

Trường Đại Học Khoa Học Tự Nhiên TP.HCM

# XÁC SUẤT THỐNG KÊ

Lưu hành nội bộ  
2020

## Dữ liệu và thống kê

### 3. Lý thuyết mẫu

3.1 Giới thiệu về  
thống kê  
3.2 Mẫu ngẫu nhiên  
3.3 Các đặc trưng  
của mẫu ngẫu nhiên  
3.4 Phân phối mẫu

- **Dữ liệu (data)** là thông tin có được từ những quan sát, những phép đếm, những đo đạc, hoặc các câu trả lời.
- **Thống kê (statistics)** là khoa học về thu thập, tổ chức, phân tích, và giải thích dữ liệu để đưa ra các quyết định.
- **Tổng thể (population)** là toàn bộ tập hợp tất cả các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hoặc định lượng nào đó.
- **Mẫu (sample)** là một tập con của một tổng thể.

## Dữ liệu và thống kê

Ví dụ:

- Trong một cuộc khảo sát gần đây, 150 sinh viên trường ĐH KHTN được hỏi rằng họ có tham gia vào một cuộc thi học thuật nào đó trong năm hay không. 35 sinh viên trả lời có. Xác định tổng thể và mẫu?
- Vào cuối ngày, một thanh tra kiểm soát chất lượng chọn 20 bóng đèn từ các bóng đèn được sản xuất trong ngày và kiểm tra chúng.

## Tham số và thống kê

### 3. Lý thuyết mẫu

3.1 Giới thiệu về  
thống kê  
3.2 Mẫu ngẫu nhiên  
3.3 Các đặc trưng  
của mẫu ngẫu nhiên  
3.4 Phân phối mẫu

- **Tham số (parameter)** là một mô tả số về một đặc trưng của một *tổng thể*.
- **Thống kê (statistic)** là một mô tả số về một đặc trưng của một *mẫu*.

Tham số → Tổng thể  
Thống kê → Mẫu

## Tham số và thống kê

**Ví dụ:** Hãy xác định giá trị số sau đây mô tả một tham số tổng thể hay một thống kê mẫu?

- a Một khảo sát về một mẫu 250 sinh viên đại học báo cáo rằng chi tiêu trung bình hàng tuần của các sinh viên là 350.000 đ.  
*Bởi vì giá trị trung bình 350.000 đ dựa trên một mẫu, nên đây là một thống kê mẫu.*
- b Chi tiêu trung bình hàng tuần của tất cả các sinh viên là 400.000 đ.  
*Bởi vì giá trị trung bình 400.000 đ dựa trên một tổng thể, nên đây là một tham số tổng thể.*

## Các nhánh của thống kê

Nghiên cứu thống kê có hai nhánh chính: **thống kê mô tả** và **thống kê suy luận**.

- **Thống kê mô tả:** Liên quan đến việc tổ chức, tổng hợp, và trình bày dữ liệu.
- **Thống kê suy luận:** Liên quan đến việc sử dụng một mẫu để rút ra kết luận về một tổng thể.

## Thống kê mô tả và suy luận

**Ví dụ:** Trong một nghiên cứu gần đây, các tình nguyện viên ngủ ít hơn 6 tiếng có khả năng trả lời sai trong một bài kiểm tra khoa học cao hơn gấp bốn lần những người ngủ ít nhất 8 tiếng. Xác định phần nào là thống kê mô tả và kết luận nào có thể được rút ra bằng cách sử dụng thống kê suy luận?

*Phát biểu “khả năng trả lời sai gấp bốn lần” là một thống kê mô tả.  
Kết luận được rút ra từ mẫu là những người ngủ ít hơn 6 tiếng có khả năng trả lời sai cao hơn những người ngủ ít nhất 8 tiếng.*

## Mẫu ngẫu nhiên

Mẫu ngẫu nhiên kích thước  $n$  là tập hợp của  $n$  biến ngẫu nhiên độc lập  $X_1, X_2, \dots, X_n$  được thành lập từ biến ngẫu nhiên  $X$  trong tổng thể nghiên cứu và có cùng quy luật phân phối xác suất với  $X$ .

**Kí hiệu:**  $W = (X_1, X_2, \dots, X_n)$

## Thống kê

Một thống kê (statistic) là một hàm bất kỳ của các quan sát trong một mẫu ngẫu nhiên.

Nếu  $X_1, X_2, \dots, X_n$  là một mẫu ngẫu nhiên kích thước  $n$  thì:

◦ **Trung bình mẫu:**  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

◦ **Phương sai mẫu:**

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

◦ **Phương sai mẫu có điều chỉnh:**

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

Ta thấy rằng:  $S^2 = \frac{n}{n-1} s^2$

◦ **Độ lệch chuẩn mẫu:**  $S = \sqrt{S^2}$   
đều là các thống kê.

Bởi vì thống kê là một mẫu ngẫu nhiên nên nó có phân phối xác suất.

**Định nghĩa**

Phân phối xác suất của một thống kê đgl một phân phối mẫu.

Ví dụ: Phân phối xác suất của  $\bar{X}$  đgl phân phối mẫu của trung bình.

**Nhận xét**

Phân phối xác suất của một thống kê phụ thuộc vào phân phối của tổng thể, kích thước mẫu và phương pháp chọn mẫu.

## Phân phối mẫu của trung bình và phương sai

Trường hợp tổng thể có phân phối xác suất chưa biết

Trong trường hợp này, định lí giới hạn trung tâm khẳng định rằng

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

Trong thực hành khi mẫu có kích thước đủ lớn ( $n \geq 30$ ), ta có các phân phối xấp xỉ chuẩn sau:  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$  và  $\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx N(0, 1)$

## Phân phối mẫu của tỉ lệ

Giả sử cần khảo sát đặc trưng  $\mathcal{A}$  của tổng thể, khảo sát  $n$  phần tử và đặt

$$X_i = \begin{cases} 1 & \text{nếu thỏa } \mathcal{A} \\ 0 & \text{nếu khác} \end{cases}$$

thu được mẫu ngẫu nhiên  $X_1, \dots, X_n$  với  $X_i \sim B(1, p)$ ,  $p$  là tỉ lệ phần tử thỏa đặc trưng  $\mathcal{A}$ .

Đặt  $X = \sum_{i=1}^n X_i$  là số phần tử thỏa đặc trưng  $\mathcal{A}$  trong mẫu

khảo sát, thì  $X \sim B(n, p)$ .

Tỉ lệ mẫu  $\hat{p}$  là một ước lượng của tỉ lệ  $p$  được xác định bởi

$$\hat{p} = \frac{X}{n}$$

3. Lý thuyết  
mẫu3.1 Giới thiệu về  
thống kê

3.2 Mẫu ngẫu nhiên

3.3 Các đặc trưng  
của mẫu ngẫu nhiên

3.4 Phân phối mẫu

## Phân phối mẫu của tỉ lệ

Kì vọng và phương sai của  $\hat{p}$  là

$$\mathbb{E}(\hat{p}) = p, \quad \mathbb{V}ar(\hat{p}) = \frac{p(1-p)}{n}$$

Theo định lí giới hạn trung tâm ta có

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} N(0, 1)$$

Vì vậy trong thực hành, khi

$np \geq 5, n(1-p) \geq 5$ , thì  $\hat{p} \approx N(p, \frac{p(1-p)}{n})$