số liệu gốc.

Xét mẫu số liệu ghép nhóm cho bởi Bảng 3.1.

Tứ phân vị thứ r là

$$Q_r = a_p + \frac{\frac{r \cdot n}{4} - (m_1 + \dots + m_{p-1})}{m_p} \cdot (a_{p+1} - a_p),$$

trong đó $[a_p; a_{p+1})$ là nhóm chứa tứ phân vị thứ r với $r = 1, 2, 3$.



Khoảng tứ phân vị của mẫu số liệu ghép nhóm, kí hiệu là Δ_Q , là hiệu số giữa tứ phân vị thứ ba Q_3 và tứ phân vị thứ nhất Q_1 của mẫu số liệu đó, tức là $\Delta_Q = Q_3 - Q_1$.

Ý nghĩa. Khoảng tứ phân vị của mẫu số liệu ghép nhóm xấp xỉ cho khoảng tứ phân vị của mẫu số liệu gốc. Khoảng tứ phân vị cũng được dùng để đo mức độ phân tán của mẫu số liệu ghép nhóm. Khoảng tứ phân vị càng lớn thì mẫu số liệu càng phân tán.

Nhận xét. Do khoảng tứ phân vị của mẫu số liệu ghép nhóm chỉ phụ thuộc vào nửa giữa của mẫu số liệu, nên không bị ảnh hưởng bởi các giá trị bất thường và có thể dùng đại lượng này để loại giá trị bất thường.

» **Ví dụ 2.** Thời gian chờ khám bệnh của các bệnh nhân tại phòng khám X được cho trong bảng sau:

Thời gian (phút)	[0; 5)	[5; 10)	[10; 15)	[15; 20)
Số bệnh nhân	3	12	15	8

- Tìm khoảng tứ phân vị của mẫu số liệu ghép nhóm này.
- Từ một mẫu số liệu về thời gian chờ khám bệnh của các bệnh nhân tại phòng khám Y người ta tính được khoảng tứ phân vị bằng 9,23. Hỏi trong hai phòng khám X và Y, thời gian chờ của bệnh nhân tại phòng khám nào phân tán hơn?

Giải

- a) Cỡ mẫu là $n = 3 + 12 + 15 + 8 = 38$. Gọi $x_1, \dots, x_{38}$ là thời gian chờ khám bệnh của 38 bệnh nhân này và giả sử rằng dãy số liệu gốc này đã được sắp xếp theo thứ tự tăng dần. Tứ phân vị thứ nhất của mẫu số liệu gốc là x_{10} nên nhóm chứa tứ phân vị thứ nhất là nhóm $[5; 10)$ và ta có:

$$Q_1 = 5 + \left[\frac{\frac{38}{4} - 3}{12} \right] \cdot 5 \approx 7,71.$$

Tứ phân vị thứ ba của mẫu số liệu gốc là x_{29} nên nhóm chứa tứ phân vị thứ ba là nhóm $[10; 15)$ và ta có:

$$Q_3 = 10 + \left[\frac{\frac{3 \cdot 38}{4} - 15}{15} \right] \cdot 5 = 14,5.$$

Vậy khoảng tứ phân vị của mẫu số liệu ghép nhóm là: $\Delta_Q = Q_3 - Q_1 \approx 14,5 - 7,71 = 6,79$.

- b) Do $\Delta_Q = 6,79 < 9,23$ nên thời gian chờ của bệnh nhân tại phòng khám Y phân tán hơn thời gian chờ của bệnh nhân tại phòng khám X.

» **Luyện tập 2.** Một người ghi lại thời gian đàm thoại của một số cuộc gọi cho kết quả như bảng sau:

Thời gian t (phút)	Số cuộc gọi
$0 \leq t < 1$	8
$1 \leq t < 2$	17
$2 \leq t < 3$	25
$3 \leq t < 4$	20
$4 \leq t < 5$	10

Tính khoảng tứ phân vị của mẫu số liệu ghép nhóm trên.

» **Vận dụng.** Hãy giải bài toán trong *tình huống mở đầu* bằng cách sử dụng khoảng biến thiên và khoảng tứ phân vị của mẫu số liệu ghép nhóm.

BÀI TẬP

3.1. Thống kê số thẻ vàng của mỗi cầu lạc bộ trong giải ngoại hạng Anh mùa giải 2021 – 2022 cho kết quả như sau:

101	79	79	78	75	73	68	67	67	63
63	61	60	59	57	55	55	50	47	42.

(Theo *premierleague.com*)

- a) Hãy ghép nhóm dãy số liệu trên thành các nhóm có độ dài bằng nhau với nhóm đầu tiên là $[40; 50)$.
- b) Tính khoảng biến thiên, khoảng tứ phân vị của mẫu số liệu gốc và mẫu số liệu ghép nhóm thu được ở câu a. Giá trị nào là giá trị chính xác? Giá trị nào là giá trị xấp xỉ?

3.2. Thu nhập theo tháng của người lao động ở hai nhà máy như sau:

Thu nhập (triệu đồng)	$[5; 8)$	$[8; 11)$	$[11; 14)$	$[14; 17)$	$[17; 20)$
Số người của nhà máy A	20	35	45	35	20
Số người của nhà máy B	17	23	30	23	17

Tính mức thu nhập trung bình của người lao động ở hai nhà máy trên. Dựa vào khoảng tứ phân vị, hãy xác định xem mức thu nhập của người lao động ở nhà máy nào biến động nhiều hơn.

3.3. Bảng sau đây cho biết chiều cao của các học sinh hai lớp 12A và 12B.

Chiều cao (cm)	$[145; 150)$	$[150; 155)$	$[155; 160)$	$[160; 165)$	$[165; 170)$	$[170; 175)$
Số học sinh của lớp 12A	1	0	15	12	10	5
Số học sinh của lớp 12B	0	0	17	10	9	6

- a) Tìm khoảng biến thiên, khoảng tứ phân vị cho các mẫu số liệu ghép nhóm về chiều cao của học sinh hai lớp 12A và 12B.
- b) Để so sánh độ phân tán về chiều cao của học sinh hai lớp này ta nên dùng khoảng biến thiên hay khoảng tứ phân vị? Vì sao?