Giới Thiệu

Linh Truong

Jan, 2024

Nội dung

- THỐNG KÊ?
- DỮ LIỆU & BIẾN
- TÓM TẮT

Mục tiêu

1. Hiểu được tư duy Thống Kê. Ý nghĩa và quy trình làm thống kê?
2. Biết được một số khái niệm căn bản trong Thống Kê

Nội dung

- Giới thiệu môn học
- Thống Kê
- Dữ liệu & Biến

Ví dụ

Sức khỏe > Tin tức

Thứ tư, 27/12/2023, 16:35 (GMT+7)

Dân số Việt Nam còn 3,6 triệu người vào năm 2500 nếu vẫn giảm sinh

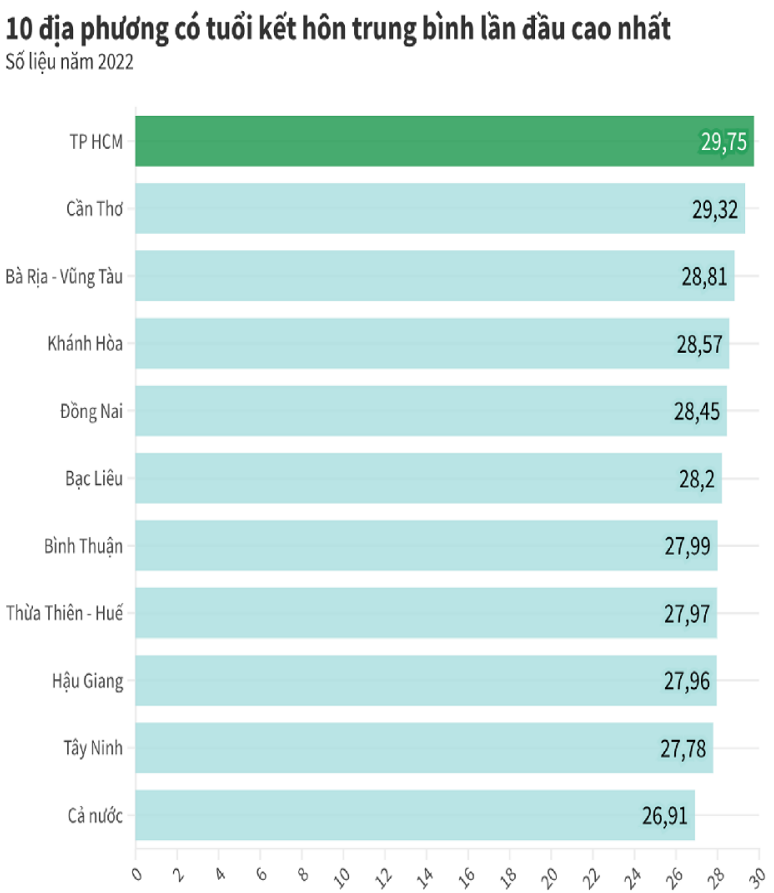
208

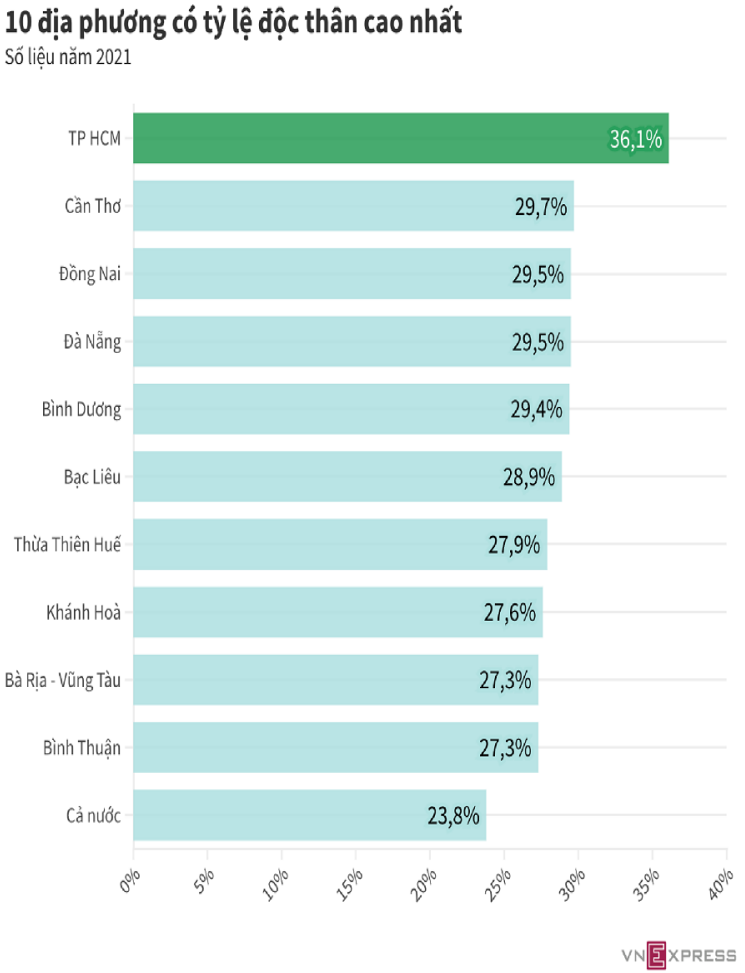
Liên Hợp Quốc cảnh báo đến năm 2500 dân số Việt Nam chỉ còn 3,6 triệu người, bằng số dân tỉnh Nghệ An hiện nay, nếu mức sinh tiếp tục giảm.

