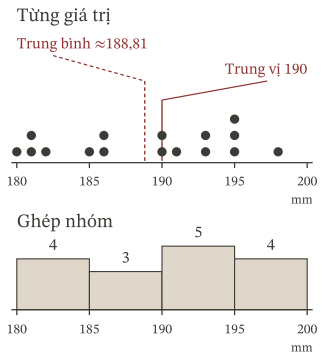
ZO MATH

ÔN THI TOÁN THPT 2027

Đọc dữ liệu và đo xu thế trung tâm

Bài học

Kiến thức, ví dụ và câu hỏi tự kiểm tra.



Mở trang Đọc dữ liệu và đo xu thế trung tâm

R2-G01

Phiên bản nội dung 0.4 · Ngày tạo PDF: 04/11/2026

© 2026 ZO Math. Bản quyền nội dung thuộc ZO Math.

Mục lục

Bài học **2**
Khởi động . . . 2
1. Mỗi dòng dữ liệu đang nói về điều gì? . . . 3
2. Những cách khác nhau để nói về vị trí trung tâm . . . 4
3. Từ từng giá trị đến các khoảng . . . 11
4. Ước lượng các số đặc trưng từ bảng nhóm . . . 13
5. Chọn số đại diện theo câu hỏi . . . 22
Khép lại Bài học . . . 27
Nguồn và cách dùng dữ liệu **27**
Tài liệu toán học . . . 27
Palmer Penguins . . . 28
Bike Sharing . . . 29
NOAA/NCEI . . . 29

Bài học

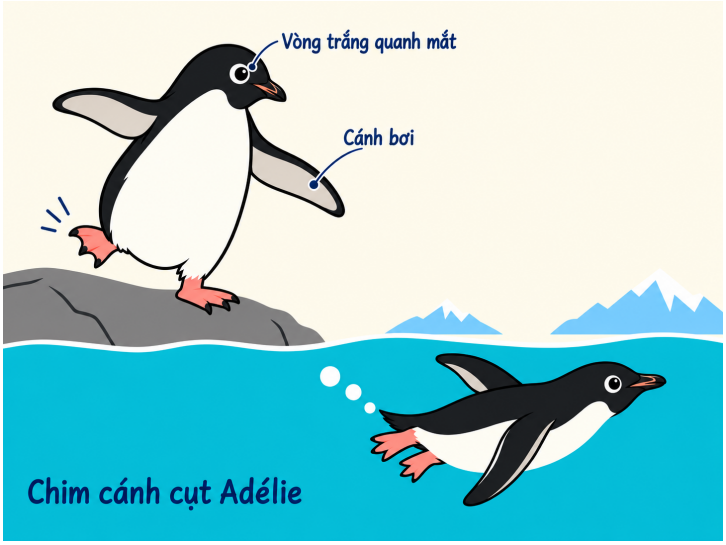
Chọn số nào để mô tả một mẫu số liệu, và có thể kết luận gì từ số ấy?

Mỗi số đặc trưng cho biết một khía cạnh của dữ liệu. Hiểu ý nghĩa và giới hạn của từng số giúp ta lựa chọn phù hợp, kể cả khi chỉ còn bảng ghép nhóm.

Khởi động

Nhiệm vụ 01 — Ba thao tác nền

Chim cánh cụt dùng hai cánh có dạng mái chèo để bơi; trong bài này, ta gọi bộ phận ấy là **cánh bơi**. Adélie là tên một loài chim cánh cụt. Bộ dữ liệu **Palmer Penguins** tập hợp các số đo cơ thể chim cánh cụt được nghiên cứu ở khu vực quần đảo Palmer, Nam Cực.



Minh họa chim cánh cụt Adélie và cánh bơi.

Ba số đo chiều dài cánh bơi dưới đây được ghi cho ba cá thể Adélie trong năm 2007: 181, 186, 195 mm. Mỗi số ứng với một cá thể; mm là viết tắt của milimét. Em sẽ dùng các số đo để ôn cách tính trung bình và đọc khoảng; ý b là một tình huống giả định riêng để ôn cách tính tỉ lệ.

a) Sắp chúng theo thứ tự không giảm và tính trung bình cộng, làm tròn đến một chữ số thập phân.

b) Giả sử một nhóm có 12 cá thể chim, trong đó 3 cá thể có chiều dài cánh bơi từ 190 mm trở lên. Ba cá thể ấy chiếm bao nhiêu phần trăm của nhóm?

c) Với các khoảng [180; 185), [185; 190), [190; 195), số đo chiều dài cánh bơi 190 mm thuộc khoảng nào? Viết bất đẳng thức giải thích.

Xem lời giải (mở trên trang học liệu).

Sau khi đối chiếu lời giải, nếu còn câu chưa làm được hoặc chưa hiểu cách làm, em chọn phần ôn tương ứng dưới đây. Ôn xong, hãy tự làm lại câu đó. Nếu đã làm đúng và hiểu cả ba ý, em học tiếp.

Em còn vướng ở đâu?	Mở phần ôn nào?	Ôn xong, làm lại
Tính trung bình cộng hoặc tỉ lệ phần trăm	Tính với đủ số quan sát (mở trên trang học liệu)	Ý a hoặc b của Nhiệm vụ 01 (trang 2)
Xác định một số thuộc khoảng nào	Đọc khoảng và giữ thứ tự (mở trên trang học liệu)	Ý c của Nhiệm vụ 01 (trang 2)

Làm đúng Khởi động cho thấy em đã làm được những thao tác mở đầu này. Các kiến thức còn lại sẽ được học và tự kiểm tra ở những phần sau.

1. Mỗi dòng dữ liệu đang nói về điều gì?

Gợi mở 01 — Chỗ trống có phải là số 0?

Đây là năm dòng đầu của bảng số đo chim cánh cụt Adélie năm 2007 trong bộ Palmer Penguins vừa dùng ở Khởi động. Mỗi dòng ghi thông tin của một cá thể; cột cuối ghi chiều dài cánh bơi của cá thể ấy. Một ô chưa có số đo được ghi là "Thiếu":

Dòng trong bảng gốc	Loài	Năm	Chiều dài cánh bơi, mm
1	Adélie	2007	181
2	Adélie	2007	186
3	Adélie	2007	195
4	Adélie	2007	Thiếu
5	Adélie	2007	193

Một bạn muốn thay ô "Thiếu" ở dòng 4 bằng 0 để tính trung bình. Khi ấy, bảng sẽ nói gì về chiều dài cánh bơi của cá thể ở dòng 4? Điều đó có giống với "chưa có số đo" không?

Số 0 sẽ nói rằng chiều dài được ghi là 0 mm. Trong khi đó, ô thiếu cho biết chưa có số đo chiều dài ở dòng ấy. Hai điều này khác nhau. Khi tính trung bình chiều dài, ta vẫn ghi nhận có một ô thiếu, nhưng chỉ cộng các số đo có sẵn và chia cho số giá trị đã cộng.

Mỗi dòng trong bảng này ghi thông tin của một cá thể chim. Cá thể ấy là **đơn vị quan sát**. Đại lượng được ghi là chiều dài cánh bơi; trong thống kê, đại lượng này gọi là **biến**, có đơn vị đo là mm. Số thứ tự dòng giúp tìm lại thông tin trong bảng gốc; nó không phải số đo hoặc mã nhận diện cá thể do nhà nghiên cứu cấp.

Phần dữ liệu về loài Adélie được ghi nhận năm 2007 có 50 dòng, trong đó có 49 số đo chiều dài dùng được. Số giá trị được đưa vào phép tính gọi là **cỡ mẫu**; ở đây $n = 49$. Nếu một dòng có số đo chiều dài nhưng thiếu thông tin giới tính, số đo chiều dài ấy vẫn được dùng. Trong bài này, **mẫu gốc** là danh sách các giá trị trước khi ghép nhóm.

Dữ liệu được thu thập ở khu vực quần đảo Palmer, Nam Cực. Việc quan sát một số cá thể trong một địa điểm và giai đoạn xác định không tự cho phép kết luận về mọi cá thể của loài.

Nguồn dữ liệu: Allison Horst, Alison Hill và Kristen Gorman, Palmer Penguins, cột flipper_length_mm ghi chiều dài cánh bơi theo mm; dữ liệu gốc do Kristen Gorman và chương trình nghiên cứu sinh thái dài hạn tại trạm Palmer cung cấp. Phần trích trong mục này là Adélie năm 2007; truy cập ngày 22/10/2026.

Nhiệm vụ 02 — Đếm đúng mẫu dùng để tính

Một bạn muốn tính chiều dài cánh bơi trung bình từ 17 dòng đầu của bảng Adélie năm 2007. Trong đó, dòng 4 thiếu số đo chiều dài; 16 dòng còn lại có số đo. Bạn cộng các số đo có sẵn được 3021 mm rồi chia cho 17. Hãy kiểm tra xem cách tính ấy có mô tả đúng những số đo đang có không.

a) Nêu đơn vị quan sát, biến và đơn vị đo.

b) Mẫu số cần dùng là bao nhiêu? Vì sao?

c) Sửa câu: "Cá thể ở dòng 4 có chiều dài cánh bơi bằng 0."

Xem lời giải (mở trên trang học liệu).

2. Những cách khác nhau để nói về vị trí trung tâm

Một số được tính đúng vẫn có thể khiến người đọc hình dung sai về mẫu. Để thấy rõ điều này, ta tạm dùng một tình huống nhỏ, trong đó chỉ một giá trị thay đổi.

2.1. Số trung bình và trung vị

Gợi mở 02 — "Ở giữa" và "chia đều" có cùng nghĩa không?

Tình huống giả định. Thời gian chờ của năm người, tính bằng phút, lần lượt là

$$2;\ 3;\ 3;\ 4;\ 8.$$

Nếu dùng cùng một thời gian để thay cho thời gian chờ của cả năm người mà tổng vẫn như cũ, thời gian ấy phải bằng bao nhiêu? Đây là cách tìm một con số tóm tắt, không phải thay đổi thời gian mà từng người thực sự đã chờ. Sau khi xếp năm thời gian từ nhỏ đến lớn, giá trị đứng giữa bằng bao nhiêu?

Bây giờ thay thời gian 8 phút bằng 28 phút, giữ nguyên bốn giá trị còn lại. Hãy dự đoán hai câu trả lời trên thay đổi thế nào. Nếu thông báo "thời gian chờ đại diện là 8 phút", người đọc có hình dung đúng trải nghiệm của phần lớn năm người ấy không?

Số trung bình cho biết tổng chia đều cho mỗi quan sát là bao nhiêu. Năm người có thời gian chờ 2; 3; 3; 4; 8 phút. Tổng thời gian chờ là 20 phút. Nếu tổng ấy được chia đều cho năm người thì mỗi người ứng với $\frac{20}{5} = 4$ phút. Đó là **thời gian chờ trung bình**.

Có thể kiểm lại bằng cách thay năm thời gian bằng 4; 4; 4; 4; 4: tổng vẫn là 20 phút. Phép thay này giúp hiểu ý nghĩa của số trung bình; thời gian chờ thực tế vẫn là 2; 3; 3; 4; 8 phút, không phải ai cũng đã chờ 4 phút.

Với một mẫu gồm n giá trị, $n \geq 1$, gọi các giá trị là $x_1, x_2, \dots, x_n$ và số trung bình là $\overline{x}$. Nếu thay mỗi giá trị bằng $\overline{x}$ thì tổng mới là $n\overline{x}$. Muốn tổng mới bằng tổng ban đầu, ta cần

$$n\overline{x} = x_1 + x_2 + \cdots + x_n.$$

Chia hai vế cho n, ta được

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Vì sao có thể xem trung bình là một mức cân bằng? So với 4 phút, thời gian 2 phút thấp hơn 2 phút; mỗi thời gian 3 phút thấp hơn 1 phút. Ba phần thiếu cộng lại là $2 + 1 + 1 = 4$ phút. Thời gian 8 phút vượt mức 4 phút đúng 4 phút; thời gian 4 phút không thiếu cũng không vượt. Phần thiếu và phần vượt bằng nhau.

Gọi $x_i - \overline{x}$ là độ lệch của giá trị thứ i so với trung bình. Độ lệch âm biểu thị thấp hơn, độ lệch dương biểu thị cao hơn. Kí hiệu $\sum_{i=1}^{n}$ có nghĩa là cộng biểu thức phía sau với i lần lượt từ 1 đến n. Cộng các độ lệch tức là lấy tổng ban đầu trừ đi n lần số trung bình:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 + \cdots + x_n) - n\overline{x} = 0.$$

Một giá trị thay đổi làm trung bình thay đổi thế nào? Nếu chỉ thay thời gian 8 phút bằng 28 phút, tổng tăng từ 20 lên 40 phút, còn số người vẫn là 5. Trung bình tăng từ $\frac{20}{5} = 4$ lên $\frac{40}{5} = 8$ phút. Bốn người còn lại không đổi thời gian chờ, nhưng trung bình vẫn tăng vì tổng đã tăng.

Tổng quát, khi chỉ một giá trị tăng thêm Δ và số giá trị vẫn là n, tổng tăng thêm Δ, nên trung bình tăng thêm $\frac{\Delta}{n}$. Nếu $\Delta < 0$, giá trị đó giảm và trung bình giảm tương ứng. Kết luận này xét việc thay một giá trị, không phải thêm một quan sát mới.

