Chương

01

GIỚI THIỆU

Các cuộc bầu cữ, các nghiên cứu, các khảo sát và các công việc liên quan đến việc thu nhập dữ liệu từ một nhóm lớn; để rồi từ dữ liệu ấy chúng ta có thể đưa ra một quyết định, một kết luận về nhóm lớn ấy.

Đó chính là mục đích của thống kê.

- Mục tiêu của thống kê đó là rút ra được một kết luận gì đó từ một nhóm lớn bằng cách kiểm tra *một số* thành phần của nhóm đó.

- Trong nội dung chương này, chúng ta sẽ xem xét một số khái niệm căn bản của thống kê, quy trình giải quyết một bài toán thống kê và cuối cùng chúng ta sẽ bắt đầu giải quyết bài toán thống kê bằng cách mô tả dữ liệu và thảo luận về cách lấy mẫu.

NỘI DUNG

- Một số khái niệm
- Tư duy thống kê
- Kiểu dữ liệu
- Thu nhập dữ liệu mẫu

NỘI DUNG

- **Một số khái niệm**
- Tư duy thống kê
- Kiểu dữ liệu
- Thu nhập dữ liệu mẫu

Dữ liệu (data)

- Dữ liệu (data): là tập hợp các quan sát/ quan trắc.

- Ví dụ: các kết quả đo lường, giới tính, hoặc các kết quả khảo sát…

Thống kê (Statistics)

- Thống kê (statistics): là môn khoa học về việc lập kế hoạch nghiên cứu và thí nghiệm và thu nhập dữ liệu sau đó tổ chức, tổng hợp, trình bày, phân tích, diễn giải để rút ra kết luận dựa trên dữ liệu thu nhập được.

Quần thể (Population)

- Quần thể (Population): là tập hợp tất cả các giá trị dữ liệu cần được xem xét

Mẫu (sample)

- Mẫu (sample): là một tập con của quần thể

Ví dụ

- Công ty Gallup khảo sát 1013 người trưởng thành ở Mỹ. Kết quả có 66% người phản hồi lo lắng về hành vì đánh cắp thông tin cá nhân.
- Quần thể bao gồm: 241,742,385 người trưởng thành ở Mỹ
- Mẫu gồm: 1013 người được khảo sát
- Mục tiêu của khảo sát dùng từ dữ liệu thu nhập được để rút ra kết luận về toàn bộ quần thể

NỘI DUNG

- Một số khái niệm
- **Tư duy thống kê**
- Kiểu dữ liệu
- Thu nhập dữ liệu mẫu

- Trong phần này chúng ta sẽ xem xét tổng quan một quy trình giải quyết một bài toán thống kê.
- Quy trình gồm 3 bước:
  - Chuẩn bị
  - Phân tích
  - Kết luận

CHUẨN BỊ - Ngữ cảnh

- Câu hỏi thống kê là gì? hoặc mục tiêu của việc điều tra, nghiên cứu là gì?

- Dữ liệu có ý nghĩa gì?

CHUẨN BỊ - nguồn của dữ liệu

- Nguồn dữ liệu thu nhập được có khách quan hay không?
- Nguồn dữ liệu thu nhập được có bị lệch hay không?
- Hãy luôn thận trọng và hoài nghi về nguồn dữ liệu có được, nguồn dữ liệu có thể bị sai lệch

CHUẨN BỊ - Phương pháp lấy mẫu

- Phương pháp lấy mẫu có ảnh hưởng đến tính hợp lý của kết luận hay không?

- Hãy cẩn thận, những người tình nguyện tham gia các khảo sát thông thường sẽ cho ra các kết quả sai lệch (những người tình nguyện có thể có cùng một lý do tham gia)

- Hãy thữ những phương pháp lấy mẫu khác nhau để cho ra kết quả tốt

PHÂN TÍCH – Trực quan hóa dữ liệu

- Công việc phân tích nên bắt đầu bằng việc trực quan hóa dữ liệu bằng cách sử dụng các biểu đồ thích hợp.

- Trực quan hóa, giúp người phân tích có cái nhìn ban đầu cũng như cảm giác về dữ liệu.

PHÂN TÍCH – Áp dụng các phương pháp thống kê

- Trong nội dung những chương sau, chúng ta sẽ thảo luận về các phương pháp thống kê được dùng để phân tích.

- Hiện nay, với công nghệ và máy tính chúng ta có thể thực hiện việc phân tích mà không cần đến kỹ năng tính toán quá nhiều; tuy nhiên để sử dụng được các công nghệ trên, chúng ta phải làm quen với các khái niệm cũng như những phương pháp thống kê căn bản.

