§ 3 CÁC SỐ ĐẶC TRƯNG CỦA MẪU SỐ LIỆU

Để nhanh chóng nắm bắt được những thông tin quan trọng chứa đựng trong mẫu số liệu, ta đưa ra một vài chỉ số gọi là ***các số đặc trưng của mẫu số liệu***.

1. Số trung bình

• Giả sử ta có một mẫu số liệu kích thước N là $x_1, x_2, ..., x_N$. Ở lớp dưới, ta đã biết *số trung bình* (hay số trung bình cộng) của mẫu số liệu này, kí hiệu là $\overline{x}$, được tính bởi công thức

$$\overline{x} = \frac{x_1 + x_2 + ... + x_N}{N}. \qquad (1)$$

Để cho gọn, ta kí hiệu tổng $x_1 + x_2 + ... + x_N$ là $\sum_{i=1}^{N} x_i$ và đọc là "tổng của các x_i với i chạy từ 1 đến N". Với kí hiệu này, công thức (1) được viết gọn là

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

• Giả sử mẫu số liệu được cho dưới dạng một bảng phân bố tần số (bảng 7) :

Giá trị	$x_1\ x_2\ \ ...\ \ x_m$	
Tần số	$n_1\ n_2\ \ ...\ \ n_m$	N

Bảng 7

Khi đó, công thức tính số trung bình (1) trở thành

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + ... + n_mx_m}{N} = \frac{1}{N}\sum_{i=1}^{m} n_ix_i, \qquad (2)$$

trong đó n_i là tần số của số liệu x_i, $(i = 1, 2, ..., m)$, $\sum_{i=1}^{m} n_i = N$.

• Giả sử mẫu số liệu kích thước N được cho dưới dạng bảng tần số ghép lớp. Các số liệu được chia thành m lớp ứng với m đoạn (bảng 7*a*) hoặc m lớp ứng với

m nửa khoảng (bảng 7*b*). Ta gọi trung điểm x_i của đoạn (hay nửa khoảng) ứng với lớp thứ i là ***giá trị đại diện*** của lớp đó.

Lớp	Giá trị đại diện	Tần số
$[a_1 ; a_2]$	x_1	n_1
$[a_3 ; a_4]$	x_2	n_2
$\vdots$	$\vdots$	$\vdots$
$[a_{2m-1} ; a_{2m}]$	x_m	n_m
		$N = \sum_{i=1}^{m} n_i$

Bảng 7a

Lớp	Giá trị đại diện	Tần số
$[a_1 ; a_2)$	x_1	n_1
$[a_2 ; a_3)$	x_2	n_2
$\vdots$	$\vdots$	$\vdots$
$[a_m ; a_{m+1})$	x_m	n_m
		$N = \sum_{i=1}^{m} n_i$

Bảng 7b

Khi đó, số trung bình của mẫu số liệu này được tính xấp xỉ theo công thức

$$\overline{x} \approx \frac{1}{N}\sum_{i=1}^{m} n_i x_i .$$

Ví dụ 1. Một nhà thực vật học đo chiều dài của 74 lá cây và thu được bảng tần số sau (đơn vị : mm) :

Lớp	Giá trị đại diện	Tần số
[5,45 ; 5,85)	5,65	5
[5,85 ; 6,25)	6,05	9
[6,25 ; 6,65)	6,45	15
[6,65 ; 7,05)	6,85	19
[7,05 ; 7,45)	7,25	16
[7,45 ; 7,85)	7,65	8
[7,85 ; 8,25)	8,05	2
		$N = 74$

Khi đó, chiều dài trung bình của 74 lá này xấp xỉ là

$$\overline{x} \approx \frac{5\,.\,5{,}65 + 9\,.\,6{,}05 + ... + 8\,.\,7{,}65 + 2\,.\,8{,}05}{74} \approx 6{,}80 \text{ (mm)}.$$ □

Ý nghĩa của số trung bình

Số trung bình của mẫu số liệu được dùng làm đại diện cho các số liệu của mẫu. Nó là một số đặc trưng quan trọng của mẫu số liệu.

Chẳng hạn, nếu biết rằng thời gian trung bình để điều trị khỏi bệnh *A* đối với bệnh nhân nam là 5,3 ngày, đối với bệnh nhân nữ là 6,2 ngày thì ta có thể cho rằng nói chung với bệnh *A* thì bệnh nhân nam chóng bình phục hơn so với bệnh nhân nữ.

Tuy nhiên, khi các số liệu trong mẫu có sự chênh lệch rất lớn đối với nhau thì số trung bình chưa đại diện tốt cho các số liệu trong mẫu.

