Chương 3

TỔNG THỂ VÀ MẪU

1. TỔNG THỂ VÀ MẪU

1.1 Tổng thể

Khi nghiên cứu về một vấn đề người ta thường khảo sát trên một dấu hiệu nào đó, các dấu hiệu này thể hiện trên nhiều phần tử. Tập hợp các phần tử mang dấu hiệu được gọi là *tổng thể* hay *đám đông* (*population*).

• **Ví dụ 1** *Nghiên cứu tập hợp gà trong một trại chăn nuôi ta quan tâm đến dấu hiệu trọng lượng. Nghiên cứu chất lượng học tập của sinh viên trong một trường đại học ta quan tâm đến dấu hiệu điểm.*

⊙ **Chú ý** Trong phần này ta sử dụng một số khái niệm và kí hiệu sau:

1. N: số phần tử của tổng thể, được gọi là kích thước của tổng thể.
2. X^*: dấu hiệu mà ta khảo sát.
3. $x_i\,(i = \overline{1,k})$: giá trị của dấu hiệu X^* đo được trên phần tử của tổng thể (x_i là thông tin mà ta quan tâm, còn các phần tử của tổng thể là vật mang thông tin).
4. $N_i\,(i = \overline{1,k})$: tần số của x_i (số phần tử có chung giá trị x_i).
5. $p_i = \frac{N_i}{N}$: tần suất của x_i.

⊙ **Bảng cơ cấu của tổng thể**

Sự tương ứng giữa các giá trị x_i và tần suất p_i được biểu diễn bởi bảng cơ cấu tổng thể theo dấu hiệu X^* như sau:

Giá trị của X^*	x_1	x_2	$\dots$	x_k
Tần suất p_i	p_1	p_2	$\dots$	p_k

• **Các đặc trưng của tổng thể**

1. Trung bình của dấu hiệu X^* (trung bình của tổng thể) $m = \sum_{i=1}^{k} x_i p_i$.

2. Phương sai của dấu hiệu X^* (phương sai của tổng thể) $\sigma^2 = \sum_{i=1}^{k} (x_i - m)^2 p_i$.

3. Độ lệch tiêu chuẩn của dấu hiệu X^* (độ lệch tiêu chuẩn của tổng thể)

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{k} (x_i - m)^2 p_i}$$

1.2 Mẫu

- Từ tổng thể lấy ra n phần tử và đo lường dấu hiệu X^* trên chúng. Khi đó n phần tử này lập nên một mẫu (*sample*). Số phần tử của mẫu được gọi là *kích thước của mẫu.*

- Vì từ mẫu suy ra kết luận cho tổng thể nên mẫu phải đại diện cho tổng thể và phải được chọn một cách khách quan.

- Việc lấy mẫu được tiến hành theo hai phương thức: lấy mẫu có hoàn lại và lấy mẫu không hoàn lại.

2. MÔ HÌNH XÁC SUẤT CỦA TỔNG THỂ VÀ MẪU

2.1 Đại lượng ngẫu nhiên gốc và phân phối gốc

Lấy tùy ý từ tổng thể ra một phần tử. Gọi X là giá trị của X^* đo được trên phần tử lấy ra thì X là đại lượng ngẫu nhiên có phân phối xác suất

X	x_1	x_2	...	x_i	...	x_k
P	p_1	p_2	...	p_i	...	p_k

Ta thấy dấu hiệu X^* được mô hình hóa bởi đại lượng ngẫu nhiên X. Khi đó X được gọi là đại lượng ngẫu nhiên gốc và phân phối xác suất của X được gọi là phân phối gốc.

2.2 Các tham số của đại lượng ngẫu nhiên gốc

$$E(X) = \sum_{i=1}^{k} x_i p_i.$$

$$Var(X) = \sum_{i=1}^{k} [x_i - E(X)]^2 p_i$$

2.3 Mẫu ngẫu nhiên

Lấy n phần tử của tổng thể theo phương pháp hoàn lại để quan sát. Gọi X_i là giá trị của X^* đo được trên phần tử thứ i ($i = \overline{1,n}$) thì $X_1, X_2, \ldots, X_n$ là các đại lượng ngẫu nhiên độc lập có cùng phân phối như X. Khi đó bộ $(X_1, X_2, \ldots, X_n)$ được gọi là một *mẫu ngẫu nhiên* kích thước n được tạo nên từ đại lượng ngẫu nhiên gốc X. Kí hiệu $W_X = (X_1, X_2, \ldots, X_n)$.

Giả sử X_i nhận giá trị x_i ($i = \overline{1,n}$). Khi đó $(x_1, x_2, \ldots, x_n)$ là một giá trị cụ thể của mẫu ngẫu nhiên W_X, được gọi là *mẫu cụ thể*. Kí hiệu $w_x = (x_1, x_2, \ldots, x_n)$.

