Chương III : LÝ THUYẾT CHỌN MẪU

3.1 LÝ THUYẾT CHỌN MẪU

3.2 BIỂU ĐỒ TẦN SỐ VÀ BIỂU ĐỒ TẦN SUẤT CỦA MẪU THỐNG KÊ

3.3 HÀM PHÂN PHỐI THỰC NGHIỆM

3.4 CÁC THAM SỐ ĐẶC TRƯNG CỦA MẪU

3.1. LÝ THUYẾT CHỌN MẪU

3.1.1 Mở đầu

Muốn nghiên cứu về chiều cao của người Việt Nam. Phương pháp chính xác nhất là đo chiều cao của tất cả mọi người, ghi lại số liệu từ đó có thể tính được chiều cao trung bình, độ phân tán, … Tuy nhiên, trên thực tế ta không thể làm được điều đó vì số liệu quá lớn.

Thống kê học đề nghị một phương pháp là quan sát ngẫu nhiên một số trường hợp gọi là mẫu, và trên cơ sở số liệu quan sát này ta suy rộng ra cho tổng thể.

Ví dụ 1:

Chúng ta cần điều tra tình hình sử dụng điện năng của một khu phố A trong một tháng. Khi đó chúng ta có thể khảo sát ngẫu nhiên 30 gia đình trong khu phố đó. Ở đây ta nhắc lại một số khái niệm sau:

- Tập hợp tất cả các gia đình trong khu phố A là tập hợp tổng thể.

- Tập hợp 30 gia đình chúng ta khảo sát gọi là tập hợp mẫu điều tra. Số 30 gọi là kích thước mẫu.

- Số kw điện mà mỗi gia đình sử dụng gọi là dấu hiệu điều tra (dấu hiệu thống kê).

3.1.2. Các phương pháp chọn mẫu

* ***Phương pháp chọn mẫu có lặp:*** Phần tử vừa quan sát được trả lại cho tổng thể trước khi quan sát lần sau.

* ***Phương pháp chọn mẫu không lặp:*** Phần tử vừa quan sát không trả lại cho tổng thể trước khi quan sát lần sau.

** Phương pháp chọn mẫu phân loại:*

Phương pháp này được sử dụng khi tập hợp tổng thể có số lượng lớn và cấu trúc không đồng đều về dấu hiệu thống kê.

Khi đó ta chia tổng thể thành các tập hợp nhỏ hơn (mỗi tập hợp con có các phần tử khá đồng đều) có số phần tử là: N_1, N_2,…,N_m. Và trong mỗi tập hợp con ta áp dụng một trong hai phương pháp chọn mẫu ở trên.

3.1.3. Mẫu ngẫu nhiên và mẫu thực nghiệm

Gọi X là biến ngẫu nhiên biểu thị đặc trưng cần nghiên cứu của tổng thể U.

Ví dụ: Để tìm hiểu về chiều dài của các chi tiết máy do một nhà máy sản xuất ra, ta gọi X là biến ngẫu nhiên chỉ chiều dài của các chi tiết máy đó.

Để cho việc xét các định lí toán học sau này được thuận lợi ta qui ước các mẫu được chọn theo phương pháp chọn mẫu có lặp. Việc chọn mỗi phần tử từ tổng thể U xem như thực hiện một phép thử

Lặp lại n lần phép thử T, gọi X_i là giá trị của X nhận được ở phép thử thứ i (i = 1, 2, …, n) thì $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên có cùng luật phân phối với X.

Giả sử sau khi thực hiện phép thử T n lần ta có giá trị của X là: $(x_1, x_2, \ldots, x_n)$. Khi đó $X_1 = x_1$, $X_2 = x_2$, …, $X_n = x_n$. Khi đó:

Bộ $(X_1, X_2, \ldots, X_n)$ gọi là mẫu ngẫu nhiên, bộ $(x_1, x_2, \ldots, x_n)$ là một giá trị của mẫu ngẫu nhiên gọi là mẫu thực nghiệm hay mẫu cụ thể.

3.1.4 Các phương pháp sắp xếp mẫu thực nghiệm

1. Sắp xếp thành một bộ số tăng dần hoặc giảm dần: $(x_1; x_2; \ldots; x_n)$, trong đó: $x_1 \leq x_2 \leq \ldots. \leq x_n$ hay $x_1 \geq x_2 \geq \ldots \geq x_n$.