Nếu không duy trì được mức sinh thay thế và tiếp diễn mức sinh thấp, đến năm 2700, dân số Việt Nam sẽ chỉ còn vài chục nghìn người, theo cảnh báo của Liên Hợp Quốc. Ông Mai Trung Sơn, Phó Vụ trưởng Quy mô dân số - Kế hoạch hóa gia đình, Cục Dân số, hôm 26/12 dẫn cảnh báo trên, trong bối cảnh mức sinh năm 2023 của Việt Nam tiếp tục giảm, tỷ suất sinh chỉ 1,95 ca sinh/phụ nữ so với con số năm 2022 là 2,01 và kế hoạch mức sinh thay thế là 2,1.

Vùng Đông Nam bộ và Đồng bằng sông Cửu Long, mức sinh tiếp tục xuống sâu (khoảng 1,5 con/phụ nữ). Mức sinh tại TP HCM là 1,27 con/phụ nữ, thấp nhất cả nước.

Ví dụ


10 địa phương có tuổi kết hôn trung bình lần đầu cao nhất
Số liệu năm 2022
TP HCM 29,75
Cần Thơ 29,32
Bà Rịa - Vũng Tàu 28,81
Khánh Hòa 28,57
Đồng Nai 28,45
Bạc Liêu 28,2
Bình Thuận 27,99
Thừa Thiên - Huế 27,97
Hậu Giang 27,96
Tây Ninh 27,78
Cả nước 26,91
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
Nguồn: Tổng cục Thống kê


10 địa phương có tỷ lệ độc thân cao nhất
Số liệu năm 2021
TP HCM 36,1%
Cần Thơ 29,7%
Đồng Nai 29,5%
Đà Nẵng 29,5%
Bình Dương 29,4%
Bạc Liêu 28,9%
Thừa Thiên Huế 27,9%
Khánh Hoà 27,6%
Bà Rịa - Vũng Tàu 27,3%
Bình Thuận 27,3%
Cả nước 23,8%
0% 5% 10% 15% 20% 25% 30% 35% 40%
VNEXPRESS