• **Ví dụ 2** *Kết quả điểm môn Toán của một lớp gồm 100 sinh viên cho bởi bảng sau*

Điểm	3	4	5	6	7
Số sinh viên có điểm tương ứng	25	20	40	10	5

Gọi X là điểm môn Toán của một sinh viên được chọn ngẫu nhiên trong danh sách lớp thì X là đại lượng ngẫu nhiên có phân phối

X	3	4	5	6	7
P	0,25	0,2	0,4	0,1	0,05

Chọn ngẫu nhiên 5 sinh viên trong danh sách lớp để xem điểm. Gọi X_i là điểm của sinh viên thứ i. Ta có mẫu ngẫu nhiên kích thước $n = 5$ được xây dựng từ đại lượng ngẫu nhiên X

$$W_X = (X_1, X_2, \ldots, X_n)$$

Giả sử sinh viên thứ nhất được 4 điểm, thứ hai được 3 điểm, thứ ba được 6 điểm thứ tư được 7 điểm và thứ năm được 5 điểm. Ta được mẫu cụ thể

$$w_x = (4, 3, 6, 7, 5)$$

3. THỐNG KÊ

Trong thống kê (*statistics*), việc tổng hợp mẫu $W_X = (X_1, X_2, \ldots, X_n)$ được thực hiện dưới dạng hàm $G = f(X_1, X_2, \ldots, X_n)$ của các đại lượng ngẫu nhiên $X_1, X_2, \ldots, X_n$. Khi đó G được gọi là một thống kê.

3.1 Trung bình mẫu ngẫu nhiên

□ **Định nghĩa 1** *Trung bình của mẫu ngẫu nhiên* $W_X = (X_1, X_2, \ldots, X_n)$ *là một thống kê, kí hiệu* $\overline{X}$, *được xác định bởi*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.1}$$

⊙ **Chú ý**

i) Vì $X_1, X_2, \ldots, X_n$ là các đại lượng ngẫu nhiên nên $\overline{X}$ cũng là đại lượng ngẫu nhiên.

ii) Nếu mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ có mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$ thì $\overline{X}$ sẽ nhận giá trị $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ và $\overline{x}$ được gọi là trung bình của mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$.

◇ **Tính chất**

Nếu đại lượng ngẫu nhiên gốc X có kỳ vọng $E(X) = m$ và phương sai $Var(X) = \sigma^2$ thì $E(\overline{X}) = m$ và $Var(X) = \dfrac{\sigma^2}{n}$.

⊙ **Phân phối xác suất của $\overline{X}$**

i) Nếu $X \in B(n, p)$ thì $\overline{X} \in B(n, p)$.

ii) Nếu $X \in \mathcal{P}(a)$ thì $\overline{X} \in \mathcal{P}(a)$.

iii) Nếu $X \in N(\mu, \sigma^2)$ thì $\overline{X} \in N(\mu, \frac{\sigma^2}{n})$.

iv) Nếu $X \in \chi^2(n)$ thì $\overline{X} \in \chi^2(n)$.

3.2 Phương sai của mẫu ngẫu nhiên

□ **Định nghĩa 2** *Phương sai của mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ là một thống kê, kí hiệu S^2, được xác định bởi*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

trong đó $\overline{X}$ là trung bình của mẫu ngẫu nhiên.

⊙ **Chú ý**

i) Vì $X_1, X_2, \ldots, X_n$ là các đại lượng ngẫu nhiên nên S^2 cũng là đại lượng ngẫu nhiên.

ii) Nếu mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ có mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$ thì S^2 nhận giá trị $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Khi đó s^2 được gọi là phương sai của mẫu cụ thể.

◇ **Tính chất** Nếu $Var(X) = \sigma^2$ thì $E(S^2) = \dfrac{n-1}{n}\sigma^2$.

⊙ **Phương sai điều chỉnh**

Đặt $S'^2 = \dfrac{n}{n-1}S^2$ thì ta có $E(S'^2) = \sigma^2$.

S'^2 được gọi là *phương sai điều chỉnh* của mẫu ngẫu nhiên W_X.

Với mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$ thì S'^2 sẽ nhận giá trị

$$s'^2 = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

s'^2 được gọi là *phương sai điều chỉnh* của mẫu cụ thể.

⊙ **Phân phối xác suất**

Giả sử $W_X = (X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên được xây dựng từ đại lượng ngẫu nhiên X có phân phối chuẩn với $E(X) = m$ và $Var(X) = \sigma^2$. Khi đó

i) $\dfrac{nS^2}{\sigma^2} = \sum_{i=1}^{n}\dfrac{(X_i - \overline{X})^2}{\sigma^2} \in \chi^2(n-1)$.

ii) $\sum_{i=1}^{n}\dfrac{(X_i - m)^2}{\sigma^2} \in \chi^2(n)$

3.3 Độ lệch tiêu chuẩn và độ lệch tiêu chuẩn điều chỉnh

i) Độ lệch tiêu chuẩn của mẫu ngẫu nhiên W_X là $S = \sqrt{S^2}$.

