CHƯƠNG V

CÁC SỐ ĐẶC TRƯNG CỦA MẪU SỐ LIỆU KHÔNG GHÉP NHÓM

"Thống kê là cơ sở của khoa học".
Karl Pearson (nhà thống kê người Anh, 1857 – 1936)

Bài 12 SỐ GẦN ĐÚNG VÀ SAI SỐ

THUẬT NGỮ
- Số gần đúng
- Sai số tuyệt đối
- Độ chính xác
- Sai số tương đối
- Số quy tròn

KIẾN THỨC, KĨ NĂNG
- Hiểu khái niệm số gần đúng, sai số tuyệt đối.
- Xác định số gần đúng của một số với độ chính xác cho trước.
- Xác định sai số tương đối của số gần đúng.
- Xác định số quy tròn của số gần đúng với độ chính xác cho trước.
- Biết sử dụng máy tính cầm tay để tính toán với các số gần đúng.

Đỉnh Everest được mệnh danh là "nóc nhà của thế giới", bởi đây là đỉnh núi cao nhất trên Trái Đất so với mực nước biển. Có rất nhiều con số khác nhau đã từng được công bố về chiều cao của đỉnh Everest:

8 848 m; 8 848,13 m; 8 844,43 m; 8 850 m;...

Vì sao lại có nhiều kết quả khác nhau như vậy và đâu là con số chính xác? Chúng ta sẽ cùng tìm câu trả lời trong bài học này, sau khi tìm hiểu về số gần đúng và sai số.

Đỉnh Everest

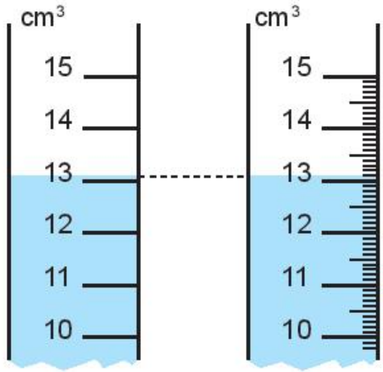
1. SỐ GẦN ĐÚNG

HĐ1. Ngày 8-12-2020, Trung Quốc và Nepal ra thông cáo chung khẳng định chiều cao mới đo được của đỉnh núi cao nhất thế giới Everest là 8 848,86 m.

(Theo *Tuoitre.vn*)

Trong các số được đưa ra ở tình huống mở đầu, số nào gần nhất với số được công bố ở trên?

HĐ2. Trang và Hoà thực hiện đo thể tích một cốc nước bằng hai ống đong có vạch chia được kết quả như Hình 5.1. Hãy cho biết số đo thể tích trên mỗi ống.



Hình 5.1

> Trong nhiều trường hợp, ta không biết hoặc khó biết số đúng (kí hiệu là $\bar{a}$) mà chỉ tìm được giá trị khác xấp xỉ nó. Giá trị này được gọi là số gần đúng, kí hiệu là a.

Chẳng hạn, các số đo khác nhau về chiều cao của đỉnh Everest trong tình huống mở đầu đều là các số gần đúng.

? Hãy lấy một ví dụ khác về số gần đúng.

Ví dụ 1. Gọi d là độ dài đường chéo của hình vuông cạnh bằng 1. Trong hai số $\sqrt{2}$ và 1,41, số nào là số đúng, số nào là số gần đúng của d?

Giải

Hình vuông có cạnh bằng 1 có độ dài của đường chéo là $d = 1 \cdot \sqrt{2} = \sqrt{2}$. Vậy $\sqrt{2}$ là số đúng; 1,41 là số gần đúng của d.

Luyện tập 1. Gọi P là chu vi của đường tròn bán kính 1 cm. Hãy tìm một giá trị gần đúng của P.

Chú ý. Ta có thể sử dụng máy tính cầm tay để tìm giá trị gần đúng của các biểu thức chứa các số vô tỉ như $\pi, \sqrt{a}, \sqrt[3]{a}, \ldots$ Chẳng hạn, dùng máy tính cầm tay để tính $2^9 \cdot \sqrt{3}$, bấm các phím như sau:

[2] [x■] [9] [▶] [×] [√■] [3] [SHIFT] [MODE] [6] [3] [=]

Kết quả nhận được có ba chữ số thập phân sau dấu phẩy là 886,810.

2. SAI SỐ TUYỆT ĐỐI VÀ SAI SỐ TƯƠNG ĐỐI

a. Sai số tuyệt đối

HĐ3. Trong HĐ2, Hoà dùng kính lúp để quan sát mực nước trên ống đo thứ hai được hình ảnh như Hình 5.2. Kí hiệu $\bar{a}$ (cm³) là số đo thể tích của nước.

Hình 5.2

Quan sát hình vẽ để so sánh $|13 - \bar{a}|$ và $|13,1 - \bar{a}|$ rồi cho biết trong hai số đo thể tích 13 cm³ và 13,1 cm³, số đo nào gần với thể tích của cốc nước hơn.

Giá trị $|a-\overline{a}|$ phản ánh mức độ sai lệch giữa số đúng $\overline{a}$ và số gần đúng a, được gọi là sai số tuyệt đối của số gần đúng a, kí hiệu là Δ_a, tức là:

$$\Delta_a = |a-\overline{a}|.$$

Chú ý

- Trên thực tế, nhiều khi ta không biết $\overline{a}$ nên cũng không biết Δ_a. Tuy nhiên, ta có thể đánh giá được Δ_a không vượt quá một số dương d nào đó.

  Chẳng hạn, trong HĐ3, ta thấy $|13,1-\overline{a}| < |13,1-13| = 0,1$ (cm³).

  Vậy với $a = 13,1$ (cm³), sai số tuyệt đối của a không vượt quá 0,1 cm³.

- Nếu $\Delta_a \le d$ thì $a-d \le \overline{a} \le a+d$, khi đó ta viết $\overline{a} = a \pm d$ và hiểu là số đúng $\overline{a}$ nằm trong đoạn $[a-d; a+d]$. Do d càng nhỏ thì a càng gần $\overline{a}$ nên d được gọi là độ chính xác của số gần đúng.

Ví dụ 2. Một công ty sử dụng dây chuyền A để đóng gạo vào bao với khối lượng mong muốn là 5 kg. Trên bao bì ghi thông tin khối lượng là 5 ± 0,2 kg. Gọi $\overline{a}$ là khối lượng thực của một bao gạo do dây chuyền A đóng gói.

a) Xác định số đúng, số gần đúng và độ chính xác.

b) Giá trị của $\overline{a}$ nằm trong đoạn nào?