Trung vị trả lời câu hỏi: sau khi sắp từ nhỏ đến lớn, giá trị ở giữa bằng bao nhiêu? Dãy 2; 3; 3; 4; 8 đã được sắp. Quan sát thứ ba đứng ở giữa: có hai quan sát đứng trước và hai quan sát đứng sau. Giá trị tại đó là 3 phút, gọi là **trung vị**, kí hiệu M_e. "Thứ ba" chỉ vị trí trong danh sách; 3 phút mới là thời gian chờ.

Khi thay 8 bằng 28, dãy mới là 2; 3; 3; 4; 28. Giá trị cuối vẫn đứng cuối, giá trị thứ ba vẫn bằng 3. Vì vậy trung vị không đổi, trong khi trung bình tăng lên 8 phút.

Để tìm trung vị, trước hết sắp tất cả các giá trị theo thứ tự không giảm, giữ đủ cả các giá trị trùng nhau. Nếu số quan sát lẻ, lấy giá trị ở vị trí chính giữa. Nếu số quan sát chẵn, không có một quan sát duy nhất đứng giữa; theo định nghĩa, trung vị được lấy bằng trung bình cộng của hai giá trị giữa. Chẳng hạn, mẫu 2; 3; 7; 10 có hai giá trị giữa là 3 và 7, nên trung vị bằng $\frac{3+7}{2} = 5$, dù mẫu không có giá trị 5.

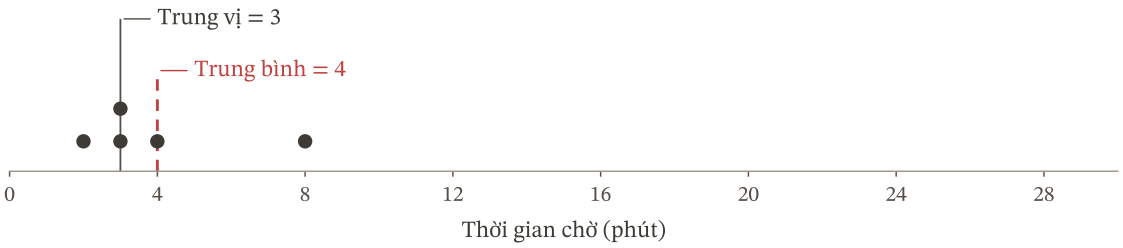
Trong mẫu 2; 3; 3; 4; 8, ta có thể thực hiện hai phép đếm riêng:

- Các quan sát có giá trị không lớn hơn trung vị 3 là 2; 3; 3: có ba quan sát.
- Các quan sát có giá trị không nhỏ hơn 3 là 3; 3; 4; 8: có bốn quan sát.

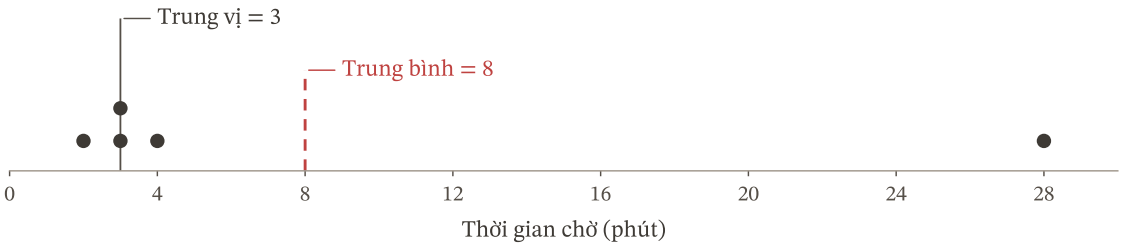
Hai quan sát bằng 3 thỏa cả hai điều kiện so sánh, nên xuất hiện trong cả hai phép đếm. Ta đang kiểm tra hai điều kiện, không chia mẫu thành hai nhóm tách biệt. Vì vậy, hai số đếm có thể cùng lớn hơn một nửa cỡ mẫu.

Tổng quát, có **ít nhất một nửa** số quan sát không lớn hơn trung vị và ít nhất một nửa không nhỏ hơn trung vị. Điều này không có nghĩa đúng một nửa nhỏ hơn và đúng một nửa lớn hơn trung vị.

Mẫu ban đầu · 2; 3; 3; 4; 8 phút



Thay 8 bằng 28 · 2; 3; 3; 4; 28 phút



Khi thay thời gian chờ 8 phút bằng 28 phút, trung bình tăng từ 4 lên 8 phút, còn trung vị vẫn bằng 3 phút.

Trong hình, mỗi chấm là một người với thời gian chờ tương ứng; các chấm trùng thời gian được xếp cao lên để vẫn nhìn thấy đủ. Hai hàng dùng cùng một trục thời gian. Đường nét đứt đỏ chỉ trung bình: từ 4 chuyển sang 8 phút. Đường liền xám chỉ trung vị: vẫn ở 3 phút. Nhãn bên cạnh mỗi đường ghi rõ số đặc trưng tương ứng.

Vì sao có lúc trung bình chưa mô tả tốt thời gian chờ của phần lớn nhóm? Trong mẫu thứ hai, bốn trong năm người chỉ chờ từ 2 đến 4 phút. Nếu chỉ thông báo "thời gian chờ đại diện là 8 phút", người đọc có thể tưởng thời gian chờ của phần lớn nhóm gần 8 phút. Số trung bình cao hơn khoảng thời gian của bốn người ấy vì thời gian 28 phút chiếm phần lớn tổng 40 phút.

Nếu muốn mô tả người ở giữa khi xếp theo thời gian chờ, trung vị 3 phút phù hợp hơn. Một báo cáo rõ hơn có thể viết: "Trung vị thời gian chờ là 3 phút; bốn trong năm người chờ từ 2 đến 4 phút, còn một người chờ 28 phút." Người chờ lâu vẫn được ghi nhận. Nếu quan tâm tổng thời gian chờ, trung bình 8 phút vẫn hữu ích vì $5 \cdot 8 = 40$ phút.

Trung vị không phải lúc nào cũng giữ nguyên. Nếu thay 2 bằng 30 trong mẫu 2; 3; 3; 4; 28, ta phải sắp lại thành 3; 3; 4; 28; 30; giá trị giữa lúc này bằng 4. Khi một giá trị thay đổi, cần xem nó có làm thay đổi giá trị ở giữa sau sắp xếp hay không.

Vì thế, không có quy tắc "hễ có một giá trị rất lớn thì bỏ trung bình". Cần xác định câu hỏi: muốn biết tổng chia đều cho mỗi người, thời gian của người ở giữa, hay muốn mô tả cả những người chờ lâu khác biệt?

Vì sao phải tính đủ số lần xuất hiện của mỗi giá trị? Trong mẫu 2; 3; 3; 4; 8, giá trị 3 xuất hiện hai lần nên phải được cộng hai lần. Do đó

$$\overline{x} = \frac{2 + 2 \cdot 3 + 4 + 8}{1 + 2 + 1 + 1} = 4.$$

Số lần xuất hiện của một giá trị gọi là **tần số**. Mẫu trên có bốn giá trị khác nhau 2, 3, 4, 8, với tần số tương ứng 1, 2, 1, 1; tổng tần số là 5, chính là số quan sát. Khi có k giá trị khác nhau $a_1, \dots, a_k$ với tần số tương ứng $f_1, \dots, f_k$, ta viết gọn

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i a_i}{\sum_{i=1}^{k} f_i}.$$

Ở tử số, mỗi giá trị được nhân với số lần nó xuất hiện. Ở mẫu số, ta cộng các tần số để đếm đủ mọi quan sát. Tần số được gọi là **trọng số** vì nó cho biết mỗi giá trị được tính bao nhiêu lần trong phép lấy trung bình.

Gộp hai nhóm có số quan sát khác nhau. Giả sử nhóm thứ nhất có 2 quan sát, trung bình 10; nhóm thứ hai có 8 quan sát, trung bình 20. Ta khôi phục tổng của từng nhóm: $2 \cdot 10 = 20$ và $8 \cdot 20 = 160$. Cả hai nhóm có tổng 180 và có 10 quan sát, nên trung bình chung là $\frac{180}{10} = 18$.

Phép tính $\frac{10+20}{2} = 15$ tính hai nhóm ngang nhau: mỗi trung bình nhóm được tính một lần, dù nhóm thứ hai có số quan sát gấp bốn lần nhóm thứ nhất. Trong mẫu gộp, hai nhóm thực sự chiếm $\frac{2}{10}$ và $\frac{8}{10}$ số quan sát. Vì vậy, để mỗi quan sát được tính một lần, ta nhân từng trung bình nhóm với số quan sát tương ứng, cộng các tích rồi chia cho tổng số quan sát. Nếu cỡ hai nhóm bằng nhau thì công thức này rút gọn thành trung bình cộng của hai trung bình; nếu hai trung bình bằng nhau thì hai cách cũng cho cùng kết quả.

Khi đổi từ phút sang giây, trung bình và trung vị đổi thế nào? Năm thời gian 2; 3; 3; 4; 8 phút được viết thành 120; 180; 180; 240; 480 giây. Mỗi giá trị nhân với 60, nên tổng cũng nhân với 60: tổng 20 phút tương ứng 1200 giây. Số người vẫn là 5, do đó trung bình là $\frac{1200}{5} = 240$ giây, đúng bằng 4 phút.

Nhân mọi giá trị với cùng số dương 60 không làm đổi thứ tự. Giá trị giữa là 180 giây, đúng bằng trung vị 3 phút. Với mẫu chẵn, cả hai giá trị giữa đều nhân với 60 nên trung bình cộng của chúng cũng nhân với 60. Đổi đơn vị làm đổi con số viết ra nhưng vẫn nói về cùng một thời gian; 4 phút không thể được ghi thành 4 giây.

Một việc khác: mọi thời gian đều tăng thêm như nhau. Nếu mỗi người phải chờ thêm 2 phút thì thời gian thực tế đã thay đổi. Tổng tăng thêm $5 \cdot 2 = 10$ phút, nên trung bình tăng $\frac{10}{5} = 2$ phút. Thứ tự không đổi và giá trị giữa cũng tăng 2 phút; với mẫu chẵn, trung bình của hai giá trị giữa tăng 2 phút. Tổng quát, khi cộng cùng một số c vào cả n giá trị, tổng tăng thêm nc, còn trung bình và trung vị đều tăng thêm c.

Căn cứ toán học: Toán 10, Tập một, Kết nối tri thức với cuộc sống, Nhà xuất bản Giáo dục Việt Nam, Bài 13, trang 78–79.

Ví dụ 01 — Một mẫu có 16 giá trị

Trở lại 17 dòng đầu của bảng Adélie năm 2007 ở Nhiệm vụ 02. Sau khi ghi nhận một ô thiếu, ta dùng 16 số đo cánh bơi còn lại để tìm số trung bình và trung vị. Mỗi số dưới đây là chiều dài cánh bơi của một cá thể, đơn vị mm:

181; 186; 195; 193; 190; 181;
195; 193; 190; 186; 180; 182;
191; 198; 185; 195.

Sau khi sắp thứ tự:

Vị trí	1	2	3	4	5	6	7	8
Giá trị, mm	180	181	181	182	185	186	186	190

Vị trí	9	10	11	12	13	14	15	16
Giá trị, mm	190	191	193	193	195	195	195	198

Tổng là 3021 mm nên

$$\overline{x} = \frac{3021}{16} = 188{,}8125 \text{ mm}.$$

Hai vị trí ở giữa là 8 và 9, cùng mang giá trị 190 mm. Do đó

$$M_e = \frac{190 + 190}{2} = 190 \text{ mm}.$$

Con số 188,8125 là mức chia đều tổng chiều dài 3021 mm cho 16 số đo. Trung vị 190 mm là trung bình cộng của hai số đo ở vị trí thứ 8 và thứ 9 trong dãy đã sắp. Hai số tương đối gần nhau trong mẫu này; sự gần nhau ấy không phải định lí đúng với mọi mẫu. Ta vừa thấy một đối chứng ở Gợi mở 02.

Các số đo trong danh sách được ghi bằng số nguyên mm, nhưng phép chia $\frac{3021}{16}$ cho kết quả 188,8125. Bốn chữ số sau dấu phẩy phát sinh từ phép chia; chúng không cho biết dụng cụ đã đo từng cá thể chính xác đến một phần mười nghìn mm. Muốn biết độ chính xác của phép đo, ta cần thông tin về dụng cụ và cách đo. Vì vậy, không thể suy độ chính xác ấy chỉ từ số chữ số thập phân của trung bình.

2.2. Tứ phân vị: chia theo thứ tự

Biết trung vị rồi, ta đã hình dung được vị trí của những giá trị còn lại chưa? Xét hai mẫu giả định đã sắp:

$$A : 1;\ 2;\ 3;\ 4;\ 5;\ 6;\ 7;\ 8,$$
$$B : 1;\ 4;\ 4;\ 4;\ 5;\ 5;\ 5;\ 8.$$

Cả hai mẫu cùng có tổng 36, trung bình 4,5 và trung vị 4,5. Chỉ biết hai số ấy, ta chưa thấy mẫu B có sáu trong tám giá trị tập trung tại 4 và 5, còn mẫu A có đủ các giá trị nguyên từ 1 đến 8.

Ta xét thêm từng nửa của dãy đã sắp. Nửa dưới của A là 1; 2; 3; 4, có trung vị $\frac{2+3}{2} = 2{,}5$. Nửa dưới của B là 1; 4; 4; 4, có trung vị $\frac{4+4}{2} = 4$. Hai kết quả ấy là **tứ phân vị thứ nhất**, kí hiệu Q_1, của từng mẫu.