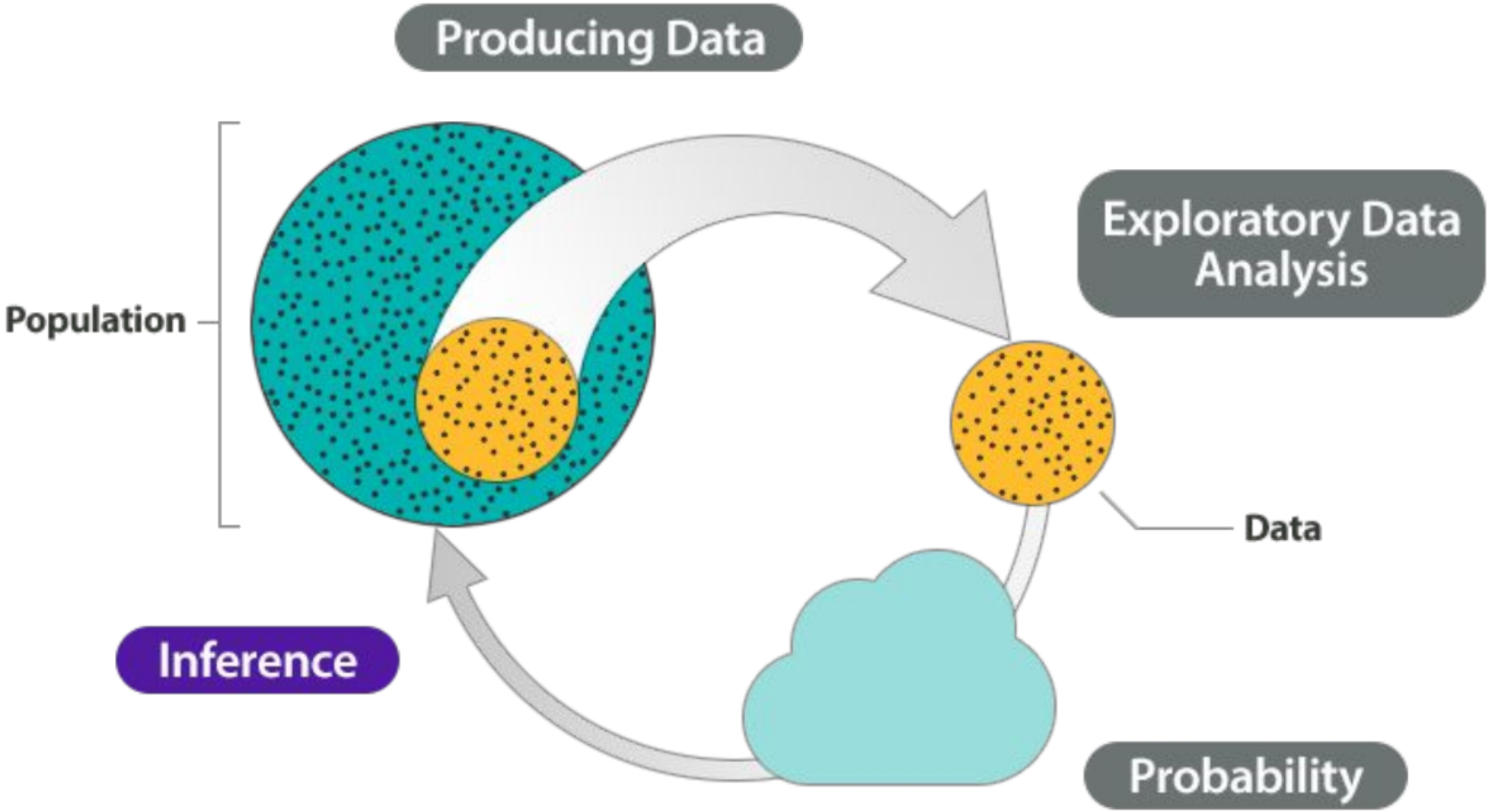
THỐNG KÊ?

Khái niệm

- **Thống Kê** là ngành khoa học liên quan đến việc phát triển các phương pháp khác nhau để lập kế hoạch nghiên cứu, thu nhập, phân tích, diễn giải và trình bày dữ liệu thực nghiệm, từ đó có thể đưa ra kết luận hoặc dự báo.
- Hiểu một cách đơn giản, thống kê là rút ra kết luận cho một nhóm lớn từ một số thành phần của nhóm đó.

Ví dụ

- Nghe bạn bè nói, học môn tự chọn lập trình phân tán với Java rất dễ. Có lẽ mình nên chọn học môn này
- Các trận đấu gần đây, Manchester United toàn thua còn Aston Villa chỉ có thắng hoặc hòa. Trận đấu đêm nay sẽ là một chiến thắng nữa dành cho Aston Villa.
- Theo khảo sát, thời gian sử dụng điện thoại trung bình trong lớp học là 3.5 giờ. Có thể thời gian sử dụng điện thoại trung bình của sinh viên trong trường cũng gần với 3.5 giờ.


Producing Data
Population
Exploratory Data Analysis
Data
Inference
Probability

Khái niệm

- **Tổng thể (Population)** là tập hợp tất cả các phần tử hoặc các sự kiện có liên đến cùng vấn đề hoặc thử nghiệm cần tìm hiểu
- **Mẫu (Sample)** là tập con của Tổng thể, được chọn để đại diện cho tổng thể trong phân tích thống kê

Khái niệm

- **Tham số (Parameter)** là đại lượng mô tả đặc trưng của tổng thể
- **Thống Kê (Statistic/ Sample Statistic)** là đại lượng mô tả đặc trưng mẫu