Giải

a) Khối lượng thực của bao gạo $\overline{a}$ là số đúng. Tuy không biết $\overline{a}$ nhưng ta xem khối lượng bao gạo là 5 kg nên 5 là số gần đúng cho $\overline{a}$. Độ chính xác là $d = 0,2$ (kg).

b) Giá trị của $\overline{a}$ nằm trong đoạn [5 − 0,2; 5 + 0,2] hay [4,8; 5,2].

Luyện tập 2. Một phép đo đường kính nhân tế bào cho kết quả là 5 ± 0,3 µm. Đường kính thực của nhân tế bào thuộc đoạn nào?

Chú ý. Trong các phép đo, độ chính xác d của số gần đúng bằng một nửa đơn vị của thước đo. Chẳng hạn, một thước đo có chia vạch đến xentimét thì mọi giá trị đo nằm giữa 6,5 cm và 7,5 cm đều được coi là 7 cm. Vì vậy, thước đo có thang đo càng nhỏ thì cho giá trị đo càng chính xác.

b. Sai số tương đối

HĐ4. Công ty (trong Ví dụ 2) cũng sử dụng dây chuyền B để đóng gạo với khối lượng chính xác là 20 kg. Trên bao bì ghi thông tin khối lượng là 20 ± 0,5 kg.

Khẳng định "Dây chuyền A tốt hơn dây chuyền B" là đúng hay sai?

Mặc dù độ chính xác của khối lượng bao gạo đóng bằng dây chuyền A nhỏ hơn nhưng do bao gạo đóng bằng dây chuyền B nặng hơn nhiều nên ta không dựa vào sai số tuyệt đối mà dựa vào sai số tương đối để so sánh.

Sai số tương đối của số gần đúng a, kí hiệu là $\boldsymbol{\delta_a}$, là tỉ số giữa sai số tuyệt đối và $|a|$, tức là $\boldsymbol{\delta_a} = \dfrac{\boldsymbol{\Delta_a}}{|a|}$.

Nhận xét. Nếu $\bar{a} = a \pm d$ thì $\Delta_a \le d$, do đó $\delta_a \le \dfrac{d}{|a|}$. Nếu $\dfrac{d}{|a|}$ càng nhỏ thì chất lượng của phép đo hay tính toán càng cao. Người ta thường viết sai số tương đối dưới dạng phần trăm.

Ví dụ 3. Trong một cuộc điều tra dân số, người ta viết dân số của một tỉnh là:

3 574 625 người ± 50 000 người.

Hãy đánh giá sai số tương đối của số gần đúng này.

Giải

Ta có a = 3 574 625 người và d = 50 000 người, do đó sai số tương đối là:

$$\delta_a \le \frac{d}{|a|} = \frac{50\ 000}{3\ 574\ 625} \approx 1,4\%.$$

Luyện tập 3. Đánh giá sai số tương đối của khối lượng bao gạo được đóng gói theo hai dây chuyền A, B ở Ví dụ 2 và HĐ4. Dựa trên tiêu chí này, dây chuyền nào tốt hơn?

3. QUY TRÒN SỐ GẦN ĐÚNG

Trong thực tế đo đạc và tính toán, nhiều khi ta chỉ cần biết giá trị gần đúng của một đại lượng với độ chính xác nào đó (kể cả khi biết được giá trị đúng của nó). Khi đó, để cho gọn, các số thường được làm tròn (còn gọi là quy tròn).

Số thu được sau khi thực hiện làm tròn số được gọi là số quy tròn. Số quy tròn là một số gần đúng của số ban đầu.

- Đối với chữ số hàng làm tròn:
  - Giữ nguyên nếu chữ số ngay bên phải nó nhỏ hơn 5;
  - Tăng 1 đơn vị nếu chữ số ngay bên phải nó lớn hơn hoặc bằng 5.
- Đối với chữ số sau hàng làm tròn:
  - Bỏ đi nếu ở phần thập phân;
  - Thay bởi các chữ số 0 nếu ở phần số nguyên.

Ví dụ 4

a) Làm tròn số 2 395,3 đến hàng chục, số 18,693 đến hàng phần trăm và số đúng $d \in [5,5; 6,5)$ đến hàng đơn vị. Đánh giá sai số tuyệt đối của phép làm tròn số đúng d.

b) Cho số gần đúng a = 2,53 với độ chính xác d = 0,01. Số đúng $\bar{a}$ thuộc đoạn nào? Nếu làm tròn số a thì nên làm tròn đến hàng nào? Vì sao?

Giải

a) Số quy tròn của số 2 395,3 đến hàng chục là 2 360; số quy tròn của số 18,693 đến hàng phần trăm là 18,69. Mọi số đúng $d \in [5,5; 6,5)$ khi làm tròn đến hàng đơn vị đều thu được số quy tròn là 6 và sai số tuyệt đối $|d - 6| \le 0,5$.

b) Số đúng $\bar{a}$ thuộc đoạn [2,53 0,01; 2,53 + 0,01] hay [2,52; 2,54]. Khi làm tròn số gần đúng a ta nên làm tròn đến hàng phần chục do chữ số hàng phần trăm của a là chữ số không chắc chắn đúng.

Nhận xét

- Khi thay số đúng bởi số quy tròn đến một hàng nào đó thì sai số tuyệt đối của số quy tròn không vượt quá nửa đơn vị của hàng làm tròn.
- Cho số gần đúng a với độ chính xác d. Khi được yêu cầu làm tròn số a mà không nói rõ làm tròn đến hàng nào thì ta làm tròn số a đến hàng thấp nhất mà d nhỏ hơn 1 đơn vị của hàng đó.

Ví dụ 5. Cho số gần đúng $a = 581\,268$ với độ chính xác $d = 200$. Hãy viết số quy tròn của số a.

Giải

Vì độ chính xác đến hàng trăm ($d = 200$) nên ta làm tròn a đến hàng nghìn theo quy tắc làm tròn ở trên. Số quy tròn của a là 581 000.

Luyện tập 4. Hãy viết số quy tròn của số gần đúng trong những trường hợp sau:

a) $11\,251\,900 \pm 300$; b) $18{,}2857 \pm 0{,}01$.

Vận dụng. Các nhà vật lí sử dụng hai phương pháp khác nhau để đo tuổi của vũ trụ (đơn vị tỉ năm) lần lượt cho hai kết quả là: $13{,}807 \pm 0{,}026$ và $13{,}799 \pm 0{,}021$.

Hãy đánh giá sai số tương đối của mỗi phương pháp. Căn cứ trên tiêu chí này, phương pháp nào cho kết quả chính xác hơn?

BÀI TẬP

5.1. Trong các số sau, những số nào là số gần đúng?

a) Cân một túi gạo cho kết quả là 10,2 kg.

b) Bán kính Trái Đất là 6 371 km.

c) Trái Đất quay một vòng quanh Mặt Trời mất 365 ngày.

