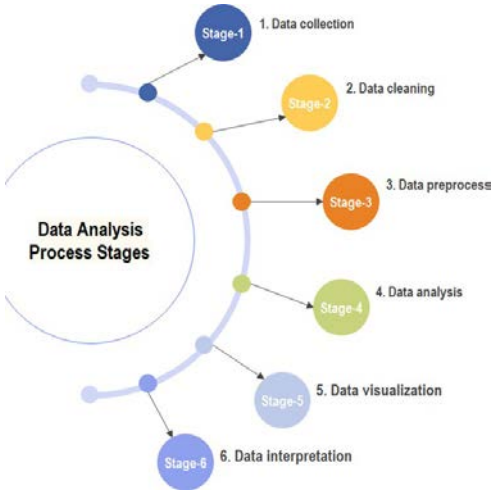
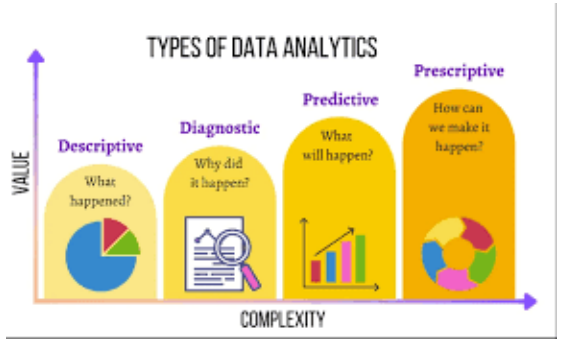
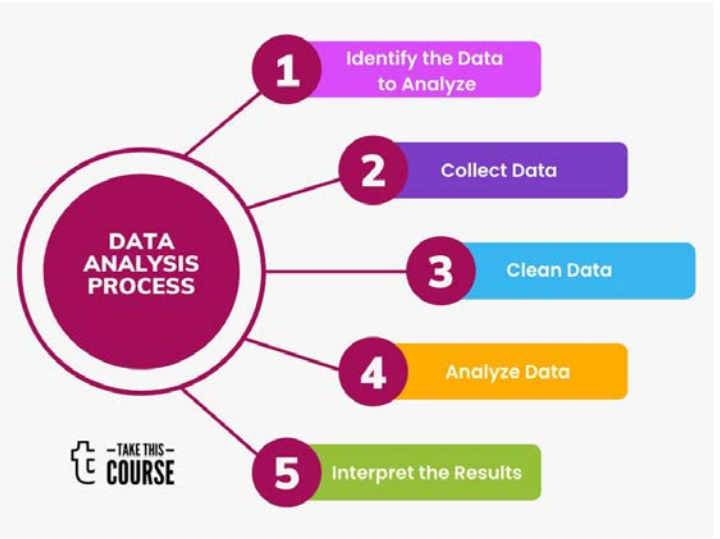




PHÂN TÍCH DỮ LIỆU

(Data Analysis)



Lê Văn Hạnh
levanhanhvn@gmail.com

NỘI DUNG MÔN HỌC

PHẦN 1 TỔNG QUAN & THU THẬP DỮ LIỆU CHO VIỆC PHÂN TÍCH

1. Khoa học dữ liệu
2. Thu thập dữ liệu
3. Tìm hiểu dữ liệu

PHẦN 2: TIỀN XỬ LÝ DỮ LIỆU *(Data Preprocessing)*

4. Nhiệm vụ chính trong tiền xử lý dữ liệu
5. PANDAS
6. Thao tác với các định dạng khác nhau của tập tin dữ liệu
7. Làm sạch và Chuẩn bị dữ liệu
8. Sắp xếp dữ liệu: nối, kết hợp và định hình lại
9. Tổng hợp dữ liệu và các tác vụ trên nhóm

PHẦN 3 TRỰC QUAN HÓA DỮ LIỆU *(Data Visualization)*

10. Đồ thị và Biểu đồ
11. Vẽ đồ thị và Trực quan hóa

PHẦN 1
TỔNG QUAN
&
THU THẬP DỮ LIỆU CHO VIỆC PHÂN TÍCH

Chương 2

THU THẬP DỮ LIỆU
(Data Collection)

Lê Văn Hạnh
levanhanhvn@gmail.com

NỘI DUNG CHƯƠNG 2

1. Những thách thức đối với các nhà khoa học dữ liệu
2. Chiến lược lấy mẫu để đảm bảo kết quả phù hợp
3. Thu thập dữ liệu định tính (*Qualitative Data*)
4. Thu thập dữ liệu định lượng (*Quantitative Data Collection*)
5. Bài thực hành

1. NHỮNG THÁCH THỨC ĐỐI VỚI CÁC NHÀ KHOA HỌC DỮ LIỆU

i. Nhiều nguồn dữ liệu

- Các loại ứng dụng và công cụ khác nhau tạo ra dữ liệu với nhiều định dạng khác nhau.
- Phải làm sạch và chuẩn bị dữ liệu để tạo sự nhất quán cho dữ liệu đó.

⇒ Hoạt động này có thể rất nhàm chán và tốn thời gian.

ii. Nắm rõ vấn đề kinh doanh

Phải làm việc với nhiều bên liên quan và các nhà quản lý doanh nghiệp để xác định vấn đề cần giải quyết.

⇒ Điều này có thể rất khó khan, đặc biệt là trong các công ty lớn với nhiều nhóm có các yêu cầu khác nhau.

iii.Loại bỏ thiên kiến

- Các công cụ máy học không hoàn toàn chính xác và do đó có thể tồn tại sự không chắc chắn hoặc thiên kiến. Thiên kiến là sự mất cân bằng trong dữ liệu đào tạo hoặc hành vi dự đoán của mô hình giữa các nhóm khác nhau, chẳng hạn như độ tuổi hoặc khung thu nhập.
- Ví dụ: nếu công cụ được đào tạo chủ yếu dựa trên dữ liệu từ các cá nhân trung niên thì công cụ này có thể kém chính xác hơn khi đưa ra các dự đoán liên quan đến những người trẻ tuổi và lớn tuổi hơn. Lĩnh vực máy học cung cấp cơ hội để giải quyết các thiên kiến bằng cách phát hiện và đo lường chúng trong dữ liệu và mô hình.

NỘI DUNG CHƯƠNG 2

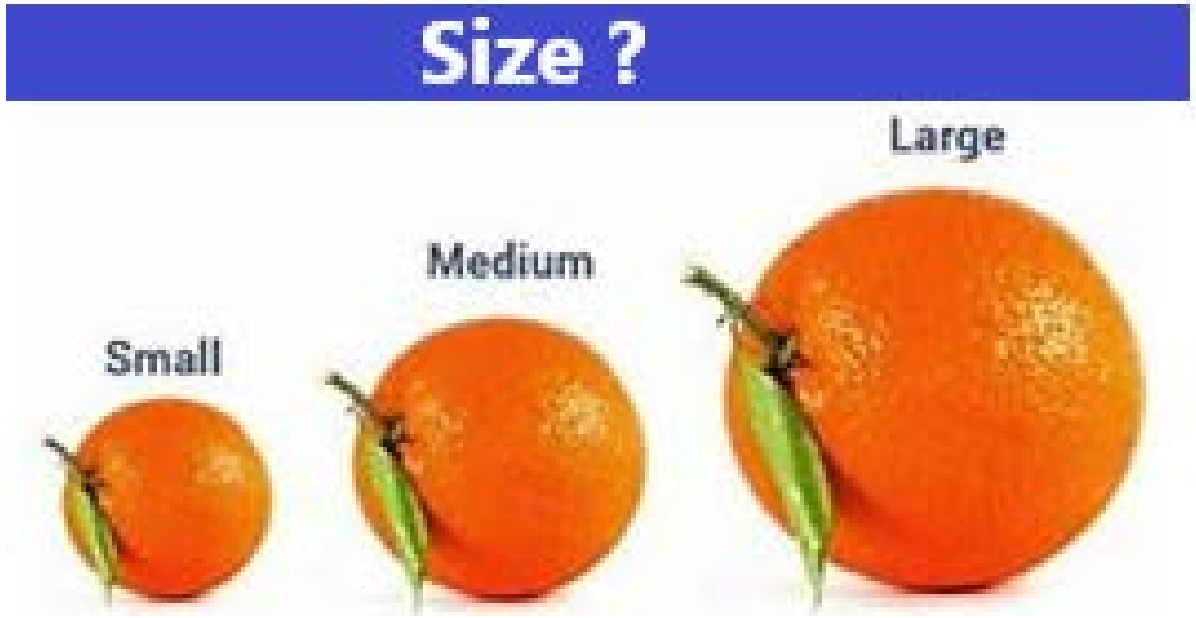
1. Những thách thức đối với các nhà khoa học dữ liệu
2. Chiến lược lấy mẫu để đảm bảo kết quả phù hợp
3. Thu thập dữ liệu định tính (*Qualitative Data*)
4. Thu thập dữ liệu định lượng (*Quantitative Data Collection*)
5. Bài thực hành

2. CHIẾN LƯỢC LẤY MẪU ĐỂ ĐẢM BẢO KẾT QUẢ PHÙ HỢP

2.1. Xác định kích thước mẫu (hay Cỡ mẫu - Sample size)?

Thực ra, không có quy tắc nghiêm ngặt nào cho việc lựa chọn cỡ mẫu. Có thể đưa ra quyết định về cỡ mẫu dựa trên:

- Mục tiêu của dự án
- Thời gian sẵn có
- Ngân sách
- Mức độ chính xác cần thiết.



2.1. Xác định kích thước mẫu (hay Cỡ mẫu - Sample size)?

2.1.1. Kích thước mẫu (hay Cỡ mẫu -Sample size)?

- Kích thước mẫu là số lượng câu trả lời hoàn chỉnh mà khảo sát nhận được. Nó phải đại diện cho nhóm đối tượng mục tiêu có ý kiến hoặc hành vi mà khảo sát quan tâm.
- Kích thước mẫu của nghiên cứu càng lớn, sai số trong các ước lượng sẽ càng thấp, khả năng đại diện cho tổng thể càng cao.
- Tuy nhiên, việc thu thập cỡ mẫu lớn sẽ làm tiêu tốn nhiều thời gian, công sức, tiền bạc ở toàn bộ các khâu từ thu thập, kiểm tra, phân tích.

  $\Rightarrow$ Việc chọn kích thước mẫu cần phải được xem xét một cách có cân nhắc để mọi thứ được cân bằng và hiệu quả.

2.1. Xác định kích thước mẫu (hay Cỡ mẫu - Sample size)?

2.1.2. Các thuật ngữ liên quan đến việc chọn cỡ mẫu

i. ***Population size*** (*kích thước của tổng thể*):

- Đại diện cho tổng số người trong nhóm cần nghiên cứu. Nếu đang khảo sát người dân ở Việt Nam, kích thước của tổng thể sẽ vào khoảng 100 triệu người (cuối năm 2024). Khi khảo sát 1 công ty, kích thước của tổng thể sẽ là tổng số nhân viên.
- Kích thước của mẫu nghiên cứu sẽ cần chiếm một tỷ lệ nhất định so với kích thước của tổng thể.

ii. ***The margin of error*** (*Biên độ sai số hay sai số cho phép*):

- Là tỷ lệ phần trăm cho thấy kết quả khảo sát phản ánh chính xác ý kiến của toàn bộ tổng thể như thế nào.
- Biên độ sai số càng thấp thì câu trả lời càng chính xác ở mức độ tin cậy nhất định.
- Thường ba tỷ lệ sai số hay sử dụng là: ±0.01 (1%), ±0.05 (5%), ±0.1 (10%), trong đó mức phổ biến nhất là ±0.05.