Nửa trên của A là 5; 6; 7; 8, có trung vị $\frac{6+7}{2} = 6{,}5$. Nửa trên của B là 5; 5; 5; 8, có trung vị 5. Đây là **tứ phân vị thứ ba**, kí hiệu Q_3. Cùng với trung vị của toàn mẫu, gọi là **tứ phân vị thứ hai** Q_2, ta có ba mốc Q_1, Q_2, Q_3.

Như vậy, Q_1 và Q_3 bổ sung thông tin về các giá trị ở nửa dưới và nửa trên của mẫu. Chúng giúp thấy một sự khác nhau mà trung vị chung 4,5 chưa cho thấy. Ta xác định các tứ phân vị bằng cách xét các

quan sát trong dãy đã sắp, rồi tìm trung vị của toàn dãy và của từng nửa. Ta không chia đoạn từ giá trị nhỏ nhất đến lớn nhất thành bốn đoạn dài bằng nhau. Hai mẫu đều có hai đầu là 1 và 8 nhưng có Q_1, Q_3 khác nhau.

Cách xác định tứ phân vị bằng trung vị của hai nửa, theo Toán 10, Tập một, bộ Kết nối tri thức với cuộc sống, trang 80–81, được thực hiện như sau:

1. Sắp các giá trị theo thứ tự không giảm, giữ đủ những giá trị trùng nhau.
2. Tìm trung vị Q_2. Nếu cỡ mẫu chẵn, chia dãy thành hai nửa có số giá trị bằng nhau. Nếu cỡ mẫu lẻ, nửa dưới gồm các quan sát đứng trước vị trí chính giữa; nửa trên gồm các quan sát đứng sau vị trí ấy.
3. Tìm trung vị của nửa dưới để được Q_1 và trung vị của nửa trên để được Q_3.

Chẳng hạn, với năm giá trị đã sắp 1; 2; 3; 4; 5, quan sát thứ ba xác định $Q_2 = 3$; hai nửa dùng tiếp là 1; 2 và 4; 5. Ta chỉ không đưa **quan sát ở vị trí chính giữa** vào hai nửa ấy, không xóa tất cả các giá trị bằng trung vị. Quan sát đó vẫn thuộc mẫu ban đầu.

Trong mẫu B, $Q_1 = 4$. Tuy nhiên, chỉ có một trong tám quan sát nhỏ hơn 4, còn có bốn trong tám quan sát không lớn hơn 4. Hai tỉ lệ lần lượt là $\frac{1}{8}$ và $\frac{4}{8}$; không tỉ lệ nào bằng 25%.

Nguyên nhân là ba quan sát cùng mang giá trị 4: nếu đếm các giá trị nhỏ hơn 4 thì không tính cả ba; nếu đếm các giá trị không lớn hơn 4 thì tính cả ba. Vì thế, không thể hiểu ba tứ phân vị là những mốc luôn chia các giá trị của mẫu thành bốn nhóm, mỗi nhóm có đúng 25% số quan sát. Khi tìm Q_1, Q_2, Q_3, ta cần thực hiện cách xác định bằng trung vị của hai nửa đã nêu; đổi dấu "nhỏ hơn" thành "không lớn hơn" không khắc phục được sự khác nhau về số đếm này.

Trong Ví dụ 01, nửa dưới gồm các vị trí từ 1 đến 8, nửa trên từ 9 đến 16. Vì thế

$$Q_1 = \frac{182 + 185}{2} = 183{,}5;$$
$$Q_2 = \frac{190 + 190}{2} = 190;$$
$$Q_3 = \frac{193 + 195}{2} = 194.$$

Cả ba tứ phân vị đều có đơn vị mm. Số 183,5 mm là trung vị của tám số đo ở nửa dưới; số 194 mm là trung vị của tám số đo ở nửa trên; 190 mm là trung vị của toàn bộ 16 số đo.

Căn cứ toán học: Toán 10, Tập một, Kết nối tri thức với cuộc sống, Bài 13, trang 80–81. Khi dùng phần mềm hoặc đọc một kết quả phân vị khác, cần kiểm tra phương pháp tính: phần mềm có thể cung cấp quy ước khác với cách lấy trung vị của hai nửa.

Ví dụ 02 — Khi cỡ mẫu lẻ

Năm số đo cánh bơi có sẵn đầu tiên của các cá thể Adélie năm 2007 là 181; 186; 195; 193; 190 mm. Sau khi sắp, ta được

$$181;\ 186;\ 190;\ 193;\ 195.$$

Trung vị là 190. Nửa dưới gồm 181; 186, nửa trên gồm 193; 195. Do đó $Q_1 = 183{,}5$ mm và $Q_3 = 194$ mm. Số đo 190 mm ở vị trí thứ ba xác định Q_2; khi tìm Q_1 và Q_3, ta chỉ xét hai số đứng trước và hai số đứng sau vị trí ấy. Không xóa số đo này khỏi mẫu ban đầu.

Ví dụ này dùng lại một phần dữ liệu vừa học để giải thích quy tắc khi số giá trị là số lẻ. Đọc hiểu lời giải giúp em nắm cách làm; để biết có tự làm được hay không, em cần thử một bài khác chưa xem lời giải.

Nhiệm vụ 03 — Vị trí và giá trị

Nhiệm vụ này củng cố hai phần vừa học: số trung bình và trung vị ở mục 2.1, tứ phân vị ở mục 2.2. Dùng bảng 16 số đo cánh bơi của chim Adélie năm 2007 đã sắp trong Ví dụ 01 (trang 7). Em sẽ phân biệt vị trí trong danh sách với chiều dài được ghi, rồi xem trung bình và trung vị thay đổi thế nào khi một giá trị thay đổi.

a) Một bạn viết "$Q_1 = 4{,}5$ vì trung vị của tám vị trí đầu nằm giữa vị trí 4 và 5". Chỉ ra sự nhầm lẫn và sửa kết quả.

b) Trong Ví dụ 02, nếu đưa quan sát 190 vào cả hai nửa, em có còn đang dùng cách xác định tứ phân vị bằng trung vị của hai nửa không? Giải thích.

c) Một mẫu thời gian chờ **giả định mới** gồm 1; 2; 2; 3; 7 phút. Thay riêng giá trị 7 bằng 27. Tính trung bình, trung vị trước và sau thay đổi. Giải thích vì sao chúng phản ứng khác nhau, rồi chọn số phù hợp nếu muốn mô tả người ở giữa theo thời gian chờ.

Xem lời giải (mở trên trang học liệu).

2.3. Mốt: giá trị xuất hiện nhiều nhất

Một cửa hàng muốn biết cỡ giày nào được hỏi mua nhiều nhất. Trong danh sách giả định 36; 36; 36; 37; 37; 38; 38; 39; 40, cỡ 36 xuất hiện ba lần, nhiều hơn từng cỡ còn lại. Đây là câu hỏi về **tần số**, khác với chia đều tổng hoặc tìm vị trí giữa. Trung vị của danh sách là 37; nó không trả lời cỡ nào xuất hiện nhiều nhất.

Cỡ 36 chỉ chiếm $\frac{3}{9}$ số lượt hỏi. Vì vậy, "nhiều nhất" không đồng nghĩa "quá một nửa". Nếu cần dự kiến nhiều cỡ hàng, cửa hàng phải xem cả bảng tần số; chỉ giữ một cỡ nổi trội sẽ bỏ qua sáu lượt hỏi còn lại. Tình huống này minh họa ý nghĩa của mốt, chưa phải một dự báo nhu cầu bán hàng.

Tần số của một giá trị là số lần giá trị ấy xuất hiện trong mẫu. **Mốt** là giá trị có tần số lớn nhất. Theo quy ước của Toán 10, Tập một, bộ Kết nối tri thức với cuộc sống, trang 81, khi các giá trị khác nhau có tần số bằng nhau thì mẫu không có mốt. Với mẫu có ít nhất hai giá trị khác nhau, ta phân biệt ba trường hợp:

- **Không có mốt:** tất cả các giá trị khác nhau đều xuất hiện cùng một số lần. Ta dùng quy ước mẫu không có mốt, vì không có giá trị nào nổi trội về tần số.
- **Một mốt:** có đúng một giá trị xuất hiện nhiều lần hơn mỗi giá trị còn lại. Giá trị ấy là mốt.
- **Nhiều mốt:** có từ hai giá trị trở lên cùng đạt tần số lớn nhất, đồng thời vẫn có giá trị khác mang tần số nhỏ hơn. Tất cả các giá trị đạt tần số lớn nhất ấy đều là mốt.

Vì vậy, khi các tần số không đồng loạt bằng nhau, tìm mốt là tìm đủ các giá trị có tần số lớn nhất. Không chỉ đếm xem một giá trị có lặp lại hay không; cần so sánh số lần xuất hiện của nó với các giá trị khác.

Ví dụ, mẫu 1; 1; 2; 2; 3; 3 không có giá trị nào xuất hiện nhiều hơn giá trị khác: các tần số đều bằng 2, nên không có mốt theo quy ước trên. Còn mẫu 1; 1; 2; 2; 3 có hai mốt là 1 và 2: chúng cùng có tần số lớn nhất 2, lớn hơn tần số 1 của giá trị 3. Mốt là các giá trị 1 và 2, không phải số lần xuất hiện 2.

Ở mẫu 16 giá trị, 195 xuất hiện ba lần, nhiều hơn các giá trị khác, nên mốt là 195 mm. Khi mở rộng đến toàn bộ 49 giá trị Adélie 2007, các số 180, 190, 195 đều xuất hiện năm lần và cùng đạt tần số lớn nhất. Mẫu lớn này có ba mốt.

Nếu muốn biết chiều dài nào được ghi nhận nhiều lần nhất trong nhóm chim này, mốt trả lời câu hỏi ấy. Nó không khẳng định phần lớn cá thể đều có chiều dài đó.

Không thể mang mốt của một phần trích sang mẫu lớn hơn mà không đếm lại.

Căn cứ toán học: Toán 10, Tập một, Kết nối tri thức với cuộc sống, Bài 13, trang 81, gồm định nghĩa và hai nhận xét về nhiều mốt, không có mốt; đối chiếu thêm Sách giáo viên Toán 10, trang 144–145. Xem lưu ý về định nghĩa và quy ước mốt (trang 28) khi mẫu chỉ có một giá trị khác nhau. Các tần số Penguins ở trên được tính từ dữ liệu Adélie 2007.

3. Từ từng giá trị đến các khoảng

Gợi mở 03 — Biết khoảng có đủ để khôi phục giá trị không?

Một bản tóm tắt chỉ ghi rằng chiều dài cánh bơi của một cá thể thuộc khoảng [190; 195) mm. Em còn biết được những giới hạn nào của số đo? Em có thể tìm lại chính xác chiều dài của cá thể ấy không?

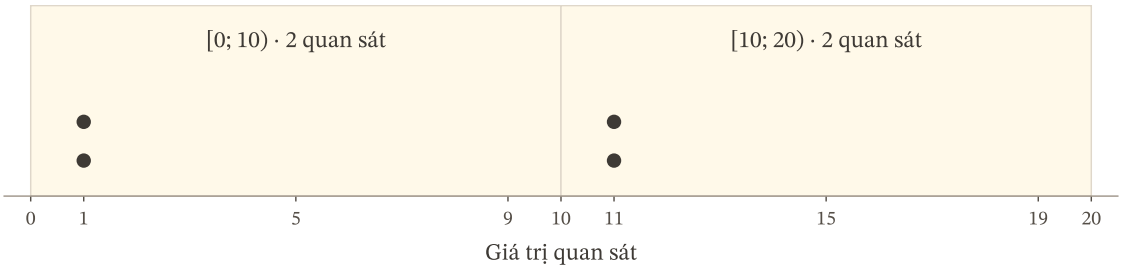
Thông tin trên cho biết số đo không nhỏ hơn 190 mm và nhỏ hơn 195 mm, nhưng chưa cho biết giá trị cụ thể. **Ghép nhóm** là xếp các số đo vào từng khoảng rồi đếm số quan sát trong mỗi khoảng. Cách này giúp nhìn sự phân bố của nhiều số đo, nhưng không giữ lại từng giá trị ban đầu.

Xét hai mẫu **giả định**, cùng đơn vị: A gồm 1; 1; 11; 11, còn B gồm 9; 9; 19; 19. Cả hai được ghép theo hai khoảng [0; 10) và [10; 20):

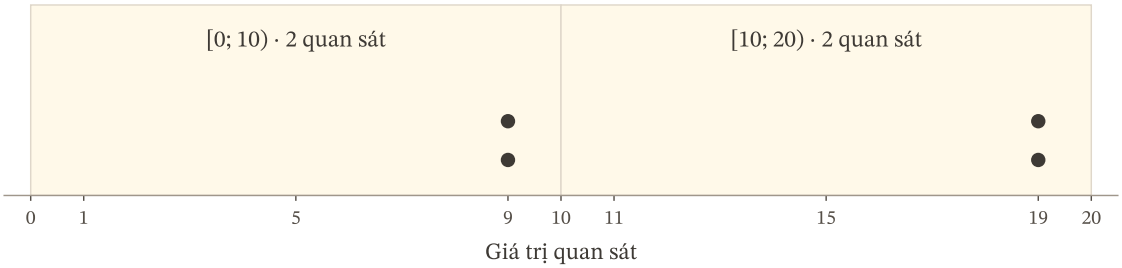
Mẫu gốc	Số giá trị trong [0; 10)	Số giá trị trong [10; 20)	Trung bình	Trung vị
A: 1; 1; 11; 11	2	2	6	6
B: 9; 9; 19; 19	2	2	14	14

Với A, trung bình là $\frac{1+1+11+11}{4} = 6$ và trung vị là $\frac{1+11}{2} = 6$. Với B, hai phép tính tương ứng cho 14. Người chỉ giữ bảng tần số 2; 2 không phân biệt được mình đã bắt đầu từ A hay B. Vì vậy, một phép tính chỉ dùng bảng ấy sẽ cho cùng một kết quả, trong khi hai mẫu gốc có kết quả khác nhau. Điều này chứng tỏ không thể luôn khôi phục chính xác trung bình hoặc trung vị của mẫu gốc từ bảng ghép nhóm.