5.2. Giải thích kết quả "Đo độ cao của một ngọn núi cho kết quả là $1\,235 \pm 5$ m" và thực hiện làm tròn số gần đúng.

5.3. Sử dụng máy tính cầm tay tìm số gần đúng cho $\sqrt[3]{7}$ với độ chính xác 0,0005.

5.4. Các nhà vật lí sử dụng ba phương pháp đo hằng số Hubble lần lượt cho kết quả như sau:

$67{,}31 \pm 0{,}96$; $67{,}90 \pm 0{,}55$; $67{,}74 \pm 0{,}46$.

Phương pháp nào chính xác nhất tính theo sai số tương đối?

5.5. An và Bình cùng tính chu vi của hình tròn bán kính 2 cm với hai kết quả như sau:

Kết quả của An: $S_1 = 2\pi R \approx 2 \cdot 3{,}14 \cdot 2 = 12{,}56$ cm;

Kết quả của Bình: $S_2 = 2\pi R \approx 2 \cdot 3{,}1 \cdot 2 = 12{,}4$ cm.

Hỏi:

a) Hai giá trị tính được có phải là các số gần đúng không?

b) Giá trị nào chính xác hơn?

5.6. Làm tròn số 8 316,4 đến hàng chục và 9,754 đến hàng phần trăm rồi tính sai số tuyệt đối của số quy tròn.

Bài 13

CÁC SỐ ĐẶC TRƯNG ĐO XU THẾ TRUNG TÂM

THUẬT NGỮ
- Số trung bình
- Trung vị
- Tứ phân vị
- Mốt

KIẾN THỨC, KĨ NĂNG
- Lựa chọn và tính các số đặc trưng đo xu thế trung tâm của một mẫu số liệu: số trung bình, trung vị, tứ phân vị, mốt.
- Giải thích ý nghĩa, vai trò của các số đặc trưng trong mẫu số liệu thực tiễn.
- Rút ra kết luận từ ý nghĩa của các số đặc trưng đo xu thế trung tâm.

Hai phương pháp học tiếng Anh khác nhau được áp dụng cho hai lớp A và B có trình độ tiếng Anh tương đương nhau. Sau hai tháng, điểm khảo sát tiếng Anh (thang điểm 10) của hai lớp được cho như hình bên.

2	7	6	3	9
8	6	7	9	2
5	7	5	9	8
8	7	4	3	5
5	4	5	7	7

Lớp A

6	7	6	4	7
9	3	8	7	5
5	6	8	7	4
5	3	10	7	9
6	7	6	7	5

Lớp B

Quan sát hai mẫu số liệu trên, có thể đánh giá được phương pháp học tập nào hiệu quả hơn không?

Để làm được điều đó, người ta thường tính toán các số đặc trưng cho mỗi mẫu số liệu rồi so sánh.

Bài học này sẽ giới thiệu về các số đặc trưng đo xu thế trung tâm, tức là các số cho ta biết thông tin về vị trí trung tâm của mẫu số liệu và được dùng làm đại diện cho mẫu số liệu.

1. SỐ TRUNG BÌNH VÀ TRUNG VỊ

a. Số trung bình

Từ mẫu số liệu về điểm số của hai lớp A, B trên, em hãy:

HĐ1. Tính trung bình cộng điểm khảo sát tiếng Anh của mỗi lớp A và B.

HĐ2. Dựa trên điểm trung bình, hãy cho biết phương pháp học tập nào hiệu quả hơn.

Số trung bình (số trung bình cộng) của mẫu số liệu $x_1, x_2, \ldots, x_n$, kí hiệu là $\overline{x}$, được tính bằng công thức:

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

Chú ý. Trong trường hợp mẫu số liệu cho dưới dạng bảng tần số thì số trung bình được tính theo công thức:

$$\overline{x} = \frac{m_1x_1 + m_2x_2 + \ldots + m_kx_k}{n},$$

trong đó m_k là tần số của giá trị x_k và $n = m_1 + m_2 + \ldots + m_k$.

Ví dụ 1. Thống kê số cuốn sách mỗi bạn trong lớp đã đọc trong năm 2021, An thu được kết quả như bảng bên. Hỏi trong năm 2021, trung bình mỗi bạn trong lớp đọc bao nhiêu cuốn sách?

Số cuốn sách	1	2	3	4	5
Số bạn	3	5	15	10	7

Giải

Số bạn trong lớp là $n = 3 + 5 + 15 + 10 + 7 = 40$ (bạn).

Trong năm 2021, trung bình mỗi bạn trong lớp đọc số cuốn sách là:

$$\frac{3 \cdot 1 + 5 \cdot 2 + 15 \cdot 3 + 10 \cdot 4 + 7 \cdot 5}{40} = 3,325 \text{ (cuốn).}$$

Ý nghĩa. Số trung bình là giá trị trung bình cộng của các số trong mẫu số liệu, nó cho biết vị trí trung tâm của mẫu số liệu và có thể dùng để đại diện cho mẫu số liệu.

Luyện tập 1. Bảng sau cho biết thời gian chạy cự li 100 m của các bạn trong lớp (đơn vị giây):

Thời gian	12	13	14	15	16
Số bạn	5	7	10	8	6

Hãy tính thời gian chạy trung bình cự li 100 m của các bạn trong lớp.

b. Trung vị

HĐ3. Một công ty nhỏ gồm 1 giám đốc và 5 nhân viên, thu nhập mỗi tháng của giám đốc là 20 triệu đồng, của nhân viên là 4 triệu đồng.

a) Tính thu nhập trung bình của các thành viên trong công ty.

b) Thu nhập trung bình có phản ánh đúng thu nhập của nhân viên công ty không?

Trong trường hợp mẫu số liệu có giá trị bất thường (rất lớn hoặc rất bé so với đa số các giá trị khác), người ta không dùng số trung bình để đo xu thế trung tâm mà dùng trung vị.

Để tìm trung vị của một mẫu số liệu, ta thực hiện như sau:

- Sắp xếp các giá trị trong mẫu số liệu theo thứ tự không giảm.
- Nếu số giá trị của mẫu số liệu là số lẻ thì giá trị chính giữa của mẫu là trung vị. Nếu là số chẵn thì trung vị là trung bình cộng của hai giá trị chính giữa của mẫu.

Ví dụ 2. Hãy tìm trung vị cho mẫu số liệu về lương của giám đốc và nhân viên công ty được cho trong HĐ3.

Giải

Để tìm trung vị của mẫu số liệu trên, ta làm như sau:

• Sắp xếp số liệu theo thứ tự không giảm:

$$4 \qquad 4 \qquad \underbrace{4 \qquad 4}_{\textit{Hai giá trị chính giữa}} \qquad 4 \qquad 20.$$

• Dãy trên có hai giá trị chính giữa cùng bằng 4. Vậy trung vị của mẫu số liệu cũng bằng 4.

