

Bài 8: MẪU NHẪU NHIÊN-MỘT SỐ THỐNG KÊ QUAN TRỌNG

Vũ Mạnh Tới

Bộ môn Toán-Trường Đại học Thủy lợi

Ngày 13 tháng 5 năm 2024

Đặt vấn đề:

- Trong thống kê, người ta quan tâm nhất là vấn đề làm thế nào để rút ra được những kết luận chính xác cho một đối tượng nghiên cứu nào đó.
- Thông thường các đối tượng nghiên cứu có số lượng rất lớn nên ta không thể quan sát trên toàn bộ tập hợp đó.
- Để thu được kết luận, người ta phải nghiên cứu trên 1 “mô hình thu nhỏ” của tập hợp rồi từ đó dẫn đến kết luận. Phương pháp làm như vậy được gọi là phương pháp mẫu.

8.1. Mẫu ngẫu nhiên

Định nghĩa

- Tập hợp gồm tất cả các đối tượng có **chung một đặc tính** nào đó mà ta đang quan tâm được gọi là một **tổng thể**.
- Việc chọn ra từ tổng thể một tập con bất kỳ nào đó gọi là phép lấy mẫu. Tập con này gọi là một mẫu, số lượng phần tử trong mẫu gọi là cỡ mẫu.
- Nếu việc chọn mẫu được tiến hành một cách ngẫu nhiên và các quan sát là độc lập thì mẫu được gọi là **mẫu ngẫu nhiên**.

- Mẫu ngẫu nhiên bao gồm n cá thể lấy ra từ tổng thể, khi đó tổng thể đều được đặc trưng bởi một đặc tính chung bởi BNN X có phân phối xác suất $f(x)$.
- BNN X có kỳ vọng μ và phương sai σ^2 , khi đó μ là trung bình của tổng thể; σ^2 là phương sai của tổng thể.
- Ta gán mỗi cá thể trong mẫu bởi một giá trị của X và ta có n BNN X_1, X_2, \dots, X_n , trong đó X_i là số đo đặc tính chung của cá thể thứ i và có giá trị x_i .
- BNN $X_i, i = 1, \dots, n$ đều có phân phối $f(x)$ và $\mu_{X_i} = \mu; \sigma_{X_i}^2 = \sigma^2$.

Ví dụ 1: Cần nghiên cứu chiều cao trung bình của người Việt Nam trưởng thành. Không thể đo hết toàn bộ người Việt Nam. Phải tiến hành lấy ngẫu nhiên n (Ví dụ $n=1000$) người. Khi đó

- **Tổng thể:** Toàn bộ người Việt Nam với đặc trưng quan tâm là BNN X đo chiều cao (cm) giả sử tuân theo phân phối chuẩn với trung bình tổng thể μ , độ lệch chuẩn tổng thể σ .
- Mỗi cá thể đặt là X_i và cho kết quả đo là x_i cm
Chiều cao người 1 là 165, tức là giá trị của X_1 là $x_1 = 165$;
:
Chiều cao người thứ 1000 là 167, tức là giá trị của X_{1000} là $x_{1000} = 167$.

Một mẫu ngẫu nhiên cỡ n có thể cho bởi dãy

$$X_1; X_2; \dots; X_n$$

Hoặc cho bởi bảng tần số

X	X_1	X_2	\dots	X_k
Tần số	n_1	n_2	\dots	n_k

Với chú ý: $n = n_1 + n_2 + \dots + n_k$

8.2. Một số thống kê quan trọng

Định nghĩa:

Thống kê là hàm của các BNN nằm trong một mẫu ngẫu nhiên.

- Thống kê được dùng nhiều nhất để đo tâm của một tập các số liệu được bố trí theo trật tự độ lớn là **trung bình mẫu, trung vị, mode**.
- Những thống kê để đo mức độ phân tán và độ biến động của tập số liệu là **phương sai mẫu, độ lệch chuẩn mẫu**.

Trung bình mẫu và phương sai mẫu

Cho X_1, X_2, \dots, X_n là một mẫu ngẫu nhiên cỡ n . Khi đó:

Trung bình mẫu. Trung bình mẫu là một thống kê, kí hiệu \bar{X} , được xác định như sau

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{giá trị trung bình mẫu } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Phương sai mẫu là một thống kê ký hiệu là S^2 được xác định bởi

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad \text{giá trị } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}.$$

Độ lệch chuẩn mẫu: $s = \sqrt{s^2}$.

Chú ý: $s^2 = \frac{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{n(n-1)}.$

Phân biệt trung bình, phương sai của tổng thể và của mẫu:

- Ký hiệu: trung bình, phương sai tổng thể: μ và σ^2 ; Trung bình, phương sai mẫu: \bar{X} ; s^2 .
- Trung bình mẫu và phương sai mẫu phải được tính toán từ mẫu.

Ví dụ 1: Một vị thanh tra thực phẩm kiểm tra một mẫu ngẫu nhiên 7 hộp cá ngừ có cùng nhãn hiệu để xác định phần trăm các tạp chất lạ. Các số liệu sau đây đã được ghi lại:

1,8; 2,1; 1,7; 1,6; 0,9; 2,7 và 1,8.

Hãy tính trung bình mẫu, phương sai mẫu

ĐS: $\bar{x} = 1,8$; $s = 0,5416$; $s^2 = 0,2933$.