Độ lệch tiêu chuẩn của mẫu cụ thể w_x là $s = \sqrt{s^2}$, trong đó s là giá trị của S.

ii) Độ lệch tiêu chuẩn điều chỉnh của mẫu ngẫu nhiên W_X là $S' = \sqrt{S'^2}$.

Độ lệch tiêu chuẩn điều chỉnh của mẫu cụ thể w_x là $s' = \sqrt{s'^2}$, trong đó s' là giá trị của S'.

4. SẮP XẾP SỐ LIỆU

Quá trình nghiên cứu thống kê thường trãi qua 2 khâu: thu thập các số liệu liên quan đến việc nghiên cứu và xử lý số liệu. Để việc xử lý được thuận lợi ta cần phải sắp xếp lại số liệu.

4.1 Trường hợp mẫu có kích thước nhỏ

Giả sử mẫu có kích thước n và đại lượng ngẫu nhiên gốc X nhận các giá trị có thể x_i $(i = \overline{1,k})$ với số lần lặp lại (tần số) n_i $(i = \overline{1,k})$. Ta thường lập bảng như sau:

x_i	n_i
x_i	n_1
x_2	n_2
$\ldots$	$\ldots$
x_k	n_k

Chú ý $\sum_{i=1}^{k} n_i = n$.

• **Ví dụ 3** *Tiến hành thu thập dữ liệu số trẻ ở lứa tuổi đến trường của 30 gia đình ở một huyện ta được kết quả cho bởi bảng*

0	3	0	0	3	0
2	2	0	1	2	1
0	0	1	2	4	0
4	2	1	0	1	0
0	2	0	1	3	2

Sắp xếp số liệu lại ta có bảng sau

Số trẻ ở lứa tuổi đến trường	n_i
0	12
1	6
2	7
3	3
4	2

4.2 Trường hợp mẫu có kích thước lớn

Ta chia mẫu thành các khoảng (lớp), trong mỗi khoảng ta chọn một giá trị đại diện. Người ta thường chia thành các khoảng đều nhau (có thể khoảng đầu hoặc cuối có độ dài khác với độ dài của các khoảng còn lại) và chọn giá trị đại diện là giá trị trung tâm của khoảng. Ta qui ước đầu mút bên phải của mỗi khoảng thuộc khoảng đó mà không thuộc khoảng tiếp theo khi tính tần số của mỗi khoảng.

• **Ví dụ 4** *Chiều cao của 400 cây sao được chia thành các khoảng được xếp trong bảng sau:*

Khoảng chiều cao	*Tần số n_i*	*Độ dài của khoảng*
5,5 − 8,5	*18*	*3*
8,5 − 12,5	*58*	*4*
12,5 − 16,5	*62*	*4*
16,5 − 20,5	*72*	*4*
20,5 − 24,5	*57*	*4*
24,5 − 28,5	*42*	*4*
28,5 − 32,5	*36*	*4*
32,5 − 36,5	*10*	*4*

5. BẢNG TÍNH $\overline{x}$, s^2

5.1 Tính trực tiếp

Ta dùng công thức

$$\boxed{\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i \\ s^2 &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - (\overline{x})^2 \end{aligned}} \tag{3.2}$$

trong đó x_i $(i = \overline{1,k})$ là các giá trị của X^*.

• **Ví dụ 5** *Số xe hơi bán được trung bình trong một tuần ở mỗi đại lý trong 45 đại lý cho bởi*

Số xe hơi được bán trong tuần / đại lý	n_i
1	15
2	12
3	9
4	5
5	3
6	1

Ta lập bảng tính như sau

x_i	n_i	$n_i x_i$	$n_i x_i^2$
1	15	15	15
2	12	24	48
3	9	27	81
4	5	20	80
5	3	15	75
6	1	6	36
$\sum$	$n = 45$	107	335

Ta có

$$\overline{x} = \tfrac{107}{45} = 2,38$$
$$s^2 = \tfrac{335}{45} - (2,38)^2 = 7,444 - 5,664 = 1,78.$$

• **Ví dụ 6** *Theo dõi 336 trường hợp tàu cập cảng, người ta thấy khoảng thời gian ngắn nhất giữa hai lần tàu vào cảng liên tiếp là 4 giờ, thời gian dài nhất là 80 giờ.*

Vì số liệu nhiều nên ta sắp xếp thành lớp có độ dài 8 và thay mỗi lớp bởi giá trị trung tâm $x_i^0 = \dfrac{x_{min} + x_{max}}{2}$.