Trong mẫu số liệu được sắp xếp trên, số phần tử ở bên trái trung vị và số phần tử ở bên phải trung vị bằng nhau và bằng 3. Lương của giám đốc cao hơn hẳn số trung bình, đây chính là *giá trị bất thường*. Nếu ta thay lương của giám đốc là 30; 40; 50;... (triệu đồng) thì trung vị vẫn không thay đổi trong khi số trung bình sẽ thay đổi.

Ý nghĩa. Trung vị là giá trị chia đôi mẫu số liệu, nghĩa là trong mẫu số liệu được sắp xếp theo thứ tự không giảm thì giá trị trung vị ở vị trí chính giữa. Trung vị không bị ảnh hưởng bởi giá trị bất thường trong khi số trung bình bị ảnh hưởng bởi giá trị bất thường.

Luyện tập 2. Chiều dài (đơn vị feet) của 7 con cá voi trưởng thành được cho như sau:

48 53 51 31 53 112 52.

Tìm số trung bình và trung vị của mẫu số liệu trên. Trong hai số đó, số nào phù hợp hơn để đại diện cho chiều dài của 7 con cá voi trưởng thành này?

2. TỨ PHÂN VỊ

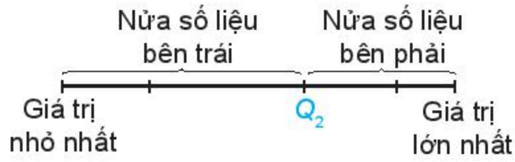
HĐ4. Điểm (thang điểm 100) của 12 thí sinh cao điểm nhất trong một cuộc thi như sau:

58 74 92 81 97 88 75 69 87 69 75 77.

Ban tổ chức muốn trao các giải Nhất, Nhì, Ba, Tư cho các thí sinh này, mỗi giải trao cho 25% số thí sinh (3 thí sinh).

Em hãy giúp ban tổ chức xác định các ngưỡng điểm để phân loại thí sinh.

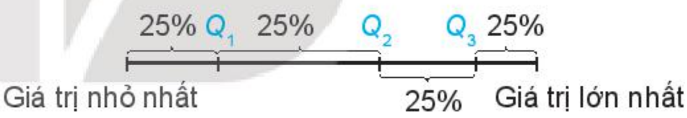
> Để tìm các tứ phân vị của mẫu số liệu có *n* giá trị, ta làm như sau:
> - Sắp xếp mẫu số liệu theo thứ tự không giảm.
> - Tìm trung vị. Giá trị này là Q_2.
> - Tìm trung vị của nửa số liệu bên trái Q_2 (không bao gồm Q_2 nếu *n* lẻ). Giá trị này là Q_1.
> - Tìm trung vị của nửa số liệu bên phải Q_2 (không bao gồm Q_2 nếu *n* lẻ). Giá trị này là Q_3.
>
> Q_1, Q_2, Q_3 được gọi là các tứ phân vị của mẫu số liệu.



Hình 5.3b

Chú ý. Q_1 được gọi là tứ phân vị thứ nhất hay tứ phân vị dưới, Q_3 được gọi là tứ phân vị thứ ba hay tứ phân vị trên.

Ý nghĩa. Các điểm Q_1, Q_2, Q_3 chia mẫu số liệu đã sắp xếp theo thứ tự từ nhỏ đến lớn thành bốn phần, mỗi phần đều chứa 25% giá trị (H.5.3a).



Hình 5.3a. Các tứ phân vị

Ví dụ 3. Hàm lượng Natri (đơn vị miligam, 1 mg = 0,001 g) trong 100 g một số loại ngũ cốc được cho như sau:

0	340	70	140	200	180	210	150	100	130
140	180	190	160	290	50	220	180	200	210.

Hãy tìm các tứ phân vị. Các tứ phân vị này cho ta thông tin gì?

Giải

- Sắp xếp các giá trị này theo thứ tự không giảm:

$$0\ 50\ 70\ 100\ 130\ 140\ 140\ 150\ 160\ \underbrace{180\ 180}_{\text{Hai giá trị chính giữa}}\ 180\ 190\ 200\ 200\ 210\ 210\ 220\ 290\ 340.$$

- Vì $n = 20$ là số chẵn nên Q_2 là trung bình cộng của hai giá trị chính giữa:

$$Q_2 = (180 + 180) : 2 = 180.$$

- Ta tìm Q_1 là trung vị của nửa số liệu bên trái Q_2:

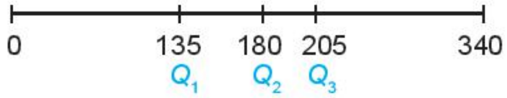
0 50 70 100 130 140 140 150 160 180

và tìm được $Q_1 = (130 + 140) : 2 = 135$.

- Ta tìm Q_3 là trung vị của nửa số liệu bên phải Q_2:

180 180 190 200 200 210 210 220 290 340

và tìm được $Q_3 = (200 + 210) : 2 = 205$.



Hình 5.4. Hình ảnh về sự phân bố của mẫu số liệu

Các tứ phân vị cho ta hình ảnh phân bố của mẫu số liệu. Khoảng cách từ Q_1 đến Q_2 là 45 trong khi khoảng cách từ Q_2 đến Q_3 là 25. Điều này cho thấy mẫu số liệu tập trung với mật độ cao ở bên phải của Q_2 và mật độ thấp ở bên trái của Q_2 (H.5.4).

Luyện tập 3. Bảng sau đây cho biết số lần học tiếng Anh trên Internet trong một tuần của một số học sinh lớp 10:

Số lần	0	1	2	3	4	5
Số học sinh	2	4	6	12	8	3

Hãy tìm các tứ phân vị cho mẫu số liệu này.

3. MỐT

HĐ5. Một cửa hàng giày thể thao đã thống kê cỡ giày của một số khách hàng nam được chọn ngẫu nhiên cho kết quả như sau:

38 39 39 38 40 41 39 39 38 39 39 39 40 39 39.

a) Tính cỡ giày trung bình. Số trung bình này có ý nghĩa gì với cửa hàng không?

b) Cửa hàng nên nhập cỡ giày nào với số lượng nhiều nhất?

> **Mốt** của mẫu số liệu là giá trị xuất hiện với tần số lớn nhất.

Ý nghĩa. Có thể dùng mốt để đo xu thế trung tâm của mẫu số liệu khi mẫu số liệu có nhiều giá trị trùng nhau.

Ví dụ 4. Thời gian truy cập Internet (đơn vị giờ) trong một ngày của một số học sinh lớp 10 được cho như sau:

0 0 1 1 1 3 4 4 5 6.

Tìm mốt cho mẫu số liệu này.