Ví dụ

- Có thể dùng **trung bình mẫu** (thống kê) để ước lượng giá trị **trung bình của tổng thể** (tham số)
- Có thể dùng ***tỷ lệ tốt nghiệp của một lớp học*** (thống kê) để ước lượng ***tỷ lệ tốt nghiệp của trường*** (tham số)

Phân loại

- **Thống kê mô tả (Descriptive Statistics)** là quá trình mô tả, tóm tắt những đặc trưng từ những thông tin thu nhận được (mẫu)
  - Một số đặc trưng : xu hướng tập trung, độ phân tán, hình dạng phân phối...
  - Có hai cách: mô tả bằng biểu đồ hoặc mô tả bằng một số độ đo
- **Thống kê suy diễn (Inferential Statistics)** là một quá trình phân tích dữ liệu để suy diễn nhằm tìm ra những thuộc tính của tổng thể.
  - Để suy diễn cần phải dựa trên *lý thuyết xác suất*
  - Một số bài toán suy diễn như: *ước lượng, kiểm định, hồi quy, phân tích phương sai...*

Quy trình thực hiện

- B1: Xác định câu hỏi: Mục tiêu của việc điều tra nghiên cứu là gì?
- B2: Xác định dữ liệu phù hợp
- B3: Lấy mẫu (Chọn dữ liệu và phương pháp phù hợp)
- B4: Phân tích dữ liệu thu được
- B5: Đưa ra kết luận hoặc dự đoán

Thu nhập dữ liệu mẫu

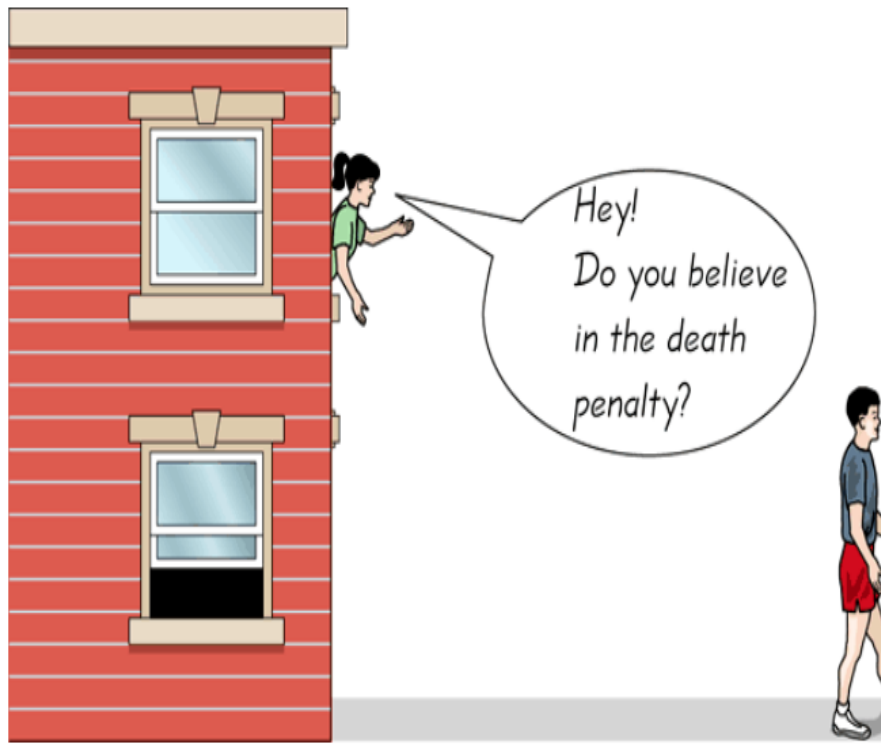
Dữ liệu mẫu thông thường được lấy từ hai nguồn:

Từ quan sát (observational study): không tác động vào đối tượng nghiên cứu VD: quan sát tình hình thời tiết

