BÀI 5: CƠ SỞ LÝ THUYẾT MẪU

Mục tiêu

Giới thiệu một số khái niệm cơ bản của Thống kê toán học, cụ thể là những vấn đề liên quan đến cặp phạm trù tổng thể và mẫu, đến các khái niệm thống kê, thống kê của đặc trưng mẫu và phân phối xác suất của thống kê đặc trưng mẫu, xem xét cụ thể các khái niệm đó trong một số trường hợp đặc biệt nhưng thường gặp trong thực hành.

Thời lượng

- 8 tiết

Các kiến thức cần có

- Khái niệm phương pháp mẫu
- Tổng thể nghiên cứu
- Định nghĩa
- Mô tả tổng thể
- Các số đặc trưng của tổng thể
- Mẫu ngẫu nhiên
- Các phương pháp lấy mẫu
- Định nghĩa mẫu ngẫu nhiên
- Mô tả mẫu ngẫu nhiên
- Thống kê (Statistics)
- Định nghĩa
- Các thống kê đặc trưng mẫu
- Mẫu ngẫu nhiên hai chiều
- Khái niệm
- Phương pháp mô tả mẫu
- Thống kê đặc trưng mẫu hai chiều
- Quy luật phân phối xác suất của một số thống kê
- Trường hợp biến ngẫu nhiên gốc có phân phối 0–1
- Trường hợp hai biến ngẫu nhiên gốc có phân phối 0–1
- Trường hợp biến ngẫu nhiên gốc có phân phối chuẩn
- Trường hợp hai biến ngẫu nhiên gốc có phân phối chuẩn

TÌNH HUỐNG KHỞI ĐỘNG BÀI

Tình huống

Điều tra mức thu nhập cá nhân trong một tháng (triệu đồng) ở huyện Đông Anh, ta có bảng số liệu mẫu sau:

Thu nhập	1-2	2-3	3-4	4-5	5-6	6-7
số người	10	8	5	7	3	2

Cần phải tính thu nhập bình quân đầu người và độ chênh lệch thu nhập để xác định mức sống của người dân và mức độ đồng đều về thu nhập trong vùng.

Câu hỏi

1. Thu nhập bình quân đầu người là bao nhiêu?

2. Độ chênh lệch thu nhập là bao nhiêu?

3. Độ chênh lệch bình quân hiệu chỉnh?

5.1. Khái niệm phương pháp mẫu

Bài toán:

Chúng ta cần nghiên cứu tính chất định tính hoặc định lượng của các phần tử trong một tập hợp nào đó. Khi đó ta có hai phương pháp thực hiện nghiên cứu

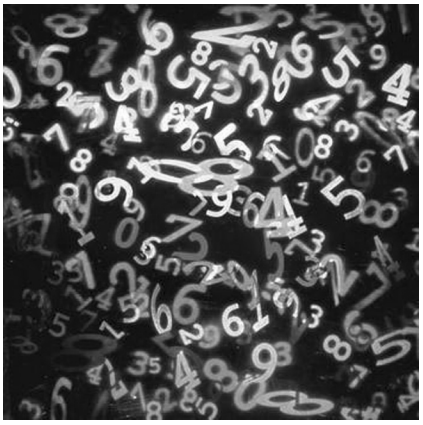
- Nghiên cứu toàn bộ các phần tử của tập hợp và ghi lại các đặc tính cần quan tâm. Khi thực hiện nghiên cứu toàn bộ ta gặp phải những hạn chế sau:
  - Phải trả chi phí lớn về kinh tế và thời gian do số lượng các phần tử trong tập toàn bộ quá lớn.
  - Có thể dẫn tới phá huỷ toàn bộ tập hợp cần nghiên cứu. Ví dụ nghiên cứu thời gian hoạt động của các thiết bị điện tử. Khi áp dụng phương pháp này sẽ dẫn tới phá huỷ toàn bộ các thiết bị điện tử.
  - Có những tập hợp mà ta không thể nghiên cứu được toàn bộ. Ví dụ như trong lĩnh vực khảo cổ học.

  Vậy ta thấy trong đa số các trường hợp nghiên cứu toàn bộ tập hợp là không khả thi.
- Nghiên cứu bộ phận, từ tập hợp nghiên cứu ta lấy ra một tập con và nghiên cứu toàn bộ các phần tử trong tập con đó và từ đó đưa ra kết luận cho các phần tử trong tập hợp nghiên cứu.

Phương pháp nghiên cứu thứ hai gọi là phương pháp nghiên cứu mẫu.