Giải

Vì số học sinh truy cập Internet 1 giờ mỗi ngày là lớn nhất (có 3 học sinh) nên mốt là 1.

Nhận xét

- Mốt có thể không là duy nhất. Chẳng hạn, với mẫu số liệu

  8 7 10 9 7 5 7 8 8

  các số 7; 8 đều xuất hiện với số lần lớn nhất (3 lần) nên mẫu số liệu này có hai mốt là 7 và 8.
- Khi các giá trị trong mẫu số liệu xuất hiện với tần số như nhau thì mẫu số liệu không có mốt.

- Mốt còn được định nghĩa cho mẫu dữ liệu định tính (dữ liệu không phải là số). Ví dụ báo Tuổi trẻ đã thực hiện thăm dò ý kiến của bạn đọc với câu hỏi "Theo bạn, VFF nên chọn huấn luyện viên ngoại hay nội dẫn dắt đội tuyển bóng đá nam Việt Nam?".

  Tại thời điểm 21 giờ ngày 27-4-2021 kết quả bình chọn như sau:

Lựa chọn	Huấn luyện viên nội	Huấn luyện viên ngoại	Ý kiến khác
Số lượt bình chọn	1 897	3 781	747

  Trong mẫu dữ liệu này, lựa chọn "huấn luyện viên ngoại" có nhiều người bình chọn nhất, được gọi là *mốt*.

Vận dụng. Hãy tính các số đặc trưng đo xu thế trung tâm cho mẫu số liệu về điểm khảo sát của lớp A và lớp B ở đầu bài học để phân tích và so sánh hiệu quả học tập của hai phương pháp này.

BÀI TẬP

5.7. Tìm số trung bình, trung vị, mốt và tứ phân vị của mỗi mẫu số liệu sau đây:

a) Số điểm mà năm vận động viên bóng rổ ghi được trong một trận đấu:

9 8 15 8 20.

b) Giá của một số loại giày (đơn vị nghìn đồng):

350 300 650 300 450 500 300 250.

c) Số kênh được chiếu của một số hãng truyền hình cáp:

36 38 33 34 32 30 34 35.

5.8. Hãy chọn số đặc trưng đo xu thế trung tâm của mỗi mẫu số liệu sau. Giải thích và tính giá trị của số đặc trưng đó.

a) Số mặt trăng đã biết của các hành tinh:

Hành tinh	Thuỷ tinh	Kim tinh	Trái Đất	Hoả tinh	Mộc tinh	Thổ tinh	Thiên Vương tinh	Hải Vương tinh
Số mặt trăng	0	0	1	2	63	34	27	13

(Theo *NASA*)

b) Số đường chuyền thành công trong một trận đấu của một số cầu thủ bóng đá:

32 24 20 14 23.

c) Chỉ số IQ của một nhóm học sinh: 60 72 63 83 68 74 90 86 74 80.

d) Các sai số trong một phép đo: 10 15 18 15 14 13 42 15 12 14 42.

5.9. Số lượng học sinh giỏi Quốc gia năm học 2018 – 2019 của 10 trường Trung học phổ thông được cho như sau:

$$0 \quad 0 \quad 4 \quad 0 \quad 0 \quad 0 \quad 10 \quad 0 \quad 6 \quad 0.$$

a) Tìm số trung bình, mốt, các tứ phân vị của mẫu số liệu trên.

b) Giải thích tại sao tứ phân vị thứ nhất và trung vị trùng nhau.

5.10. Bảng sau đây cho biết số chỗ ngồi của một số sân vận động được sử dụng trong Giải Bóng đá Vô địch Quốc gia Việt Nam năm 2018 (số liệu gần đúng).

Sân vận động	Cẩm Phả	Thiên Trường	Hàng Đẫy	Thanh Hoá	Mỹ Đình
Số chỗ ngồi	20 120	21 315	23 405	20 120	37 546

(Theo *vov.vn*)

Các giá trị số trung bình, trung vị, mốt bị ảnh hưởng thế nào nếu bỏ đi số liệu chỗ ngồi của Sân vận động Quốc gia Mỹ Đình?

Em có biết?

John Graunt (1620 – 1674) là một nhà buôn người Anh. Ông được xem là người đầu tiên đưa ra suy luận về tổng thể dựa trên thông tin của một phần (mẫu). Năm 1662, khi điều tra nhân khẩu, ông nhận ra rằng trung bình mỗi năm trong 11 gia đình có 3 người mất. Với giả thiết tỉ lệ này không đổi trong toàn bộ dân cư London và biết rằng trung bình trong một năm ở London có 13 000 người mất, ông đã ước lượng được số hộ gia đình ở London khoảng 48 000. Và với giả thiết trung bình mỗi gia đình có 8 người, ông ước lượng được dân số của London khoảng 384 000 người.

John Graunt (1620 – 1674)

Bài 14

CÁC SỐ ĐẶC TRƯNG ĐO ĐỘ PHÂN TÁN

THUẬT NGỮ
- Khoảng biến thiên
- Khoảng tứ phân vị
- Phương sai
- Độ lệch chuẩn

KIẾN THỨC, KĨ NĂNG
- Tính các số đặc trưng đo độ phân tán.
- Biết ý nghĩa của các số đặc trưng đo độ phân tán.
- Phát hiện các giá trị bất thường sử dụng các công cụ toán học.

Dưới đây là điểm trung bình môn học kì I của hai bạn An và Bình:

	Toán	Vật lí	Hoá học	Ngữ văn	Lịch sử	Địa lí	Tin học	Tiếng Anh
An	9,2	8,7	9,5	6,8	8,0	8,0	7,3	6,5
Bình	8,2	8,1	8,0	7,8	8,3	7,9	7,6	8,1

Điểm trung bình môn học kì của An và Bình đều là 8,0 nhưng rõ ràng Bình "học đều" hơn An. Có thể dùng những số đặc trưng nào để đo mức độ "học đều"?

Bài này sẽ giới thiệu một vài số đặc trưng như vậy.

1. KHOẢNG BIẾN THIÊN VÀ KHOẢNG TỨ PHÂN VỊ

HĐ1. Một cổ động viên của câu lạc bộ Everton, Anh đã thống kê điểm số mà hai câu lạc bộ Leicester City và Everton đạt được trong năm mùa giải Ngoại hạng Anh gần đây, từ mùa giải 2014 – 2015 đến mùa giải 2018 – 2019 như sau:

Leicester City: 41 81 44 47 52.

Everton: 47 47 61 49 54.

Cổ động viên đó cho rằng, Everton thi đấu ổn định hơn Leicester City. Em có đồng ý với nhận định này không? Vì sao?