KẾT LUẬN

- Sau khi thực hiện phân tích bằng thống kê, việc cuối cùng là đưa ra kết luận. Tuy nhiên, cần phải chú ý rằng kết luận rút ra có ý nghĩa trên thực tế hay không?

NỘI DUNG

- Một số khái niệm
- Tư duy thống kê
- **Kiểu dữ liệu**
- Thu nhập dữ liệu mẫu

DỮ LIỆU ĐỊNH LƯỢNG (Quantitative data)

- Dữ liệu định lượng (quantitative data or numerical data): là dữ liệu có thể đo đếm được.

- *Ví dụ:* tuổi, cân nặng, chiều cao, thu nhập…

- Dữ liệu định lượng có thể phân biệt 2 loại đó là dữ liệu có giá trị **rời rạc** hay dữ liệu có giá trị **liên tục**.

DỮ LIỆU ĐỊNH LƯỢNG (Quantitative data)

- Dữ liệu định lượng có thể phân biệt 2 loại đó là dữ liệu có giá trị rời rạc hay dữ liệu có giá trị liên tục.
  - Dữ liệu rời rạc (discrete data): giá trị dữ liệu là các số nguyên.
  - Dữ liệu liên tục (continuous data): giá trị dữ liệu là các số thực

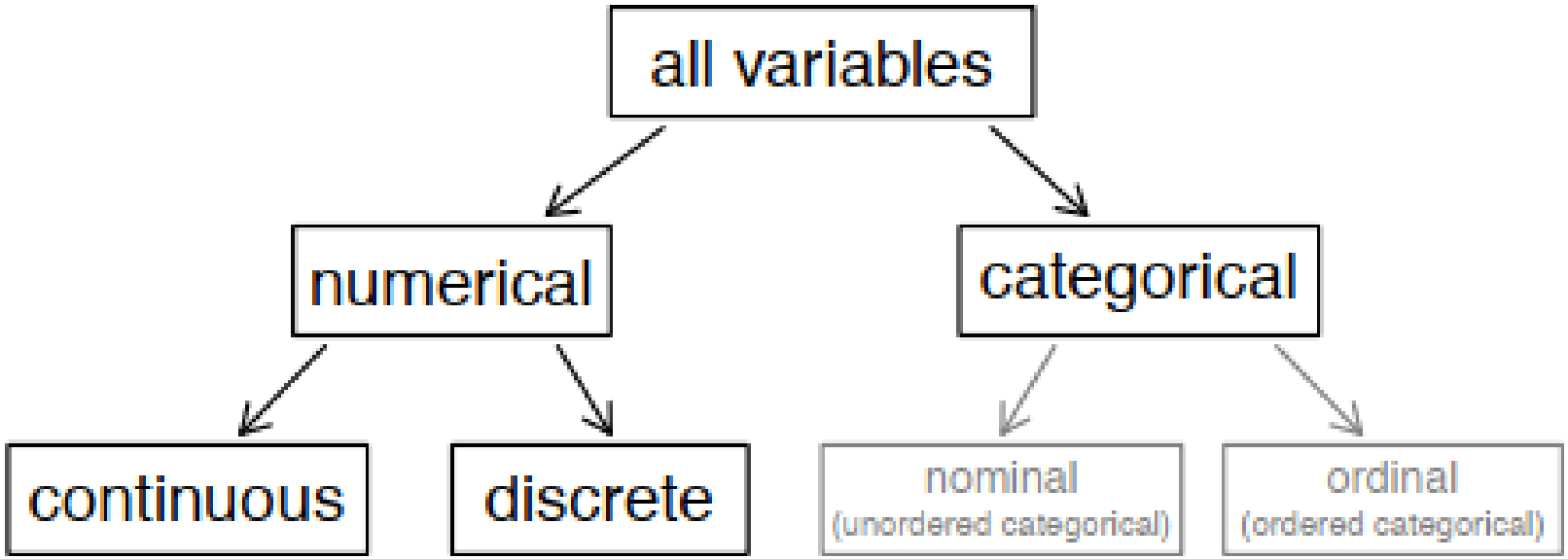
DỮ LIỆU ĐỊNH TÍNH (Qualitative data)

- Dữ liệu định tính hay dữ liệu phân loại (qualitative data or categorical data): là dữ liệu sử dụng để phân loại, giá trị của dữ liệu được sử dụng để đại diện cho một phân loại nào đó

- *Ví dụ:* giới tính, màu sắc, xếp hạng…

- Dữ liệu định tính có thể chia làm hai loại đó là: dữ liệu định tính **có thứ tự** và dữ liệu định tính **không có thứ tự**

KIỂU DỮ LIỆU


all variables
numerical
categorical
continuous
discrete
nominal
(unordered categorical)
ordinal
(ordered categorical)

NỘI DUNG

- Một số khái niệm
- Tư duy thống kê
- Kiểu dữ liệu
- **Thu nhập dữ liệu mẫu**

- Nếu dữ liệu mẫu không được thu nhập một cách phù hợp thì dù cho chúng ta thu nhập nhiều đến đâu đi chăng nữa nó cũng không có ý nghĩa.