5.2. Tổng thể nghiên cứu

5.2.1. Định nghĩa



Tổng thể (population) là tập hợp các phần tử cần nghiên cứu tính chất định tính hoặc định lượng, số phần tử trong tổng thể gọi là cỡ của tổng thể, ký hiệu là N.

Ví dụ:

- Thu nhập của toàn bộ dân cư của một nước.
- Chất lượng sản phẩm của một nhà máy.
- Nhu cầu tiêu dùng điện của các hộ gia đình.

Khi nghiên cứu tổng thể thì các phần tử có thể có hai loại tính chất định tính hoặc định lượng cần quan tâm, do đó ta có hai loại biến:

- Biến định lượng là các số đo của phần tử;

  Ví dụ: Cân nặng, chiều cao, tuổi, thu nhập,…
- Biến định tính là tính chất nào đó của đối tượng nghiên cứu.

  Ví dụ: Giới tính, chất lượng, dân tộc, tôn giáo,…

Đối với các biến ta có các cách mã hoá như sau:

- Kỹ thuật mã hoá

- Mã hoá biến định lượng: Ta lấy giá trị của biến định lượng làm mã của biến
- Mã hoá biến định tính: Ta gán tính chất định tính của biến ứng với các số nguyên.

Ví dụ:

Đối tượng là thu nhập của hộ gia đình ta có các mức: Nghèo, trung bình, giàu. Ta mã hoá các biến như sau:

Nghèo → -1; Trung bình → 0; Giàu → 1

Vậy khi nghiên cứu tổng thể ta luôn có thể giả sử là các các phần tử của tổng thể có dấu hiệu định lượng.

5.2.2. Mô tả tổng thể

Cho tổng thể với các phần tử $\{x_1, x_2, \ldots x_N\}$, ta có thể thu gọn bằng cách gộp các giá trị giống nhau lại và biểu diễn như dạng.

x_i	x_1	$x_2 \ldots\ldots\ldots\ldots x_k$
N_i	N_1	$N_2 \ldots\ldots\ldots\ldots N_k$

trong đó N_i $(i = 1,...,k)$ là số lần giá trị x_i xuất hiện trong tổng thể, ta có

$$N_1 + N_2 + \ldots + N_k = N$$

Đặt $f_1 = \frac{N_i}{N}$ $(i = 1,\ldots,k)$, f_i được gọi là tần suất của x_i trong tổng thể và ta có bảng tần suất.

x_i	x_1	$x_2 \ldots\ldots\ldots\ldots x_k$
f_i	f_2	$f_2 \ldots\ldots\ldots\ldots f_k$

Hiển nhiên ta có: $f_1 + f_2 + \ldots + f_k = 1$

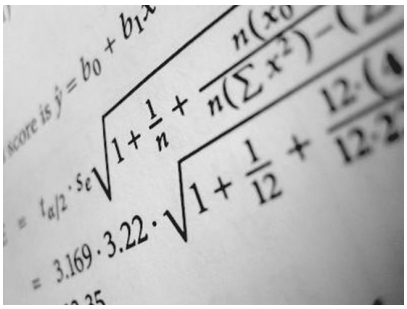
Bảng tần suất giống như một bảng phân phối xác suất của biến ngẫu nhiên, do đó ta có thể đồng nhất tổng thể nghiên cứu với một biến ngẫu nhiên X nào đó với hàm phân phối F. Vậy, thay vì nghiên cứu tổng thể thì ta quy về nghiên cứu biến ngẫu nhiên X.

5.2.3. Các số đặc trưng của tổng thể

- **Trung bình tổng thể**

Trung bình tổng thể là đại lượng ký hiệu là m được xác định bởi:

$$m = \frac{1}{N}\sum_{i=1}^{N} N_i x_i = \sum_{i=1}^{N} f_i x_i$$



Ta thấy m có thể xem là kỳ vọng của biến ngẫu nhiên X.

- **Phương sai tổng thể**

Phương sai tổng thể là đại lượng ký hiệu là s được xác định bởi:

$$s = \frac{1}{N}\sum_{i=1}^{N} (x_i - m)^2 = \sum_{i=1}^{N} f_i x_i^2 - (m)^2.$$

Ta thấy s có thể xem là phương sai của biến ngẫu nhiên X.

