BÀI GIẢNG: KHOẢNG BIẾN THIÊN VÀ KHOẢNG TỨ PHÂN VỊ TIẾT 1
CHUYÊN ĐỀ: CÁC SỐ ĐẶC TRƯNG ĐO MỨC ĐỘ PHÂN TÁN CỦA MẪU SỐ LIỆU GHÉP NHÓM
MÔN: TOÁN 12 (KẾT NỐI TRI THỨC VỚI CUỘC SỐNG)

GIÁO VIÊN: NGUYỄN CÔNG CHÍNH

MỤC TIÊU

- ✓ *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 (^{0}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!

I. KHOẢNG BIẾN THIÊN

HĐ1 Trong tình huống mở đầu, gọi $x_1, x_2, \ldots, x_{30}$ là nhiệt độ cao nhất trong ngày của 30 ngày tháng Sáu năm 2021 (mẫu số liệu gốc).

a) Có thể tính chính xác khoảng biến thiên cho mẫu số liệu gốc hay không?

b) Giá trị lớn nhất, giá trị nhỏ nhất x_i có thể nhận là gì?

c) Hãy đưa ra một giá trị xấp xỉ cho khoảng biến thiên của mẫu số liệu gốc.

Cho mẫu số liệu ghép nhóm:

Nhóm	$[a_1;a_2)$	$\cdots$	$[a_i;a_{i+1})$	$\cdots$	$[a_k;a_{k+1})$
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 $n = m_1 + \ldots + m_k$ là cỡ mẫu.

Cách giải:

a) Không thể tính chính xác khoảng biến thiên cho mẫu số liệu gốc.

b) Giá trị nhỏ nhất có thể là 30°C, giá trị lớn nhất là giá trị nhiệt độ lớn nhất có thể là $39,9^0C$.

c) Một giá trị xấp xỉ cho khoảng biến thiên của mẫu số liệu gốc là: $39,9-30=9,9\left(^{\circ}\mathrm{C}\right)$

Kết luận:

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.

Cách giải:

Khoảng biến thiên của mẫu số liệu ghép nhóm trong bảng 3.1 là: $R=a_{k+1}-a_1$.

Gọi giá trị nhỏ nhất của mẫu số liệu gốc là a_1' thì $a_1'\geq a_1$.

Gọi giá trị lớn nhất của mẫu số liệu gốc là a_k' thì $a_{k+1}'<a_{k+1}$.

Khi đó, khoảng biến thiên của mẫu số liệu gốc là: $R'=a_{k+1}'-a_1'$.

Do đó, $R>R'$

Vậy 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.

Cách 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	14	4	2

a) Tính khoảng biến thiên R cho mẫu số liệu ghép nhóm trên.

b) 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?

Cách giải:

a) Khoảng biến thiên R cho mẫu số liệu ghép nhóm trên là: $45-25=20$

b) Khoảng biến thiên của mẫu số liệu gốc là: $43-27=16$

II. KHOẢNG TỨ PHÂN VỊ

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

a) Có thể tính chính xác khoảng tứ phân vị của mẫu số liệu gốc hay không?

b) Tìm tứ phân vị thứ nhất Q_1 và tứ phân vị thứ ba Q_3 cho mẫu số liệu ghép nhóm.

c) 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

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

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



Cách giải:

a) Không thể tính chính xác khoảng tứ phân vị của mẫu số liệu gốc.

b) Cỡ mẫu $n=30$.

Giả sử $y_1, y_2, \ldots, y_{30}$ là nhiệt độ cao nhất trong ngày của 30 ngày tháng Sáu năm 2022 và giả sử dãy số liệu gốc này đã được sắp xếp theo thứ tự tăng dần.

Vì $\frac{n}{4}=\frac{30}{4}=7,5$ và $2+3<7,5<2+3+4$ nên nhóm chứa tứ phân vị thứ nhất là nhóm $[32;34)$ và tứ phân vị thứ nhất là: $Q_1 = 32 + \frac{\frac{30}{4}-(2+3)}{4}\cdot(34-32) = 33,25$

Vì $\frac{3n}{4}=\frac{3.30}{4}=22,5$ và $2+3+4+11<22,5<2+3+4+11+8$ nên nhóm chứa tứ phân vị thứ ba là nhóm $[36;38)$ và tứ phân vị thứ ba là: $Q_3 = 36 + \frac{\frac{3.30}{4}-(2+3+4+11)}{8}\cdot(38-36) = 36,625$

c) Một giá trị xấp xỉ cho khoảng tứ phân vị của mẫu số liệu gốc là: $36,625-33,25=3,375$

Kết luận: 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

a) Tìm khoảng tứ phân vị của mẫu số liệu ghép nhóm này.

b) 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?