- Phương pháp lấy mẫu ảnh hưởng đến chất lượng của phân tích thống kê

- Trong nội dung phần này, chúng ta sẽ thảo luận một số phương pháp lấy mẫu.

THU NHẬP DỮ LIỆU MẪU

- Dữ liệu mẫu thông thường được lấy từ hai nguồn:
  - Từ quan sát
  - Từ thử nghiệm

Từ quan sát (observational study)

- Từ quan sát (observational study): là dữ liệu được bằng cách quan sát và đo lường được mà tác động hay tìm cách thay đổi đối tượng nghiên cứu

- VD: Một khảo sát về thời gian sử dụng facebook của sinh viên IUH được tiến hành online cho kết quả rằng thời gian sử dụng trung bình facebook là 3.2 giờ.

→ Đây là dữ liệu thu được từ quan sát, vì người được khảo sát không chịu sự tác động nào

Từ thử nghiệm (experiment)

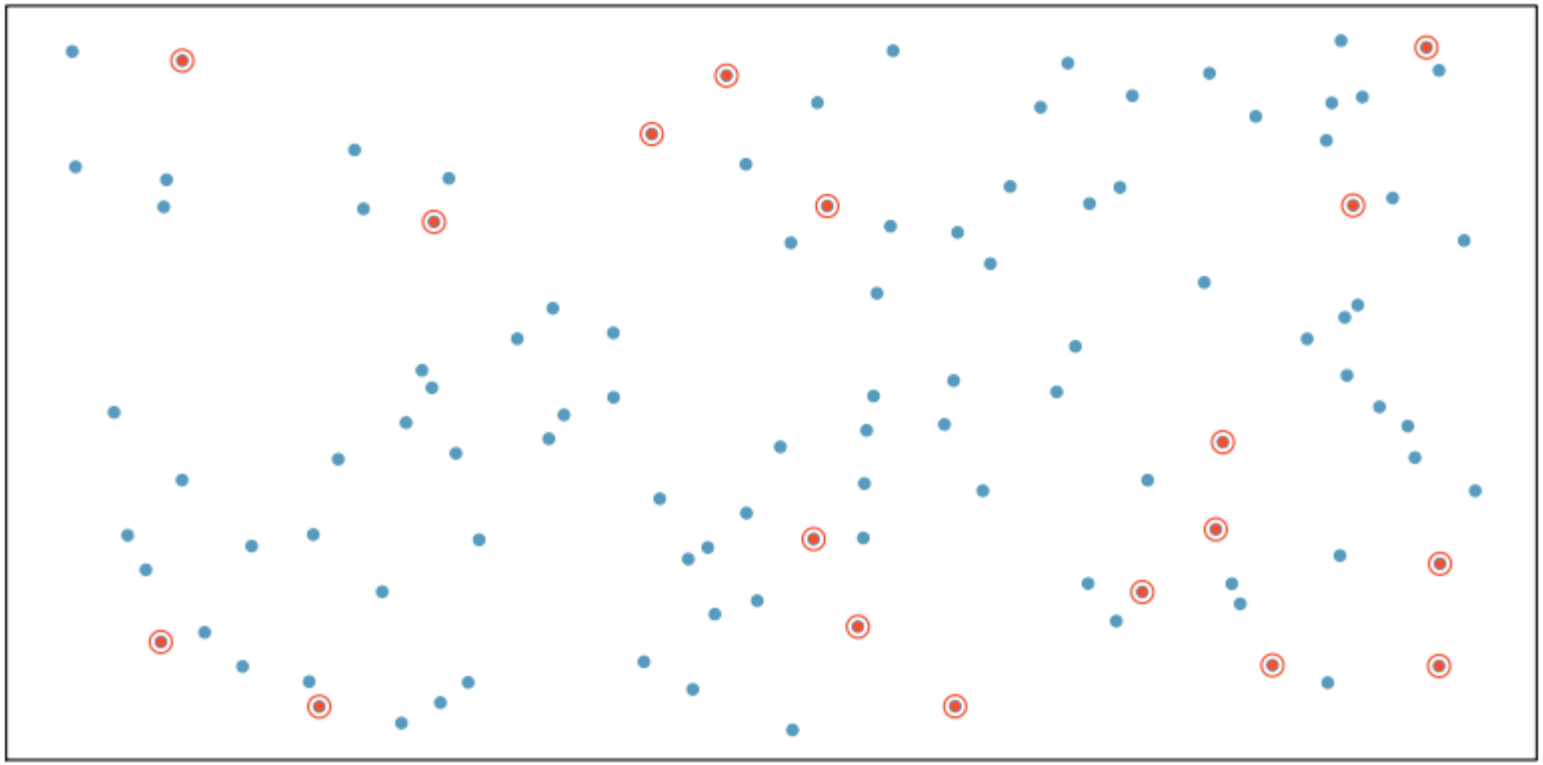
- Từ thử nghiệm (experiment): là dữ liệu được bằng cách áp dụng một số phương pháp tác động lên đối tượng cần nghiên cứu sau đó quan sát ảnh hưởng của phương pháp lên đối tượng đó.
- VD: Trong một thử nghiệm y tế cộng đồng, người ta tiêm cho 200.745 trẻ em loại vaccine X, và tiêm cho 201.229 trẻ em một loại vaccine giả dược (vaccine giả dược này không gây ảnh hưởng gì đến sức khỏe)