Ví dụ 2. Một nhóm 11 học sinh tham gia một kì thi. Số điểm thi của 11 học sinh đó được sắp xếp từ thấp đến cao như sau (thang điểm 100) :

0 ; 0 ; 63 ; 65 ; 69 ; 70 ; 72 ; 78 ; 81 ; 85 ; 89.

Số trung bình là

$$\frac{0 + 0 + 63 + ... + 85 + 89}{11} \approx 61{,}09.$$

Quan sát dãy điểm trên, ta thấy hầu hết các em (9 em) trong nhóm có số điểm vượt số trung bình. Như vậy, số trung bình này không phản ánh đúng trình độ trung bình của nhóm. Trong trường hợp này, có một số đặc trưng khác thích hợp hơn đó là *số trung vị*. □

2. Số trung vị

Giả sử ta có một mẫu số liệu kích thước N được sắp xếp theo thứ tự không giảm. Nếu N là một số lẻ thì số liệu đứng thứ $\frac{N+1}{2}$ *(số liệu đứng chính giữa) gọi là* ***số trung vị****. Nếu N là một số chẵn, ta lấy trung bình cộng của hai số liệu đứng thứ* $\frac{N}{2}$ và $\frac{N}{2}+1$ *làm số trung vị.*

Số trung vị được kí hiệu là M_e.

Ví dụ 3. Điều tra về số học sinh trong 28 lớp học, ta được mẫu số liệu sau (sắp xếp theo thứ tự tăng dần) :

38	39	39	40	40	40	40	40	40	41	41	41	42	42
43	43	43	43	44	44	44	44	44	45	45	46	47	47

Số liệu đứng thứ 14 là 42, đứng thứ 15 là 43. Do vậy, số trung vị là

$$M_e = \frac{42+43}{2} = 42,5.$$ □

H1

a) Tính số trung vị của mẫu số liệu trong ví dụ 2.

b) Tính số trung bình của mẫu số liệu trong ví dụ 3 và so sánh nó với số trung vị.

CHÚ Ý

Khi các số liệu trong mẫu không có sự chênh lệch quá lớn thì số trung bình và số trung vị xấp xỉ nhau.

H2 *Đo chiều cao của 36 học sinh của một trường, ta có mẫu số liệu sau, sắp xếp theo thứ tự tăng (đơn vị : cm) :*

160	161	161	162	162	162	163	163	163	164	164	164
164	165	165	165	165	165	166	166	166	166	167	167
168	168	168	168	169	169	170	171	171	172	172	174

Tìm số trung vị của mẫu số liệu này.

3. Mốt

Cho một mẫu số liệu dưới dạng bảng phân bố tần số. Ta đã biết giá trị có tần số lớn nhất được gọi là ***mốt*** của mẫu số liệu và kí hiệu là M_o.

Ví dụ 4. Một cửa hàng bán quần áo thống kê số áo sơ mi nam đã bán ra trong một quý theo các cỡ khác nhau và có được bảng tần số sau

Cỡ áo (x)	36	37	38	39	40	41	42
Số áo bán được (n)	13	45	110	184	126	40	5

Điều mà cửa hàng quan tâm nhất là *cỡ áo nào được khách hàng mua nhiều nhất*. Bảng thống kê trên cho thấy cỡ áo mà khách hàng mua nhiều nhất là cỡ 39 (tức là giá trị 39 có tần số lớn nhất). Vậy giá trị 39 là mốt của mẫu số liệu này. □

CHÚ Ý

Một mẫu số liệu có thể có một hay nhiều mốt.

Ví dụ 5. Một cửa hàng bán 6 loại quạt với giá tiền là 100, 150, 300, 350, 400, 500 (đơn vị là nghìn đồng). Số quạt cửa hàng bán ra trong mùa hè vừa qua được thống kê trong bảng tần số sau

Giá tiền (x)	100	150	300	350	400	500
Số quạt bán được (n)	256	353	534	300	534	175

Ta thấy mẫu số liệu trên có hai mốt là 300 nghìn đồng và 400 nghìn đồng. Đó là giá tiền của hai loại quạt được khách hàng mua nhiều nhất. □

4. Phương sai và độ lệch chuẩn

Ví dụ 6. Điểm trung bình từng môn học của hai học sinh An và Bình trong năm học vừa qua được cho trong bảng sau :

Môn	**Điểm của An**	**Điểm của Bình**
Toán	8	8,5
Vật lí	7,5	9,5
Hoá học	7,8	9,5
Sinh học	8,3	8,5
Ngữ văn	7	5
Lịch sử	8	5,5
Địa lí	8,2	6
Tiếng Anh	9	9
Thể dục	8	9
Công nghệ	8,3	8,5
Giáo dục công dân	9	10

H3 *Tính điểm trung bình (không kể hệ số) của tất cả các môn học của An và của Bình. Theo em, bạn nào học khá hơn ?*

Nhìn vào bảng điểm, ta có ngay nhận xét là An học đều các môn, còn Bình thì không. Sự chênh lệch, biến động giữa các điểm của An thì ít, của Bình thì nhiều.