2.1. Xác định kích thước mẫu (hay Cỡ mẫu - Sample size)?

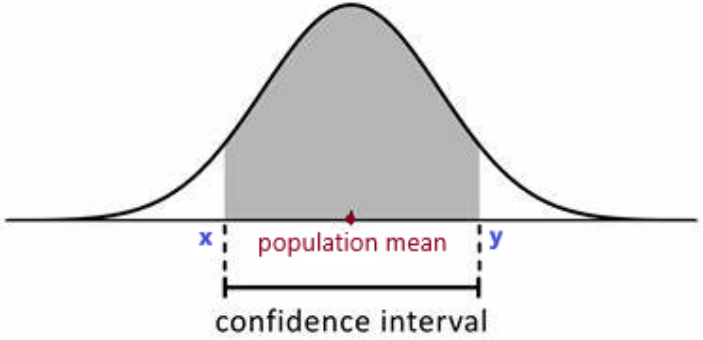
2.1.2. Các thuật ngữ liên quan đến việc chọn cỡ mẫu

iii. ***Confidence level*** (*Mức độ tin cậy*)

- Nghĩa là mức độ chắc chắn rằng các đặc điểm của cỡ mẫu được chọn phải khái quát được cho đặc điểm tổng thể.
- Ví dụ: với mức độ tin cậy về sự đồng thuận của người dân là 95% về 1 chủ trương của chính quyền, có nghĩa là khi chủ trương này được thực hiện, sẽ chắc chắn có 95% người dân ủng hộ.

2.1. Xác định kích thước mẫu (hay Cỡ mẫu - Sample size)?

2.1.2. Các thuật ngữ liên quan đến việc chọn cỡ mẫu



iv. ***Confidence interval*** (*Khoảng tin cậy*):

- Là mức độ chắc chắn rằng các đặc điểm của cỡ mẫu được chọn phải khái quát được cho đặc điểm tổng thể. Ví dụ: khoảng tin cậy 95% có nghĩa là có thể chắc chắn 95% kết quả nằm giữa số x và y. Khoảng tin cậy thường dùng là 95% hoặc 99%.
- Ví dụ: khi nghiên cứu chiều cao của các cầu thủ bóng rổ, các nhà nghiên cứu lấy một mẫu ngẫu nhiên từ tổng thể và thiết lập chiều cao trung bình là 188cm. Giá trị trung bình 188cm là giá trị được ước tính từ trung bình. Giá trị ước tính này có hạn chế là không cho biết giá trị trung bình 188cm này có thể cách xa giá trị trung bình của tổng thể như thế nào.

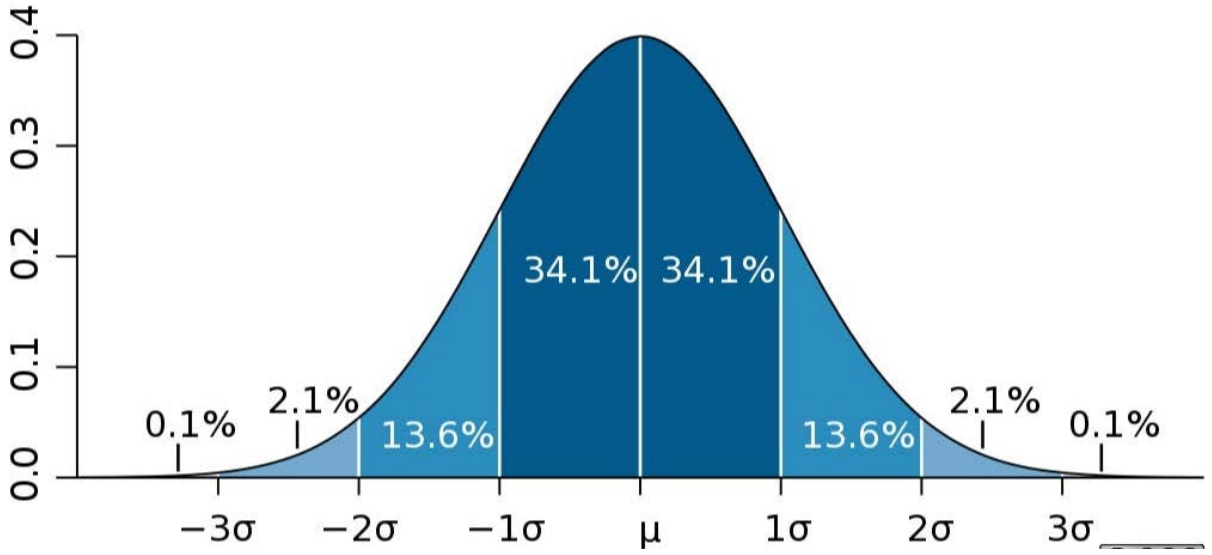
Giả sử tìm ra được 95 cầu thủ trên 100 cầu thủ được lấy mẫu có chiều cao nằm trong khoảng 183cm và 193cm. Khi đó, khoảng tin cậy là 95% giữa 2 giá trị 183cm và 193cm.

Nếu muốn độ tin cậy cao hơn nữa, có thể mở rộng khoảng tin cậy lên 99%. Làm như vậy sẽ luôn tạo ra một khoảng tin cậy lớn hơn (ví dụ từ 178cm đến 198cm).

2.1. Xác định kích thước mẫu (hay Cỡ mẫu - Sample size)?

2.1.2. Các thuật ngữ liên quan đến việc chọn cỡ mẫu

v. ***Standard deviation*** (*Độ lệch chuẩn*): đo lường mức độ phân tán dữ liệu so với mức trung bình (*average* hoặc *mean*). Nếu độ lệch chuẩn thấp thì hầu hết các điểm dữ liệu đều gần với mức trung bình; nếu nó cao, dữ liệu sẽ được trải rộng hơn.



2.2. Xác định cỡ mẫu theo ước lượng tổng thể

2.2.1. Trường hợp KHÔNG BIẾT quy mô tổng thể

- Bảng phân phối của giá trị z dựa trên độ tin cậy mong muốn:
- Công thức

$$n = \frac{z^2 \times p(1-p)}{e^2}$$

Mức độ tin cậy mong muốn (Desired confidence level)	*z-score*
80%	1.28
85%	1.44
90%	1.65
95%	1.96
99%	2.58

Trong đó:

- ***n***: kích thước mẫu cần xác định.
- ***z***: giá trị tra bảng phân phối ***z*** dựa vào độ tin cậy lựa chọn. Thông thường, độ tin cậy được sử dụng là 95% tương ứng với ***z*** = 1.96.
- ***p***: tỷ lệ ước lượng cỡ mẫu ***n*** thành công. Thường p được chọn = 0.5 để tích số ***p(1-p)*** là lớn nhất, điều này đảm bảo an toàn cho mẫu ***n*** ước lượng.
- ***e***: sai số cho phép. Thường ba tỷ lệ sai số hay sử dụng là: ±01 (1%), ±0.05 (5%), ±0.1 (10%), trong đó mức phổ biến nhất là ±0.05.

2.2. Xác định cỡ mẫu theo ước lượng tổng thể

2.2.1. Trường hợp KHÔNG BIẾT quy mô tổng thể

- Ví dụ: Khảo sát nhân viên giao hàng

Giả sử muốn khảo sát nhân viên giao hàng ở một thành phố có dân số 500.000 người. Độ tin cậy mong muốn (*confidence level*) là 95% và sai số (*margin of error*) 5%. Sử dụng công thức ở trên, có thể xác định được cỡ mẫu cần dùng là 385.

Mức độ tin cậy mong muốn (Desired confidence level)	*z-score*
80%	1.28
85%	1.44
90%	1.65
95%	1.96
99%	2.58

$$n = \frac{z^2 \times p(1-p)}{e^2}$$

$$n = \frac{z^2 \times p(1-p)}{e^2} = \frac{1.96^2 \times 0.5(1-0.5)}{0.05^2} = \frac{3.8416 \times 0.25}{0.0025} = 384.16 \approx 385$$

2.2. Xác định cỡ mẫu theo ước lượng tổng thể

2.2.1. Trường hợp BIẾT quy mô tổng thể

- Công thức

$$n = \frac{P}{1 + (P \times e^2)}$$

Trong đó:

- ***n***: kích thước mẫu cần xác định.
- ***P***: *Population size* (kích thước của tổng thể).
- ***e***: sai số cho phép. Thường ba tỷ lệ sai số hay sử dụng là: ±0.01 (1%), ±0.05 (5%), ±0.1 (10%), trong đó mức phổ biến nhất là ±0.05.

- Ví dụ: Nghiên cứu sự hài lòng của 3527 hộ dân trong Phường về việc hiến đất mở rộng hẻm. Như vậy cỡ mẫu tối thiểu cần có của nghiên cứu nếu sai số e = ±0.05 sẽ là 360 hộ:

$$n = \frac{P}{1 + (P \times e^2)} = \frac{3527}{1 + (3527 \times 0.05^2)} = \frac{3527}{1 + 8.8175} = 359.26 = 360$$

2.3. Xác định cỡ mẫu theo định lượng

- Các phương pháp này thường yêu cầu cỡ mẫu lớn. Tuy nhiên, phân tích viên lại có quỹ thời gian giới hạn và nếu không có nguồn tài chính tài trợ thì khả năng lấy mẫu theo ước lượng tổng thể sẽ khó có thể thực hiện.
- Do đó, các phân tích viên thường sử dụng công thức lấy mẫu dựa vào phương pháp định lượng được sử dụng để phân tích dữ liệu.
- Hai phương pháp thường dùng là
  - Phân tích nhân tố khám phá (EFA)
  - Hồi quy.

2.3. Xác định cỡ mẫu theo định lượng

2.3.1. Kích thước mẫu theo EFA *(Exploratory Factor Analysis- Phân tích nhân tố khám phá)*

- Phân tích nhân tố khám phá dùng để rút gọn một tập hợp k biến quan sát thành một tập F (với F < k) các nhân tố có ý nghĩa hơn. Trong nghiên cứu, thường người ta sẽ thu thập được một số lượng biến khá lớn và rất nhiều các biến quan sát trong đó có liên hệ tương quan với nhau.

Ví dụ, thay vì đi nghiên cứu **20** đặc điểm nhỏ của một đối tượng, nhưng ta nhận thấy dữ liệu có **4** đặc điểm lớn, mỗi đặc điểm lớn lại gồm **5** đặc điểm nhỏ, nên ta chỉ nghiên cứu **4** đặc điểm lớn.

2.3. Xác định cỡ mẫu theo định lượng

2.3.1. Kích thước mẫu theo EFA *(Exploratory Factor Analysis- Phân tích nhân tố khám phá)*

- Kích thước mẫu tối thiểu để sử dụng EFA là **50**, tốt hơn là từ **100** trở lên. Tỷ lệ "*số quan sát*" trên một "*biến phân tích*" *(hay "biến đo lường")* là **5:1** hoặc **10:1** (thậm chí nên là **20:1** nếu có thể). "*Số quan sát*" hiểu một cách đơn giản là số phiếu khảo sát hợp lệ cần thiết; "*biến phân tích*" là một câu hỏi đo lường trong bảng khảo sát.

Ví dụ, nếu bảng khảo sát có 30 câu hỏi (tương ứng với 30 biến quan sát thuộc các nhân tố khác nhau), mỗi câu hỏi lại có thang đo 5 mức độ. 30 câu này được sử dụng để phân tích trong EFA.

Áp dụng tỷ lệ 5:1, cỡ mẫu tối thiểu sẽ là 30 × 5 = 150,

nếu tỷ lệ 10:1, cỡ mẫu tối thiểu là là 30 × 10 = 300.

Kích thước mẫu này lớn hơn kích thước tối thiểu 50 (hoặc 100), vì vậy ta cần cỡ mẫu tối thiểu để thực hiện phân tích EFA là 150 hoặc 300 tùy tỷ lệ lựa chọn dựa trên khả năng có thể khảo sát được.

2.3. Xác định cỡ mẫu theo định lượng

2.3.2. Kích thước mẫu theo hồi quy

Ví dụ: khảo sát có 4 biến độc lập (4 thang đo, được đặt là biến m), mỗi thang đo được đo lường bằng 5 câu hỏi, như vậy tổng cộng có 20 (=4 x 5) 5 biến quan sát.