5.3. Mẫu ngẫu nhiên

Trong phần trước ta đã biết rằng không thể nghiên cứu cặn kẽ từng phần tử của tổng thể, do đó ta phải nghiên cứu hạn chế trên một nhóm nhỏ được rút ra từ tổng thể gọi là mẫu và từ đó rút ra kết luận cho tổng thể, do vậy ta mong muốn mẫu đại diện tốt nhất cho tổng thể. Nói chung, để có được một mẫu đại diện tốt nhất cho tổng thể người ta thường phải tiến hành xây dựng mẫu theo một quy trình chọn ngẫu nhiên các phần tử của mẫu. Một mẫu như vậy được gọi là *mẫu ngẫu nhiên* (random sample).

5.3.1. Các phương pháp lấy mẫu

Có rất nhiều phương pháp chọn mẫu ngẫu nhiên để thoả mãn tính đại diện tốt nhất cho tổng thể và phù hợp với mục tiêu nghiên cứu. Sau đây ta chỉ nghiên cứu những phương pháp chủ yếu.

- Cách chọn mẫu ngẫu nhiên đơn giản
  - Chọn mẫu ngẫu nhiên có hoàn lại: Từ tổng thể ta rút ngẫu nhiên một phần tử và ghi lại các đặc trưng cần quan tâm, sau đó trả lại phần tử đó về tổng thể và làm tương tự ở các lần tiếp theo cho tới khi ta được một mẫu cỡ n.
  - Chọn mẫu ngẫu nhiên không hoàn lại: Làm tương tự như trên, chỉ khác là sau mỗi lần rút các phần tử ta loại phần tử đó ra khỏi tổng thể.
- Chọn mẫu phân cấp

  Ở những tổng thể lớn có thể có những yêu cầu phải chọn một mẫu phân cấp chẳng hạn như điều tra phân tích mức sống của dân cư trong nước thường có những yêu cầu kết luận cho các vùng, các miền.
  - Mẫu phân cấp đơn giản có thể được thành lập như sau: Chia tổng thể ra thành k tổng thể bộ phận và ta thực hiện cách lấy mẫu ngẫu nhiên đơn giản trên mỗi tổng thể thành phần rồi tổng hợp lại để có mẫu của toàn bộ tổng thể.

Ta cũng có thể tiến hành lấy mẫu phân cấp theo những quy trình phức tạp hơn. Chẳng hạn như sau khi chia tổng thể ra thành k tổng thể bộ phận, ta chọn ngẫu nhiên trong số k tổng thể bộ phận đó ra m tổng thể rồi tiếp tục thực hiện lấy mẫu ngẫu nhiên trên từng tổng thể được chọn để tổng hợp thành mẫu của toàn bộ tổng thể.

5.3.2. Định nghĩa mẫu ngẫu nhiên

Một mẫu ngẫu nhiên cỡ n của biến ngẫu nhiên X là một bộ n các biến ngẫu nhiên $X_1, X_2, \ldots X_n$ độc lập và có cùng phân phối với biến ngẫu nhiên X, trong đó mỗi X_k là một *quan sát* về biến ngẫu nhiên X.

Ta ký hiệu x_k là kết quả quan sát được ở lần thứ k, tức là quan sát X_k nhận giá trị x_k (k = 1,2,…, n). Khi đó bộ giá trị $(x_1, x_2, \ldots, x_n)$ gọi là giá trị cụ thể của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$.

Ví dụ 1:

Khi gieo con xúc xắc 5 lần ta được một mẫu ngẫu nhiên (X_1, X_2, X_3, X_4, X_5) trong một lần lấy mẫu nào đó, chẳng hạn ta được giá trị của mẫu là (3, 5, 2, 3, 1).

Ví dụ 2:

Nghiên cứu thời gian hoạt động của các thiết bị điện tử do một công ty sản xuất, ta lấy ngẫu nhiên n thiết bị, khi đó ta được một mẫu ngẫu nhiên (X_1, X_2, ….,X_n), theo dõi thời gian hoạt động của n thiết bị điện tử này ta được các giá trị mẫu là (x_1, x_2, …,x_n).

5.3.3. Mô tả mẫu ngẫu nhiên

Cho biến ngẫu nhiên X và một mẫu ngẫu nhiên (X_1, X_2, ….,X_n) với các giá trị mẫu (x_1, x_2, …,x_n). Để mô tả mẫu ngẫu nhiên ta có hai cách như sau:

- Biểu đồ tần suất

  Ta có thể thu gọn bằng cách gộp các giá trị giống nhau trong mẫu và biểu diễn dưới dạng bảng sau:

x_i	x_1	x_2	…	x_n
n_i	n_1	n_2	…	n_k

trong đó n_i là số lần giá trị x_i xuất hiện trong mẫu. Ta có:

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

Ví dụ:

Giá trị mẫu quan sát là (5; 1; 8; 5; 3; 8; 9; 7; 5; 1; 8; 3), cỡ mẫu n = 12, số liệu được thu gọn lại có dạng:

x_i	1	3	5	7	8	9
n_i	2	2	3	1	3	1

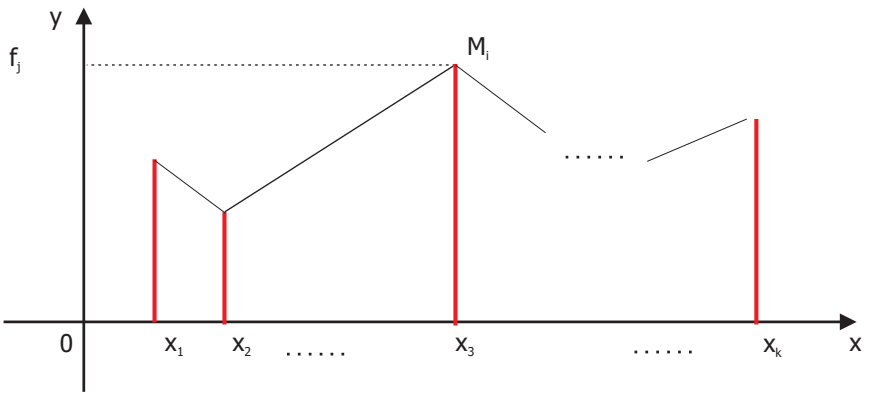
Đặt $f_i = \frac{n_i}{n}$ và gọi đó là tần suất của x_i trong mẫu, khi đó ta có bảng biểu diễn tần suất mẫu.

x_i	x_1	x_2	…	x_n
n_i	f_1	f_2	…	f_k

Ta có:

$$f_1 + f_2 + \ldots + f_k = (n_1 + n_2 + \ldots + n_k)/n = 1.$$

Trên trục tọa độ 0xy ta biểu diễn các điểm $M_i(x_i, f_i)$ và nối các điểm M_i với nhau ta được một biểu đồ tần suất trong Hình 1.



Hình 1: Trình bày mẫu bằng biểu đố tần suất

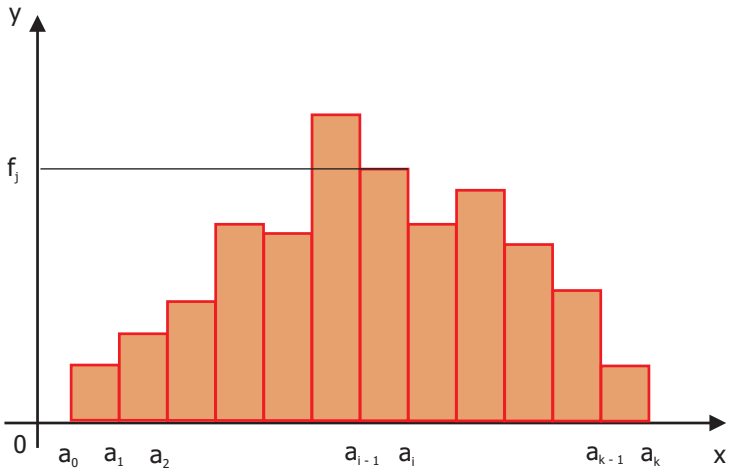
- Tổ chức đồ (biểu đồ tần số)

  Chia miền giá trị của mẫu thành k khoảng $(a_0; a_1]$, $(a_1; a_2]$, ... , $(a_{k-1}; a_k]$, ký hiệu n_i là số các giá trị mẫu rơi vào khoảng $(a_{i-1}; a_i]$, (i=1,2,..,k). Ta biểu diễn mẫu dưới dạng:

Khoảng	$[a_0 - a_1]$	$[a_1 - a_2]$	…	$[a_{k-1} - a_k]$
n_i	n_1	n_2	…	n_k

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

n_i là số giá trị mẫu rơi vào khoảng $(a_{i-1}; a_i]$. Trong mặt phẳng Oxy, trên trục Ox biểu diễn các khoảng $(a_{i-1}; a_i]$, trên trục Oy biểu diễn các giá trị $y_i = n_i /(n.h_i)$, trong đó h_i là độ dài khoảng $(a_{i-1}; a_i]$, i =1,2,..k . Ta dựng các hình chữ nhật có chiều cao là y_i và độ dài đáy là h_i. Hình được tạo bởi các hình chữ nhật trên được gọi là tổ chức đồ (biểu đồ tần số).



Hình 2: Tổ chức đồ