Trong 5 mùa giải, điểm thấp nhất, cao nhất của Leicester City lần lượt là 41; 81 trong khi của Everton là 47; 61. Về trực quan, thành tích của Everton ổn định hơn Leicester City. Người ta có nhiều cách để đo sự ổn định này. Cách đơn giản nhất là dùng hiệu số (Điểm cao nhất – Điểm thấp nhất). Giá trị này được gọi là khoảng biến thiên.

Khoảng biến thiên, kí hiệu là R, là hiệu số giữa giá trị lớn nhất và giá trị nhỏ nhất trong mẫu số liệu.

Ý nghĩa. Khoảng biến thiên dùng để đo độ phân tán của mẫu số liệu. Khoảng biến thiên càng lớn thì mẫu số liệu càng phân tán.

Ví dụ 1. Điểm kiểm tra học kì môn Toán của các bạn Tổ 1, Tổ 2 lớp 10A được cho như sau:

Tổ 1: 7 8 8 9 8 8 8.

Tổ 2: 10 6 8 9 9 7 8 7 8.

a) Điểm kiểm tra trung bình của hai tổ có như nhau không?

b) Tính các khoảng biến thiên của hai mẫu số liệu. Căn cứ trên chỉ số này, các bạn tổ nào học đồng đều hơn?

Giải

a) Điểm kiểm tra trung bình của hai tổ đều bằng 8.

b) Đối với Tổ 1: Điểm kiểm tra thấp nhất, cao nhất tương ứng là 7; 9. Do đó khoảng biến thiên là: $R_1 = 9 - 7 = 2$.

Đối với Tổ 2: Điểm kiểm tra thấp nhất, cao nhất tương ứng là 6; 10. Do đó khoảng biến thiên là: $R_2 = 10 - 6 = 4$.

Do $R_2 > R_1$ nên ta nói các bạn Tổ 1 học đều hơn các bạn Tổ 2.

Luyện tập 1. Mẫu số liệu sau cho biết chiều cao (đơn vị cm) của các bạn trong tổ:

163 159 172 167 165 168 170 161.

Tính khoảng biến thiên của mẫu số liệu này.

Nhận xét. Sử dụng khoảng biến thiên có ưu điểm là đơn giản, dễ tính toán song khoảng biến thiên chỉ sử dụng thông tin của giá trị lớn nhất và giá trị nhỏ nhất mà bỏ qua thông tin từ tất cả các giá trị khác. Do đó, khoảng biến thiên rất dễ bị ảnh hưởng bởi các giá trị bất thường.

HĐ2. Trong một tuần, nhiệt độ cao nhất trong ngày (đơn vị °C) tại hai thành phố Hà Nội và Điện Biên được cho như sau:

Hà Nội: 23 25 28 28 32 33 35.

Điện Biên: 16 24 26 26 26 27 28.

a) Tính các khoảng biến thiên của mỗi mẫu số liệu và so sánh.

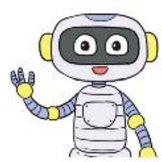
b) Em có nhận xét gì về sự ảnh hưởng của giá trị 16 đến khoảng biến thiên của mẫu số liệu về nhiệt độ cao nhất trong ngày tại Điện Biên?

c) Tính các tứ phân vị và hiệu $Q_3 - Q_1$ cho mỗi mẫu số liệu. Có thể dùng hiệu này để đo độ phân tán của mẫu số liệu không?

Khoảng tứ phân vị, kí hiệu là Δ_Q, là hiệu số giữa tứ phân vị thứ ba và tứ phân vị thứ nhất, tức là:

$$\Delta_Q = Q_3 - Q_1.$$

Về bản chất, khoảng tứ phân vị là khoảng biến thiên của 50% số liệu chính giữa của mẫu số liệu đã sắp xếp.



Ý nghĩa. Khoảng tứ phân vị cũng là một số đo độ phân tán của mẫu số liệu. Khoảng tứ phân vị càng lớn thì mẫu số liệu càng phân tán.

Chú ý. Một số tài liệu gọi khoảng biến thiên là *biên độ* và khoảng tứ phân vị là *độ trải giữa*.

Ví dụ 2. Mẫu số liệu sau cho biết số ghế trống tại một rạp chiếu phim trong 9 ngày:

7 8 22 20 15 18 19 13 11.

Tìm khoảng tứ phân vị cho mẫu số liệu này.

Giải

Trước hết, ta sắp xếp mẫu số liệu theo thứ tự không giảm:

7 8 11 13 15 18 19 20 22.

Mẫu số liệu gồm 9 giá trị nên trung vị là số ở vị trí chính giữa $Q_2 = 15$.

Nửa số liệu bên trái là 7, 8, 11, 13 gồm 4 giá trị, hai phần tử chính giữa là 8, 11.

Do đó, $Q_1 = (8 + 11) : 2 = 9{,}5$.

Nửa số liệu bên phải là 18, 19, 20, 22 gồm 4 giá trị, hai phần tử chính giữa là 19, 20.

Do đó, $Q_3 = (19 + 20) : 2 = 19{,}5$.

Vậy khoảng tứ phân vị cho mẫu số liệu là $\Delta_Q = 19{,}5 - 9{,}5 = 10$.

Luyện tập 2. Mẫu số liệu sau đây cho biết số bài hát ở mỗi album trong bộ sưu tập của An:

12 7 10 9 12 9 10 11 10 14.

Hãy tìm khoảng tứ phân vị cho mẫu số liệu này.

2. PHƯƠNG SAI VÀ ĐỘ LỆCH CHUẨN

Khoảng biến thiên chỉ sử dụng thông tin của giá trị lớn nhất và nhỏ nhất của mẫu số liệu (bỏ qua thông tin của tất cả các giá trị khác), còn khoảng tứ phân vị chỉ sử dụng thông tin của 50% số liệu chính giữa. Có một vài số đặc trưng khác đo độ phân tán sử dụng thông tin của tất cả các giá trị trong mẫu số liệu. Hai trong số đó là *phương sai* và *độ lệch chuẩn*.

Cụ thể là với mẫu số liệu $x_1, x_2, \ldots, x_n$, nếu gọi số trung bình là $\overline{x}$ thì với mỗi giá trị x_i, độ lệch của nó so với giá trị trung bình là $x_i - \overline{x}$.

- Phương sai là giá trị $s^2 = \dfrac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n}$.
- Căn bậc hai của phương sai, $s = \sqrt{s^2}$, được gọi là độ lệch chuẩn.

Chú ý. Người ta còn sử dụng đại lượng để đo độ phân tán của mẫu số liệu:

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

Ý nghĩa. Nếu số liệu càng phân tán thì phương sai và độ lệch chuẩn càng lớn.

Ví dụ 3. Mẫu số liệu sau đây cho biết sĩ số của 5 lớp khối 10 tại một trường Trung học:

43 45 46 41 40.