- Theo *Green* (i), đưa ra hai trường hợp.
  - *Trường hợp 1*: nếu mục đích phép hồi quy chỉ đánh giá mức độ phù hợp tổng quát của mô hình thì cỡ mẫu tối thiểu là **50 + 8m**.
  - *Trường hợp 2*: nếu mục đích muốn đánh giá các yếu tố của từng biến độc lập thì cỡ mẫu tối thiểu nên là **104 + m**. Với m là số lượng biến độc lập được đưa vào phân tích hồi quy, không phải là số biến quan sát hay số câu hỏi của nghiên cứu.
- Theo *Harris* (ii), cỡ mẫu phù hợp để chạy hồi quy đa biến thì cỡ mẫu tối thiểu phải là **50 + m**.
- Còn theo *Hair* và cộng sự (iii) cho rằng cỡ mẫu tối thiểu nên theo tỷ lệ **5:1**, tức là 5 quan sát cho một biến độc lập. Như vậy, nếu m=4, cỡ mẫu tối thiểu sẽ là 5 x m = 5 x 4 = 20.

Tuy nhiên, **5:1** chỉ là cỡ mẫu tối thiểu cần đạt, để kết quả hồi quy có ý nghĩa thống kê cao hơn, cỡ mẫu lý tưởng nên theo tỷ lệ **10:1** hoặc **15:1**, thậm chí là **50:1** (khi sử dụng phương pháp hồi quy).

i. *Green & Salkind, Using SPSS for Windows and Macintosh: Analyzing and Understanding Data, Prentice Hall, New Jersey, 2003*
ii. *Harris, A primer of multivariate statistics, New York: Academic Press,1985*
iii. *Multivariate Data Analysis, Pearson, New Jersey, 2009*

2.3. Xác định cỡ mẫu theo định lượng

2.3.3. Sử dụng nhiều phương pháp để xác định kích thước mẫu

- Nếu một nghiên cứu sử dụng kết hợp nhiều phương pháp xử lý thì sẽ lấy **kích thước mẫu cần thiết lớn nhất trong các phương pháp**.

  Ví dụ, nếu nghiên cứu vừa sử dụng phân tích EFA và vừa phân tích hồi quy. Kích thước mẫu cần thiết của EFA là 200, kích thước mẫu cần thiết của hồi quy là 100, kích thước mẫu cần thiết của nghiên cứu là 200 (hoặc từ 200 trở lên).

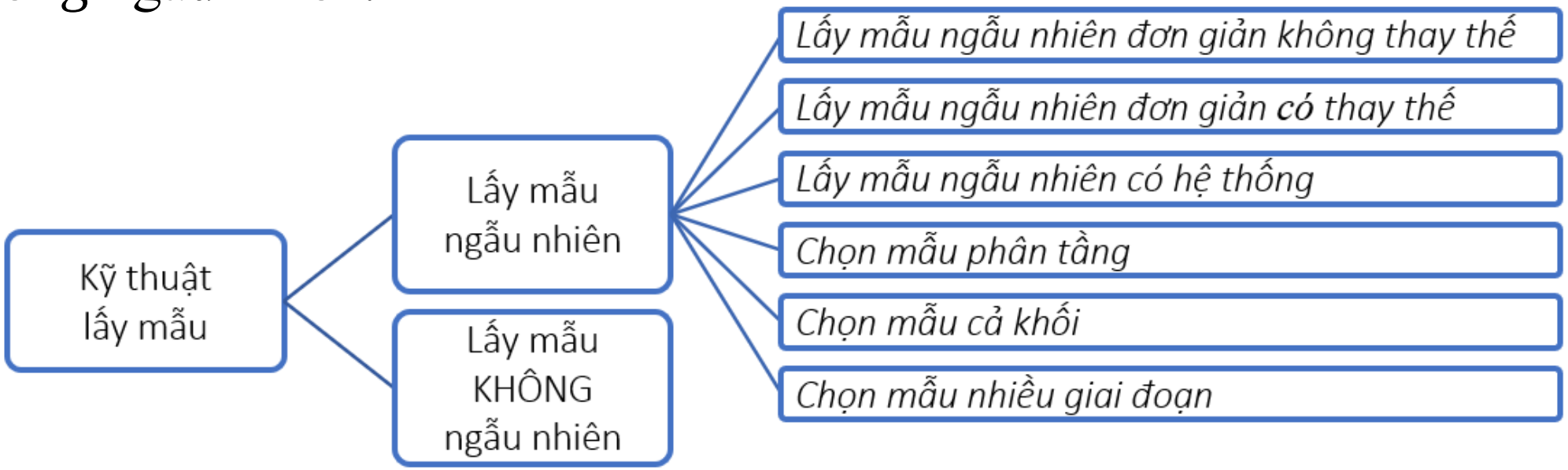
- Thường EFA luôn đòi hỏi cỡ mẫu lớn hơn rất nhiều so với hồi quy, chính vì vậy có thể sử dụng công thức tính kích thước mẫu tối thiểu cho EFA làm công thức tính kích thước mẫu cho nghiên cứu.

- Lưu ý rằng, đây là cỡ mẫu tối thiểu, nên nếu sử dụng cỡ mẫu lớn hơn kích thước tối thiểu, nghiên cứu sẽ càng có giá trị.

2.4. Kỹ thuật lấy mẫu

Sau khi đã chọn cỡ mẫu cho khảo sát, phân tích viên cần xác định kỹ thuật lấy mẫu nào sẽ sử dụng để chọn mẫu từ nhóm đối tượng mục tiêu. Kỹ thuật lấy mẫu được xem là phù hợp sẽ tùy thuộc vào tính chất và mục tiêu của dự án.

Kỹ thuật lấy mẫu có thể được chia thành hai loại: lấy mẫu ngẫu nhiên và lấy mẫu không ngẫu nhiên.



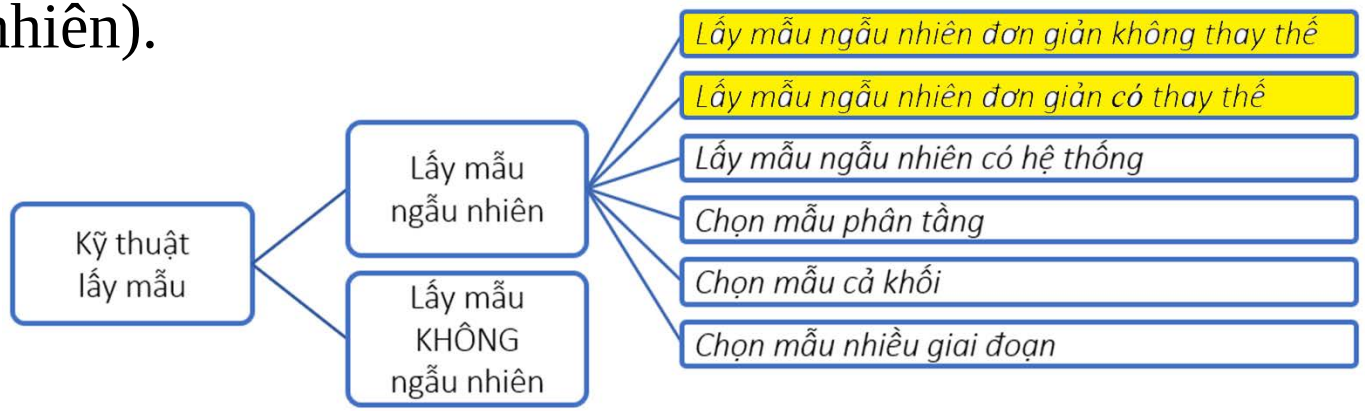
2.4. Kỹ thuật lấy mẫu

2.4.1. Lấy mẫu ngẫu nhiên

- Là chọn mẫu ngẫu nhiên từ một quần thể, không có bất kỳ điều kiện cụ thể nào. Điều này có thể được thực hiện bằng cách chọn mẫu từ danh sách hoặc tại địa điểm khảo sát.
- Trong lấy mẫu ngẫu nhiên lại chia thành một số loại như sau và các loại này lại có thể lồng ghép nhau trong quá trình lấy mẫu:

i. *Lấy mẫu ngẫu nhiên đơn giản **không** thay thế*: Nếu muốn đảm bảo rằng một hộ gia đình cụ thể không được chọn nhiều lần, có thể xóa hộ gia đình đó khỏi danh sách ngay sau khi nhận được mẫu của họ.

ii. *Lấy mẫu ngẫu nhiên đơn giản **có** thay thế*: Nếu cho phép 1 hộ được phép tham gia lấy mẫu nhiều lần (mặc dù vẫn lấy ngẫu nhiên).

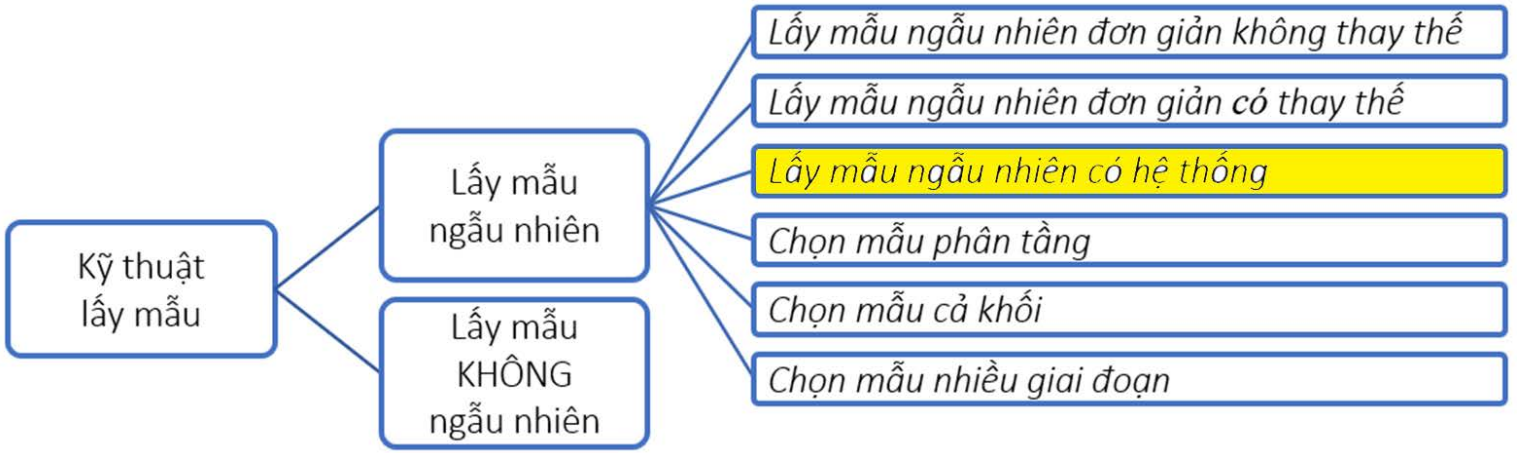


2.4. Kỹ thuật lấy mẫu

2.4.1. Lấy mẫu ngẫu nhiên

iii.*Lấy mẫu ngẫu nhiên có hệ thống*: là phương pháp lấy mẫu ngẫu nhiên được sử dụng phổ biến nhất, theo đó sẽ chia số lượng tổng thể cho cỡ mẫu và đưa ra con số X trở thành khoảng cách giữa các mẫu để lựa chọn.

Ví dụ trong nghiên cứu về sự hài lòng của 3527 hộ dân (*Population size* - kích thước của tổng thể) trong Phường về việc hiến đất mở rộng hẻm. Cỡ mẫu (sample size) tối thiểu là 360 hộ. Như vậy X= 3527/360 = 9.79 ≈ 9. Do đó, mọi hộ gia đình có số thứ tự là bội số của 9 sẽ được chọn..