→ Trong ví dụ này, người ta đã chia đối tượng cần nghiên cứu ra làm hai nhóm cho nên đây là dữ liệu thu được từ thử nghiệm

Lấy mẫu ngẫu nhiên đơn giản (simple random sample)

- **Lấy mẫu ngẫu nhiên đơn giản (simple random sample):** là cách lấy mẫu mà mỗi giá trị dữ liệu được lấy từ quần thể theo *cùng một cách* và *cơ hội được chọn* của mỗi giá trị dữ liệu là như nhau.

Lấy mẫu ngẫu nhiên đơn giản (Simple Random Sample)

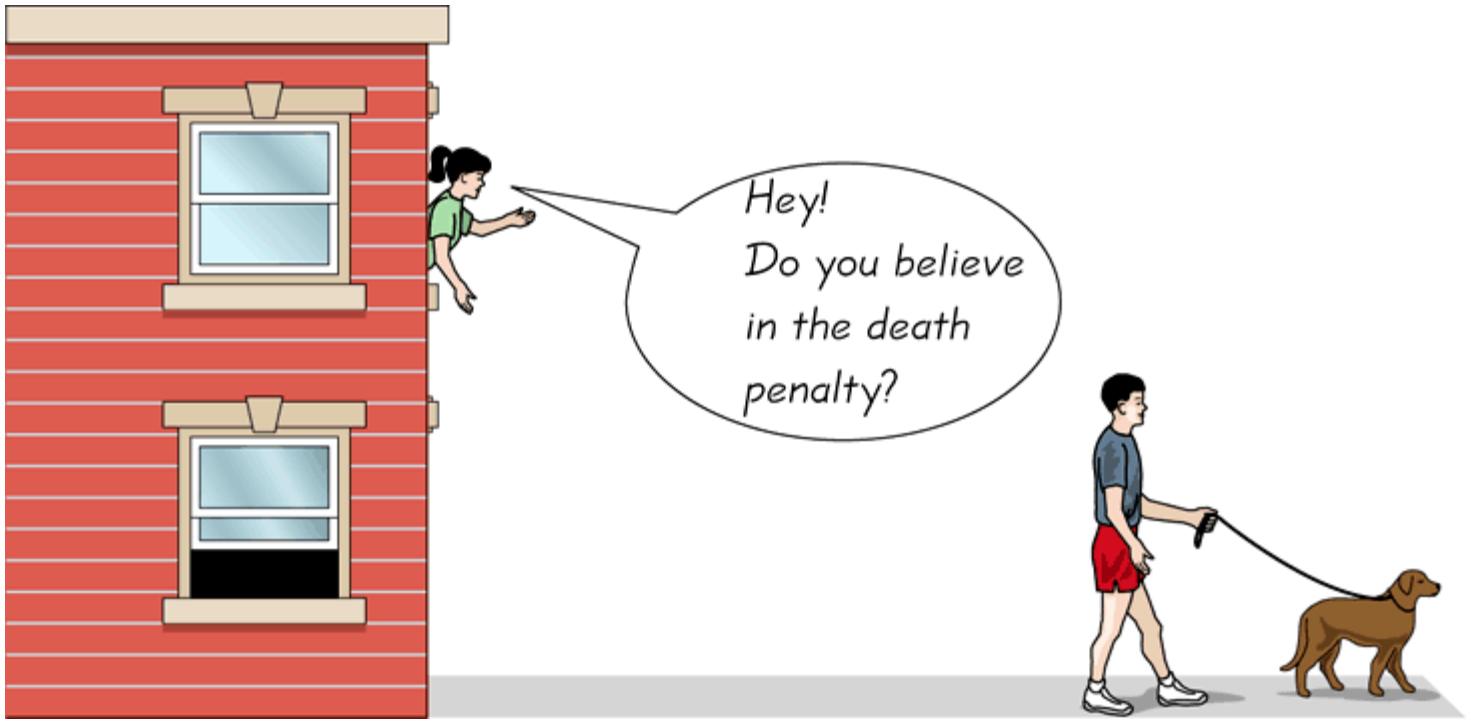


Lấy mẫu có hệ thống (Systematic Sampling)

- Lấy mẫu có hệ thống (Systematic Sampling): là lấy mẫu bằng cách chọn một điểm bắt đầu và điểm kết thúc, sau đó lần lượt chọn phần tử thứ k từ quần thể.