Mẫu A: 1; 1; 11; 11 · **Trung bình = trung vị = 6**



Mẫu B: 9; 9; 19; 19 · **Trung bình = trung vị = 14**



Cùng có hai quan sát trong mỗi nhóm, nhưng mẫu A có trung bình và trung vị bằng 6, còn mẫu B có cả hai số bằng 14.

Mỗi chấm trong hình biểu thị một giá trị gốc; hai hàng dùng cùng một trục số. Các giá trị của A nằm gần đầu trái các nhóm, còn các giá trị của B nằm gần đầu phải. Bảng chỉ ghi mỗi nhóm có hai giá trị nên không lưu được sự khác nhau đó.

Bảng nhóm vẫn giữ những điều có ích: cỡ mẫu, số quan sát trong từng khoảng và số quan sát nhỏ hơn một biên nhóm. Cần phân biệt thông tin đọc được chắc chắn từ bảng với những giá trị chỉ có thể ước lượng.

Ta chia trục thành các khoảng không chồng lấn và bao phủ mọi giá trị của mẫu. Ta dùng các khoảng lấy đầu mút trái và không lấy đầu mút phải (gọi là đóng trái–mở phải). Vì thế 190 thuộc [190; 195), không thuộc [185; 190).

Để đọc bảng, cần phân biệt số quan sát trong một nhóm với số quan sát cộng dồn từ nhóm đầu. Trung điểm của nhóm được dùng làm số thay thế khi ước lượng; nó không phải một số đo đã biết.

Ví dụ 03 — Ghép nhóm mẫu nhỏ

Để tóm tắt 16 số đo cánh bơi Adélie ở Ví dụ 01 (trang 7), ta xếp chúng vào bốn khoảng. Bảng cho biết mỗi khoảng có bao nhiêu cá thể và số cá thể cộng dồn đến khoảng ấy:

Khoảng, mm	Tần số	Giá trị đại diện, mm	Tần số tích lũy
[180; 185)	4	182,5	4
[185; 190)	3	187,5	7
[190; 195)	5	192,5	12
[195; 200)	4	197,5	16

Xét hàng [190; 195): **tần số** 5 cho biết có năm số đo thuộc khoảng này. Hai nhóm trước có tổng tần số $4 + 3 = 7$, nên có bảy số đo nhỏ hơn 190 mm. Tính đến hết nhóm đang xét, có $7 + 5 = 12$ số đo

nhỏ hơn 195 mm; đó là **tần số tích lũy** ở hàng này. Tần số tích lũy trước nhóm là 7, còn đến hết nhóm là 12: hai số này có vai trò khác nhau.

Giá trị đại diện của nhóm là trung điểm $\frac{190+195}{2} = 192{,}5$ mm. Khi ước lượng trung bình từ bảng, ta dùng số ấy thay tạm cho từng số đo chưa biết trong nhóm. Điều này không khẳng định cả năm cá thể, hay một cá thể nào, có đúng chiều dài 192,5 mm. Mục tiếp theo giải thích vì sao chọn trung điểm.

Tổng tần số $4 + 3 + 5 + 4 = 16$, bằng cỡ mẫu. Chiều dài là đại lượng đo liên tục, dù các số liệu đang dùng được ghi bằng số nguyên mm. Các khoảng đã được chọn trực tiếp trên trục độ dài; không tự đổi biên khi chuyển từ bảng sang công thức.

Căn cứ toán học: Toán 11, Tập một, Kết nối tri thức với cuộc sống, Bài 8, trang 58–59, và Bài 9, trang 62. Trường hợp nhóm rời rạc được ghi bằng hai giá trị nguyên bao gồm cả hai đầu cần đọc quy tắc hiệu chỉnh riêng ở trang 63; các khoảng độ dài đang xét đã được cho trực tiếp dưới dạng đóng trái–mở phải.

Nhiệm vụ 04 — Thông tin còn lại sau ghép nhóm

Dùng bảng ghép nhóm 16 số đo cánh bơi trong Ví dụ 03 (trang 12). Em cần xác định điều còn đọc được từ bảng và điều không thể suy ra khi không có danh sách từng số đo.

a) Có bao nhiêu chiều dài nhỏ hơn 195 mm?

b) Có thể khẳng định cả năm cá thể ở nhóm [190; 195) đều có chiều dài 192,5 mm không?

c) Nếu đếm 195 vào cả hai nhóm kề nhau, phép kiểm nào giúp em phát hiện sai sót?

Xem lời giải (mở trên trang học liệu).

4. Ước lượng các số đặc trưng từ bảng nhóm

Gợi mở 04 — Làm thế nào khi không còn từng giá trị?

Giả sử báo cáo chỉ còn bảng số cá thể theo từng khoảng chiều dài cánh bơi, không còn danh sách các số đo. Em có tính lại chính xác số trung bình và trung vị của mẫu gốc không? Nếu chỉ cần một kết quả gần đúng, em có thể dùng thông tin nào trong bảng?

Hai mẫu A và B ở mục trước có cùng bảng nhóm nhưng trung bình và trung vị khác nhau. Khi không còn từng giá trị, muốn đưa ra một kết quả gần đúng, ta phải nói rõ mình đã thay thế hoặc giả sử điều gì.

Để ước lượng **trung bình**, ta dùng trung điểm thay cho mỗi giá trị chưa biết trong nhóm. Để ước lượng **trung vị và tứ phân vị**, ta giả sử tần số được phân bố đều theo độ rộng của nhóm. Các ví dụ dưới đây sẽ giải thích từng cách ước lượng và chỉ rõ những thông tin về dữ liệu gốc mà bảng nhóm chưa cho biết.

Bảng sau mở rộng từ mẫu nhỏ 16 số đo sang **toàn bộ** 49 **số đo cánh bơi có sẵn của loài Adélie năm 2007** trong phần dữ liệu đang học. Các số đo đều có đơn vị mm:

Khoảng, mm	Tần số f_i	Trung điểm c_i, mm	Tần số tích lũy
[170; 175)	2	172,5	2
[175; 180)	3	177,5	5
[180; 185)	15	182,5	20
[185; 190)	10	187,5	30

Khoảng, mm	Tần số f_i	Trung điểm c_i, mm	Tần số tích lũy
[190; 195)	10	192,5	40
[195; 200)	9	197,5	49
[200; 205)	0	202,5	49
[205; 210)	0	207,5	49
[210; 215)	0	212,5	49

Các khoảng cuối có tần số 0 vẫn cho thấy trục chia đều. Chúng không thêm quan sát vào mẫu.

4.1. Số trung bình của mẫu ghép nhóm

Giả sử bảng chỉ cho biết một giá trị thuộc [0; 10). Ta biết giá trị đó từ 0 đến dưới 10, nhưng chưa biết chính xác là bao nhiêu. Muốn tính gần đúng trung bình từ bảng nhóm, ta cần chọn một số để thay tạm cho các giá trị chưa biết trong khoảng này.

Quy tắc tính trung bình ghép nhóm chọn **trung điểm** 5. Số 5 cách mỗi đầu mút một khoảng bằng 5. Nếu chọn 2 thì gần đầu 0 hơn, nhưng cách đầu 10 đến 8; một giá trị gần 10 có thể khác số thay thế 2 khá nhiều. Chọn trung điểm bảo đảm khoảng cách từ bất kì giá trị nào trong [0; 10) tới số thay thế không vượt quá 5.

Điều này **không cho biết trung bình thật của nhóm bằng** 5. Nếu hai giá trị thật là 1; 1 thì trung bình là 1; nếu là 9; 9 thì trung bình là 9. Cả hai trường hợp đều được ghi thành "hai giá trị thuộc [0; 10)". Bảng không cho biết trường hợp nào đã xảy ra. Trung điểm là số thay thế theo quy tắc, không phải giá trị được xác nhận xuất hiện nhiều nhất hoặc có khả năng xảy ra lớn nhất.

Với khoảng $[a; b)$, $a < b$, trung điểm là $c = \frac{a+b}{2}$. Giả sử nhóm thứ i có f_i quan sát và trung điểm c_i. Thay mỗi quan sát của nhóm bằng c_i nghĩa là thay tổng của nhóm bằng $f_i c_i$. Cộng các tổng thay thế rồi chia cho tổng số quan sát, ta được **số trung bình của mẫu ghép nhóm**:

$$\overline{x}_{\text{nhóm}} = \frac{\sum_i f_i c_i}{n}, \qquad n = \sum_i f_i.$$

Tần số f_i giữ vai trò trọng số: nhóm có bao nhiêu quan sát thì trung điểm được tính bấy nhiêu lần. Nếu chỉ cộng các trung điểm rồi chia cho số nhóm, ta đã tính mỗi nhóm một lần, dù số quan sát của các nhóm khác nhau.

Với hai mẫu A và B, bảng nhóm cho cùng kết quả $\frac{2\cdot 5+2\cdot 15}{4} = 10$. Kết quả này khác trung bình gốc của A là 6 và của B là 14, ngay cả khi không làm tròn. Sự khác nhau phát sinh từ việc thay các giá trị thật bằng trung điểm.

Sai khác giữa hai số trung bình có thể lớn đến đâu? Xét một nhóm rộng 5 mm, chẳng hạn [180; 185), có trung điểm 182,5 mm. Mọi số đo trong nhóm đều cách trung điểm không quá 2,5 mm. Vì vậy, khi thay một số đo bằng trung điểm, số đó có thể tăng hoặc giảm, nhưng lượng thay đổi không vượt quá 2,5 mm.

Nếu cả n số đo đều thuộc các nhóm rộng 5 mm, tổng mới có thể lệch tổng cũ không quá $n \cdot 2{,}5$ mm. Hai số trung bình đều được tính bằng cách chia tổng tương ứng cho cùng số n, nên chúng chênh nhau không quá $\frac{n\cdot 2{,}5}{n} = 2{,}5$ mm.

Với độ rộng chung h, lập luận tương tự cho giới hạn $\frac{h}{2}$: mỗi giá trị thay đổi không quá $\frac{h}{2}$; tổng thay đổi không quá $n \cdot \frac{h}{2}$; chia cho n thì hai trung bình chênh nhau không quá $\frac{h}{2}$.

Đây là **mức sai khác không thể vượt quá**, không phải sai khác thực tế của mọi mẫu. Chẳng hạn, hai số đo 181 mm và 184 mm cùng thuộc $[180; 185)$. Thay cả hai bằng 182,5 mm làm số thứ nhất tăng 1,5 mm và số thứ hai giảm 1,5 mm. Tổng không đổi, nên trung bình cũng không đổi.

Nhóm hẹp hơn làm giới hạn $\frac{h}{2}$ nhỏ hơn. Nhưng sai khác thực tế còn tùy những lượng tăng và giảm có bù lại cho nhau hay không; vì vậy, nó không nhất thiết giảm sau mỗi lần chia nhóm hẹp hơn. Nhiệm vụ 11 (mở trên trang học liệu) giúp đối chiếu hai điều này bằng dữ liệu thật.

Ví dụ 04 — Trung điểm giúp tính, nhưng không khôi phục dữ liệu gốc

Ta cần ước lượng chiều dài cánh bơi trung bình của 49 cá thể Adélie năm 2007 từ bảng ghép nhóm trên:

$$\begin{aligned}\overline{x}_{\text{nhóm}} &= \frac{\begin{aligned}&2 \cdot 172{,}5 + 3 \cdot 177{,}5\\ &+ 15 \cdot 182{,}5 + 10 \cdot 187{,}5\\ &+ 10 \cdot 192{,}5 + 9 \cdot 197{,}5\end{aligned}}{49}\\ &= \frac{9192{,}5}{49}\\ &\approx 187{,}60 \text{ mm}.\end{aligned}$$

Tính trực tiếp từ 49 giá trị gốc cho

$$\overline{x} = \frac{9143}{49} \approx 186{,}59 \text{ mm}.$$

Hai kết quả khác nhau vì phép tính từ bảng đã thay từng giá trị bằng trung điểm của nhóm. Chênh lệch không chỉ do làm tròn.

Căn cứ công thức: Toán 11, Tập một, Kết nối tri thức với cuộc sống, Bài 9, trang 62–63. Các tổng và kết quả ở Ví dụ 04 do ZO Math tính từ phần Adélie 2007.