2.4. Kỹ thuật lấy mẫu

2.4.1. Lấy mẫu ngẫu nhiên

iii. Lấy mẫu ngẫu nhiên có hệ thống

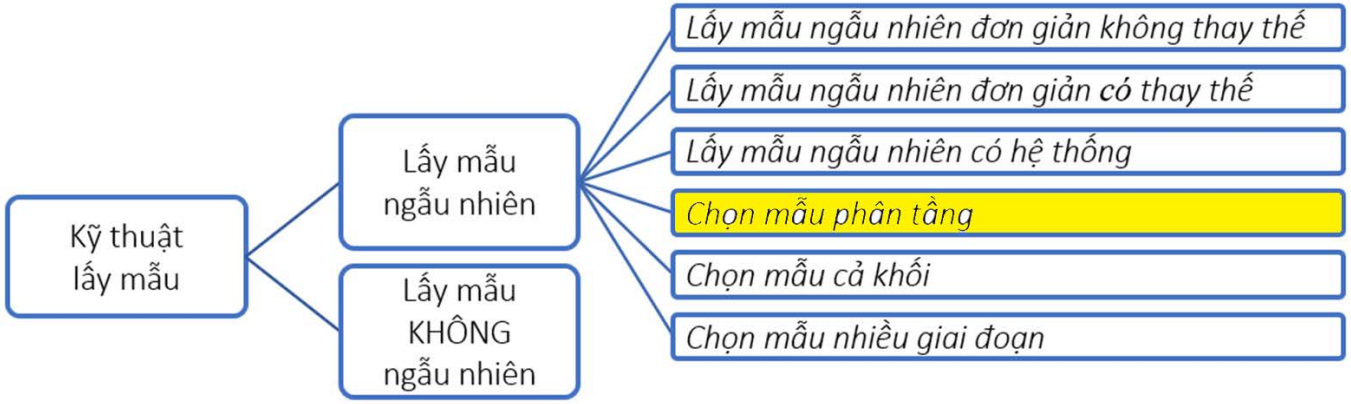
Lấy lại ví dụ về 3527 hộ dân của Phường ở trên. Giả sử, Phường gồm 6 Khu phố. Khi đó cần thực hiện:

- Cột 2: Thống kê số hộ trong mỗi Khu phố.
- Cột 3: Tính tỷ lệ phần trăm giữa số hộ của mỗi Khu phố so với cả Phường.
- Cột 3: Xác định số mẫu cần lấy đối với từng khu phố dựa trên tỷ lệ phần trăm.

Khu phố	*Số hộ*	*Tỷ lệ (%)*	*Số mẫu cần lấy*
(1)	*(2)*	*(3)=(2)/2527*	*(4)=360*(3)*
1	627	17.78	64
2	501	14.20	51
3	529	15.00	54
4	687	19.48	70
5	574	16.27	59
6	609	17.27	62
CỘNG	***3527***	***100***	***360***

2.4. Kỹ thuật lấy mẫu

2.4.1. Lấy mẫu ngẫu nhiên

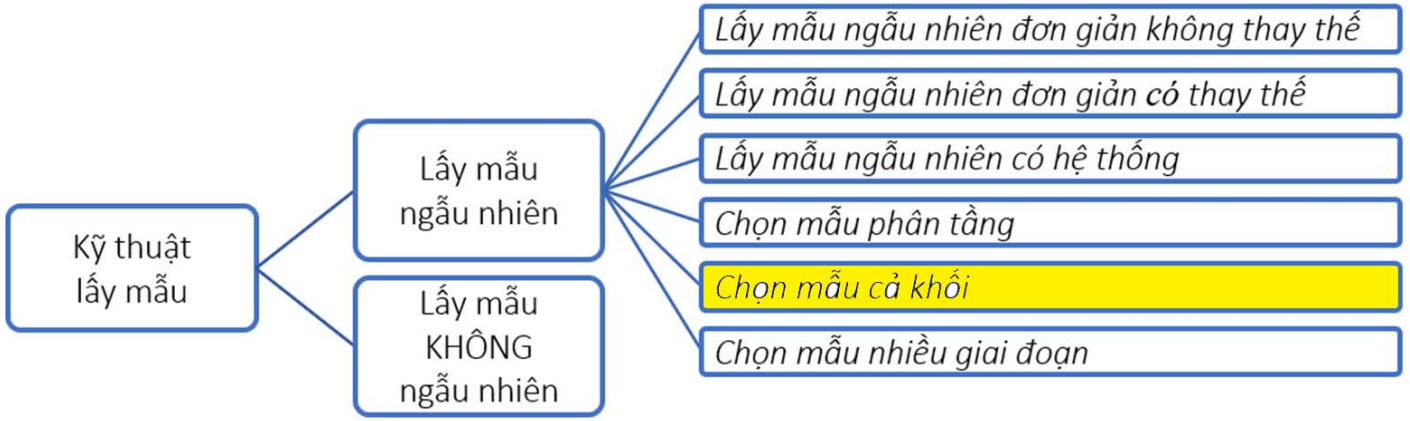


iv. Chọn mẫu phân tầng

- Tuy nhiên, dân số nói chung là hỗn hợp và không đồng nhất. Để đảm bảo bao gồm đầy đủ tất cả các loại dân số, cần xác định các tầng lớp hoặc đặc điểm khác nhau và sự đại diện thực tế của họ (tức là tỷ lệ) trong dân số.
- Việc phân tầng có thể dựa theo vùng, khu vực, loại hình, quy mô, trình độ văn hóa, giới tính, lứa tuổi, tình trạng hôn nhân, thu nhập, …
- Số lượng mẫu cần lấy của mỗi tầng có thể dựa trên dùng cách chọn mẫu ngẫu nhiên đơn giản hay chọn mẫu hệ thống.

2.4. Kỹ thuật lấy mẫu

2.4.1. Lấy mẫu ngẫu nhiên

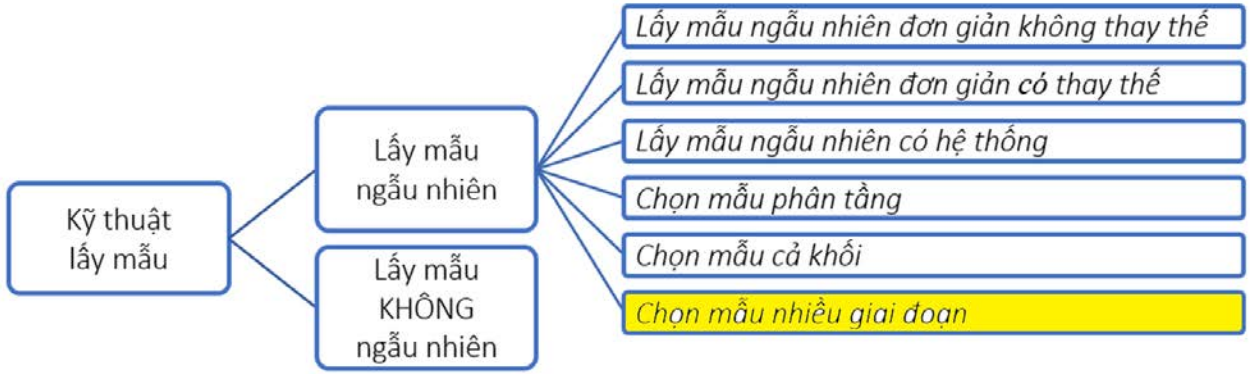


v. *Chọn mẫu cả khối*

- Trước tiên lập danh sách tổng thể chung theo từng khối (như làng, xã, phường, lượng sản phẩm sản xuất trong 1 khoảng thời gian…). Sau đó, chọn ngẫu nhiên một số khối và điều tra tất cả các đơn vị trong khối đã chọn.
- Phương pháp này thường được dùng khi không có sẵn danh sách đầy đủ của các đơn vị trong tổng thể cần nghiên cứu. Ví dụ : Tổng thể chung là sinh viên của một trường đại học. Khi đó sẽ lập danh sách sinh viên theo từng lớp, sau đó chọn ra một số lớp để lấy mẫu trên từng sinh viên của các lớp đã chọn.

2.4. Kỹ thuật lấy mẫu

2.4.1. Lấy mẫu ngẫu nhiên



v. Chọn mẫu nhiều giai đoạn

- Phương pháp này thường áp dụng đối với tổng thể chung có quy mô quá lớn và địa bàn nghiên cứu quá rộng. Việc chọn mẫu phải trải qua nhiều giai đoạn (nhiều cấp).
- Trước tiên phân chia tổng thể chung thành các đơn vị cấp I, rồi chọn các đơn vị mẫu cấp I. Tiếp đến phân chia mỗi đơn vị mẫu cấp I thành các đơn vị cấp II, rồi chọn các đơn vị mẫu cấp II, …
- Trong mỗi cấp có thể áp dụng các cách chọn mẫu ngẫu nhiên đơn giản, chọn mẫu hệ thống, chọn mẫu phân tầng, chọn mẫu cả khối để chọn ra các đơn vị mẫu.
- Ví dụ :Muốn chọn ngẫu nhiên 50 hộ từ một Phường có 10 Khu phố, mỗi khu phố có 50 hộ. Cách tiến hành như sau: Trước tiên đánh số thứ tự các khu phố từ 1 đến 10, chọn ngẫu nhiên trong đó 5 khu phố. Đánh số thứ tự các hộ trong từng khu phố được chọn. Chọn ngẫu nhiên ra 10 hộ trong mỗi khu phố ta sẽ có đủ mẫu cần thiết.

2.4. Kỹ thuật lấy mẫu

2.4.1. Lấy mẫu ngẫu nhiên

vi. Chọn mẫu nhiều giai đoạn

- Phương pháp này thường áp dụng đối với tổng thể chung có quy mô quá lớn và địa bàn nghiên cứu quá rộng. Việc chọn mẫu phải trải qua nhiều giai đoạn (nhiều cấp).
- Trước tiên phân chia tổng thể chung thành các đơn vị cấp I, rồi chọn các đơn vị mẫu cấp I. Tiếp đến phân chia mỗi đơn vị mẫu cấp I thành các đơn vị cấp II, rồi chọn các đơn vị mẫu cấp II, …
- Trong mỗi cấp có thể áp dụng các cách chọn mẫu ngẫu nhiên đơn giản, chọn mẫu hệ thống, chọn mẫu phân tầng, chọn mẫu cả khối để chọn ra các đơn vị mẫu.
- Ví dụ :Muốn chọn ngẫu nhiên 50 hộ từ một Phường có 10 Khu phố, mỗi khu phố có 50 hộ. Cách tiến hành như sau: Trước tiên đánh số thứ tự các khu phố từ 1 đến 10, chọn ngẫu nhiên trong đó 5 khu phố. Đánh số thứ tự các hộ trong từng khu phố được chọn. Chọn ngẫu nhiên ra 10 hộ trong mỗi khu phố ta sẽ có đủ mẫu cần thiết.

2.5. Các bước cần thực hiện khi lấy mẫu để đảm bảo kết quả phù hợp

Nếu không có sẵn chiến lược lấy mẫu, có thể việc thu thập dữ liệu sai lệch hoặc không mang tính đại diện, khiến dữ liệu của không hợp lệ hoặc kết quả nghiên cứu sẽ không được công nhận.

- ***B1: Xác định quy mô của khảo sát***
  - *Population size* (kích thước của tổng thể) là toàn thể đối tượng mục tiêu của nghiên cứu. Đôi khi, cuộc khảo sát có thể yêu cầu phải bao quát toàn bộ đối tượng mục tiêu, như trường hợp nghiên cứu về dân số (thường được gọi là khảo sát điều tra dân số).
  - Khi lựa chọn tất cả các mẫu là không khả thi, phân tích viên muốn chọn một mẫu nhỏ hơn có thể đại diện cho toàn thể và phản ánh các đặc điểm của toàn thể đó. Khi đó được gọi là khảo sát mẫu.
- ***B2: Xác định cỡ mẫu của dữ liệu***
- ***B3: Xác định kỹ thuật lấy mẫu sẽ sử dụng*** *(lấy mẫu ngẫu nhiên hay không ngẫu nhiên)*

2.5. Các bước cần thực hiện khi lấy mẫu để đảm bảo kết quả phù hợp

- *B4: Giảm thiểu lỗi lấy mẫu*

Việc mắc sai lầm trong quá trình chọn mẫu là điều bình thường. Do đó, nỗ lực cần luôn nỗ lực trong việc giảm sai số lấy mẫu để giúp cho mẫu được chọn mang tính đại diện cho tổng thể nhất có thể. Độ tin cậy của mẫu phụ thuộc vào cách giảm thiểu lỗi lấy mẫu. Mức độ sai sót trong quá trình lấy mẫu thay đổi tùy theo kỹ thuật hoặc phương pháp được dùng để chọn mẫu.