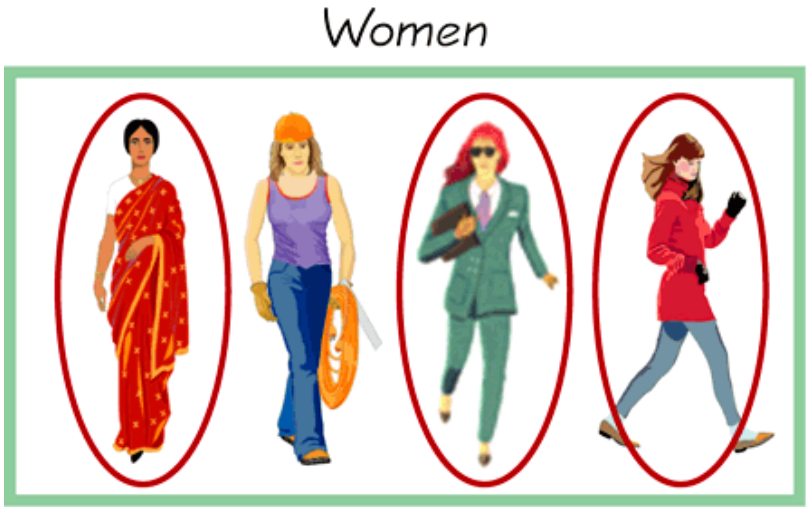
Lấy mẫu tiện lợi (Convenience Sampling)

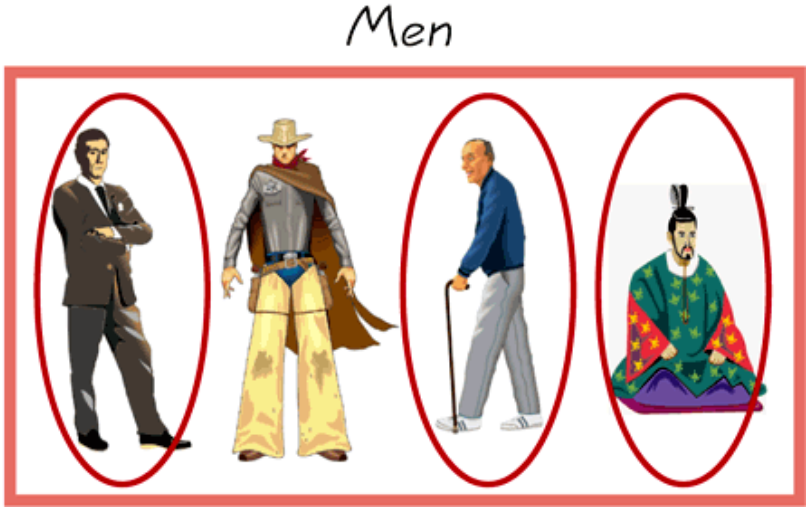
- **Lấy mẫu tiện lợi**: là cách lấy mẫu mà kết quả được thu nhập một cách dễ dàng