4.2. Trung vị và tứ phân vị của mẫu ghép nhóm

Ta có một mẫu **giả định** gồm 26 giá trị và muốn ước lượng trung vị khi chỉ còn bảng ghép nhóm. Cách ước lượng sau đây tìm một mốc sao cho số giá trị ở phía dưới mốc ấy được tính gần đúng bằng một nửa mẫu, tức là 13 giá trị.

Bảng nhóm cho biết:

- Có 8 giá trị nhỏ hơn 20.
- Có 10 giá trị thuộc khoảng $[20; 30)$.

Nếu lấy mốc 20, phía dưới mốc mới có 8 giá trị, chưa đủ 13. Nếu lấy mốc 30, phía dưới mốc đã có $8 + 10 = 18$ giá trị, vượt quá 13. Vì vậy, ta tìm mốc ước lượng bên trong nhóm $[20; 30)$.

Bảng chưa cho biết mười giá trị trong nhóm tập trung gần 20, gần 30, hay phân bố theo cách nào khác. Nếu còn danh sách đã sắp xếp, ta sẽ lấy trung bình cộng của giá trị thứ 13 và thứ 14 để tính trung vị. Nhưng bảng nhóm chỉ cho biết cả hai giá trị ấy thuộc $[20; 30)$, chưa cho biết chúng bằng bao nhiêu.

Để tính một giá trị ước lượng, ta giả sử tần số được phân bố đều theo độ rộng của nhóm. Theo cách tính gần đúng này:

- Nửa khoảng đầu $[20; 25)$, từ 20 đến dưới 25, được tính là chứa một nửa tần số của nhóm, tức 5.
- Nửa khoảng sau $[25; 30)$, từ 25 đến dưới 30, được tính là chứa tần số 5 còn lại.

Hai phần không chồng lấn: giá trị bằng 25 thuộc phần sau.

Khi đó, tần số phía dưới mốc 25 được ước tính là $8 + 5 = 13$: 8 ở các nhóm trước và 5 trong nửa đầu nhóm đang xét. Vì vậy, theo cách ước lượng này, ta chọn 25.

Có thể viết lại cách tính ấy thành công thức. Sau 8 giá trị ở các nhóm trước, ta còn cần thêm $13-8 = 5$. Phần cần thêm chiếm

$$\frac{13-8}{10} = \frac{1}{2}$$

tần số của nhóm. **Theo giả thiết phân bố đều**, ta đi thêm cùng tỉ lệ ấy trên độ rộng $30 - 20$. Bắt đầu từ đầu trái 20, ta được

$$\begin{aligned} M_{e,\text{nhóm}} &= 20 + \frac{13-8}{10} \cdot (30-20) \\ &= 20 + 5 \\ &= 25. \end{aligned}$$

Trong công thức, số 10 dưới mẫu là **tần số của nhóm**; còn $30 - 20 = 10$ là **độ rộng của nhóm**. Tỉ số cho biết cần lấy bao nhiêu phần tần số; nhân với độ rộng cho biết cần đi thêm bao xa từ đầu trái.

Kết quả 25 **là trung vị ước lượng từ bảng nhóm.** Trung vị chính xác của mẫu gốc vẫn là

$$M_e = \frac{x_{13} + x_{14}}{2},$$

nhưng hai giá trị x_{13} và x_{14} chưa được biết. Mức 13 trong phép ước lượng biểu thị một nửa tổng tần số; ta không lấy riêng giá trị thứ 13 làm trung vị.

Bước chuyển từ **tỉ lệ tần số** sang **tỉ lệ độ rộng** dựa vào giả thiết phân bố đều. Bảng không tự bảo đảm giả thiết ấy đúng, cũng không bảo đảm có đúng năm giá trị thật nhỏ hơn 25 trong nhóm đang xét. Vì thế, kết quả tính từ bảng có thể khác trung vị tính trực tiếp từ dữ liệu gốc.

Cách ước lượng vừa dùng được gọi là **nội suy tuyến tính**: biết tần số tích lũy tại hai biên nhóm, ta tạm coi tần số ấy tăng đều giữa hai biên để tìm mốc cần đạt. Với trung vị, mốc cần đạt là 50% tổng tần số; với tứ phân vị thứ nhất và thứ ba, các mốc tương ứng là 25% và 75%. Ta sẽ viết công thức chung cho cách tính này rồi áp dụng vào bảng số đo cánh bơi.

Gọi:

- a là đầu trái của nhóm và $h > 0$ là độ rộng nhóm;
- F là tần số tích lũy trước nhóm, còn $f > 0$ là tần số của nhóm;
- r là mức tần số tích lũy cần đạt;
- q là giá trị cần ước lượng, có cùng đơn vị với dữ liệu.

Trước nhóm đã tích lũy được F, nên phần còn thiếu là $r - F$. Phần ấy chiếm tỉ lệ $\frac{r-F}{f}$ của tần số nhóm. Theo giả thiết tăng đều, ta lấy cùng tỉ lệ ấy của độ rộng h, rồi cộng vào đầu trái a. Do đó

$$q = a + \frac{r - F}{f}\,h.$$

Cũng có thể diễn đạt bằng phương trình: tại giá trị q, phần độ rộng đã xét là $q - a$, chiếm tỉ lệ $\frac{q-a}{h}$ của cả nhóm. Tần số tích lũy ước lượng tại đó là $F + f\frac{q-a}{h}$. Yêu cầu đạt mức r cho

$$F + f\frac{q-a}{h} = r \quad\Longrightarrow\quad q - a = \frac{r-F}{f}\,h \quad\Longrightarrow\quad q = a + \frac{r-F}{f}\,h.$$

Chọn mức tích lũy và chọn nhóm. Với mẫu gồm n giá trị, các mức tần số tích lũy cần đạt là $r = \frac{n}{2}$ cho trung vị, $r = \frac{n}{4}$ cho Q_1 và $r = \frac{3n}{4}$ cho Q_3. Từ bảng cộng dồn, chọn nhóm có mức trước nhóm F và mức đến hết nhóm $F + f$ thỏa $F \le r \le F + f$, với $f > 0$. Nếu $F < r < F + f$, kết quả nằm bên trong nhóm.

Nếu mức cần đạt r đúng bằng tần số tích lũy ở biên chung của hai nhóm kề nhau có tần số dương, ta có thể tính từ nhóm đứng trước hoặc nhóm đứng sau. Hai phép tính cùng cho giá trị ở biên ấy.

Chẳng hạn, bảng ở Nhiệm vụ 27 (mở trên trang học liệu) có nhóm $[215; 220)$ với $F = 10$, $f = 7$ và nhóm $[220; 225)$ với $F = 17$, $f = 11$. Khi cần đạt $r = 17$:

- Tính từ nhóm $[215; 220)$, ta cần thêm đủ $17 - 10 = 7$ quan sát, bằng toàn bộ tần số của nhóm. Vì vậy, ta đi hết độ rộng nhóm và được $215 + \frac{17-10}{7} \cdot 5 = 220$ mm.
- Tính từ nhóm $[220; 225)$, tần số tích lũy trước nhóm đã bằng 17. Ta không cần đi thêm vào nhóm, nên được $220 + \frac{17-17}{11} \cdot 5 = 220$ mm.

Đây là hai cách tìm cùng một giá trị biên. Quy tắc xếp số đo vào nhóm vẫn giữ nguyên: một số đo bằng 220 thuộc $[220; 225)$, không thuộc $[215; 220)$.

Nếu có nhóm tần số bằng 0**.** Nhóm ấy không chứa quan sát nào, nên không được chọn để thay vào công thức có mẫu số f. Cần phân biệt trường hợp này với hai nhóm có tần số dương ở ngay cạnh nhau. Nếu giữa chúng có một hoặc nhiều nhóm không chứa quan sát, tần số tích lũy giữ nguyên trên cả khoảng trống ấy. Khi mức cần đạt đúng bằng mức đang giữ nguyên, điều kiện về tần số tích lũy chưa chọn được một giá trị duy nhất trong khoảng trống. Vì vậy, không áp dụng kết luận "hai cách tính cho cùng một biên" qua khoảng trống này; muốn chọn một giá trị cụ thể cần thêm quy tắc cho tình huống đó.

Vì sao mẫu lẻ dùng vị trí giữa, còn bảng nhóm dùng $\frac{n}{2}$? Với danh sách 49 số đo gốc đã sắp, ta có thể tìm chính xác giá trị ở vị trí thứ 25. Khi chỉ còn bảng nhóm, ta không biết giá trị ấy. Phép ước lượng tìm giá trị tại đó tần số tích lũy tăng đều đạt $\frac{49}{2} = 24{,}5$, tức 50% tổng tần số. Mức 24,5 thuộc cách mô tả liên tục này; nó không phải số thứ tự của một cá thể hay "nửa cá thể" thật. Hai cách trả lời hai việc khác nhau — tìm giá trị từ danh sách đã biết và ước lượng từ bảng — nên có thể cho kết quả khác nhau.

Ví dụ 05 — Xác định nhóm trước khi thay số

Trở lại bảng 49 số đo cánh bơi Adélie năm 2007, mức tần số tích lũy cần đạt để ước lượng trung vị là $r = \frac{49}{2} = 24{,}5$. Tần số tích lũy đến nhóm $[180; 185)$ là 20, đến nhóm $[185; 190)$ là 30. Vì vậy dùng nhóm $[185; 190)$, với $a = 185$, $F = 20$, $f = 10$, $h = 5$:

$$M_{e,\text{nhóm}} = 185 + \frac{24{,}5 - 20}{10} \cdot 5$$
$$= 187{,}25 \text{ mm}.$$

Trước nhóm đã có 20 quan sát; nhóm này có 10 quan sát. Theo giả thiết tần số được phân bố đều theo độ rộng của nhóm, ta đi thêm $\frac{24{,}5-20}{10} = 0{,}45$ độ rộng nhóm, tính từ đầu mút trái. Nhóm rộng 5 mm nên kết quả là $185 + 0{,}45 \cdot 5 = 187{,}25$ mm.

Với Q_1, mức cần đạt là 25% tổng tần số: $r = \frac{49}{4} = 12{,}25$. Mức này nằm giữa hai tần số tích lũy 5 và 20, vì vậy ta dùng nhóm chiều dài $[180; 185)$ mm. Trước nhóm có 5 quan sát, trong nhóm có 15 quan sát, nên

$$Q_{1,\text{nhóm}} = 180 + \frac{12{,}25 - 5}{15} \cdot 5$$
$$\approx 182{,}42 \text{ mm}.$$

Nếu nhầm F thành tần số tích lũy đến hết nhóm đang xét, phần đi thêm sẽ âm. Việc hiểu F giúp kiểm phép tính trước khi bấm máy.

Căn cứ toán học: Toán 11, Tập một, Kết nối tri thức với cuộc sống, Bài 9, trung vị ở trang 64 và tứ phân vị ở trang 64–65.

Nhiệm vụ 05 — Tự nội suy và giải thích kết quả

Dùng bảng ghép nhóm 49 số đo cánh bơi Adélie năm 2007 ở đầu mục 4. Em sẽ ước lượng Q_3 và kiểm tra một lời khẳng định về trung vị, để phân biệt kết quả từ bảng nhóm với kết quả tính từ từng số đo.

a) Tìm nhóm dùng để tính Q_3. Ghi r, a, F, f, h, rồi tính $Q_{3,\text{nhóm}}$.

b) Một bạn nói: "Trung vị thật của 49 giá trị phải bằng 187,25 mm vì em đã tính đúng công thức." Em đồng ý đến đâu? Biết trung vị tính trực tiếp từ dữ liệu gốc là 186 mm.

Xem lời giải (mở trên trang học liệu).

4.3. Mốt của mẫu ghép nhóm

Khi còn danh sách từng số đo, ta tìm mốt bằng cách đếm xem giá trị nào xuất hiện nhiều nhất. Nhưng khi các số đo đã được ghép thành từng khoảng, bảng chỉ cho biết **mỗi khoảng chứa bao nhiêu giá trị**, không còn cho biết **mỗi giá trị xuất hiện bao nhiêu lần**.

Trở lại bảng 49 số đo chiều dài cánh bơi của chim Adélie năm 2007. Nhóm $[180; 185)$ mm có tần số lớn nhất là 15. Hai nhóm liền trước và liền sau có tần số như sau:

Khoảng chiều dài cánh bơi, mm	Tần số
[175; 180)	3
[180; 185)	15
[185; 190)	10

Bảng cho biết có 15 cá thể mang số đo từ 180 mm đến dưới 185 mm. Tuy nhiên, ta chưa biết các số đo ấy là bao nhiêu và số đo nào lặp lại nhiều nhất. Vì vậy, **biết nhóm đông nhất chưa đủ để tìm lại mốt của mẫu gốc**.