• Để đo mức độ chênh lệch giữa các giá trị của mẫu số liệu so với số trung bình, người ta đưa ra hai số đặc trưng là *phương sai* và *độ lệch chuẩn*.

Giả sử ta có một mẫu số liệu kích thước N là $\{x_1, ..., x_N\}$. ***Phương sai*** *của mẫu số liệu này, kí hiệu là s^2, được tính bởi công thức sau*

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2, \qquad (3)$$

trong đó $\overline{x}$ là số trung bình của mẫu số liệu.
Căn bậc hai của phương sai được gọi là ***độ lệch chuẩn****, kí hiệu là s.*

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

Ý nghĩa của phương sai và độ lệch chuẩn

Trong công thức (3), ta thấy phương sai là trung bình cộng của bình phương khoảng cách từ mỗi số liệu tới số trung bình. Như vậy, *phương sai và độ lệch chuẩn đo mức độ phân tán của các số liệu trong mẫu quanh số trung bình. Phương sai và độ lệch chuẩn càng lớn thì độ phân tán càng lớn.*

CHÚ Ý

Có thể biến đổi công thức (3) thành

$$\boxed{s^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N}x_i\right)^2.} \qquad (4)$$

Sử dụng công thức (4) thuận tiện hơn trong tính toán.

Trở lại ví dụ ở trên, ta hãy tính phương sai và độ lệch chuẩn điểm các môn học của An và Bình. Trước hết, ta tính các tổng $\sum_{i=1}^{N}x_i$ và $\sum_{i=1}^{N}x_i^2$.

Từ số liệu ở cột điểm của An, ta có

$$\sum_{i=1}^{11}x_i = 89{,}1 \ ; \ \sum_{i=1}^{11}x_i^2 = 725{,}11.$$

Từ số liệu ở cột điểm của Bình, ta có

$$\sum_{i=1}^{11}x_i = 89 \ ; \ \sum_{i=1}^{11}x_i^2 = 750{,}5.$$

Tiếp theo, ta thế các kết quả này vào công thức (4) để tìm s^2.

Phương sai và độ lệch chuẩn điểm các môn học của An là

$$s_A^2 = \frac{725,11}{11} - \left(\frac{89,1}{11}\right)^2 \approx 0,309 \ ; \quad s_A \approx \sqrt{0,3091} \approx 0,556.$$

Phương sai và độ lệch chuẩn điểm các môn học của Bình là

$$s_B^2 = \frac{750,5}{11} - \left(\frac{89}{11}\right)^2 \approx 2,764 \ ; \quad s_B \approx \sqrt{2,764} \approx 1,663.$$

Ta thấy mức độ "học lệch" của Bình so với An được thể hiện qua việc so sánh hai phương sai : Phương sai điểm các môn học của Bình gấp gần 9 ($\approx$ 8,945) lần phương sai điểm các môn học của An. Điều đó phù hợp với nhận xét Bình học lệch hơn An.

Ta cũng có thể so sánh độ học lệch của Bình và An thông qua việc so sánh hai độ lệch chuẩn. □

Việc tính các tổng $\sum_{i=1}^{N} x_i$ và $\sum_{i=1}^{N} x_i^2$ sẽ nhanh chóng nếu dùng máy tính bỏ túi. (Xem bài đọc thêm để được hướng dẫn chi tiết về cách sử dụng máy tính bỏ túi trong tính toán thống kê).

• Nếu số liệu được cho dưới dạng bảng phân bố tần số (bảng 7) thì phương sai được tính bởi công thức

$$\boxed{s^2 = \frac{1}{N}\sum_{i=1}^{m} n_i x_i^2 - \frac{1}{N^2}\left(\sum_{i=1}^{m} n_i x_i\right)^2.} \qquad (5)$$

Ví dụ 7. Sản lượng lúa (đơn vị là tạ) của 40 thửa ruộng thí nghiệm có cùng diện tích được trình bày trong bảng tần số sau đây.