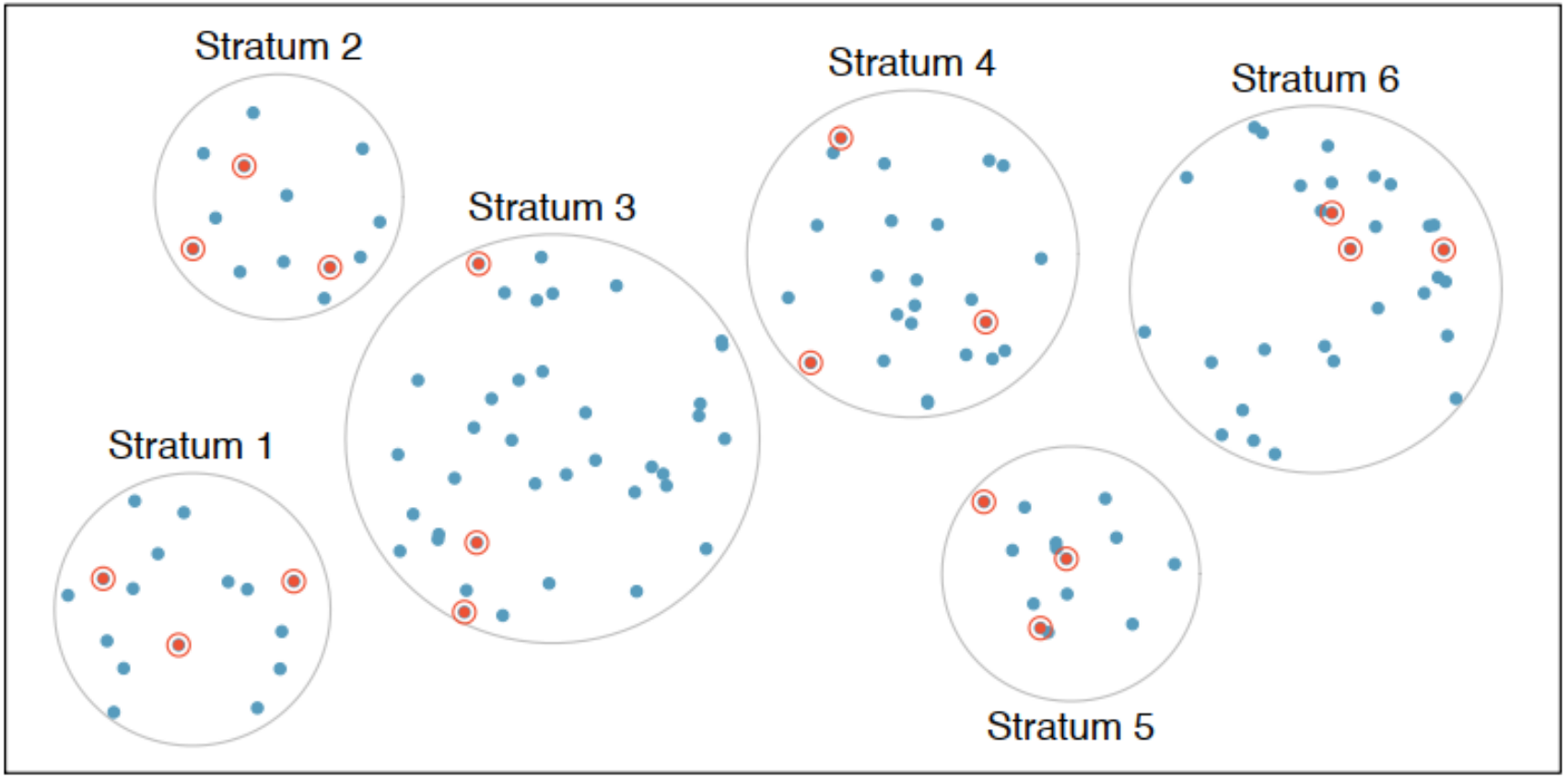
Lấy mẫu phân tầng (Stratified sampling)

- Lấy mẫu phân tầng (Straified sampling): chia quần thể thành nhiều nhóm nhỏ, mỗi nhóm có cùng đặc tính, sau đó lấy mẫu bằng cách chọn ngẫu nhiên từ các nhóm nhỏ đó.



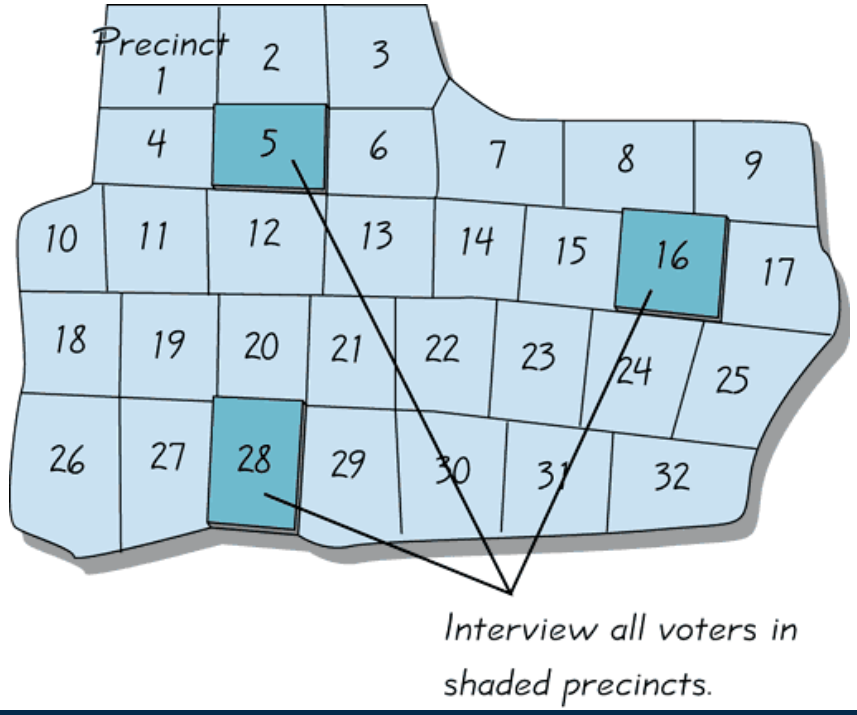


Lấy mẫu phân tầng (Stratified sampling)