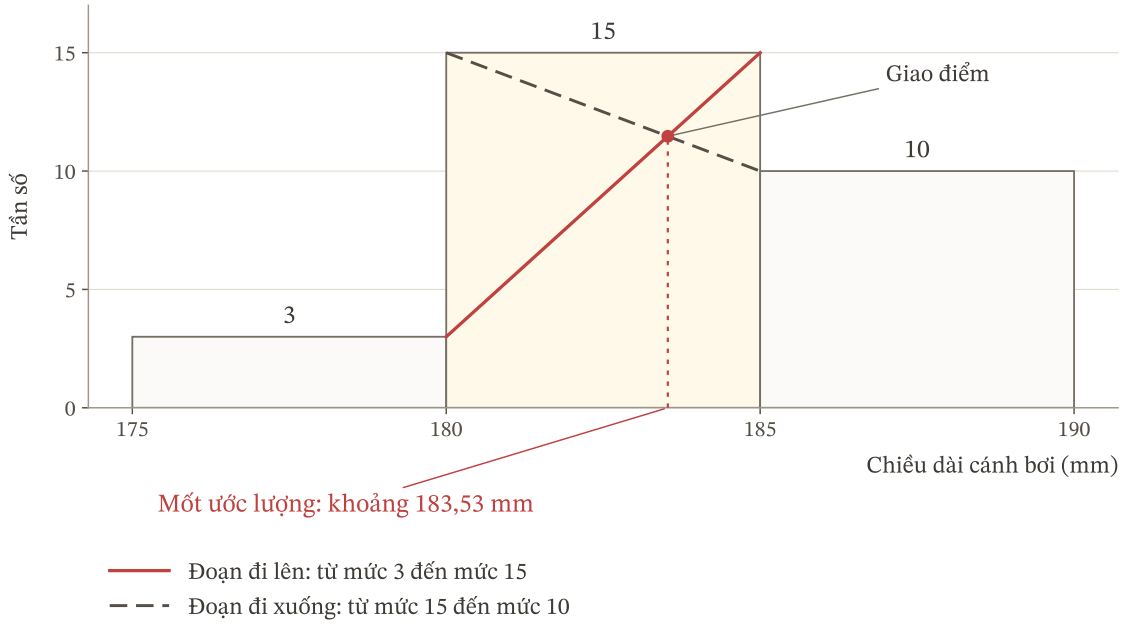
Để tính mốt của mẫu ghép nhóm, trước hết ta chọn nhóm có tần số lớn nhất. Nhóm này được gọi là **nhóm chứa mốt**. Sau đó, ta dùng một quy tắc ước lượng để chọn một giá trị từ nhóm ấy, có xét thêm tần số của hai nhóm liền kề.

Từ ba cột tần số đến một giá trị ước lượng. Ba nhóm đang xét đều rộng 5 mm. Ta biểu diễn chúng bằng ba hình chữ nhật có cùng chiều rộng và có chiều cao lần lượt là 3, 15, 10.

Chỉ riêng cột cao nhất chưa giúp ta chọn một điểm cụ thể từ 180 đến 185: toàn bộ cạnh trên của cột đều cao như nhau. Cách dựng dưới đây sử dụng thêm chiều cao của hai cột bên cạnh:

- Vẽ đoạn thẳng đi lên, từ mức 3 ở cạnh trái của cột giữa đến mức 15 ở cạnh phải.
- Vẽ đoạn thẳng đi xuống, từ mức 15 ở cạnh trái của cột giữa đến mức 10 ở cạnh phải.

Trong hình, đoạn đi lên được vẽ bằng nét liền đỏ; đoạn đi xuống bằng nét đứt xám. Hai đoạn cắt nhau bên trong cột giữa được tô vàng nhạt. Từ giao điểm, kẻ đường vuông góc xuống trục chiều dài cánh bơi. Giá trị đọc được trên trục là mốt ước lượng theo cách dựng này.



Chiếu giao điểm của hai đoạn thẳng xuống trục chiều dài để đọc mốt ước lượng. Đây là kết quả của phép dựng, không phải một số đo được xác định từ dữ liệu gốc.

Đây là **một quy tắc ước lượng từ bảng nhóm**. Bảng không cho biết cách phân bố cụ thể của các số đo bên trong mỗi nhóm, nên không thể tự nó xác nhận vị trí nào là mốt của mẫu gốc. Phép dựng hai đoạn thẳng giúp xác định giá trị mà công thức mốt ghép nhóm sẽ cho.

Tính vị trí giao điểm. Gọi t là khoảng cách theo phương ngang từ đầu trái 180 đến vị trí cần tìm, tính bằng mm. Cả nhóm rộng 5 mm, nên phần độ rộng đã đi qua là

$$\frac{t}{5}.$$

Đoạn thẳng đi lên bắt đầu ở mức 3 và kết thúc ở mức 15, tức tăng tổng cộng

$$15 - 3 = 12.$$

Vì đây là đoạn thẳng, đi qua bao nhiêu phần chiều rộng thì độ cao tăng thêm bấy nhiêu phần của 12. Do đó, tại vị trí cách đầu trái t mm, độ cao của đoạn này là

$$3 + 12 \cdot \frac{t}{5}.$$

Tương tự, đoạn thẳng đi xuống bắt đầu ở mức 15 và kết thúc ở mức 10, tức giảm tổng cộng

$$15 - 10 = 5.$$

Tại cùng vị trí ấy, độ cao của đoạn này là

$$15 - 5 \cdot \frac{t}{5}.$$

Tại giao điểm, hai đoạn thẳng có cùng độ cao. Vì vậy,

$$3 + 12 \cdot \frac{t}{5} = 15 - 5 \cdot \frac{t}{5}.$$

Chuyển các số hạng và thu gọn, ta được

$$(12 + 5) \cdot \frac{t}{5} = 12,$$

suy ra

$$\frac{t}{5} = \frac{12}{12 + 5}, \qquad t = \frac{12}{17} \cdot 5.$$

Như vậy, từ đầu trái 180, ta đi thêm $\frac{12}{17}$ độ rộng nhóm. Mốt ước lượng là

$$\begin{aligned} M_{o,\text{nhóm}} &= 180 + \frac{12}{17} \cdot 5 \\ &\approx 183{,}53 \text{ mm}. \end{aligned}$$

Hai số 12 và 5 trong tỉ số có nguồn gốc cụ thể: $12 = 15 - 3$ là mức tăng của đoạn đi lên; $5 = 15 - 10$ là mức giảm của đoạn đi xuống. Chúng xuất hiện trong công thức khi ta tìm vị trí mà hai đoạn có cùng độ cao.

Công thức tổng quát. Xét bảng có ít nhất hai nhóm liên tiếp cùng độ rộng và có đúng một nhóm đạt tần số lớn nhất. Gọi:

- a là đầu trái của nhóm có tần số lớn nhất;
- $h > 0$ là độ rộng nhóm;
- f là tần số của nhóm ấy;
- f_- và f_+ lần lượt là tần số của nhóm liền trước và nhóm liền sau.

Trong cách dựng vừa xét, đoạn đi lên tăng $f - f_-$, còn đoạn đi xuống giảm $f - f_+$. Vì vậy, phần độ rộng cần lấy từ đầu trái là

$$\frac{f - f_-}{(f - f_-) + (f - f_+)}.$$

Nhân tỉ số ấy với độ rộng h, rồi cộng vào đầu trái a, ta được

$$M_{o,\text{nhóm}} = a + \frac{f - f_-}{(f - f_-) + (f - f_+)}\, h.$$

Nếu nhóm có tần số lớn nhất đứng đầu bảng, ta dùng quy ước $f_- = 0$; nếu đứng cuối bảng, ta dùng $f_+ = 0$.

Khi chỉ có một nhóm đạt tần số lớn nhất, cả hai độ chênh $f - f_-$ và $f - f_+$ đều dương. Tỉ số trong công thức vì thế nằm giữa 0 và 1, nên giá trị tính được nằm bên trong nhóm đã chọn.

Công thức cũng cho thấy vai trò của hai nhóm liền kề:

- Nếu hai nhóm ấy có cùng tần số, hai độ chênh bằng nhau. Tỉ số bằng $\frac{1}{2}$, nên kết quả là trung điểm của nhóm chứa mốt.
- Nếu nhóm bên phải có tần số lớn hơn nhóm bên trái, độ chênh về phía phải nhỏ hơn. Tỉ số lớn hơn $\frac{1}{2}$, nên kết quả nằm về bên phải trung điểm.

Trong ví dụ trên, nhóm bên phải có tần số 10, lớn hơn tần số 3 của nhóm bên trái. Kết quả 183,53 mm vì thế lớn hơn trung điểm 182,5 mm, đúng với vị trí giao điểm trên hình.

Điều kiện cần kiểm trước khi tính. Công thức trên dùng cho các nhóm cùng độ rộng. Nếu một nhóm rộng hơn nhóm khác, nhóm ấy có thể chứa nhiều quan sát hơn chỉ vì khoảng của nó rộng hơn. Khi đó, so sánh riêng tần số chưa đủ để xác định nơi dữ liệu tập trung nhiều hơn; không thay trực tiếp vào công thức này. Ví dụ 09 (trang 25) sẽ minh họa trường hợp các nhóm khác độ rộng.

Nếu tất cả các nhóm có cùng tần số, quy ước đang dùng là mẫu ghép nhóm không có mốt.

Nếu có nhiều nhóm cùng đạt tần số lớn nhất. Trước hết, xác định những nhóm ấy và xem chúng có kề nhau không. Số nhóm có tần số lớn nhất chưa cho biết sẽ có bao nhiêu giá trị mốt khác nhau khi tính bằng công thức.

Chẳng hạn, bốn nhóm $[0;10)$, $[10;20)$, $[20;30)$, $[30;40)$ có tần số lần lượt là $3; 8; 8; 2$. Hai nhóm ở giữa cùng có tần số lớn nhất. Nếu xét nhóm $[10;20)$, hai độ chênh là $8-3=5$ và $8-8=0$, nên công thức cho $10 + \frac{5}{5+0} \cdot 10 = 20$. Nếu xét nhóm $[20;30)$, hai độ chênh là $8-8=0$ và $8-2=6$, nên công thức cho $20 + \frac{0}{0+6} \cdot 10 = 20$. Hai nhóm được xét nhưng chỉ có một giá trị tính được: biên chung 20.

Nếu một nhóm và cả hai nhóm liền kề đều cùng đạt tần số lớn nhất, cả hai độ chênh đều bằng 0. Công thức khi ấy có mẫu số bằng 0, nên không tính được giá trị mốt từ nhóm đang xét. Trong bài làm, cần chỉ ra các nhóm cùng đạt tần số lớn nhất và nói rõ công thức không xác định ở nhóm nào. Nếu đề chưa cho quy tắc xử lí thêm, ta dừng ở kết luận đó, không tự chọn một trung điểm làm mốt. Trường hợp này khác với quy ước không có mốt khi **tất cả** các nhóm có cùng tần số.

Ví dụ 06 — Mốt của bảng nhóm và mốt của mẫu gốc

Với bảng ghép nhóm của 49 số đo chiều dài cánh bơi Adélie năm 2007, ta vừa tính được

$$M_{o,\text{nhóm}} \approx 183{,}53 \text{ mm}.$$

Kết quả này được tính từ tần số của ba nhóm liền nhau và cách ước lượng đã giải thích. Nó không cho biết số đo 183,53 mm đã xuất hiện bao nhiêu lần trong danh sách ban đầu, cũng không chứng minh đây là số đo xuất hiện nhiều nhất.

Khi đếm trực tiếp trong dữ liệu gốc, ta có kết quả khác: các số đo 180, 190, 195 mm đều xuất hiện 5 lần, cùng đạt tần số lớn nhất. Vì vậy, mẫu gốc có ba mốt là

$$180 \text{ mm}, \qquad 190 \text{ mm}, \qquad 195 \text{ mm}.$$

Hai kết quả được xác định từ hai dạng thông tin khác nhau. Với mẫu gốc, ta đếm số lần xuất hiện của từng số đo. Với mẫu ghép nhóm, ta chỉ biết số quan sát trong từng khoảng và dùng công thức để ước lượng. Do đó, mốt tính từ bảng nhóm có thể khác các mốt tìm được từ danh sách ban đầu.

Căn cứ công thức: Toán 11, Tập một, Kết nối tri thức với cuộc sống, Bài 9, trang 66; Bài tập Toán 11, Tập một, trang 48, bài 3.15 ở trang 51 và đáp án trang 129. Cách dựng giao điểm hai đoạn thẳng được đối chiếu với SCERT Kerala, Statistics, Class XI (2016), mục 5.3.1, trang 129. Hình và phép biến đổi trên được ZO Math dựng lại cho bảng Adélie. Xem thêm lưu ý về định nghĩa và quy ước (trang 28).

5. Chọn số đại diện theo câu hỏi

Các phần trước đã giải thích cách tính và ý nghĩa của trung bình, trung vị, tứ phân vị và mốt. Bây giờ, ta dùng những hiểu biết ấy để quyết định nên báo cáo số nào trước một câu hỏi cụ thể.

Gợi mở 05 — Đại diện cho dữ liệu nào và trả lời câu hỏi gì?

Một mẫu **giả định** gồm tám giá trị:

$$2;\ 2;\ 3;\ 3;\ 14;\ 14;\ 15;\ 15.$$

Hãy tính trung bình và trung vị của mẫu. Nếu chỉ được biết hai kết quả ấy, người đọc có nhận ra rằng các giá trị tập trung thành hai cụm tách biệt không?

Tiếp theo, một báo cáo muốn so sánh chiều dài cánh bơi của hai loài chim cánh cụt, nhưng chỉ đưa ra số trung bình tính chung cho cả hai loài. Con số ấy đã đủ để thực hiện việc so sánh chưa? Em cần thêm thông tin gì?

Với mẫu giả định trên, ta tính được

$$\overline{x} = \frac{2+2+3+3+14+14+15+15}{8} = 8{,}5.$$

Dãy đã được sắp theo thứ tự tăng dần. Hai giá trị ở giữa là 3 và 14, nên

$$M_e = \frac{3+14}{2} = 8{,}5.$$

Hai phép tính đều đúng. Tuy nhiên, số 8,5 chưa cho thấy một đặc điểm đáng chú ý: bốn giá trị nằm ở 2 hoặc 3, còn bốn giá trị nằm ở 14 hoặc 15. Không có giá trị nào nằm trong khoảng giữa 3 và 14.