Ví dụ 2: Đo trọng lượng (gam) của các quả trứng gà cho số liệu

Khối lượng	150	160	165	170	180	185
Số quả	4	16	25	30	15	10

Tính \bar{x} ; s

ĐS: $\bar{x} = 169,35$; $s = 8,7229$

8.3. Phân phối xác suất của một số thống kê

Phân phối lấy mẫu của trung bình mẫu

- **1 mẫu:** Giả sử một mẫu ngẫu nhiên cỡ n (mọi n) được lấy từ **tổng thể chuẩn** với kỳ vọng là μ và phương sai là σ^2 .

Khi đó $\bar{X} \sim N\left(\bar{x}; \mu; \frac{\sigma}{\sqrt{n}}\right)$. Tức là

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(z; 0; 1).$$

- **2 mẫu:** Mẫu 1 cỡ n_1 và mẫu 2 cỡ n_2 lấy ra từ hai **tổng thể chuẩn** có trung bình μ_1, μ_2 , phương sai σ_1^2, σ_2^2 tương ứng. Khi đó:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(z; 0; 1).$$

Định lí giới hạn trung tâm:

- (1 mẫu): Nếu \bar{X} là trung bình của một mẫu ngẫu nhiên cỡ n được lấy từ một tổng thể với trung bình tổng thể μ và phương sai σ^2 , thì

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(z; 0; 1) \text{ với } n \geq 30 \text{ (đủ lớn)}.$$

- (2 mẫu): Mẫu 1 cỡ n_1 và mẫu 2 cỡ n_2 lấy ra từ hai tổng thể có trung bình μ_1, μ_2 , phương sai σ_1^2, σ_2^2 tương ứng. Khi đó:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(z; 0; 1) \text{ với } n_1 \geq 30, n_2 \geq 30 \text{ (đủ lớn)}.$$

Chú ý: Nếu các tổng thể là phân phối chuẩn, hoặc xấp xỉ chuẩn, thì cỡ mẫu không cần đủ lớn.

Định lí giới hạn trung tâm rất quan trọng trong thống kê. Nó là cơ sở cho các bài toán thống kê.

Mẫu ngẫu nhiên cỡ n lấy ra từ tổng thể xấp xỉ chuẩn (hoặc chuẩn) thì

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \text{ có phân phối } t \text{ bậc } v = n - 1.$$

Giá trị $t_\alpha, \alpha \in (0, 1)$ với bậc v được cho bởi bảng A_4 thỏa mãn: $P(T > t_\alpha) = \alpha$.

VD: Tìm $t_{\alpha/2}$ với bậc $v = 15$ thỏa mãn

$$P(T > t_{\alpha/2}) = \alpha/2 \text{ với } \alpha = 0.05.$$

Đáp số: $t_{\alpha/2} = 2,131$

HƯỚNG DẪN TÍNH CÁC THỐNG KÊ TRÊN MÁY TÍNH CASIO

Đối với máy Casio f(x)-500MS hoặc 570MS

Bước 1: Xóa toàn bộ dữ liệu cũ : SHIFT MODE 3 =

Bước 2: Vào được chế độ SD (chế độ thống kê): MODE 2 (500MS), hoặc MODE MODE 2 (570MS)

Bước 3: Nhập số liệu:

Nếu số liệu cho bởi dãy $X_1; X_2; \dots; X_n$:

Ta nhập: X_1 M+ X_2 M+ ... X_n M+

Nếu số liệu cho bởi bảng tần số

Ta nhập: $X_1 ; n_1$ M+ $X_2 ; n_2$ M+... $X_k ; n_k$ M+

Bước 4: Lấy kết quả:

SHIFT S-VAR 1 = được \bar{x}

SHIFT S-VAR 3 = được s ; tiếp x^2 = được s^2

SHIFT S-SUM 1 = được $\sum x_i^2$

SHIFT S-SUM 2 = được $\sum x_i$

SHIFT S-SUM 3 = được n

Đối với máy Casio f(x)-500 ES hoặc 570 ES hoặc vn-plus

Bước 1: Bật hoặc tắt chế độ có tần số hay không có tần số: SHIFT MODE ∇ 4 (4 : STAT)

Sau đó chọn 1 (1 : ON) nếu dữ liệu cho bởi bảng tần số; chọn 2 (2 : OFF) nếu dữ liệu cho bởi dãy.

Bước 2: Vào được chế độ STAT (chế độ thống kê):
MODE 3 (3 : STAT rồi chọn 1-VAR)

Bước 3: Nhập số liệu:

Nếu số liệu cho bởi dãy $X_1; X_2; \dots; X_n$: Ta nhập :

$$X_1 = X_2 = \dots X_n =$$

Nếu số liệu cho bởi bảng tần số.

Ta nhập lần lượt hết giá trị cột X rồi sang cột tần số nhập tương ứng.

Chú ý nhập xong thì phải ấn AC

Bước 4: Lấy kết quả:

SHIFT STAT 5 (5 : Var) chọn 1 = được n

SHIFT STAT 5 (5 : Var) chọn 2 = được \bar{x} ;

SHIFT STAT 5 (5 : Var) chọn 4 = được s

SHIFT STAT 4 (4 : Sum) chọn 1 = được $\sum x_i^2$

SHIFT STAT 4 (4 : Sum) chọn 2 = được $\sum x_i$