HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

o0o

Chương 4: Lý thuyết mẫu

XÁC SUẤT THỐNG KÊ

Bộ môn Toán, Khoa Cơ bản 1

Hà Nội - 2023

Chương 4: Lý thuyết mẫu

1. 4.1 Tổng thể và mẫu
2. 4.2 Thống kê và các đặc trưng của mẫu ngẫu nhiên

4.1.1 Tổng thể

Tổng thể

- Tổng thể là tập hợp tất cả các phần tử mà ta cần nghiên cứu dấu hiệu định tính hoặc định lượng nào đó. Mỗi phần tử của tổng thể được gọi là một cá thể.
- Số lượng các phần tử của tổng thể được gọi là kích thước của tổng thể, ký hiệu N.

Ví dụ 1. Nghiên cứu thu nhập của kỹ sư CNTT ở Việt Nam.
Tổng thể: tất cả các kỹ sư CNTT ở Việt Nam.
Dấu hiệu nghiên cứu: thu nhập.

Mỗi dấu hiệu định lượng có thể mô hình hóa bằng một biến ngẫu nhiên X. Khi đó X được gọi là biến ngẫu nhiên gốc.

4.1.2 Mẫu

Mẫu

- Mẫu là một tập hợp các phần tử được lấy ra từ tổng thể.
- Số lượng các phần tử của mẫu được gọi là kích thước của mẫu, ký hiệu n $(n \ll N)$.

Yêu cầu của việc lấy mẫu: Mẫu phải mang tính đại diện cho tổng thể, tức là phản ánh đúng đặc điểm của tổng thể theo dấu hiệu nghiên cứu.

4.1.3. Các phương pháp lấy mẫu

- Lấy mẫu ngẫu nhiên $\nearrow$ không hoàn lại; $\searrow$ có hoàn lại
- Lấy mẫu theo lớp
- Lấy mẫu theo chu kỳ

4.1.4 Mẫu ngẫu nhiên

- Việc chọn mẫu kích thước n có thể xem như tiến hành n quan sát độc lập về biến ngẫu nhiên X nào đó. Gọi X_i là giá trị của X ở lần quan sát thứ i $(i = 1, 2, \ldots, n)$, ta có $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên độc lập có cùng quy luật phân bố xác suất với X. $(X_1, X_2, \ldots, X_n)$ được gọi là một mẫu ngẫu nhiên kích thước n.
- Gọi x_i là kết quả quan sát được ở lần thứ i, $i = 1, 2, \ldots, n$. Khi đó $(x_1, x_2, \ldots, x_n)$ là một giá trị cụ thể mà mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ nhận, nó còn được gọi là một mẫu cụ thể.
- Tập hợp tất cả các mẫu cụ thể được gọi là không gian mẫu.

4.1.5 Mô tả số liệu mẫu

- Giả sử cần nghiên cứu dấu hiệu định lượng X ở tổng thể. Bằng các phương pháp lấy mẫu, ta nhận được một mẫu cụ thể có kích thước n gồm n số liệu: $x_1, x_2, \ldots, x_n$.
- Mẫu có thể được biểu diễn bởi bảng tần số:

X	x_1	x_2	$\ldots$	x_k
Tần số	n_1	n_2	$\ldots$	n_k

trong đó:

$$x_1 < x_2 < \ldots < x_k$$

n_i là số lần xuất hiện giá trị x_i trong mẫu

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

- Nếu ký hiệu $f_i = \dfrac{n_i}{n}$ là tần suất xuất hiện giá trị x_i trong mẫu thì ta có bảng tần suất:

X	x_1	x_2	$\dots$	x_k
Tần suất	f_1	f_2	$\dots$	f_k

- Nếu n lớn thì có thể chia mẫu ra làm k lớp, sau đó lấy điểm giữa của mỗi lớp làm điểm đại diện cho lớp. Khi đó mẫu được biểu diễn bởi bảng sau:

Lớp thứ i	$[a_0, a_1]$	$(a_1, a_2]$	$\ldots$	$(a_{k-1}, a_k]$
Tần số	n_1	n_2	$\ldots$	n_k

trong đó n_i là số cá thể trong mẫu có dấu hiệu X thỏa mãn $a_{i-1} < X \leq a_i$.

Chú ý:

- Điểm đại diện của lớp thứ i là $x_i = \dfrac{a_{i-1} + a_i}{2}$.
- Thông thường độ dài các lớp bằng nhau.

4.2.1 Thống kê

Định nghĩa

Hàm $T = T(X_1, X_2, \ldots, X_n)$ với $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên được gọi là một thống kê.

Nhận xét:

- Thống kê cũng là một biến ngẫu nhiên tuân theo một quy luật phân bố xác suất nhất định và có các tham số đặc trưng như kỳ vọng ET, phương sai DT, ...