Trung bình và trung vị không bắt buộc phải trùng với một giá trị của mẫu. Điều cần xét ở đây là **ta muốn người đọc biết gì về dữ liệu**. Nếu muốn mô tả hai cụm giá trị, chỉ báo cáo 8,5 là chưa đủ. Đổi từ trung bình sang trung vị cũng không giúp người đọc nhận ra hai cụm, vì hai kết quả bằng nhau. Ta cần giữ thêm danh sách, bảng, hình biểu diễn hoặc một câu mô tả rõ hai cụm ấy.

Vì vậy, trước khi chọn số đại diện, cần xác định câu hỏi đang đặt ra.

Nếu hỏi "Tổng số lượt thuê xe trong tháng chia đều cho mỗi ngày là bao nhiêu?", ta dùng **trung bình**. Trung bình được tính bằng tổng chia cho số ngày; nhân kết quả chưa làm tròn với số ngày sẽ thu lại tổng.

Nếu hỏi "Khi xếp các ngày từ ít đến nhiều lượt thuê, mức ở giữa là bao nhiêu?", ta dùng **trung vị**. Cách tính này dựa vào vị trí trong dãy đã sắp. Nó không dùng tổng số lượt thuê, nên không thể lấy trung vị thay cho trung bình khi cần tính lại tổng.

Những câu hỏi khác có thể cần thông tin khác. Muốn biết giá trị nào xuất hiện nhiều nhất, ta xét **mốt và bảng tần số**. Muốn biết thêm các mốc ở nửa dưới và nửa trên của dãy đã sắp, ta xét **tứ phân vị**. Muốn so sánh hai loài chim, ta cần các kết quả tính riêng theo loài. Muốn biết ngày nào đông khách nhất, ta phải tìm ngày có số lượt thuê lớn nhất.

Như vậy, chọn số đại diện gồm hai việc gắn liền với nhau: **chọn số trả lời đúng câu hỏi và bổ sung thông tin mà con số ấy chưa thể hiện**.

Ví dụ 07 — Hai loài và một số trung bình chung

Adélie và Gentoo là hai loài chim cánh cụt có trong bộ dữ liệu Palmer Penguins. Xét các số đo chiều dài cánh bơi năm 2007 được trích từ bộ dữ liệu này.

Bảng dưới đây tóm tắt 49 số đo hợp lệ của loài Adélie và 34 số đo hợp lệ của loài Gentoo. "Tổng chiều dài" là tổng các số đo đã ghi; chiều dài được đo bằng mm.

Loài	Số giá trị hợp lệ	Tổng chiều dài, mm	Trung bình, mm	Trung vị, mm
Adélie	49	9143	186,59	186
Gentoo	34	7314	215,12	215

Nếu câu hỏi là "Chiều dài cánh bơi trung bình của tất cả các cá thể trong hai phần dữ liệu này là bao nhiêu?", ta gộp tổng chiều dài và số cá thể:

$$\bar{x} = \frac{9143 + 7314}{49 + 34} = \frac{16457}{83} \approx 198{,}28 \text{ mm}.$$

Đây là trung bình của 83 số đo, trong đó mỗi số đo được tính một lần. Ta không cộng hai trung bình riêng rồi chia cho 2, vì cách ấy cho hai nhóm mức đóng góp như nhau, trong khi một nhóm có 49 số đo và nhóm kia chỉ có 34 số đo.

Nhưng nếu câu hỏi là "Chiều dài cánh bơi của hai loài khác nhau như thế nào trong những số liệu đã thu thập?", số 198,28 mm chưa trả lời được. Nó không cho biết trung bình của loài nào lớn hơn, cũng không cho biết hai trung bình cách nhau bao nhiêu.

Lúc này, cần đọc kết quả theo từng loài. Trung bình của Adélie khoảng 186,59 mm, còn của Gentoo khoảng 215,12 mm. Các trung vị tương ứng là 186 mm và 215 mm. Những kết quả riêng ấy thể hiện sự khác biệt mà số trung bình chung đã không thể hiện.

Bảng cho phép so sánh các số trung bình và trung vị của hai phần dữ liệu. Bảng chưa cho biết nguyên nhân sinh học của sự khác biệt, cũng không chứng minh rằng mọi cá thể Gentoo đều có cánh bơi dài hơn mọi cá thể Adélie.

Ví dụ 08 — Cùng một tháng, hai cách mô tả số lượt thuê

Capital Bikeshare là một hệ thống xe đạp dùng chung ở Hoa Kỳ. Bộ dữ liệu **Bike Sharing** ghi số lượt thuê xe của hệ thống theo ngày. Một lượt thuê là một lần sử dụng dịch vụ; cùng một người có thể thuê nhiều lượt.



Minh họa hoạt động xe đạp dùng chung.

Trong tháng 01/2012, bảng có 31 giá trị, mỗi giá trị là số lượt thuê của một ngày. Tổng số lượt thuê trong tháng là 96744.

Câu hỏi thứ nhất: Nếu chia đều tổng số lượt thuê cho 31 **ngày, mỗi ngày được bao nhiêu lượt?**

Ta lấy tổng chia cho số ngày:

$$\overline{x} = \frac{96744}{31} \approx 3120{,}77 \text{ lượt mỗi ngày.}$$

Kết quả này mô tả mức trung bình tính từ tổng của cả tháng. Nó không khẳng định ngày nào cũng có số lượt thuê bằng mức ấy. Khi nhân trung bình **chưa làm tròn** với 31, ta thu lại tổng 96744 lượt.

Câu hỏi thứ hai: Khi xếp 31 **ngày từ ít đến nhiều lượt thuê, ngày ở giữa có bao nhiêu lượt?**

Với 31 giá trị đã sắp, giá trị ở giữa nằm tại vị trí thứ 16: có 15 giá trị đứng trước và 15 giá trị đứng sau. Giá trị tại vị trí này là 3243, nên trung vị bằng 3243 lượt.

Trung vị được xác định bằng vị trí trong dãy đã sắp, không phải bằng cách chia tổng cho số ngày. Vì vậy, lấy 3243 nhân với 31 không cho lại tổng số lượt thuê của tháng.

Hai kết quả khoảng 3120,77 và 3243 khá gần nhau, nhưng điều đó chưa cho biết các ngày còn lại có bao nhiêu lượt thuê. Chẳng hạn, ta chưa biết ngày ít khách nhất và ngày đông khách nhất cách nhau bao nhiêu. Cũng không thể chỉ dựa vào sự gần nhau của trung bình và trung vị để kết luận dữ liệu đối xứng hoặc không có ngày khác biệt rõ rệt. Những câu hỏi ấy cần được xem xét bằng danh sách số liệu hoặc hình biểu diễn.

Mốt lại trả lời một câu hỏi khác: **Có số lượt thuê nào được ghi nhận nhiều lần hơn các số còn lại không?** Trong 31 ngày đang xét, các số lượt thuê đều khác nhau, mỗi giá trị xuất hiện một lần. Theo quy ước đã nêu đối với mẫu số liệu gốc, mẫu này không có mốt.

Nếu ghép các số lượt thuê vào bốn nhóm

$$[1000; 2000), \quad [2000; 3000), \quad [3000; 4000), \quad [4000; 5000),$$

thì tần số lần lượt là 3; 10; 10; 8. Bảng nhóm cho biết hai khoảng ở giữa cùng có nhiều ngày nhất. Điều đó không có nghĩa là có một số lượt thuê cụ thể xuất hiện nhiều lần: nhiều giá trị khác nhau vẫn có thể cùng thuộc một khoảng.

Khi viết báo cáo, cần giữ đúng đối tượng được đếm là **lượt thuê**. Trung bình khoảng 3121 lượt mỗi ngày không có nghĩa là trung bình khoảng 3121 người khác nhau thuê xe mỗi ngày. Các kết quả cũng chỉ mô tả tháng đang xét; chúng chưa đủ để dự báo chính xác số lượt thuê của một ngày trong tương lai.

Nguồn: Hadi Fanaee-T và João Gama, Bike Sharing, UCI Machine Learning Repository, tệp day.csv, cột ngày (dteday) và cột số lượt thuê (cnt), phần trích tháng 01/2012. Dữ liệu CC BY 4.0; ZO Math chọn phần trích và tính bảng tóm tắt, không thay giá trị gốc.

Ví dụ 09 — Đọc số 0 cùng ký hiệu quan trắc

Chọn đúng số đại diện còn đòi hỏi hiểu đúng những giá trị dùng để tính số ấy. Ví dụ sau cho thấy vì sao đôi khi cần đọc cả ký hiệu đi kèm số đo.

Một trạm khí tượng tại sân bay Seattle–Tacoma, bang Washington, Hoa Kỳ, ghi lượng giáng thủy theo ngày. **Giáng thủy** là nước từ mưa, tuyết… rơi xuống. Để hiểu đơn vị mm, hãy hình dung lượng nước ấy được giữ lại và phủ đều trên một mặt phẳng nằm ngang: độ sâu của lớp nước là lượng giáng thủy được báo cáo. Tuyết được quy đổi thành lượng nước sau khi tan. Vì vậy, số mm ở đây chỉ độ sâu lớp nước tương đương, không phải độ sâu lớp tuyết.

Bảy ngày đầu tháng 01/2023 được ghi như sau:

Ngày	Lượng giáng thủy được ghi, mm	Ký hiệu vết giáng thủy
01/01/2023	0,3	Không có ký hiệu T
02/01/2023	0,0	Không có ký hiệu T
03/01/2023	0,0	T
04/01/2023	0,5	Không có ký hiệu T
05/01/2023	0,0	T
06/01/2023	3,0	Không có ký hiệu T
07/01/2023	7,1	Không có ký hiệu T

Ngày 03 và ngày 05 đều có giá trị số là 0,0, nhưng còn kèm ký hiệu **T**. Ký hiệu này chỉ **vết giáng thủy**: đã ghi nhận một lượng giáng thủy rất nhỏ. Do đó, đọc riêng số 0,0 rồi kết luận hai ngày ấy hoàn toàn không có giáng thủy sẽ bỏ qua thông tin của ký hiệu đi kèm.

Xét toàn bộ 31 ngày trong tháng 01/2023 tại cùng trạm, tổng các lượng giáng thủy được ghi là 78,5 mm. Trung bình được tính bằng

$$\overline{x} = \frac{78{,}5}{31} \approx 2{,}53 \text{ mm mỗi ngày.}$$

Con số này trả lời câu hỏi: “Nếu chia đều tổng lượng giáng thủy được ghi trong tháng cho 31 ngày, mỗi ngày được bao nhiêu?”

Nếu xếp 31 giá trị từ nhỏ đến lớn, giá trị ở vị trí thứ 16 là 0,8 mm. Vì vậy, trung vị bằng 0,8 mm. Con số này cho biết mức ở giữa dãy đã sắp.

Để hiểu vì sao trung bình lớn hơn trung vị, ta xem thêm bảng tần số:

Lượng giáng thủy được ghi, mm	**Số ngày**
{0}	13, trong đó 6 ngày có ký hiệu T
(0; 1)	4
[1; 5)	8
[5; 10)	3
[10; 20)	3
[20; 40)	0

Có 17 ngày ghi lượng giáng thủy dưới 1 mm, nhưng cũng có 6 ngày ghi từ 5 mm trở lên. Những ngày có lượng giáng thủy lớn đóng góp nhiều vào tổng của tháng, từ đó làm tăng trung bình. Trung vị được xác định bằng vị trí thứ 16, nên vẫn nằm trong các giá trị dưới 1 mm.

Hai kết quả mô tả hai khía cạnh khác nhau. Trung bình gắn với tổng lượng được ghi trong tháng; trung vị cho biết mức ở giữa khi xếp các ngày theo lượng giáng thủy. Muốn thấy được cả những ngày ghi ít và những ngày ghi nhiều, cần đọc thêm bảng tần số.

Ký hiệu T vẫn cần được giữ khi diễn giải kết quả. Trong 13 ngày có giá trị số bằng 0, có 6 ngày được ghi nhận vết giáng thủy. Vì vậy, "có 13 ngày được ghi bằng 0" và "có 13 ngày hoàn toàn không có giáng thủy" là hai phát biểu khác nhau. Tổng 78,5 mm và trung bình khoảng 2,53 mm mỗi ngày đều được tính từ các giá trị số đã ghi; ta không tự gán thêm một lượng nước cho những ngày có ký hiệu T.

Cũng cần chú ý cách lập bảng: hàng {0} chỉ chứa đúng giá trị 0, còn những hàng khác là các khoảng có độ rộng khác nhau. Các nhóm không thỏa mãn điều kiện rộng bằng nhau của công thức mốt ghép nhóm vừa học, nên không áp dụng công thức ấy cho bảng này. Nếu cần trung vị hoặc tứ phân vị, ta có thể tính trực tiếp từ 31 giá trị gốc còn được lưu giữ.