- Đối với các mẫu được chọn ngẫu nhiên:
  - Thường được chọn khi khảo sát cần suy ra tỷ lệ một số đặc điểm nhất định của dân số mục tiêu.
  - Thường cho phép kết quả có sai số lấy mẫu +/- **X%**.
- Đối với các mẫu được chọn không ngẫu nhiên:
  - Có thể rất hữu ích trong các tình huống khi cần nhanh chóng tiếp cận mẫu mục tiêu với các đặc điểm được chỉ định.
  - Lỗi lấy mẫu của loại hình này vẫn chưa được xác định.
  - Thường được chọn khi muốn nắm bắt càng nhiều nhận thức càng tốt. Do đó không quan tâm đến lỗi lấy mẫu hoặc lấy mẫu theo tỷ lệ.

NỘI DUNG CHƯƠNG 2

1. Những thách thức đối với các nhà khoa học dữ liệu
2. Chiến lược lấy mẫu để đảm bảo kết quả phù hợp
3. Thu thập dữ liệu định tính (*Qualitative Data*)
4. Thu thập dữ liệu định lượng (*Quantitative Data Collection*)
5. Bài thực hành

3. THU THẬP DỮ LIỆU ĐỊNH TÍNH (QUALITATIVE DATA)

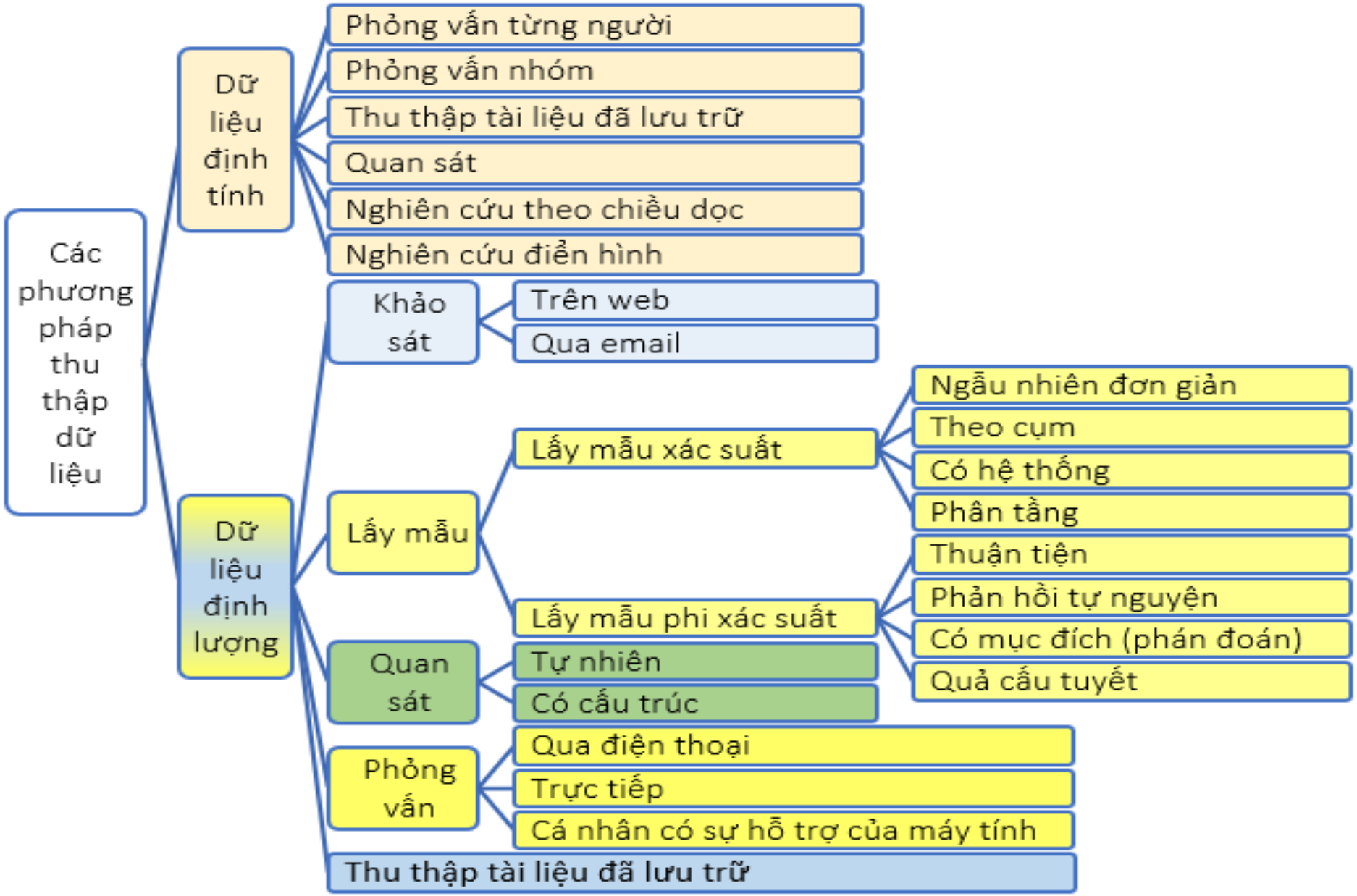
3.1. Một số ví dụ phân biệt giữa định tính và định lượng

- Một số ví dụ về dữ liệu định tính
  - Tóc có thể có các màu vàng, đen, trắng (bạc), nâu.
  - Động vật được chia thành 2 loại: động vật có xương sống và động vật không có xương sống.
- Dữ liệu định lượng:
  - Là bất kỳ thông tin định lượng nào có thể được sử dụng để tính toán hoặc phân tích thống kê. Dạng dữ liệu này giúp đưa ra các quyết định thực tế dựa trên các dẫn xuất toán học. Dữ liệu định lượng được sử dụng để trả lời các câu hỏi như Bao nhiêu? Bao lâu? ..
  - Một số ví dụ về định lượng:
    - Có bốn gói kẹo và ba cái bánh nướng xốp được đựng trong giỏ.
    - Một ly nước có ga có thể cung cấp cho người dùng 97,5 calo.

3.2. Thu thập dữ liệu định tính

- ***Giới thiệu***
  - Dữ liệu định tính còn được gọi là dữ liệu phân loại (*categorical data*) vì có thể được nhóm/sắp xếp theo danh mục dựa trên các thuộc tính, tính chất của một sự vật hoặc hiện tượng.
  - Dữ liệu định tính không bao gồm các giá trị số.
- ***Ưu điểm của dữ liệu định tính***
  - Giúp phân tích chuyên sâu
  - Hiểu khách hàng nghĩ gì
  - Dữ liệu phong phú
- ***Nhược điểm của dữ liệu định tính***
  - Tốn thời gian
  - Không dễ khái quát hóa
  - Phụ thuộc vào kỹ năng của phân tích viên

3.3. Các phương pháp thu thập dữ liệu định tính



3.3. Các phương pháp thu thập dữ liệu định tính

i. Phỏng vấn từng người một

- Đây là một trong những công cụ thu thập dữ liệu được sử dụng phổ biến nhất cho các câu hỏi nghiên cứu định tính, chủ yếu là do cách tiếp cận của nó.
- Phương pháp phỏng vấn có thể không chính thức và không có cấu trúc - đàm thoại.
- Các câu hỏi mở hầu hết được hỏi một cách tự nhiên và người phỏng vấn sẽ để diễn biến cuộc phỏng vấn quyết định các câu hỏi sẽ được hỏi.

ii. Phỏng vấn nhóm

- Được thực hiện trong bối cảnh thảo luận nhóm. Nhóm được giới hạn ở 6-10 người và một người điều hành được chỉ định để kiểm duyệt cuộc thảo luận đang diễn ra.
- Tùy thuộc vào dữ liệu được sắp xếp, các thành viên trong nhóm có thể có điểm chung. Ví dụ, một phân tích viên tiến hành nghiên cứu về các vận động viên chạy điền kinh sẽ chọn những vận động viên là vận động viên chạy điền kinh hoặc đã từng là vận động viên chạy điền kinh và có đủ kiến thức về chủ đề này.

3.3. Các phương pháp thu thập dữ liệu định tính

iii. Thu thập qua hồ sơ lưu trữ trước đó

- Phương pháp này sử dụng các tài liệu đáng tin cậy hiện có và các nguồn thông tin tương tự làm nguồn dữ liệu. Nó tương tự như việc đi đến một thư viện. Ở đó, người ta có thể xem qua sách và tài liệu tham khảo khác để thu thập dữ liệu liên quan có thể sử dụng trong nghiên cứu.
- Dữ liệu này có thể được sử dụng trong nghiên cứu mới.

iv. Quan sát (observation)

- Trong phương pháp này, phân tích viên hòa mình vào bối cảnh nơi người trả lời đang ở, để ý đến những người tham gia và ghi chép.
- Bên cạnh việc ghi chú, có thể sử dụng các phương pháp ghi chép khác, chẳng hạn như ghi video và ghi âm, chụp ảnh và các phương pháp tương tự.

3.3. Các phương pháp thu thập dữ liệu định tính

v. Nghiên cứu theo chiều dọc

- Phương pháp này được thực hiện nhiều lần trên cùng một nguồn dữ liệu trong một khoảng thời gian dài (trong vài năm và đôi khi có thể kéo dài hàng thập kỷ).
- Các phương pháp thu thập dữ liệu như vậy nhằm mục đích tìm ra mối tương quan thông qua các nghiên cứu thực nghiệm về các đối tượng có đặc điểm chung.

vi. Nghiên cứu điển hình

- Phương pháp này thu thập dữ liệu từ việc phân tích chuyên sâu các nghiên cứu điển hình. Tính linh hoạt của phương pháp này được thể hiện ở cách phương pháp này có thể được sử dụng để phân tích cả các đối tượng đơn giản và phức tạp.
- Điểm mạnh của phương pháp này là sử dụng sự kết hợp của một hoặc nhiều phương pháp định tính một cách thận trọng để rút ra kết luận.

3.4. Cách tiếp cận để phân tích dữ liệu định tính

Việc phân tích dữ liệu định tính tốn rất nhiều thời gian và tiền bạc để thu thập dữ liệu. Tuy nhiên, không có quy tắc cơ bản nào được đặt ra để phân tích dữ liệu mà chỉ có hai cách tiếp cận chính để phân tích dữ liệu định tính:

- ***Phương pháp suy diễn***
  - Phương pháp suy diễn bao gồm việc phân tích dữ liệu định tính dựa trên cấu trúc mà phân tích viên xác định trước.
  - Cách tiếp cận này nhanh chóng, dễ dàng và có thể được sử dụng khi phân tích viên có ý tưởng rõ ràng về những phản hồi có thể xảy ra mà họ sẽ nhận được từ tổng thể mẫu.
- ***Phương pháp quy nạp***
  - Cách tiếp cận quy nạp không dựa trên một cấu trúc/quy tắc/khuôn khổ cơ bản được định trước.
  - Đó là một cách tiếp cận tốn nhiều thời gian và kỹ lưỡng hơn cho quá trình phân tích định tính.
  - Phương pháp quy nạp thường được sử dụng khi phân tích viên có rất ít hoặc không có ý tưởng gì về hiện tượng nghiên cứu.