2. Sắp xếp thành bảng phân phối tần số, tần suất:

a. Bảng không chia lớp:

Giả sử điều tra một mẫu kích thước n thu được bộ giá trị $(x_1,\ x_2; \ldots ;x_m)$. Khi đó ta có hai bảng phân bố sau:

Bảng phân phối tần số:

x_i	x_1	x_2	…	x_m
n_i	n_1	n_2	…	n_m

Trong đó:

+ n_i là tần số của giá trị x_i;

+ $x_1 \le x_2 \le \ldots. \le x_m$

+ $n_1 + n_2 + \ldots + n_m = n$

Bảng phân phối tần suất:

Tần suất:

$$f_i = \frac{n_i}{n}, \; i = 1, 2, \ldots, m$$

x_i	x_1	x_2	…	x_m
f_i	f_1	f_2	…	f_m

Trong đó: $f_1 + f_2 + \ldots + f_n = 1$

Ví dụ 2:

Điều tra số kW điện của 30 hộ gia đình ở một khu phố A ta thu được các số liệu như sau:

165	85	65	65	70	50	45	100	45	100
100	100	100	90	50	70	150	40	50	150
40	70	85	50	75	75	165	45	65	75

Bảng phân phối tần số:

x_i	40	45	50	65	70	75	85	90	100	150	165
n_i	2	3	4	3	3	3	2	1	5	2	2

Bảng phân phối tần suất:

x_i	40	45	50	65	70	75	85	90	100	150	165
f_i	$\frac{1}{15}$	$\frac{1}{10}$	$\frac{2}{15}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{15}$	$\frac{1}{30}$	$\frac{1}{6}$	$\frac{1}{15}$	$\frac{1}{15}$

b. Bảng chia lớp:

$[a_{i-1};a_i)$	$[a_0;a_1)$	$[a_1;a_2)$	…	$[a_{m-1};a_m)$
n_i	n_1	n_2	…	n_m

Trong đó: $n_1 + n_2 + \ldots + n_m = n$

và các lớp ghép thoả mãn:

$$a_1 - a_0 = a_2 - a_1 = \ldots = a_m - a_{m-1}$$

- Khi chia lớp ta phải căn cứ vào mục đích nghiên cứu và đặc tính của các số liệu thống kê; số lượng các lớp phải vừa đủ, thông thường số lớp thoả mãn công thức sau là vừa đủ.

$$m \approx \log_2 n + 1; m \in N$$

- Khi chia lớp ta thường lấy *giá trị đại diện cho lớp* là:

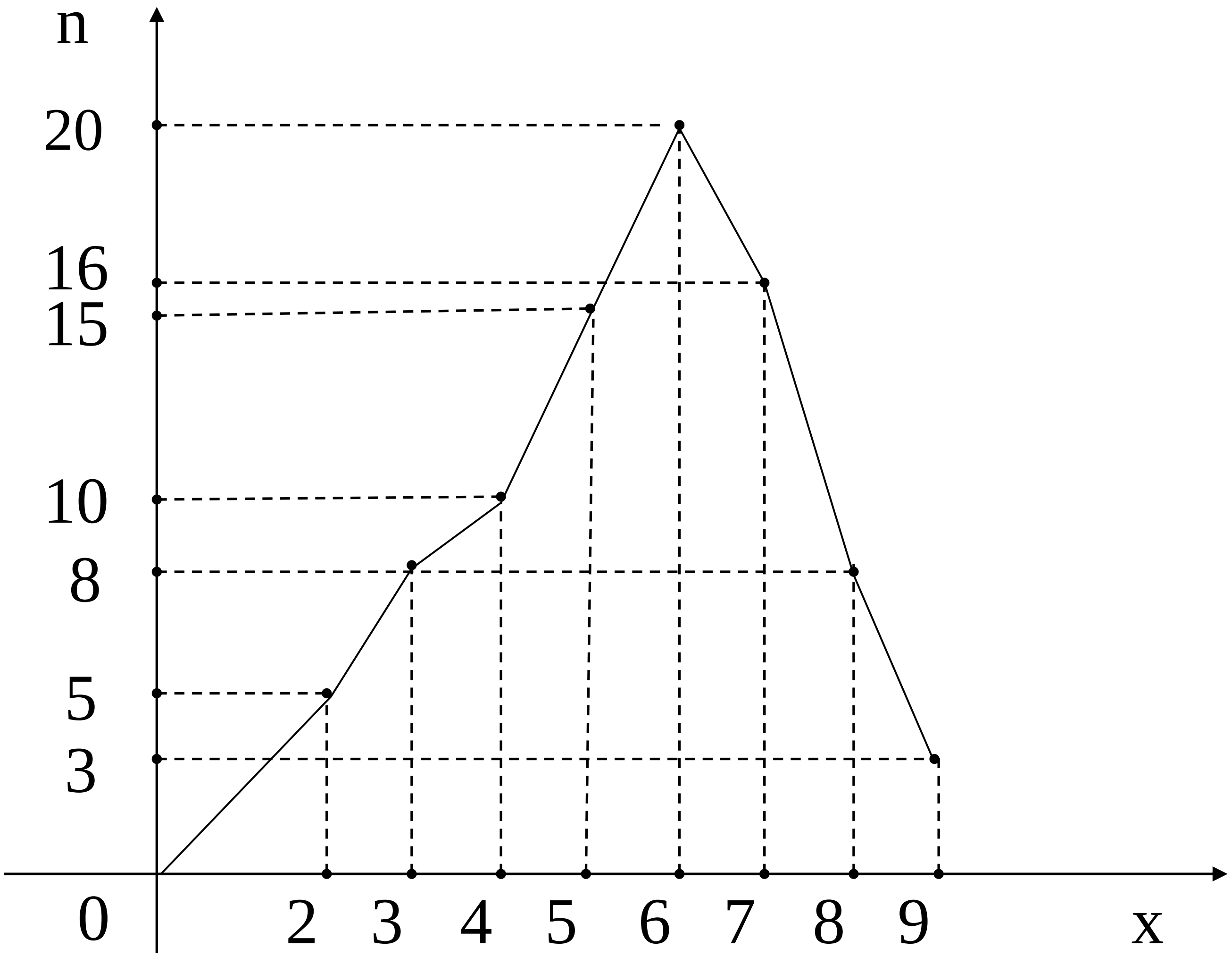
$$x_i = \frac{a_{i-1} + a_i}{2}; i = \overline{1; m}$$

3.2. BIỂU ĐỒ TẦN SỐ VÀ BIỂU ĐỒ TẦN SUẤT CỦA MẪU THỰC NGHIỆM

3.2.1 Biểu đồ của bảng phân phối không chia lớp

Ví dụ 3: Vẽ biểu đồ tần số của mẫu thực nghiệm của X cho theo bảng sau:

x_i	2	3	4	5	6	7	8	9
n_i	5	8	10	15	20	16	8	3


n
20
16
15
10
8
5
3
0
2
3
4
5
6
7
8
9
x

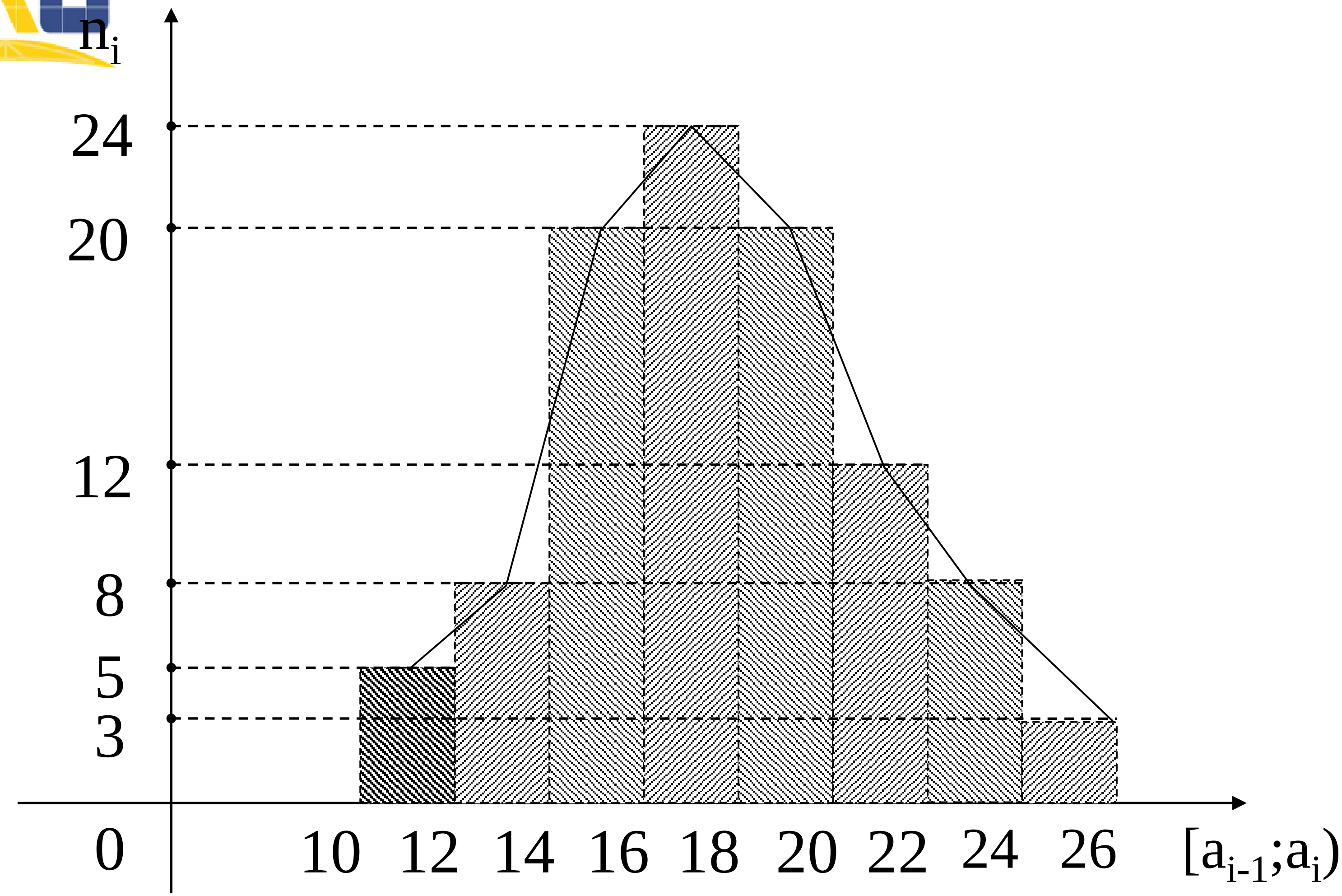
3.2.2 . Biểu đồ của bảng phân phối chia lớp

Ví dụ 4:

Vẽ biểu đồ tần số hình chữ nhật của mẫu thực nghiệm cho theo bảng sau:

$[a_{i-1};a_i)$	[10;12)	[12;14)	[14;16)	[16;18)
n_i	5	8	20	24

[18;20)	[20;22)	[22;24)	[24;26)
20	12	8	3



n_i
24
20
12
8
5
3
0
10 12 14 16 18 20 22 24 26
$[a_{i-1};a_i)$

3.3. HÀM PHÂN PHỐI THỰC NGHIỆM

Định nghĩa

Cho mẫu thực nghiệm của X như sau:

x_i	x_1	x_2	x_3	…	x_m
n_i	n_1	n_2	n_3	…	n_m
f_i	f_1	f_2	f_3	…	f_m

Trong đó $\sum_{i=1}^{m} n_i = n$

Khi đó hàm số được xác định như sau gọi là ***hàm phân phân phối thực nghiệm*** của X:

$$f(x)=\begin{cases} 0 & \text{nếu } x \le 0 \\ \dfrac{1}{n}\sum_{j=1}^{i} n_j = \sum_{j=1}^{i} f_j & \text{nếu } x_i < x \le x_{i+1}; i=\overline{1;m-1} \\ 1 & \text{nếu } x_m < x \end{cases}$$

Ví dụ 5:

Lấy mẫu kích thước n = 9 ta được các giá trị:

2, 1, 2, 3, 1, 1, 2, 2, 1.

a) Hãy lập bảng phân phối tần số, tần suất.

b) Xác định hàm phân phối thống kê F(x).