Ta có bảng tính sau

$x_i - x_{i+1}$	x_i^0	n_i	$n_i x_i^0$	$n_i {x_i^0}^2$
4 – 12	8	143	1144	9152
12 – 20	16	75	1200	19200
20 – 28	24	53	1272	30528
28 – 36	32	27	864	27648
36 – 44	40	14	560	22400
44 – 52	48	9	432	20736
52 – 60	56	5	280	15680
60 – 68	64	4	256	16384
68 – 76	72	3	216	15552
76 – 80	78	3	234	18252
$\sum$		336	6458	195532

Ta có

$\overline{x} = \frac{6458}{336} = 19,22$

$s^2 = \frac{195532}{336} - (19,22)^2 = 212,532.$

5.2 Tính theo phương pháp đổi biến

Ta dùng phương pháp này khi x_i hoặc giá trị trung tâm x_i^0 của khoảng khá lớn.

Đặt $\boxed{u_i = \dfrac{x_i - x_0}{h}}$

trong đó x_i là giá trị của dấu hiệu X^*; x_0 và h là những giá trị tùy ý.

Ta thường chọn x_0 là giá trị x_i (hoặc x_i^0) ứng với tần số lớn nhất và h là độ dài của khoảng.

Khi đó

$$\boxed{\begin{aligned} \overline{x} &= x_0 + h\overline{u} \\ s^2 &= h^2\left[\frac{1}{n}\sum_{i=1}^{k} n_i u_i^2 - (\overline{u})^2\right] \end{aligned}}$$

• **Ví dụ 7** *Tính $\overline{x}$ và s^2 từ số liệu cho ở bảng của ví dụ trước.*

Ta chọn

$x_0 = 8$ (ứng với tần số $n_i = 143$ lớn nhất)
$h = 8$ (độ dài của lớp)

$x_i - x_{i+1}$	x_i^0	n_i	u_i	$n_i u_i$	$n_i u_i^2$
4 – 12	8	143	0	0	0
12 – 20	16	75	1	75	75
20 – 28	24	53	2	106	212
28 – 36	32	27	3	81	243
36 – 44	40	14	4	56	224
44 – 52	48	9	5	45	225
52 – 60	56	5	6	30	180
60 – 68	64	4	7	28	196
68 – 76	72	3	8	24	192
76 – 80	78	3	8,75	26,25	229,6875
$\sum$		336		471,25	1176,6875

Áp dụng công thức ta có

$\overline{x} = 8.\frac{471,25}{336} + 8 = 19,22$

$s^2 = 8^2.[\dfrac{1776,6875}{336} - (\dfrac{471,25}{336})^2] = 212,5229$

6. BÀI TẬP

1. Chiều cao của 40 sinh viên nam ở một trường đại học cho bởi bảng dưới đây. Hãy sắp xếp các số liệu trên thành bảng bằng cách chia số liệu thành các khoảng thích hợp.

52	68	60	48	55	45	59	61
57	64	54	55	49	58	60	66
70	48	52	73	67	51	62	69
56	73	53	57	51	61	54	59
66	57	49	64	60	70	73	67

2. Theo dõi năng suất của 100 hecta lúa ở một vùng, người ta thu được kết quả cho ở bảng sau:

Năng suất *(tạ/ha)*	Diện tích *(ha)*
30 − 35	7
35 − 40	12
40 − 45	18
45 − 50	27
50 − 55	20
55 − 60	8
60 − 65	5
65 − 70	3

Tính giá trị trung bình, phương sai và phương sai điều chỉnh của mẫu cụ thể này.

3. Quan sát về thời gian cần thiết để sản xuất một chi tiết máy ta thu được các số liệu cho ở bảng sau:

Khoảng thời gian *(phút)*	Số quan sát
20 − 25	2
25 − 30	14
30 − 35	26
35 − 40	32
40 − 45	14
45 − 50	8
50 − 55	4

Tính giá trị trung bình, phương sai và phương sai điều chỉnh của mẫu.

4. Thống kê số hàng bán được trong một ngày và số ngày bán được số lượng hàng tương ứng, ta có bảng số liệu sau:

Lượng hàng bán trong 1 ngày kg	Số ngày (n_i)
100 − 200	5
200 − 250	12
250 − 300	56
300 − 350	107
350 − 400	75
400 − 450	70
450 − 500	35
500 − 550	30
550 − 700	10

Tính giá trị trung bình mẫu và nêu ý nghĩa của nó.

◙ TRẢ LỜI BÀI TẬP

2. $\overline{x} = 47,5$ tạ/ha, $s^2 = 68,5$, $s'^2 = 69,192$.

3. $\overline{x} = 36,6$ phút, $s^2 = 44,69$, $s'^2 = 45,14$.

4. $\overline{x} = 375,3kg$