Cách giải:

a) Cỡ mẫu là $n=3+12+15+8=38$. Gọi $x_1,\ldots,x_{38}$ 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
$0\le t<1$	8
$1\le t<2$	17
$2\le t<3$	25
$3\le t<4$	20

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

Cách giải:

Hiệu chỉnh lại bảng số liệu ta có:

Thời gian (t phút)	Số cuộc gọi
$[0:1)$	8
$[1;2)$	17
$[2;3)$	25
$[4;5)$	20

Cỡ mẫu: n = 80

Giả sử $x_1, x_2, \ldots, x_{80}$ là thời gian đàm thoại của 80 cuộc gọi và giả sử dãy số liệu gốc này đã được sắp xếp theo thứ ự tăng dần.

Vì $\frac{n}{4} = 20$ và $8 < 20 < 8+17$ nên nhóm chứa tứ phân vị thứ nhất là nhóm $[1;2)$ và tứ phân vị thứ nhất là:

$$Q_1 = 1 + \frac{\frac{80}{4} - 8}{17} \cdot 1 = \frac{29}{17}$$

Vì $\frac{3n}{4} = 60$ và $8+17+25 < 20 < 8+17+25+20$ nên nhóm chứa tứ phân vị thứ ba là nhóm $[3;4)$ và tứ phân vị thứ ba là: $Q_3 = 3 + \frac{\frac{3.30}{4} - (8+17+25)}{20} \cdot 1 = 3,5$

Vậy khoảng tứ phân vi của mẫu số liệu ghép nhóm là: $3,5 - \frac{29}{17} = \frac{61}{34}$

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.

Cách giải:

Năm 2021: Khoảng biến thiên của nhiệt độ là: $R_1 = 40 - 30 = 10$

Cỡ mẫu $n = 30$. Giả sử $y_1, y_2, \ldots, y_{30}$ là nhiệt độ cao nhất trong ngày của 30 ngày tháng Sáu năm 2021 và giả sử dãy số liệu gốc này đã được sắp xếp theo thứ tự không giảm.

Vì $\frac{n}{4} = \frac{30}{4} = 7,5$ và $2 < 7,5 < 2+8$ nên tứ phân vị thứ nhất thuộc nhóm $[32;34)$ và tứ phân vị thứ nhất là:

$$Q_1 = 32 + \frac{\frac{30}{4} - 2}{8} \cdot 2 = 33,375$$

Vì $\frac{3n}{4}=\frac{3.30}{4}=22,5$ và $2+8+5+6<22,5<2+8+5+6+9$ nên tứ phân vị thứ ba thuộc nhóm $[38;40)$ và tứ phân vị thứ ba là: $Q_3=38+\frac{\frac{3.30}{4}-(2+8+5+6)}{9}.2=\frac{115}{3}$

Khoảng biến thiên của mẫu số liệu ghép nhóm là: $\Delta_{Q_1}=\frac{115}{3}-33,375=\frac{119}{24}$

Năm 2022: Khoảng biến thiên của nhiệt độ là: $R_2=40-28=12$

Cỡ mẫu $n=30$. Giả sử $z_1,z_2,\ldots,z_{30}$ là nhiệt độ cao nhất trong ngày của 30 ngày tháng Sáu năm 2022 và giả sử dãy số liệu gốc này đã được sắp xếp theo thứ tự tăng dần.

vi $\frac{n}{4}=\frac{30}{4}=7,5$ và $2+3<7,5<2+3+4$ nên nhóm chứa tứ phân vị thứ nhất là nhóm $[32;34)$ và tứ phân vị thứ nhất là:

$$Q_1^{'}=32+\frac{\frac{30}{4}-(2+3)}{4}\cdot(34-32)=33,25$$

Vì $\frac{3n}{4}=\frac{3.30}{4}=22,5$ và $2+3+4+11<22,5<2+3+4+11+8$ nên nhóm chứa tứ phân vị thứ ba là nhóm $[36;38)$ và tứ phân vị thứ ba là:

$$Q_3^{'}=36+\frac{\frac{3.30}{4}-(2+3+4+11)}{8}\cdot(38-36)=36,625$$

Khoảng biến thiên của mẫu số liệu ghép nhóm là: $\Delta_{Q_2}=36,625-33,25=3,375$

Theo khoảng biến thiên: Vì $R_2>R_1$ nên nhiệt độ cao nhất trong ngày vào tháng 6 năm 2022 biến đổi nhiều hơn nhiệt độ cao nhất trong ngày vào tháng 6 năm 2021.

Theo khoảng tứ phân vị: vi $\Delta_{Q_1}>\Delta_{Q_2}$ nên nhiệt độ cao nhất trong ngày vào tháng 6 năm 2021 biến đổi nhiều hơn nhiệt độ cao nhất trong ngày vào tháng 6 năm 2022.