Giải:

a) Bảng phân phối tần số, tần suất:

X_i	1	2	3
n_i	4	4	1
f_i	4/9	4/9	1/9

b) Từ bảng phân phối tần số, tần suất ta có hàm phân phối thực nghiệm F(x):

$$F(x)=\begin{cases} 0 & \text{nếu } x \le 1 \\ \dfrac{4}{9} & \text{nếu } 1 < x \le 2 \\ \dfrac{8}{9} & \text{nếu } 2 < x \le 3 \\ 1 & \text{nếu } 3 < x \end{cases}$$

3.4 . CÁC THAM SỐ ĐẶC TRƯNG CỦA MẪU THỰC NGHIỆM

3.4.1. Số trung vị của mẫu thống kê

Cho mẫu thực nghiệm sắp xếp thành dãy tăng dần $x_1 \leq x_2 \leq \ldots \leq x_n$ hay giảm dần $x_1 \geq x_2 \geq \ldots \geq x_n$

Số trung vị của mẫu là số được ký hiệu và xác định như sau:

$$M_e(n) = \begin{cases} x_k & \text{nếu } n = 2k - 1 \text{ (n là số lẻ)} \\ \dfrac{x_k + x_{k+1}}{2} & \text{nếu } n = 2k \quad \text{(n là số chẵn)} \end{cases}$$

Ví dụ 6:

Tìm các số trung vị của mẫu thống kê của X cho dưới đây:

a) 2, 3, 5, 5, 8, 9, 10, 12

b)

x_i	0	1	2	3	4	5
n_i	1	3	10	7	2	2

Giải

a) Mẫu đã cho có kích thước n = 8 = 2.4 (số chẵn)

$$m_e(8) = \frac{x_4 + x_5}{2} = \frac{8+5}{2} = 6{,}5$$

b) Ta có: n = 25 (số lẻ) $\Rightarrow$ k =13

$$M_e\ (25) = x_{13} = 2$$

Ý nghĩa của số trung vị: số trung vị đặc trưng cho giá trị trung tâm của dãy các giá trị của mẫu thống kê, ta có thể hiểu rằng số trung vị có 50% của mẫu nhỏ hơn số trung vị và 50% giá trị lớn hơn số trung vị.

3.4.2. Số mốt của mẫu thực nghiệm

Giá trị của mẫu thực nghiệm của X có tần số lớn nhất gọi là **Mod** của mẫu thực nghiệm đó. Ký hiệu là M_o.

a) Trường hợp mẫu cho dưới dạng bảng phân phối tần số không chia lớp:

Nếu x_k là số có tần suất lớn nhất thì Mod của X:

$$M_o = x_k$$

b) Trường hợp mẫu cho dưới dạng bảng chia lớp:

Nếu lớp thứ k $[a_{k-1}, a_k]$ có tần suất n_k lớn nhất thì Mod của X:

$$M_o = \frac{a_{k-1} + a_k}{2}$$

Lớp chứa số Mod gọi là ***lớp Mod*** của mẫu thực nghiệm

Ví dụ 7:

a) Tìm Mod của các mẫu thực nghiệm của X ở ví dụ 6.

b) Tìm mốt của mẫu thống kê X:

$[a_{i-1}, a_i)$	20-22	22-24	24-26	26-28	28-30
n_i	2	6	8	7	2

a) - Dễ dàng thấy được Mod của X ở câu a: $M_o = 5$;

- câu b: $M_o = 2$.

b) Tần số lớn nhất là 8 nên: $M_o = (24 + 26)/2 = 25$

Lớp Mod: [24; 26)

3.4.2. Kỳ vọng, phương sai của mẫu thực nghiệm

a) Trường hợp mẫu cho dưới dạng bảng không chia lớp:

x_i	x_1	x_2	....	x_m
n_i	n_1	n_2	....	n_m
f_i	f_1	f_2	....	f_m

Kỳ vọng mẫu (trung bình mẫu):

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_mx_m}{n} = \frac{1}{n}\sum_{i=1}^{m} n_ix_i$$

hay

$$\overline{x} = f_1x_1 + f_2x_2 + \ldots + f_mx_m$$

b. Trường hợp mẫu cho dưới dạng bảng chia lớp thì khi áp dụng các công thức trên x_i chính là giá trị đại diện cho lớp $[a_{i-1}, a_i)$.