- Khi mẫu ngẫu nhiên nhận một giá trị cụ thể $(x_1, x_2, \dots, x_n)$ thì T cũng nhận một giá trị cụ thể còn gọi là giá trị quan sát của thống kê
$$T_{qs} = T(x_1, x_2, \dots, x_n).$$
- Các thống kê cùng với quy luật phân bố xác suất của chúng là cơ sở để suy rộng các thông tin của mẫu cho dấu hiệu nghiên cứu của tổng thể.

Xét mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$

- Trung bình mẫu:

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

Với mẫu cụ thể $(x_1, x_2, \ldots, x_n)$, trung bình mẫu là

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

- Phương sai mẫu:

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

Với mẫu cụ thể $(x_1, x_2, \ldots, x_n)$, phương sai mẫu là

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

Chú ý: Trong tính toán ta thường dùng công thức

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

- Độ lệch chuẩn mẫu:
$$S = \sqrt{S^2}$$
Với mẫu cụ thể $(x_1, x_2, \ldots, x_n)$, độ lệch chuẩn mẫu là $s = \sqrt{s^2}$.
- Tần suất mẫu: Giả sử cần nghiên cứu một dấu hiệu định tính A nào đó mà mỗi cá thể của tổng thể có thể có hoặc không.
Ta có thể coi dấu hiệu nghiên cứu này là một biến ngẫu nhiên X, nhận giá trị 1 nếu cá thể có dấu hiệu A và nhận giá trị 0 trong trường hợp ngược lại.
Từ tổng thể lấy mẫu ngẫu nhiên kích thước n: $(X_1, X_2, \ldots, X_n)$.
Khi đó tần suất mẫu, kí hiệu f, được xác định bởi
$$f = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\text{số cá thể có dấu hiệu } A \text{ trong mẫu}}{n}$$
là một thống kê.

- Nếu mẫu cho dưới dạng bảng tần số:

x_i	x_1	x_2	$\ldots$	x_k
n_i	n_1	n_2	$\ldots$	n_k

thì

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i,$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} n_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)^2\right],$$

trong đó $n = n_1 + n_2 + \ldots n_k$.

Ví dụ 2. Theo dõi số xe hơi bán được trong một tuần ở một số đại lý ta được bảng số liệu sau:

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

Tính $\overline{x}, s^2, s$.

- Nếu các giá trị của mẫu cụ thể không gọn (quá lớn hoặc quá bé hoặc phân tán) ta có thể thu gọn mẫu bằng cách đổi biến $u_i = \dfrac{x_i - a}{h}$. Khi đó

$$\overline{x} = a + h\overline{u};\ s^2 = h^2 s_u^2,$$

trong đó

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{k} n_i u_i,$$

$$s_u^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} n_i u_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i u_i\right)^2\right],$$

Ví dụ 3. Cân 25 bao gạo ta thu được các số liệu ở bảng sau

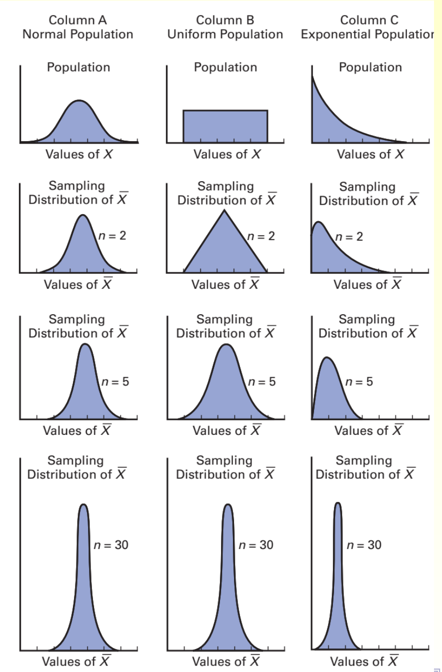
Trọng lượng	48-48,5	48,5 - 49	49-49,5	49,5 - 50	50-50,5
Số bao	2	5	10	6	2

Tính $\overline{x}, s$.

4.2.3 Phân bố xác suất của trung bình mẫu

Giả sử biến ngẫu nhiên gốc X có $E(X) = \mu, D(X) = \sigma^2$. Xét mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ từ tổng thể.

- Trường hợp 1: Nếu X có phân bố chuẩn $N(\mu, \sigma^2)$ thì:
  - i) $\overline{X} \sim N(\mu, \dfrac{\sigma^2}{n})$
  - ii) $\dfrac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0, 1)$
- Trường hợp 2 (Định lý giới hạn trung tâm): Nếu X có phân bố tùy ý thì với n đủ lớn, $\overline{X}$ có phân bố xấp xỉ $N(\mu, \dfrac{\sigma^2}{n})$, và do đó phân bố của $\dfrac{\overline{X} - \mu}{\sigma}\sqrt{n}$ xấp xỉ $N(0, 1)$.



Column A
Normal Population
Column B
Uniform Population
Column C
Exponential Population
Population
Values of X
Sampling Distribution of $\overline{X}$
$n = 2$
$n = 5$
$n = 30$
Values of $\overline{X}$