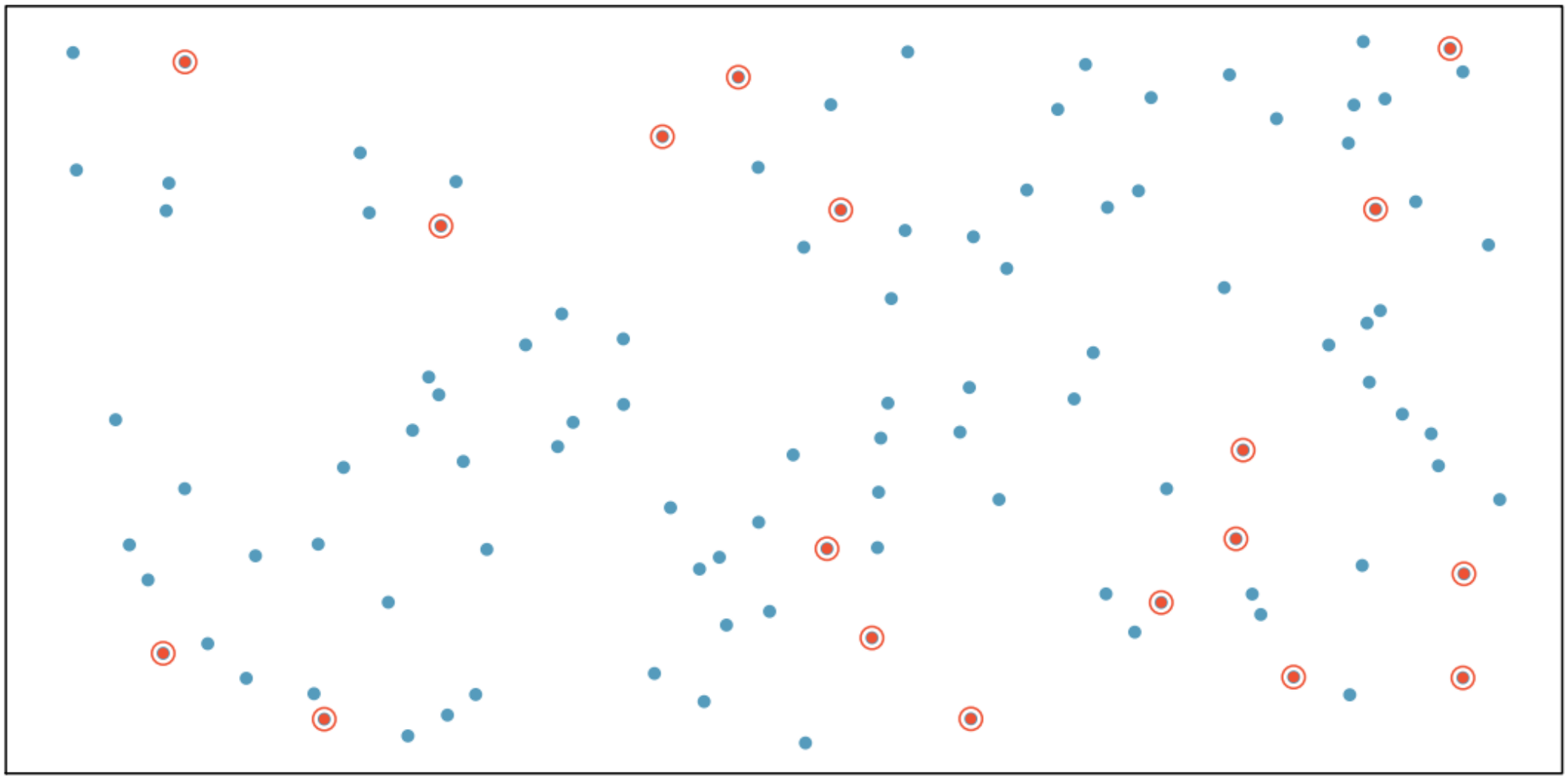
Từ thực nghiệm (experiment): có tác động vào đối tượng nghiên cứu VD: để kiểm tra tính hiệu quả của một loại thuốc, người ta thường chia nhóm người tình nguyện ra làm hai nhóm: một nhóm dùng thuốc thật, và một nhóm dùng giả dược

Một số phương pháp lấy mẫu

- Lấy mẫu ngẫu nhiên đơn giản (**Simple Random Sampling**)
- Lấy mẫu có hệ thống (**Systematic Sampling**)
- Lấy mẫu theo phân lớp (**Stratified Sampling**)
- Lấy mẫu phân cụm (**Clustered Sampling**)
- Lấy mẫu nhiều giai đoạn (**Multistage Sampling**)