Tìm phương sai và độ lệch chuẩn cho mẫu số liệu này.

Giải

Số trung bình của mẫu số liệu là: $\overline{x} = \dfrac{43 + 45 + 46 + 41 + 40}{5} = 43$.

Ta có bảng sau:

Giá trị	Độ lệch	Bình phương độ lệch
43	43 – 43 = 0	0
45	45 – 43 = 2	4
46	46 – 43 = 3	9
41	41 – 43 = –2	4
40	40 – 43 = –3	9
Tổng		26

Bạn có thể sử dụng máy tính cầm tay, phần mềm bảng tính hay phần mềm thống kê để tính các số đặc trưng.

Mẫu số liệu gồm 5 giá trị nên $n = 5$. Do đó phương sai là: $s^2 = \dfrac{26}{5} = 5{,}2$.

Độ lệch chuẩn là: $s = \sqrt{5{,}2} \approx 2{,}28$.

Luyện tập 3. Dùng đồng hồ đo thời gian có độ chia nhỏ nhất đến 0,001 giây để đo 7 lần thời gian rơi tự do của một vật bắt đầu từ điểm A ($v_A = 0$) đến điểm B. Kết quả đo như sau:

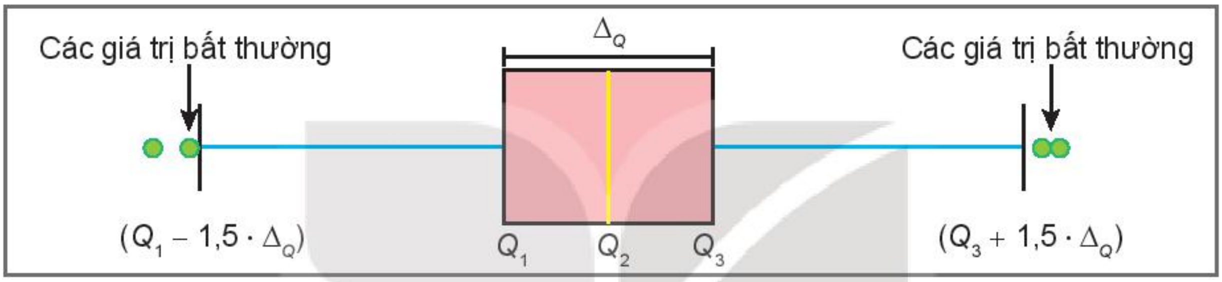
0,398 0,399 0,408 0,410 0,406 0,405 0,402.

(Theo *Bài tập Vật lí 10, Nhà xuất bản Giáo dục Việt Nam, 2018*)

Hãy tính phương sai và độ lệch chuẩn cho mẫu số liệu này. Qua các đại lượng này, em có nhận xét gì về độ chính xác của phép đo trên?

3. PHÁT HIỆN SỐ LIỆU BẤT THƯỜNG HOẶC KHÔNG CHÍNH XÁC BẰNG BIỂU ĐỒ HỘP

Trong mẫu số liệu thống kê, có khi gặp những giá trị quá lớn hoặc quá nhỏ so với đa số các giá trị khác. Những giá trị này được gọi là giá trị bất thường. Chúng xuất hiện trong mẫu số liệu có thể do nhầm lẫn hay sai sót nào đó. Ta có thể dùng biểu đồ hộp để phát hiện những giá trị bất thường này.



Hình 5.5. Biểu đồ hộp

Các giá trị lớn hơn $Q_3 + 1,5 \cdot \Delta_Q$ hoặc bé hơn $Q_1 - 1,5 \cdot \Delta_Q$ được xem là giá trị bất thường.

Ví dụ 4. Hàm lượng Natri (đơn vị mg) trong 100 g một số loại ngũ cốc được cho như sau:

0	340	70	140	200	180	210	150	100	130
140	180	190	160	290	50	220	180	200	210.

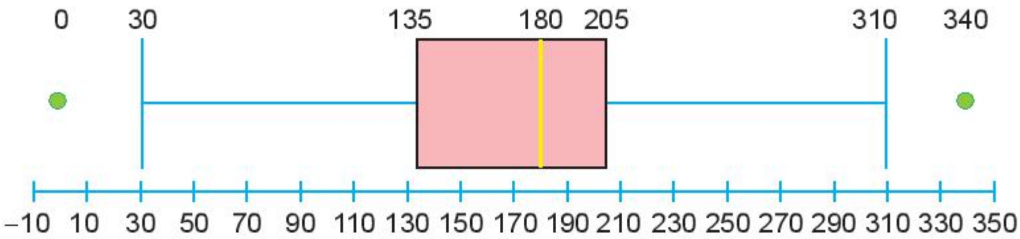
Tìm giá trị bất thường trong mẫu số liệu trên bằng cách sử dụng biểu đồ hộp.

Giải

Từ mẫu số liệu ta tính được $Q_1 = 135$ và $Q_3 = 205$. Do đó, khoảng tứ phân vị là:

$$\Delta_Q = 205 - 135 = 70.$$

Biểu đồ hộp cho mẫu số liệu này là:



Ta có $Q_1 - 1,5 \cdot \Delta_Q = 30$ và $Q_3 + 1,5 \cdot \Delta_Q = 310$ nên trong mẫu số liệu có hai giá trị được xem là bất thường là 340 mg (lớn hơn 310 mg) và 0 mg (bé hơn 30 mg).

Luyện tập 4. Một mẫu số liệu có tứ phân vị thứ nhất là 56 và tứ phân vị thứ ba là 84. Hãy kiểm tra xem trong hai giá trị 10 và 100 giá trị nào được xem là giá trị bất thường.

BÀI TẬP

5.11. Mỗi khẳng định sau đúng hay sai?

(1) Nếu các giá trị của mẫu số liệu càng tập trung quanh giá trị trung bình thì độ lệch chuẩn càng lớn.

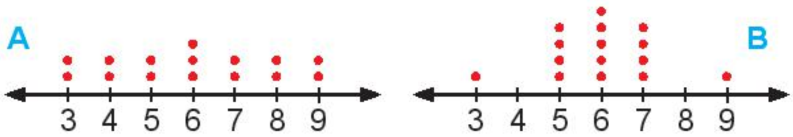
(2) Khoảng biến thiên chỉ sử dụng thông tin của giá trị lớn nhất và bé nhất, bỏ qua thông tin của các giá trị còn lại.

(3) Khoảng tứ phân vị có sử dụng thông tin của giá trị lớn nhất, giá trị bé nhất.

(4) Khoảng tứ phân vị chính là khoảng biến thiên của nửa dưới mẫu số liệu đã sắp xếp.

(5) Các số đo độ phân tán đều không âm.