Sản lượng (x)	20	21	22	23	24	
Tần số (n)	5	8	11	10	6	$N = 40$

a) Tìm sản lượng trung bình của 40 thửa ruộng.

b) Tính phương sai và độ lệch chuẩn.

Giải. Trước hết, ta tính

$$\sum_{i=1}^{5} n_i x_i = 884, \ \sum_{i=1}^{5} n_i x_i^2 = 19598.$$

a) Sản lượng trung bình của 40 thửa ruộng là

$$\overline{x} = \frac{884}{40} = 22,1 \text{ (tạ)}.$$

b) Theo công thức (5), ta có phương sai là

$$s^2 = \frac{19\,598}{40} - \left(\frac{884}{40}\right)^2 = 1,54.$$

Độ lệch chuẩn là $s = \sqrt{1,54} \approx 1,24$ (tạ). □

• Giả sử mẫu số liệu được cho dưới dạng bảng phân bố tần số ghép lớp. Các số liệu được chia thành m lớp ứng với m đoạn (hoặc nửa khoảng). Gọi x_i là giá trị đại diện của lớp thứ i (xem bảng 7*a*, 7*b*).

Khi đó, phương sai của mẫu số liệu này có thể tính xấp xỉ theo công thức (5).

Ví dụ 8. Tính phương sai và độ lệch chuẩn của mẫu số liệu cho ở ví dụ 1.

Giải. Ta có

$$\sum_{i=1}^{7} n_i x_i = 502,9,$$

$$\sum_{i=1}^{7} n_i x_i^2 = 3443,385.$$

Vậy $s^2 \approx \dfrac{3443,385}{74} - \dfrac{502,9^2}{74^2} \approx 0,347.$

Độ lệch chuẩn là $s \approx \sqrt{0,347} \approx 0,589$ (mm). □

Câu hỏi và bài tập

Trong các bài tập dưới đây, yêu cầu tính số trung bình, số trung vị, phương sai, độ lệch chuẩn (chính xác đến hàng phần trăm).

9. Có 100 học sinh tham dự kì thi học sinh giỏi Toán (thang điểm là 20). Kết quả được cho trong bảng sau đây.

Điểm	9	10	11	12	13	14	15	16	17	18	19	
Tần số	1	1	3	5	8	13	19	24	14	10	2	$N = 100$

a) Tính số trung bình.

b) Tính số trung vị và mốt. Nêu ý nghĩa của chúng.

c) Tính phương sai và độ lệch chuẩn.

10. Người ta chia 179 củ khoai tây thành chín lớp căn cứ trên khối lượng của chúng (đơn vị là gam). Ta có bảng phân bố tần số ghép lớp sau đây.

Lớp	Tần số
[10 ; 19]	1
[20 ; 29]	14
[30 ; 39]	21
[40 ; 49]	73
[50 ; 59]	42
[60 ; 69]	13
[70 ; 79]	9
[80 ; 89]	4
[90 ; 99]	2
	$N = 179$

Tính khối lượng trung bình của một củ khoai tây. Tìm phương sai và độ lệch chuẩn.

11. Bảng sau đây trích từ sổ theo dõi bán hàng của một cửa hàng bán xe máy.

Số xe bán trong ngày	0	1	2	3	4	5
Tần số	2	13	15	12	7	3

a) Tìm số xe trung bình bán được trong một ngày.

b) Tìm phương sai và độ lệch chuẩn.

Luyện tập

Trong các bài tập dưới đây, yêu cầu tính số trung bình, số trung vị, phương sai, độ lệch chuẩn chính xác đến hàng phần trăm.

12. Số liệu sau đây cho ta lãi (quy tròn) hàng tháng của một cửa hàng trong năm 2005. Đơn vị là triệu đồng.

Tháng	1	2	3	4	5	6	7	8	9	10	11	12
Lãi	12	15	18	13	13	16	18	14	15	17	20	17

a) Tìm số trung bình, số trung vị.

b) Tìm phương sai và độ lệch chuẩn.

13. Một cửa hàng vật liệu xây dựng thống kê số bao xi măng bán ra trong 23 ngày cuối năm 2005. Kết quả như sau :

47 54 43 50 61 36 65 54 50 43 62 59 36 45 45 33 53 67
21 45 50 36 58.

a) Tìm số trung bình, số trung vị.

b) Tìm phương sai và độ lệch chuẩn.

14. Số lượng khách đến tham quan một điểm du lịch trong mỗi tháng được thống kê trong bảng sau đây.

Tháng	1	2	3	4	5	6	7	8	9	10	11	12
Số khách	430	560	450	550	760	430	525	110	635	450	800	950

a) Tìm số trung bình, số trung vị.

b) Tìm phương sai và độ lệch chuẩn.