Simple Random Sampling

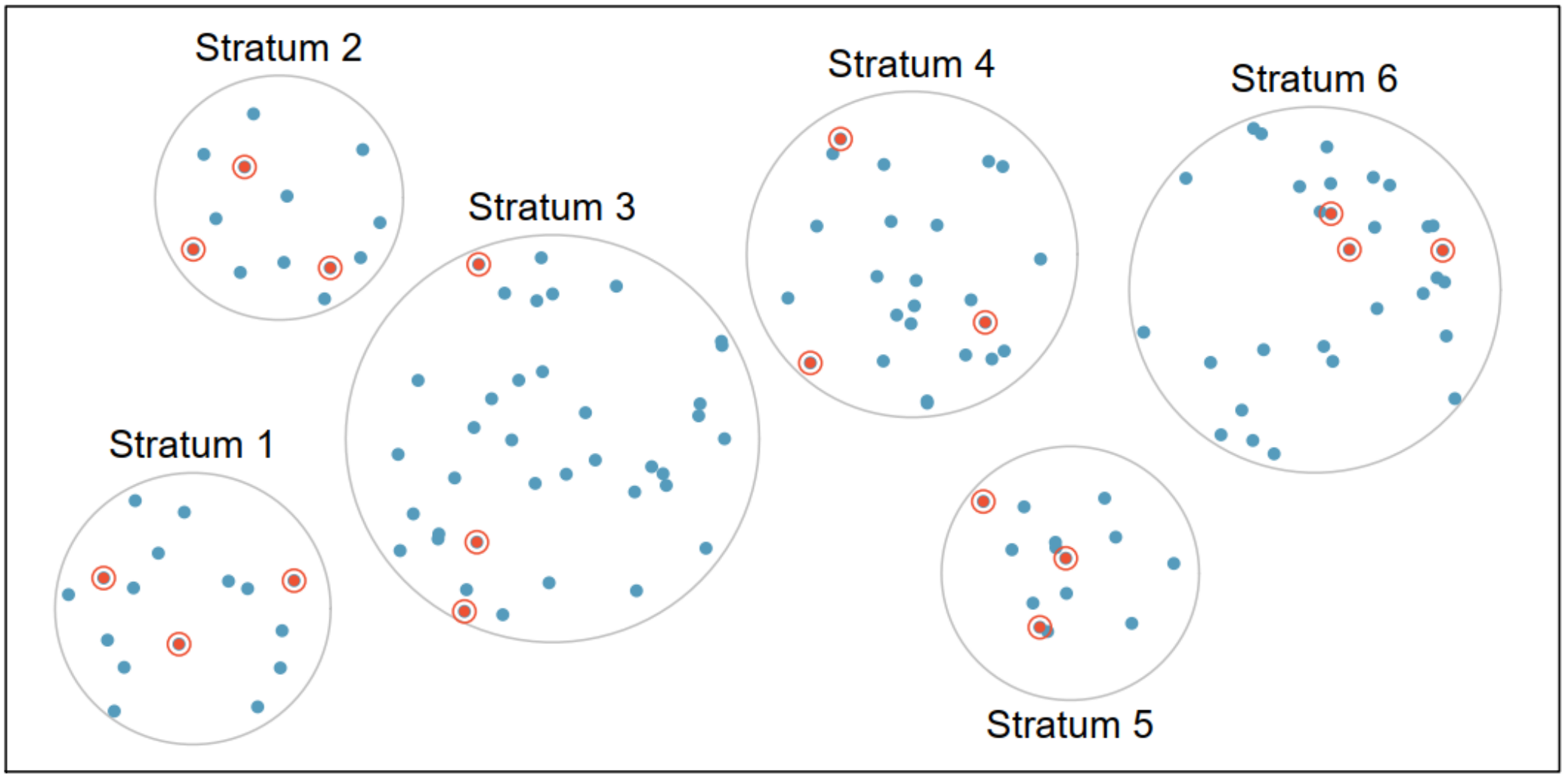


Lấy mẫu ngẫu nhiên đơn giản

Systematic Sampling

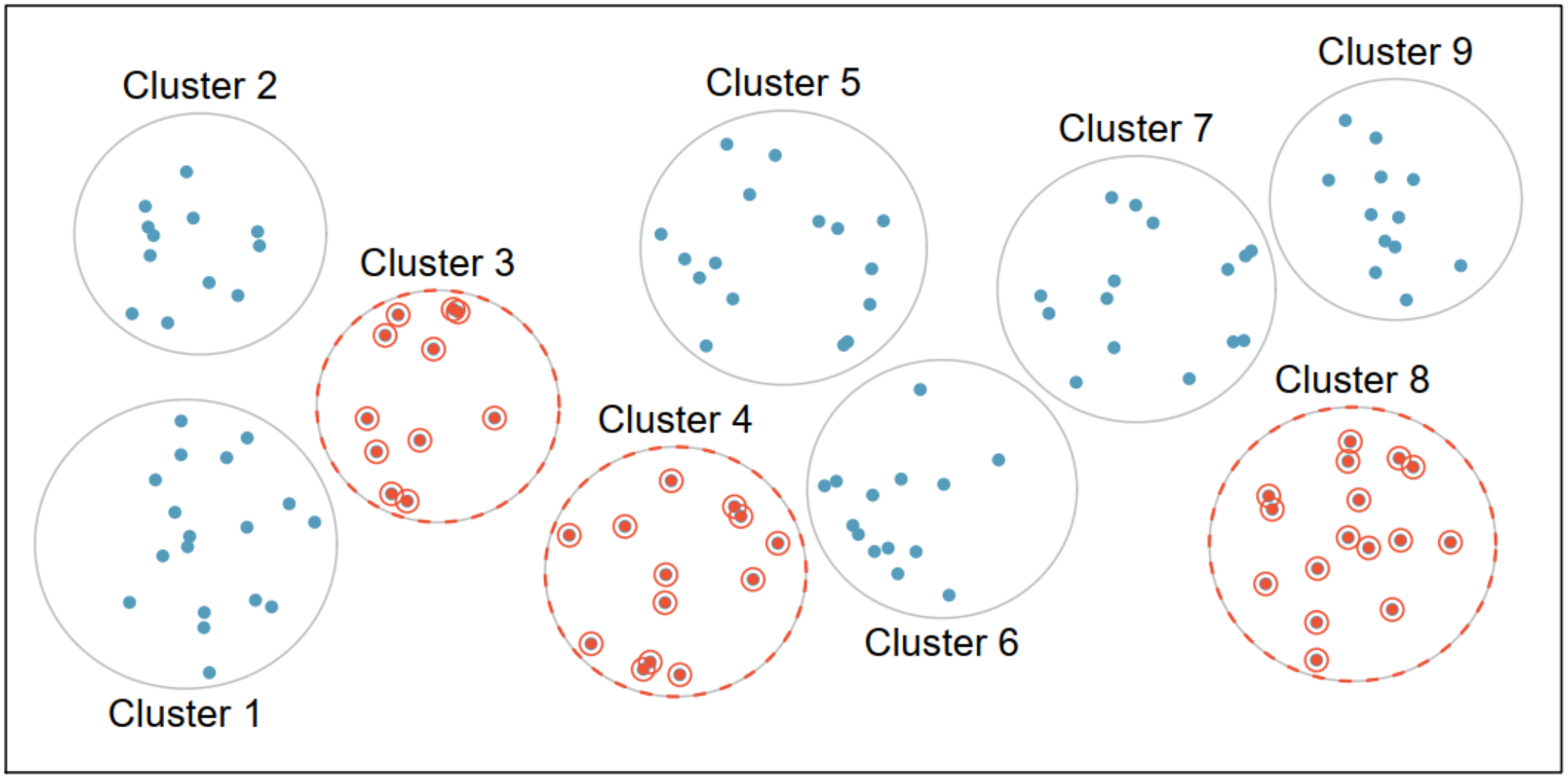
Lấy mẫu hệ thống

Stratified Sampling



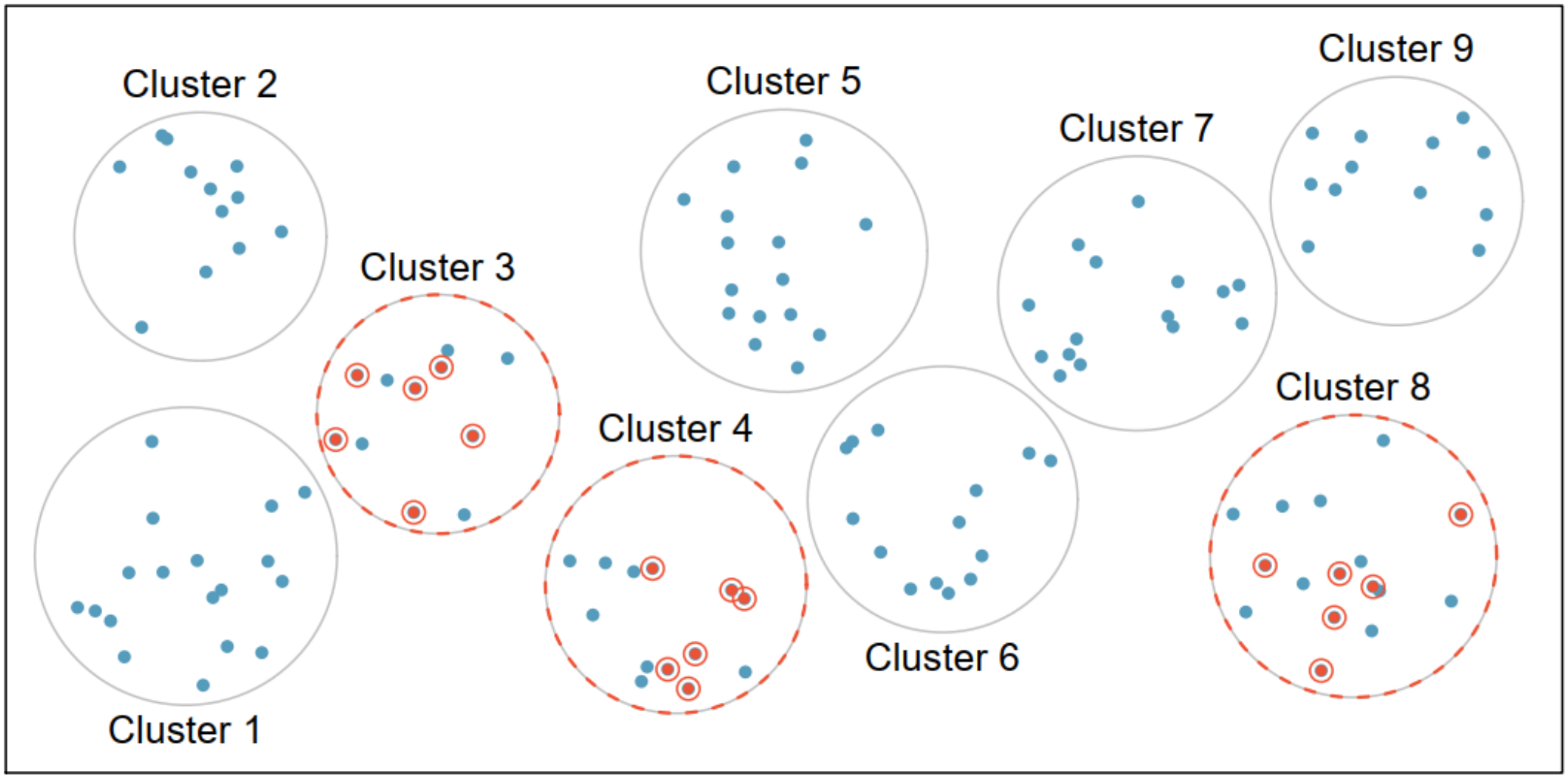
Lấy mẫu phân lớp

Clustered Sampling



Lấy mẫu phân cụm

Multistage Sampling



Lấy mẫu theo nhiều giai đoạn

DỮ LIỆU & BIẾN

Dữ liệu

Dữ liệu (data) được thu nhập từ tổng thể hoặc mẫu, thường được tổ chức ở dạng bảng

- Mỗi dòng là một quan sát (**observation**) hoặc một bản ghi (**record**)
- Mỗi cột là một biến (**variable**)

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Figure 1.3: Four rows from the `loan50` data matrix.

Biến