5.12. Cho hai biểu đồ chấm điểm biểu diễn hai mẫu số liệu A, B như sau:



Không tính toán, hãy cho biết:

a) Hai mẫu số liệu này có cùng khoảng biến thiên và số trung bình không?

b) Mẫu số liệu nào có phương sai lớn hơn?

5.13. Cho mẫu số liệu gồm 10 số dương không hoàn toàn giống nhau. Các số đo độ phân tán (khoảng biến thiên, khoảng tứ phân vị, độ lệch chuẩn) sẽ thay đổi như thế nào nếu:

a) Nhân mỗi giá trị của mẫu số liệu với 2.

b) Cộng mỗi giá trị của mẫu số liệu với 2.

5.14. Từ mẫu số liệu về thuế thuốc lá của 51 thành phố tại một quốc gia, người ta tính được:

Giá trị nhỏ nhất bằng 2,5; $Q_1 = 36$; $Q_2 = 60$; $Q_3 = 100$; giá trị lớn nhất bằng 205.

a) Tỉ lệ thành phố có thuế thuốc lá lớn hơn 36 là bao nhiêu?

b) Chỉ ra hai giá trị sao cho có 50% giá trị của mẫu số liệu nằm giữa hai giá trị này.

c) Tìm khoảng tứ phân vị của mẫu số liệu.

5.15. Mẫu số liệu sau đây cho biết cân nặng của 10 trẻ sơ sinh (đơn vị kg):

2,977	3,155	3,920	3,412	4,236
2,593	3,270	3,813	4,042	3,387.

Hãy tính khoảng biến thiên, khoảng tứ phân vị và độ lệch chuẩn cho mẫu số liệu này.

5.16. Tỉ lệ thất nghiệp ở một số quốc gia vào năm 2007 (đơn vị %) được cho như sau:

7,8	3,2	7,7	8,7	8,6	8,4	7,2	3,6
5,0	4,4	6,7	7,0	4,5	6,0	5,4.	

Hãy tìm các giá trị bất thường (nếu có) của mẫu số liệu trên.

BÀI TẬP CUỐI CHƯƠNG V

A – TRẮC NGHIỆM

5.17. Khi cân một bao gạo bằng một cân treo với thang chia 0,2 kg thì độ chính xác *d* là

A. 0,1 kg. B. 0,2 kg. C. 0,3 kg. D. 0,4 kg.

5.18. Trong hai mẫu số liệu, mẫu nào có phương sai lớn hơn thì có độ lệch chuẩn lớn hơn, đúng hay sai?

A. Đúng. B. Sai.

5.19. Có 25% giá trị của mẫu số liệu nằm giữa Q_1 và Q_3, đúng hay sai?

A. Đúng. B. Sai.

5.20. Số đặc trưng nào sau đây đo độ phân tán của mẫu số liệu?

A. Số trung bình. B. Mốt. C. Trung vị. D. Độ lệch chuẩn.

5.21. Điểm trung bình môn học kì I một số môn học của bạn An là 8; 9; 7; 6; 5; 7; 3. Nếu An được cộng thêm mỗi môn 0,5 điểm chuyên cần thì các số đặc trưng nào sau đây của mẫu số liệu không thay đổi?

A. Số trung bình. B. Trung vị. C. Độ lệch chuẩn. D. Tứ phân vị.

B – TỰ LUẬN

5.22. Lương khởi điểm của 5 sinh viên vừa tốt nghiệp tại một trường đại học (đơn vị triệu đồng) là:

3,5 9,2 9,2 9,5 10,5.

a) Giải thích tại sao nên dùng trung vị để thể hiện mức lương khởi điểm của sinh viên tốt nghiệp từ trường đại học này.

b) Nên dùng khoảng biến thiên hay khoảng tứ phân vị để đo độ phân tán? Vì sao?

5.23. Điểm Toán và điểm Tiếng Anh của 11 học sinh lớp 10 được cho trong bảng sau:

Học sinh	A	B	C	D	E	F	G	H	I	J	K
Toán	62	91	43	31	57	63	80	37	43	5	78
Tiếng Anh	65	57	55	37	62	70	73	49	65	41	64

Hãy so sánh mức độ học đều của học sinh trong môn Tiếng Anh và môn Toán thông qua các số đặc trưng: khoảng biến thiên, khoảng tứ phân vị, độ lệch chuẩn.

5.24. Bảng sau cho biết dân số của các tỉnh/thành phố Đồng bằng Bắc Bộ năm 2018 (đơn vị triệu người).

Tỉnh/thành phố	Dân số
Hà Nội	7,52
Vĩnh Phúc	1,09
Bắc Ninh	1,25
Quảng Ninh	1,27
Hải Dương	1,81
Hải Phòng	2,01

Tỉnh/thành phố	Dân số
Hưng Yên	1,19
Thái Bình	1,79
Hà Nam	0,81
Nam Định	1,85
Ninh Bình	0,97

(Theo *Tổng cục Thống kê*)

a) Tìm số trung bình và trung vị của mẫu số liệu trên.

b) Giải thích tại sao số trung bình và trung vị lại có sự sai khác nhiều.

c) Nên sử dụng số trung bình hay trung vị để đại diện cho dân số của các tỉnh thuộc Đồng bằng Bắc Bộ?

5.25. Hai mẫu số liệu sau đây cho biết số lượng trường Trung học phổ thông ở mỗi tỉnh/thành phố thuộc Đồng bằng sông Hồng và Đồng bằng sông Cửu Long năm 2017:

Đồng bằng sông Hồng: 187 34 35 46 54 57 37 39 23 57 27.

Đồng bằng sông Cửu Long: 33 34 33 29 24 39 42 24 23 19 24 15 26.

(Theo *Tổng cục Thống kê*)

a) Tính số trung bình, trung vị, các tứ phân vị, mốt, khoảng biến thiên, khoảng tứ phân vị, độ lệch chuẩn cho mỗi mẫu số liệu trên.

b) Tại sao số trung bình của hai mẫu số liệu có sự sai khác nhiều trong khi trung vị thì không?

c) Tại sao khoảng biến thiên và độ lệch chuẩn của hai mẫu số liệu khác nhau nhiều trong khi khoảng tứ phân vị thì không?

5.26. Tỉ lệ trẻ em suy dinh dưỡng (tính theo cân nặng ứng với độ tuổi) của 10 tỉnh thuộc Đồng bằng sông Hồng được cho như sau:

5,5 13,8 10,2 12,2 11,0 7,4 11,4 13,1 12,5 13,4.

(Theo *Tổng cục Thống kê*)

a) Tính số trung bình, trung vị, khoảng biến thiên và độ lệch chuẩn của mẫu số liệu trên.

b) Thực hiện làm tròn đến hàng đơn vị cho các giá trị trong mẫu số liệu. Sai số tuyệt đối của phép làm tròn này không vượt quá bao nhiêu?