15. Trên hai con đường A và B, trạm kiểm soát đã ghi lại tốc độ (km/h) của 30 chiếc ô tô trên mỗi con đường như sau :

Con đường A : 60 65 70 68 62 75 80 83 82 69 73 75 85 72 67
88 90 85 72 63 75 76 85 84 70 61 60 65 73 76.

Con đường B : 76 64 58 82 72 70 68 75 63 67 74 70 79
80 73 75 71 68 72 73 79 80 63 62 71 70
74 69 60 63.

a) Tìm số trung bình, số trung vị, phương sai và độ lệch chuẩn của tốc độ ô tô trên mỗi con đường A, B.

b) Theo em thì xe chạy trên con đường nào an toàn hơn ?

Bài đọc thêm

SỬ DỤNG MÁY TÍNH BỎ TÚI TRONG THỐNG KÊ

Máy tính bỏ túi (MTBT) là công cụ hỗ trợ rất đắc lực cho việc học Thống kê. Nhờ MTBT, Thống kê đã trở nên dễ học và dễ ứng dụng.

Chẳng hạn, đối với máy CASIO *fx - 500MS*, để tính số trung bình, phương sai và độ lệch chuẩn, chúng ta cần làm trình tự theo các bước sau :

1) Đầu tiên, để vào chế độ tính toán thống kê, ta ấn

$\boxed{\text{MODE}}$ $\boxed{2}$

2) Giả sử mẫu số liệu là $x_1, x_2, ..., x_n$. Để nhập số liệu, ta ấn

x_1 $\boxed{\text{DT}}$ x_2 $\boxed{\text{DT}}$ ... x_n $\boxed{\text{DT}}$

Để nhập mẫu số liệu $x_1, x_2, ..., x_n$, trong đó x_i có tần số n_i, $(i = 1, 2, ..., m)$, ta ấn

x_1 [SHIFT] [;] n_1 [DT] x_2 [SHIFT] [;] n_2 [DT] ...

x_m [SHIFT] [;] n_m [DT]

3) Nhập dữ liệu xong, để tính số trung bình $\overline{x}$, ta ấn

[SHIFT] [S – VAR] [1] [=]

4) Để tính độ lệch chuẩn s, ta ấn

[SHIFT] [S – VAR] [2] [=]

5) Để tính phương sai s^2 (lấy bình phương của độ lệch chuẩn), ta ấn tiếp

[x^2] [=]

Ví dụ 1. Tính số trung bình, phương sai, độ lệch chuẩn điểm các môn học của An ở ví dụ 6, §3.

Sau khi thực hiện bước 1, để nhập dữ liệu, ta ấn

2) 8 [DT] 7,5 [DT] ... 9 [DT]

3) Để tính trung bình $\overline{x}$, ta ấn

[SHIFT] [S – VAR] [1] [=]

Kết quả $\overline{x} = 8,1$, đó là số trung bình cần tìm.

4) Để tính độ lệch chuẩn s, ta ấn

[SHIFT] [S – VAR] [2] [=]

Kết quả $s \approx 0,555959449$, đó là độ lệch chuẩn cần tìm.

5) Để tính phương sai s^2, ta ấn tiếp

[x^2] [=]

Kết quả $s^2 \approx 0,309090909$, đó là phương sai cần tìm. □

Ví dụ 2. Tính số trung bình, phương sai, độ lệch chuẩn của mẫu số liệu trong ví dụ 7, §3.

Sau khi thực hiện bước 1, để nhập dữ liệu, ta ấn

2) 20 [SHIFT] [;] 5 [DT] 21 [SHIFT] [;] 8 [DT] ... 24 [SHIFT] [;] 6 [DT]

3) Để tính trung bình $\overline{x}$, ta ấn

[SHIFT] [S – VAR] [1] [=]

Kết quả $\overline{x} = 22,1$, đó là số trung bình cần tìm.

4) Để tính độ lệch chuẩn s, ta ấn

$$\boxed{\text{SHIFT}}\ \boxed{\text{S – VAR}}\ \boxed{2}\ \boxed{=}$$

Kết quả $s \approx 1{,}240967365$, đó là độ lệch chuẩn cần tìm.

5) Để tính phương sai s^2, ta ấn tiếp

$$\boxed{x^2}\ \boxed{=}$$

Kết quả $s^2 \approx 1{,}54$, đó là phương sai cần tìm. □