Biến (variable) có thể xem là một thuộc tính hoặc một đặc trưng của dữ liệu

variable	description
loan_amount	Amount of the loan received, in US dollars.
interest_rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always set as a whole number of months.
grade	Loan grade, which takes a values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total_income	Borrower's total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

Figure 1.4: Variables and their descriptions for the loan50 data set.

Phân loại

Các giá trị dữ liệu có thể được phân làm 2 loại:

- **Dữ liệu định tính (qualitative data/ categorical data)**
  - **Có thứ tự**: thứ hạng, mức độ hài lòng...
  - **Không có thứ tự**: màu sắc, giới tính, có hút thuốc?, ...
- **Dữ liệu định lượng (quatitative data/ numerical data)**
  - **Dữ liệu rời rạc (discrete)**: vị trí công việc, xếp loại học lực
  - **Dữ liệu liên tục (continuos)**: chiều cao, cân nặng, mức lương...

Cho biết các biến dưới đây thuộc loại nào?

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Figure 1.3: Four rows from the `loan50` data matrix.

- Định tính: ...
- Định lượng: ...

TÓM TẮT

- Biết được vai trò và quy trình giải quyết vấn đề bằng tư duy duy **Thống Kê**
- Phân biệt được một số khái niệm căn bản trong **Thống Kê** (population & sample, parameter & statistic, data, variables)
- Biết được một số phương pháp lấy mẫu