Ý nghĩa của số trung bình mẫu:

Số trung bình mẫu thực nghiệm là số dùng làm giá trị đại diện cho tất cả các giá trị của mẫu thực nghiệm của X.

Phương sai mẫu

$$s^2 = \frac{n_1(x_1 - \overline{x})^2 + \cdots n_m(x_m - \overline{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{m} n_i(x_i - \overline{x})^2$$

hay

$$s^2 = f_1(x_1 - \overline{x})^2 + \cdots + f_m(x_m - \overline{x})^2 = \sum_{i=1}^{m} f_i(x_i - \overline{x})^2$$

Chú ý: Phương sai còn được tính theo công thức sau:

$$s^2 = \overline{x^2} - \left(\overline{x}\right)^2$$

Phương sai mẫu hiệu chỉnh

$$s_1^2 = \frac{n}{n-1}s^2$$

Độ lệch chuẩn và độ lệch chuẩn hiệu chỉnh

- Độ lệch chuẩn: $s = \sqrt{s^2}$, với s^2 là phương sai mẫu.
- Độ lệch chuẩn hiệu chỉnh: $s_1 = \sqrt{s_1^2}$

  với s_1^2 là phương sai mẫu hiệu chỉnh.

Ý nghĩa của phương sai và độ lệch chuẩn:

Các giá trị của phương sai và độ lệch chuẩn đặc trưng cho sự đồng đều của các giá trị của mẫu thực nghiệm, nếu độ lệch chuẩn nhỏ thì các giá trị của mẫu thực nghiệm tương đối đồng đều, tập trung quanh và rất gần với giá trị trung bình.

Bài tập:

Bài tập 5.7. Từ bảng các số ngẫu nhiên người ta lấy ra 150 số. Các số đó được phân thành 10 khoảng như sau:

x_i	1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100
n_i	16	15	19	13	14	19	14	11	13	16

Xác định trung bình mẫu và phương sai mẫu

Bài tập 5.8. Khảo sát thu nhập của công nhân ở một công ty, cho bởi bảng sau (đơn vị ngàn đồng).

Thu nhập	$[500, 600]$	$[600, 700]$	$[700, 800]$	$[800, 900]$	$[900, 1000]$	$[1000, 1100]$	$[1100, 1200]$
Số người	2	10	15	30	25	14	4

Xác định thu nhập trung bình, độ lệch chuẩn.

Bài tập 5.11. Đo độ dài của một loại trục xe, ta có kết quả

Nhóm	18.4-18.6	18.6-18.8	18.8-19	19-19.2	19.2-19.4	19.4-19.6	19.6-19.8
n_i	1	4	20	41	19	8	4

Hãy tính độ dài trung bình và phương sai mẫu.

Ví dụ:

Để nghiên cứu về thâm niên công tác (tính tròn năm) của nhân viên ở một công ty lớn, người ta khảo sát thâm niên của 100 nhân viên được chọn ngẫu nhiên trong công ty. Kết quả như sau:

Thâm niên	5 - 7	8 - 10	11 - 13	14 - 16	17 -19
Số nhân viên	8	21	36	25	10

a) Hãy tính giá trị trung bình mẫu và giá trị độ lệch chuẩn mẫu.

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_mx_m}{n} = \frac{1}{n}\sum_{i=1}^{m} n_ix_i$$

$$s_1^2 = \frac{n}{n-1}s^2$$

Ví dụ 8:

Mẫu thống kê của X cho ở các bảng phân phối sau đây:

x_i	3	4	5	6	8
n_i	1	2	10	4	3

Tìm số trung vị, mốt, kỳ vọng, phương sai, phương sai mẫu hiệu chỉnh, độ lệch chuẩn, độ lệch chuẩn hiệu chỉnh.

Kích thước mẫu n = 20.

Số trung vị: $M_e(n) = 5$

Mốt: $M_o = 5$

Kỳ vọng:

$$\bar{x} = \frac{1}{20}(3 + 8 + 50 + 24 + 24) = 5{,}45$$