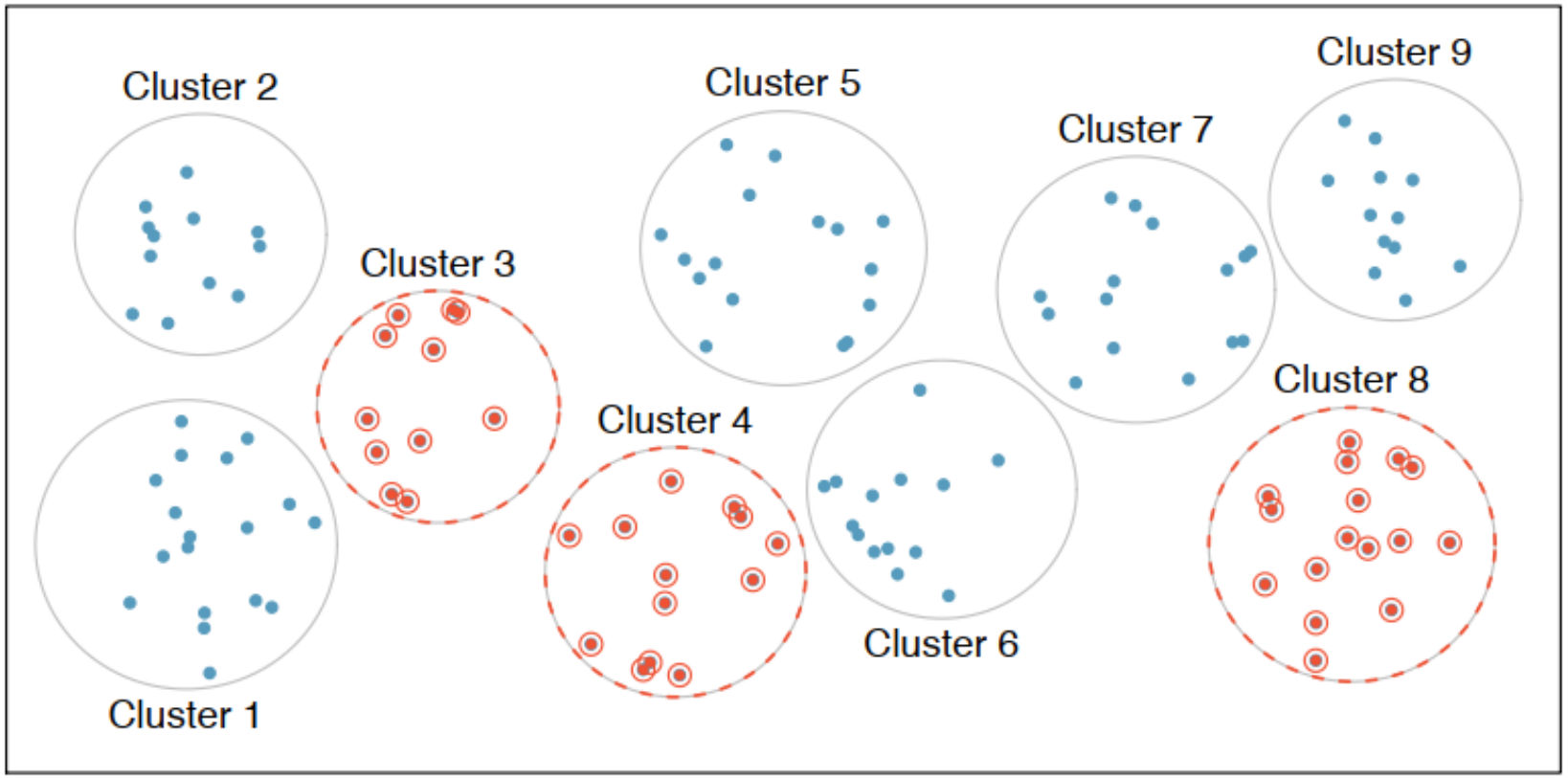


Lấy mẫu theo cụm (Cluster Sampling)

- Lấy mẫu theo cụm (Cluster Sampling): chia quần thể thành nhiều cụm nhỏ, chọn ngẫu nhiên một số cụm và lấy tất cả các dữ liệu của các cụm được chọn.