Nguồn: NOAA (Cơ quan Quản lý Khí quyển và Đại dương Quốc gia Hoa Kỳ), qua NCEI (Trung tâm Thông tin Môi trường Quốc gia), bộ dữ liệu khí tượng theo ngày GHCN-Daily. Trạm Seattle–Tacoma có mã USW00024233; PRCP là mã cột lượng giáng thủy. Phần trích gồm tháng 01/2023 và các ký hiệu ghi kèm số đo. Phần dữ liệu được lấy từ NCEI theo hệ mét; PRCP trong phần trích này đã có đơn vị mm. Ý nghĩa ký hiệu T được nêu trong hướng dẫn dữ liệu GHCN-Daily của NOAA. Cách hiểu lượng giáng thủy theo độ sâu lớp nước tương đương được trình bày trong hướng dẫn đo của WMO, Chương 6.

Nhiệm vụ 06 — Sửa câu kết luận

Một người viết bản tin từ hai ví dụ vừa học: số lượt thuê xe của Capital Bikeshare trong tháng 01/2012 và lượng giáng thủy tại trạm Seattle–Tacoma trong tháng 01/2023.

Hãy kiểm tra các phát biểu dưới đây. Khi sửa, cần nói rõ dữ liệu ghi điều gì, thuộc hệ thống hoặc địa điểm nào và trong thời gian nào.

a) Sửa câu: "Tháng 01/2012, trung bình có khoảng 3121 người khác nhau thuê xe mỗi ngày."

b) Sửa câu: "Tháng 01/2023 có 13 ngày chắc chắn hoàn toàn không có giáng thủy tại Seattle–Tacoma."

c) Từ dữ liệu của một tháng tại một trạm, có thể kết luận khí hậu của cả thành phố đã thay đổi dài hạn không? Giải thích và nêu những thông tin còn cần thu thập.

Xem lời giải (mở trên trang học liệu).

Khép lại Bài học

Qua các ví dụ vừa học, ta thấy một phép tính đúng chỉ trở thành lời mô tả có ích khi kết quả được gắn với đúng câu hỏi và đúng dữ liệu. Trước khi báo cáo, hãy nói rõ các quan sát ghi điều gì, thuộc những đối tượng nào và kết quả có đơn vị gì. Sau đó giải thích vì sao số đã chọn trả lời được câu hỏi, đồng thời nêu thêm thông tin cần thiết mà số ấy chưa thể hiện.

Nếu vẫn nhầm vị trí với giá trị, dữ liệu thiếu với số 0, hoặc ước lượng từ bảng với kết quả từ mẫu gốc, hãy trở lại đúng mục ấy.

Trước khi sang Luyện tập, thử trả lời các câu hỏi sau bằng lập luận của mình:

- Vì sao chỉ một thời gian chờ tăng mà trung bình của cả nhóm tăng?
- Xét lại mẫu thời gian chờ 2; 3; 3; 4; 8 phút. Nếu chỉ thay 8 bằng 28, vì sao trung vị không đổi? Trở lại mẫu ban đầu, nếu chỉ thay 2 bằng 30, trung vị thay đổi thế nào và vì sao?
- Vì sao hai mẫu cùng trung bình và trung vị vẫn có thể có tứ phân vị khác nhau?
- Vì sao hai mẫu cùng cho một bảng ghép nhóm mà có thể có trung bình khác nhau?

Khi trả lời, nêu bước nối giữa dữ kiện và kết luận. Các nhiệm vụ ở phần Luyện tập sẽ yêu cầu em vận dụng những lập luận ấy vào các mẫu mới.

Nguồn và cách dùng dữ liệu

Mục này cho biết kiến thức toán dựa vào tài liệu nào, số liệu đến từ đâu và ZO Math đã chọn, trình bày số liệu ra sao. Em có thể dùng các liên kết để tra cách ghi số đo hoặc đọc thêm; không cần học thuộc tên tác giả, mã cột hay mã trạm.

Tài liệu toán học

- **Toán 10, Tập một — Kết nối tri thức với cuộc sống**, Nhà xuất bản Giáo dục Việt Nam, Bài 13, trang 78–81: trung bình, trung vị, tứ phân vị và mốt của mẫu chưa ghép nhóm.
- **Toán 11, Tập một — Kết nối tri thức với cuộc sống**, Nhà xuất bản Giáo dục Việt Nam, Bài 8, trang 58–60; Bài 9, trang 62–66: mẫu ghép nhóm, các số đặc trưng và điều kiện áp dụng.
- **Chương trình giáo dục phổ thông môn Toán 2018**, Bộ Giáo dục và Đào tạo, trang 85 và 101–102: yêu cầu tính, giải thích ý nghĩa và sử dụng các số đặc trưng của mẫu không ghép nhóm và ghép nhóm, phần lớp 10 và lớp 11.
- **Toán 10 — Sách giáo viên, Kết nối tri thức với cuộc sống**, Nhà xuất bản Giáo dục Việt Nam, trang 142–145: ý nghĩa, cách lựa chọn các số đặc trưng, hướng dẫn dạy mốt và lời giải liên quan.
- **Toán 11 — Sách giáo viên, Kết nối tri thức với cuộc sống**, Nhà xuất bản Giáo dục Việt Nam, trang 103–105: cách dẫn dắt về mốt ghép nhóm, giải thích kết quả và phân biệt nhóm chứa mốt với giá trị tính được.
- **Bài tập Toán 10, Tập một — Kết nối tri thức với cuộc sống**, Nhà xuất bản Giáo dục Việt Nam, trang 75, 77, 126: phần nhắc kiến thức về mốt, bài 5.10 và lời giải.

- **Bài tập Toán 11, Tập một — Kết nối tri thức với cuộc sống**, Nhà xuất bản Giáo dục Việt Nam, trang 48, 50–52, 128–129: quy tắc mốt ghép nhóm, các bài 3.8, 3.15, 3.20–3.21, 3.25 và lời giải.

Chương trình xác định yêu cầu cần đạt; sách giáo khoa cung cấp căn cứ về định nghĩa, quy tắc và công thức; sách giáo viên cung cấp gợi ý về ý nghĩa và cách dẫn dắt. Các bộ dữ liệu dưới đây cung cấp số đo, đơn vị và phạm vi quan sát.

Lời giảng, nhiệm vụ, lời giải và hình minh họa do ZO Math biên soạn; không sao chép bài tập hoặc hình sách.

Lưu ý về định nghĩa và quy ước mốt. Toán 10, Tập một, bộ Kết nối tri thức với cuộc sống, trang 81, định nghĩa mốt theo tần số lớn nhất và nêu quy ước không có mốt khi các giá trị xuất hiện với tần số như nhau. Với mẫu có nhiều giá trị khác nhau, cần đọc cả định nghĩa lẫn quy ước này: việc các giá trị đều lặp lại chưa đủ để kết luận có mốt.

Nếu mẫu chỉ có một giá trị khác nhau, cần phân biệt điều đếm được với cách gọi tên: chẳng hạn, trong 5; 5; 5, chắc chắn giá trị 5 xuất hiện ba lần. Việc gọi đó là "mốt" hay áp dụng lời quy ước "các tần số như nhau" cần được làm rõ trong định nghĩa mà để sử dụng; không suy ra một quy tắc mới chỉ từ cách đếm. Trang sách nêu trên không giải thích riêng trường hợp này.

Với mẫu ghép nhóm, công thức ở Toán 11, Tập một, trang 66, áp dụng cho các nhóm cùng độ rộng; sách cũng nêu quy ước không có mốt khi tất cả tần số nhóm bằng nhau. Khi có nhiều nhóm cùng đạt tần số lớn nhất, xét từng nhóm cùng tần số của hai nhóm liền kề: nếu mẫu số khác 0 thì tính và đối chiếu các kết quả; nếu mẫu số bằng 0 thì ghi rõ công thức không xác định ở nhóm đó. Hai nhóm kề có thể cho cùng một giá trị tại biên, như ví dụ trong phần mốt ghép nhóm (trang 18). Không suy số giá trị mốt từ số nhóm có tần số lớn nhất và không tự dùng công thức cho các nhóm khác độ rộng.

Palmer Penguins

Đây là bộ dữ liệu về chim cánh cụt được nghiên cứu ở khu vực quần đảo Palmer, Nam Cực. Adélie, Chinstrap và Gentoo là tên ba loài có trong bảng. Các bài tập sử dụng chiều dài cánh bơi — phần cánh có dạng mái chèo giúp chim bơi — đo bằng mm; trong bảng gốc, cột này mang tên flipper_length_mm. Mỗi dòng ghi thông tin một cá thể.

Trang dữ liệu và mô tả biến · Tệp bảng số liệu CSV · CC0 1.0.

Ghi công Allison Horst, Alison Hill, Kristen Gorman. Dữ liệu gốc do Kristen Gorman và chương trình nghiên cứu sinh thái dài hạn tại trạm Palmer (Palmer Station LTER) cung cấp; trang nguồn dẫn tới các bộ dữ liệu lưu tại Environmental Data Initiative (EDI). Nghiên cứu liên quan: Gorman, Williams và Fraser (2014), Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins.

ZO Math chọn các số đo cánh bơi theo loài và năm nghiên cứu để lập bài học. Ô thiếu vẫn được giữ trong hồ sơ; khi tính với chiều dài cánh bơi, chỉ dùng các ô có số đo phù hợp, không thay ô thiếu bằng 0. Các bảng ghép nhóm được lập từ những số đo ấy; kết quả gần đúng chỉ làm tròn ở bước cuối. Những nhóm nhỏ dùng trong bài được chọn để học thao tác, không phải mẫu lấy ngẫu nhiên để đại diện cho toàn bộ loài.

Bike Sharing

Bộ dữ liệu này ghi số lượt thuê xe đạp của hệ thống Capital Bikeshare tại Hoa Kỳ trong các năm 2011–2012. Một lượt thuê là một lần sử dụng dịch vụ, không phải một người dùng khác nhau. UCI Machine Learning Repository là kho lưu trữ và giới thiệu bộ dữ liệu, được liên kết dưới đây.

Fanaee-T, H. và Gama, J., Bike Sharing — UCI Machine Learning Repository, DOI dữ liệu, CC BY 4.0.

ZO Math dùng tệp day.csv, trong đó mỗi dòng ứng với một ngày, chọn các cột ngày, số lượt thuê và thông tin cho biết ngày đó có phải ngày làm việc hay không; trích theo tháng, tính tổng và các số đặc trưng, dịch nhãn sang tiếng Việt. Không thay các giá trị quan sát. Việc ghi công không hàm ý tác giả hoặc UCI bảo trợ học liệu.

Dữ liệu mô tả một hệ thống thuê xe và giai đoạn cụ thể. Không dùng lượt thuê để suy trực tiếp số người khác nhau, năng lực phục vụ cần có ở mọi thời điểm hoặc nguyên nhân của sự khác biệt.

NOAA/NCEI

NOAA là Cơ quan Quản lý Khí quyển và Đại dương Quốc gia Hoa Kỳ; NCEI là Trung tâm Thông tin Môi trường Quốc gia thuộc NOAA. Bộ GHCN-Daily tập hợp số liệu khí tượng theo ngày. Các số liệu được sử dụng ở đây ghi lượng giáng thủy tại trạm sân bay Seattle–Tacoma, bang Washington, Hoa Kỳ; USW00024233 là mã giúp tìm đúng trạm trong nguồn.

Menne và cộng sự, GHCN-Daily, Version 3, giải thích các cột và ký hiệu trong dữ liệu, thông tin trạm USW00024233.

PRCP là mã cột lượng giáng thủy: nước từ mưa, tuyết… được quy đổi thành lượng nước tương đương. Các số liệu được chọn thuộc năm 2023, được lấy theo hệ mét và trình bày bằng mm. Ký hiệu T cho biết vết giáng thủy — một lượng rất nhỏ chưa thể hiện bằng số đo dương ở độ chính xác đang dùng.

ZO Math giữ các ký hiệu đi kèm số đo. Tổng và trung bình được tính từ những con số đã ghi; không tự đặt một lượng nước cụ thể thay cho T. Khi diễn giải kết quả, vẫn phải đọc ký hiệu này, không coi mọi ngày ghi 0 là hoàn toàn không có giáng thủy.

Chính sách dữ liệu mở của NCEI phân biệt dữ liệu chính phủ Hoa Kỳ thuộc phạm vi công cộng với dữ liệu bên thứ ba có thể có điều kiện riêng. Không suy rằng mọi dữ liệu trên NCEI đều mang CC0. Học liệu chỉ sử dụng phần dữ liệu trạm Hoa Kỳ đã nêu, không sử dụng logo hoặc ảnh của cơ quan.

Các nguồn dữ liệu được truy cập ngày 22/10/2026. Khi diễn giải số liệu, cần giữ đúng đơn vị, mẫu và thời điểm quan sát.

HỌC TIẾP CÙNG ZO MATH

Chọn điều em muốn học tiếp.

Em có thể bắt đầu với một nội dung đang học, muốn ôn lại hoặc còn vướng khi tự giải bài.

Mở học liệu của khóa học

Bảo trợ ZO Math

Nếu học liệu hữu ích với bạn, bạn có thể bảo trợ ZO Math tiếp tục biên soạn và chia sẻ.

Ngân hàng
Vietcombank

Số tài khoản
0601000137768

Chủ tài khoản
Nguyễn Tấn Nhựt

Nội dung chuyển khoản
BAO TRO ZO MATH

Xem thông tin bảo trợ ZO Math

Phiên bản nội dung 0.4 · Ngày tạo PDF: 04/11/2026

© 2026 ZO Math. Bản quyền nội dung thuộc ZO Math.