3.5. Các bước thực hiện giúp phân tích dữ liệu định tính hiệu quả

Bước 1: Sắp xếp dữ liệu

- Phần lớn dữ liệu sau thu thập sẽ không có cấu trúc và đôi khi không có ý nghĩa gì khi xem qua. Vì vậy, bước đầu tiên trong việc phân tích dữ liệu là sắp xếp nó một cách có hệ thống.
- Sắp xếp dữ liệu là chuyển đổi tất cả dữ liệu sang định dạng văn bản.
- Có thể xuất dữ liệu vào bảng tính hoặc từ những định dạng dữ liệu nào (JSON, XML, CSV, …) mà bất kỳ công cụ phân tích dữ liệu định tính nào được máy tính hỗ trợ.

Bước 2: Tổ chức lại dữ liệu theo mục tiêu nghiên cứu

- Sau khi chuyển đổi và sắp xếp dữ liệu, lúc này có một lượng lớn thông tin vẫn cần được tổ chức lại một cách có trật tự. Một trong những cách tốt nhất để tổ chức dữ liệu là quay lại mục tiêu nghiên cứu và sau đó sắp xếp dữ liệu dựa trên các câu hỏi được đặt ra.
- Sắp xếp mục tiêu nghiên cứu vào một bảng sao cho rõ ràng về mặt trực quan. Nếu không thực hiện bước này, sẽ lãng phí thời gian và sẽ không thu được kết quả cuối cùng.

3.5. Các bước thực hiện giúp phân tích dữ liệu định tính hiệu quả

Bước 3: Đặt mã cho dữ liệu được thu thập

- Mã hóa dữ liệu có nghĩa là phân loại và gán các thuộc tính và mẫu cho dữ liệu được thu thập. Quá trình mã hóa là một trong những cách tốt nhất để nén một lượng lớn thông tin được thu thập.
- Mã hóa rất quan trọng trong phân tích dữ liệu vì có thể rút ra lý thuyết từ các kết quả nghiên cứu có liên quan.

Bước 4: Xác thực dữ liệu định tính

- Vì dữ liệu là điều cần thiết cho nghiên cứu nên việc đảm bảo rằng dữ liệu không có sai sót là điều bắt buộc.
- Lưu ý rằng việc xác thực dữ liệu không chỉ là một bước trong phân tích này mà là bước cần được thực hiện định kỳ trong suốt quá trình nghiên cứu.
- Có hai mặt để xác thực dữ liệu:
  - Tính chính xác của thiết kế hoặc phương pháp nghiên cứu đang thực hiện.
  - Độ tin cậy là mức độ mà các phương pháp tạo ra dữ liệu chính xác một cách nhất quán.

3.5. Các bước thực hiện giúp phân tích dữ liệu định tính hiệu quả

Bước 5: Kết thúc quá trình phân tích

Kết thúc quá trình phân tích, kết quả phân tích phải đạt các yêu cầu sau:

- Dữ liệu phải được trình bày một cách có hệ thống để dễ dàng sử dụng.
- Phải nêu rõ:
  - Phương pháp đã sử dụng để tiến hành nghiên cứu
  - Những mặt tích cực, tiêu cực và những hạn chế của nghiên cứu.
- Nên nêu những gợi ý/suy luận về những phát hiện của phân tích viên và bất kỳ lĩnh vực liên quan nào cho nghiên cứu trong tương lai.
- Nêu lên những hiểu biết có giá trị và những phát hiện cho các bên liên quan.

NỘI DUNG CHƯƠNG 2

1. Những thách thức đối với các nhà khoa học dữ liệu
2. Chiến lược lấy mẫu để đảm bảo kết quả phù hợp
3. Thu thập dữ liệu định tính (*Qualitative Data*)
4. Thu thập dữ liệu định lượng (*Quantitative Data Collection*)
5. Bài thực hành

4. THU THẬP DỮ LIỆU ĐỊNH LƯỢNG *(Quantitative Data)*

4.1. Giới thiệu

- Thu thập dữ liệu định lượng là tất cả mọi thứ về số liệu và con số có thể được phân tích bằng phương pháp thống kê chẳng hạn như tần số phản hồi, giá trị trung bình và độ lệch chuẩn và có thể được phân tích bằng phần mềm thống kê.
- Các phân tích viên thường dựa vào dữ liệu định lượng khi họ có ý định định lượng các thuộc tính, thái độ, hành vi và các biến số được xác định khác

$\Rightarrow$ để ủng hộ (hoặc phản đối) giả thuyết về một hiện tượng cụ thể bằng cách bối cảnh hóa dữ liệu thu được thông qua khảo sát hoặc phỏng vấn mẫu nghiên cứu.

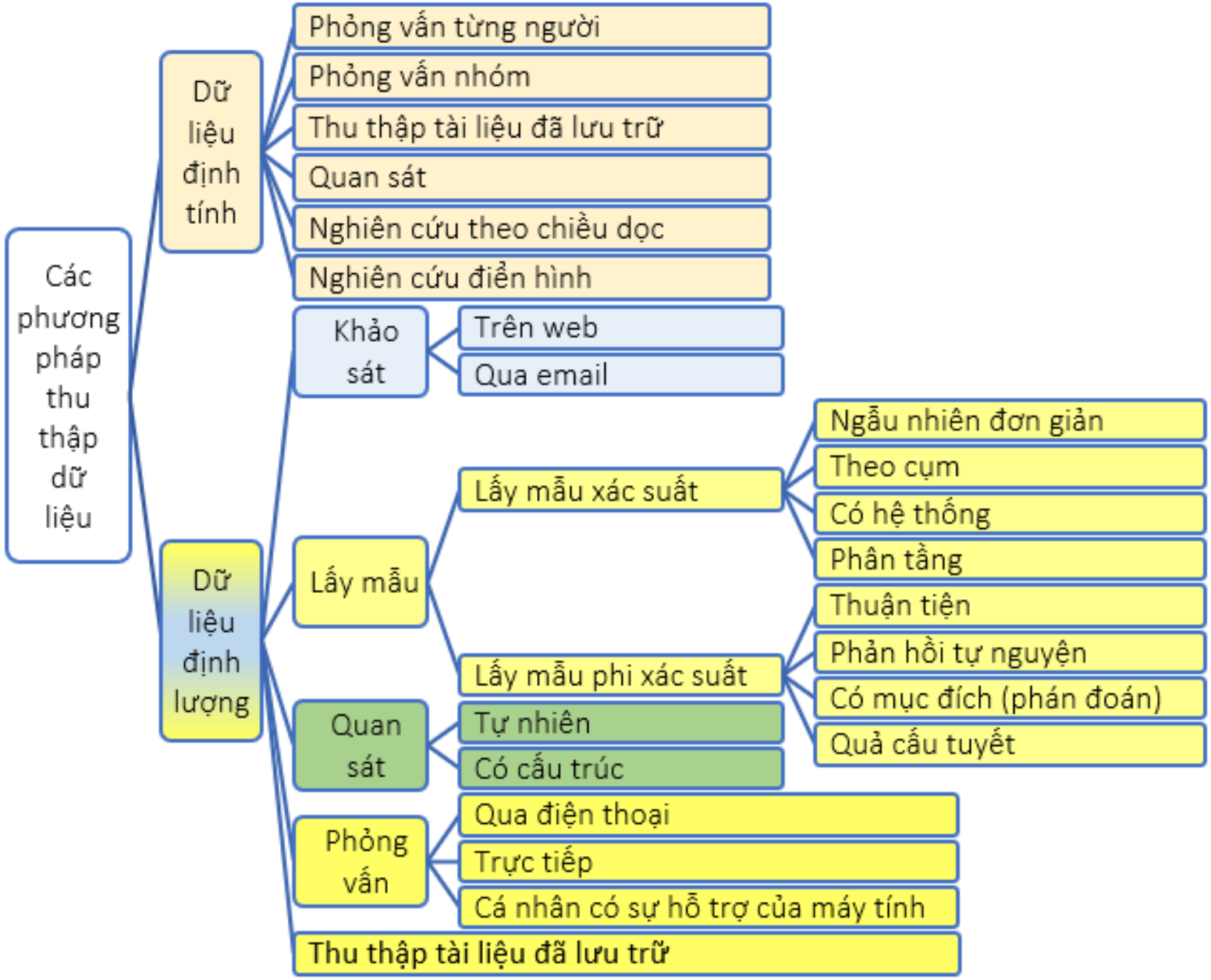
4.2. Phân loại Dữ liệu định lượng

Dữ liệu định lượng đươc chia thành 2 loại:

- *Dữ liệu định lượng Rời rạc*:
  - Là giá trị số nguyên
  - Ví dụ: nghiên cứu tìm ra số lượng xe gắn máy mà một hộ gia đình sở hữu.
- *Dữ liệu định lượng liên tục*:
  - Có khả năng là phân số hoặc số thập phân
  - Ví dụ khi nghiên cứu các phép đo vật lý của dân số như chiều cao, cân nặng, hoặc nghiên cứu về khoảng cách di chuyển.

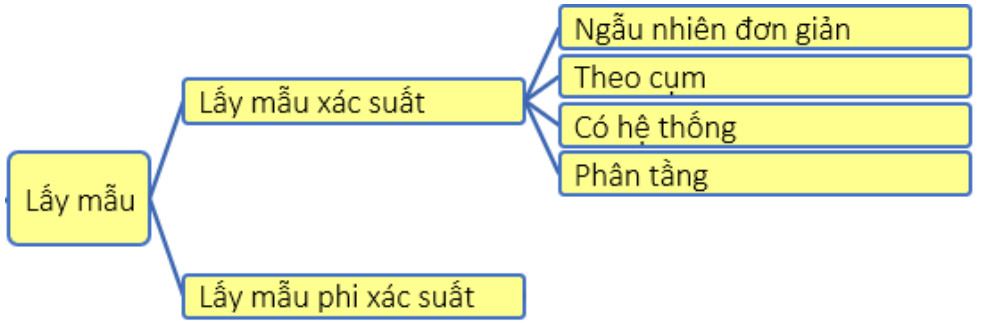
4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng

4.3.1. Phương pháp lấy mẫu



4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng

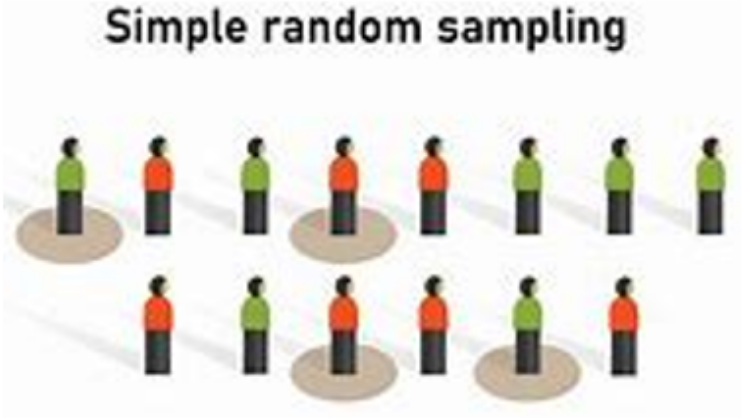
4.3.1. Phương pháp lấy mẫu



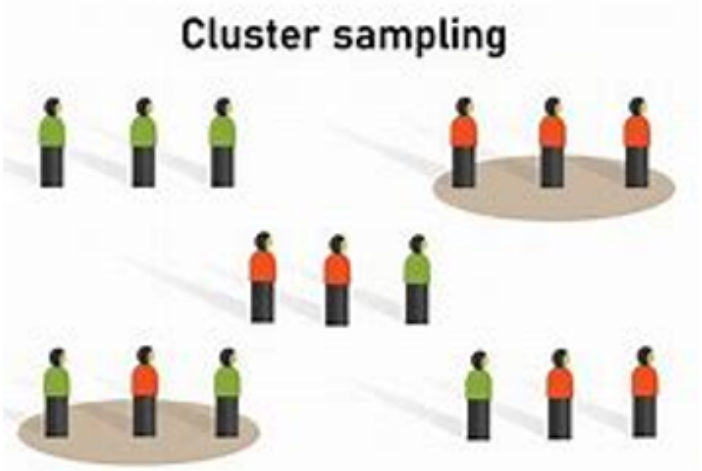
4.3.1.1. Lấy mẫu xác suất (Probability Sampling)

- Gồm 4 phương pháp

- *Lấy mẫu ngẫu nhiên đơn giản (Simple random sampling)* : Thông thường, đối tượng nhân khẩu học mục tiêu sẽ được chọn để đưa vào mẫu.