Lấy mẫu theo cụm (Cluster Sampling)

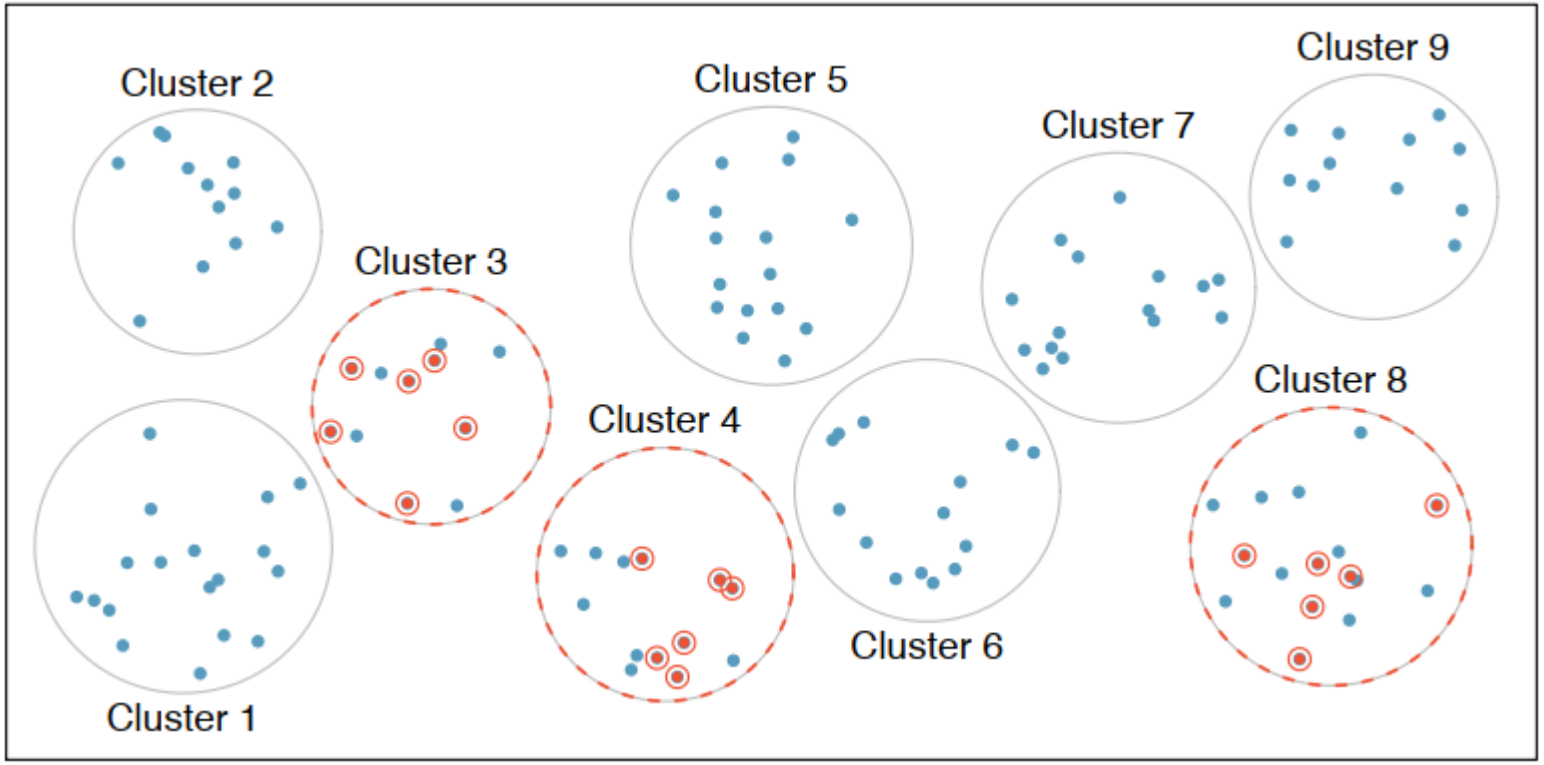


Lấy mẫu nhiều giai đoạn (Multistage Sampling)

- Lấy mẫu nhiều giai đoạn (Multisatage Sampling): là phương pháp lấy mẫu bằng cách kết hợp nhiều phương pháp lấy mẫu đơn giản với nhau

Lấy mẫu nhiều giai đoạn (Multistage Sampling)

- *Ví dụ:* một ví dụ về lấy mẫu nhiều giai đoạn



THU NHẬP DỮ LIỆU MẪU – CÁC PHƯƠNG PHÁP

- Lấy mẫu ngẫu nhiên đơn giản
- Lấy mẫu có hệ thống
- Lấy mẫu tiện lợi
- Lấy mẫu phân tầng
- Lấy mẫu theo cụm
- Lấy mẫu nhiều giai đoạn

THANK YOU