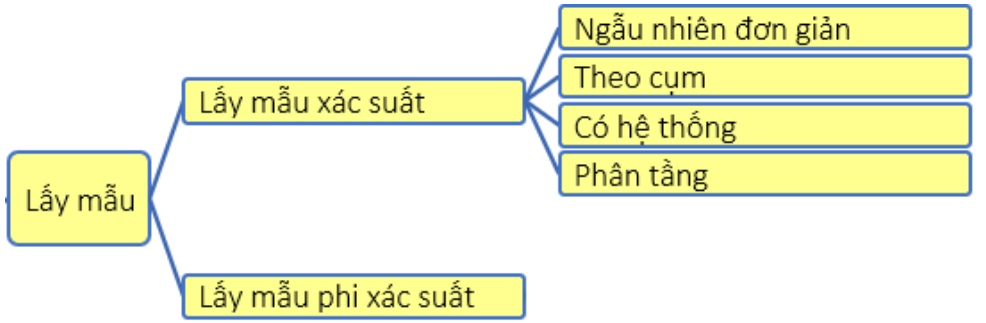
- *Lấy mẫu theo cụm (Cluster sampling)* : Lấy mẫu theo cụm là một kỹ thuật trong đó một quần thể được chia thành các nhóm hoặc cụm nhỏ hơn và chọn mẫu ngẫu nhiên từ các cụm này. Phương pháp này được sử dụng khi việc lấy mẫu ngẫu nhiên từ toàn bộ dân số là không thực tế hoặc tốn kém.



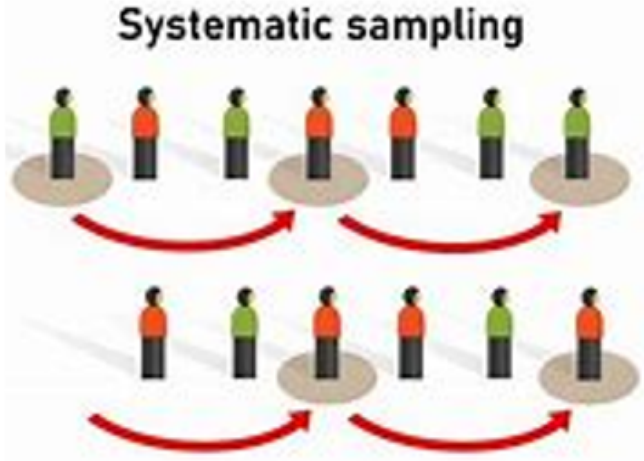
4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng

4.3.1. Phương pháp lấy mẫu

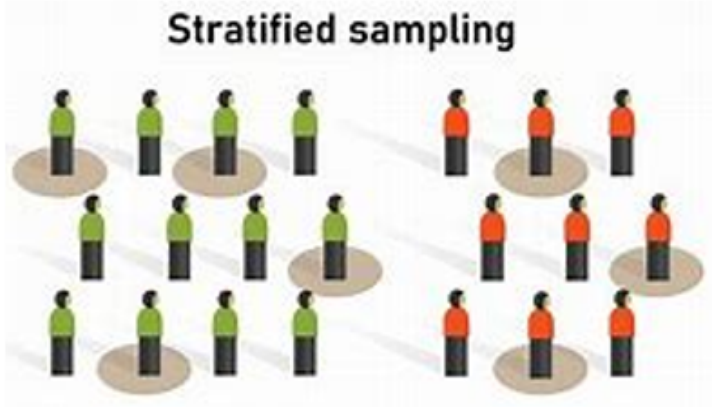
4.3.1.1. Lấy mẫu xác suất (Probability Sampling)

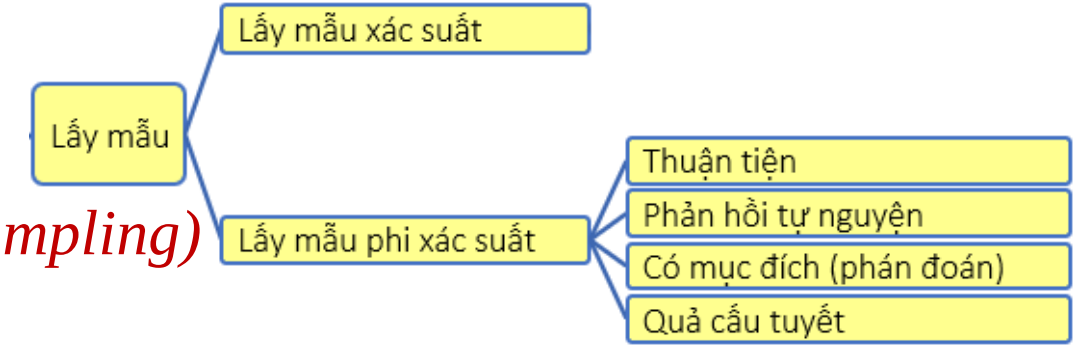


- *Lấy mẫu có hệ thống (Systematic sampling)* : Bất kỳ đối tượng nhân khẩu học mục tiêu nào cũng sẽ được đưa vào mẫu, nhưng chỉ đơn vị đầu tiên đưa vào mẫu được chọn ngẫu nhiên, phần còn lại được chọn theo thứ tự như thể cứ mười người thì có một người trong danh sách.



- *Lấy mẫu phân tầng (Stratified sampling)* : Nó cho phép chọn từng đơn vị từ một nhóm đối tượng mục tiêu cụ thể trong khi tạo mẫu. Sẽ rất hữu ích khi các phân tích viên chọn lọc việc đưa một nhóm người cụ thể vào mẫu, tức là chỉ nam hoặc nữ, người quản lý hoặc giám đốc điều hành, những người làm việc trong một ngành cụ thể.





4.3.1.2. Lấy mẫu phi xác suất (non-Probability Sampling)

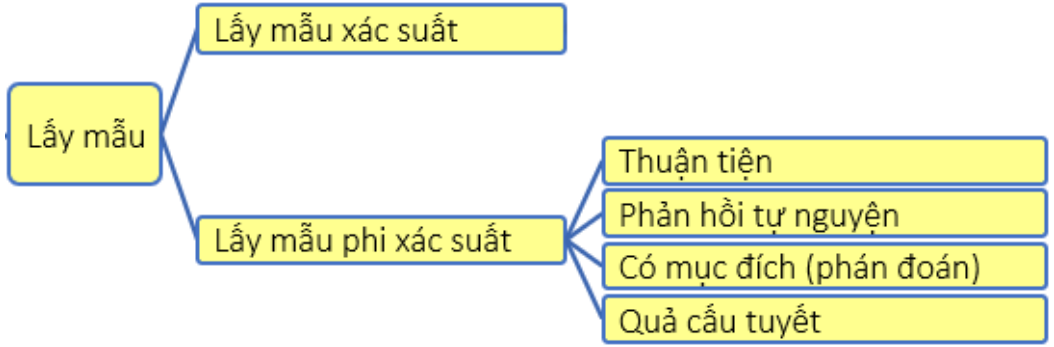
- Gồm 4 phương pháp

- *Lấy mẫu thuận tiện(Convenience sampling)*

  - Trong loại kỹ thuật lấy mẫu này, phân tích viên chọn những người tham gia mà điều tra viên có thể dễ dàng tiếp cận. Điều này có thể là do sự gần gũi về mặt địa lý, tính sẵn có tại một thời điểm nhất định hoặc sự sẵn sàng tham gia nghiên cứu.
  - Một trong những ưu điểm lớn nhất là nó ít tốn kém hơn và có thể thu thập thông tin một cách dễ dàng.
  - Hạn chế chính của phương pháp này là không thể đảm bảo rằng những người được chọn làm người tham gia đại diện cho toàn bộ các mẫu cần điều tra.
  - Ví dụ: sử dụng bạn bè hoặc gia đình làm một phần của mẫu sẽ dễ dàng hơn việc nhắm mục tiêu vào các cá nhân chưa biết. Đây là một cách thuận tiện để thu thập dữ liệu.

4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng

4.3.1. Phương pháp lấy mẫu

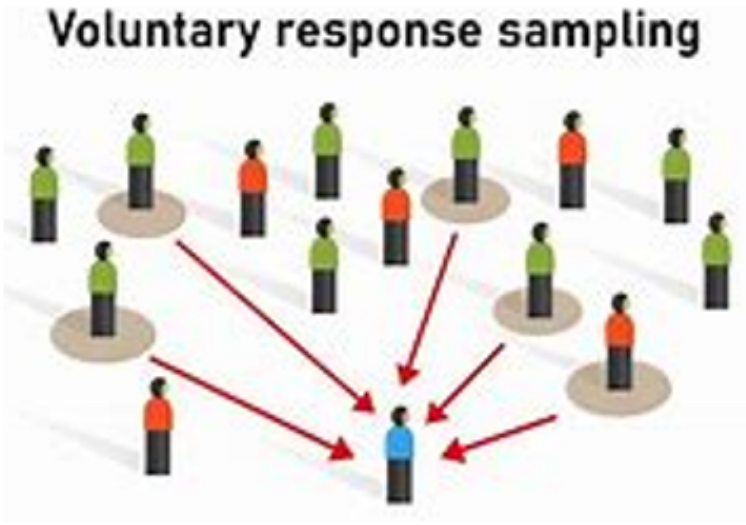


4.3.1.2. Lấy mẫu phi xác suất (non-Probability Sampling)

- Gồm 4 phương pháp

  - *Lấy mẫu phản hồi tự nguyện (Voluntary response sampling)*

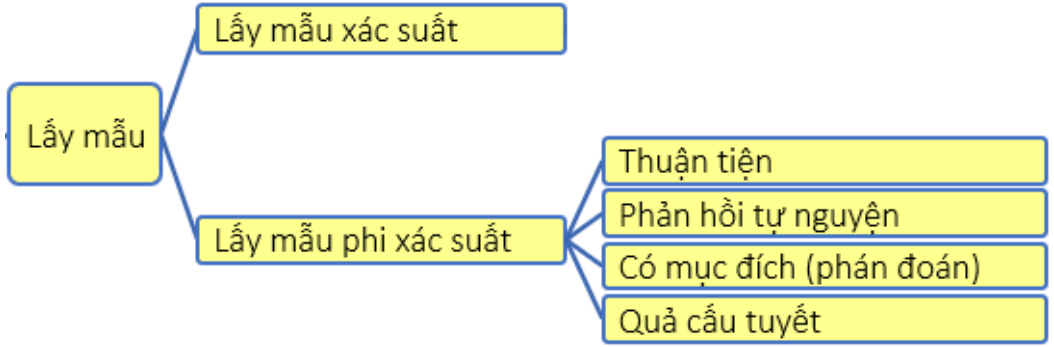
    - Kỹ thuật này có phần giống phương pháp *Lấy mẫu thuận tiện(Convenience sampling)*, thay vì dựa vào sự may rủi hay lựa chọn ngẫu nhiên, người được chủ động tự nguyện tham gia vào các nghiên cứu, khảo sát.
    - Nhược điểm lớn nhất của phương pháp này là có khả năng sai lệch cao.



    - Ví dụ: Giả sử thực hiện điều tra và thu thập thông tin về các dịch vụ hỗ trợ sinh viên. Có thể chia sẻ bảng câu hỏi khảo sát với tất cả sinh viên tại trường đại học và rất nhiều sinh viên đã quyết định hoàn thành nó. Nhưng khi áp dụng kỹ thuật này, sẽ không thể xác nhận rằng câu trả lời của một số học sinh đại diện cho quan điểm của những học sinh khác.

4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng

4.3.1. Phương pháp lấy mẫu



4.3.1.2. Lấy mẫu phi xác suất (non-Probability Sampling)

- Gồm 4 phương pháp

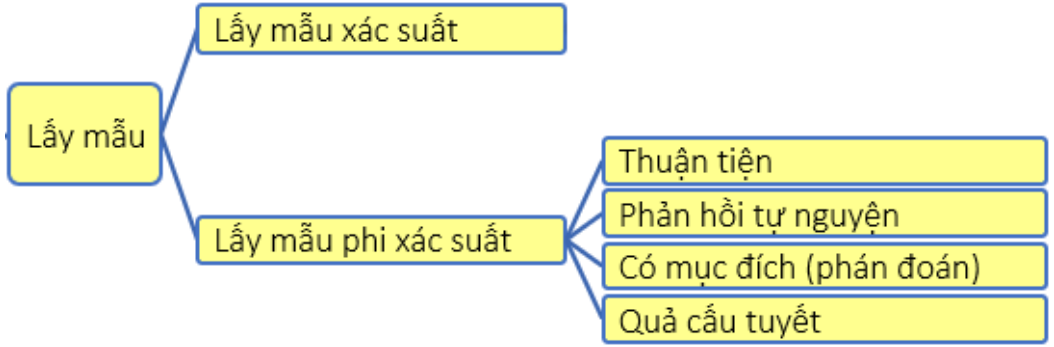
  - *Phương pháp lấy mẫu có mục đích hoặc phán đoán (Purposive or Judgemental Sampling Method)*

    - Phương pháp này cần sử dụng phán đoán cá nhân của phân tích viên để chọn mẫu.
    - Kỹ thuật lấy mẫu này chủ yếu sử dụng để thực hiện nghiên cứu định tính hoặc khi có ý định phát triển sự hiểu biết sâu sắc về một tình huống (hay 1 nhóm đối tượng) cụ thể. Khi đó, cần đặt ra các tiêu chí và lý do phù hợp để đưa vào để việc lấy mẫu có chủ đích có hiệu quả.

    - Ví dụ: muốn thực hiện một cuộc điều tra để phân tích kết quả học tập của những người bị khuyết tật đặc biệt. Có thể chọn những người khuyết tật một cách có mục đích để thu thập một lượng lớn thông tin về kết quả học tập của họ.

4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng

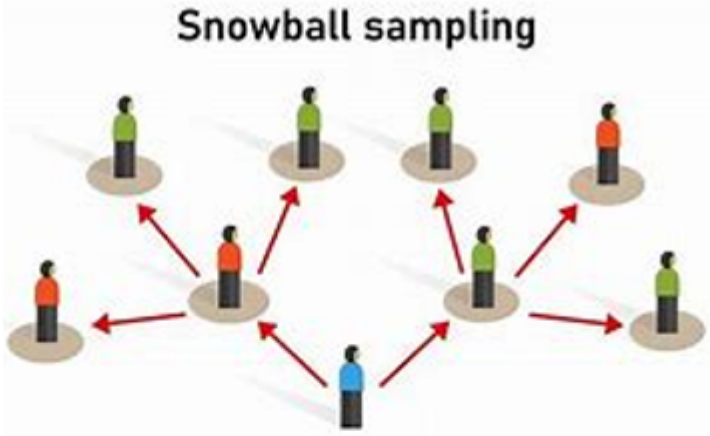
4.3.1. Phương pháp lấy mẫu



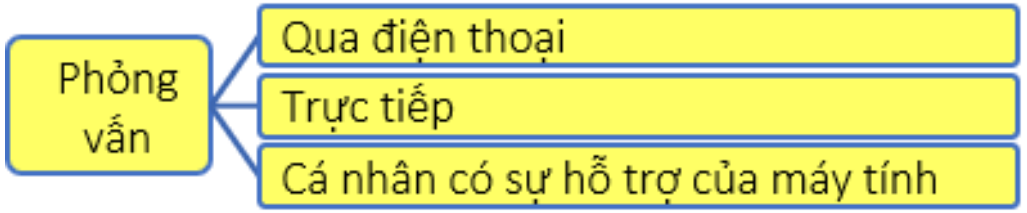
4.3.1.2. Lấy mẫu phi xác suất (non-Probability Sampling)

- Gồm 4 phương pháp

  - *Lấy mẫu quả cầu tuyết (Snowball sampling)*
    - Nếu dân số khó tiếp cận, có thể sử dụng phương pháp lấy mẫu quả cầu tuyết để thu hút nhiều người hơn.
    - Kỹ thuật lấy mẫu này bắt đầu với những người bạn có và sau đó những người ban đầu này sẽ liên lạc và chọn thêm những người khác tham gia.



    - Ví dụ: phân tích viên đang cố gắng thu thập thông tin về từng thành viên trong hộ gia đình, nhưng không thể biết được số lượng thành viên trong hộ. Khi một thành viên trong hộ sẵn sàng đóng vai trò là người tham gia cuộc điều tra. Thông qua người này, phân tích viên sẽ dễ dàng tiếp cận với tất cả các thành viên còn lại trong hộ.



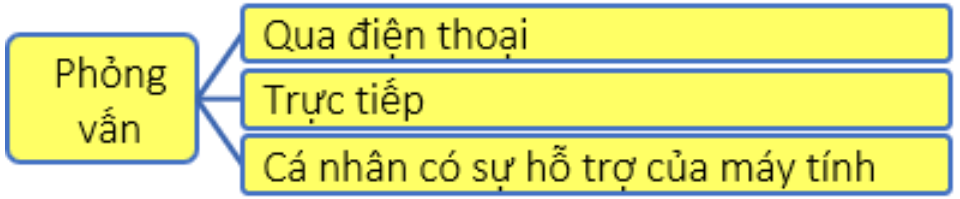
4.3.2. Phỏng vấn (Interviews)

- Là một phương pháp tiêu chuẩn được sử dụng để thu thập dữ liệu.
- Tuy nhiên, các cuộc phỏng vấn để thu thập dữ liệu định lượng có cấu trúc chặt chẽ hơn, trong đó các phân tích viên chỉ hỏi một bộ câu hỏi tiêu chuẩn và không yêu cầu gì hơn thế.
- Có ba loại phỏng vấn chính được thực hiện để thu thập dữ liệu:

(i).- Phỏng vấn qua điện thoại (Telephone interviews):

- Hình thức này đã thống trị trong bảng xếp hạng các phương pháp thu thập dữ liệu trong nhiều năm qua và dự kiến cả trong tương lai.
- Gồm các hình thức phỏng vấn:
  - Qua điện thoại.
  - Qua video bằng internet, Skype hoặc các nền tảng gọi điện video trực tuyến.

4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng



4.3.2. Phỏng vấn (Interviews)

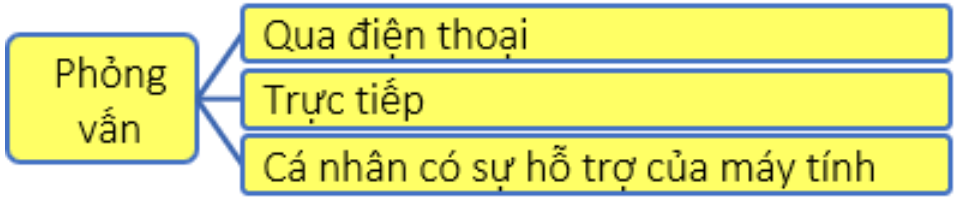
- Có ba loại phỏng vấn chính được thực hiện để thu thập dữ liệu:

(ii).- Phỏng vấn trực tiếp (Face-to-Face interviews – F2F):

- Là một kỹ thuật thu thập dữ liệu trực tiếp từ những người tham gia.
- Ưu điểm:
  - Giúp thu thập dữ liệu chất lượng vì cung cấp phạm vi để đặt các câu hỏi chi tiết và thăm dò sâu hơn để thu thập dữ liệu phong phú và nhiều thông tin.
  - Yêu cầu về trình độ đọc viết của người tham gia là không liên quan vì hình thức khảo sát này mang lại nhiều cơ hội để thu thập dữ liệu phi ngôn ngữ thông qua quan sát hoặc khám phá các vấn đề phức tạp và chưa biết.
- Mặc dù đây có thể là một phương pháp tốn kém và tốn thời gian nhưng tỷ lệ phản hồi cho các cuộc phỏng vấn *F2F* thường cao hơn.

4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng

4.3.2. Phỏng vấn (Interviews)



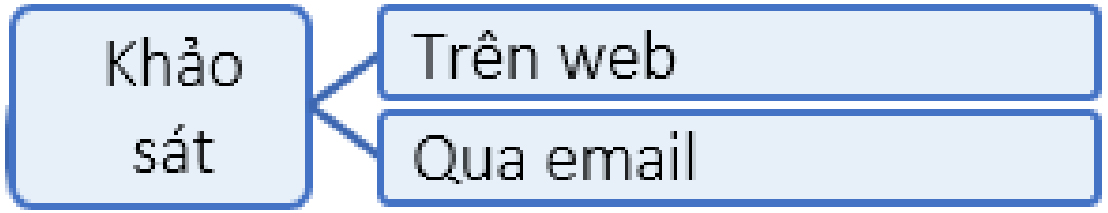
- Có ba loại phỏng vấn chính được thực hiện để thu thập dữ liệu:

(iii).- Phỏng vấn cá nhân có sự hỗ trợ của máy tính (Computer-Assisted Personal Interviewing - CAPI):

- Tương tự của cuộc phỏng vấn trực tiếp (F2F), trong đó người phỏng vấn mang theo máy tính để bàn hoặc máy tính xách tay bên mình tại thời điểm phỏng vấn để tải trực tiếp dữ liệu thu được từ cuộc phỏng vấn lên vào cơ sở dữ liệu.
- Ưu điểm:
  - Tiết kiệm rất nhiều thời gian trong việc cập nhật và xử lý dữ liệu.
  - Toàn bộ quá trình phỏng vấn gần như không cần giấy tờ bảng câu hỏi.

4.3.3. Khảo sát/Bảng câu hỏi (Surveys/Questionnaires)

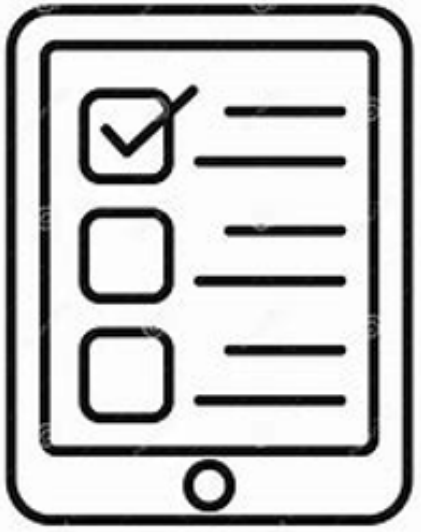
- Các cuộc khảo sát hoặc bảng câu hỏi được tạo bằng phần mềm khảo sát trực tuyến đang đóng một vai trò then chốt trong việc thu thập dữ liệu trực tuyến.
- Thông thường, danh sách kiểm tra và loại câu hỏi thang đánh giá chiếm phần lớn trong các cuộc khảo sát định lượng vì nó giúp đơn giản hóa và định lượng thái độ hoặc hành vi của người trả lời.
- Có hai loại bảng câu hỏi khảo sát quan trọng được sử dụng để thu thập dữ liệu trực tuyến cho nghiên cứu thị trường định lượng.



4.3.3. Khảo sát/Bảng câu hỏi (Surveys/Questionnaires)

- Hai loại bảng câu hỏi khảo sát được sử dụng để thu thập dữ liệu trực tuyến cho dữ liệu định lượng.

(i).- Bảng câu hỏi trên web (Web-based questionnaire)



- Trong bảng câu hỏi dựa trên web, người nhận sẽ nhận được email chứa liên kết khảo sát, nhấp vào liên kết sẽ đưa người trả lời đến công cụ khảo sát trực tuyến an toàn từ đó họ có thể thực hiện khảo sát hoặc điền vào bảng câu hỏi khảo sát.
- Đây là một trong những phương pháp thống trị và đáng tin cậy nhất cho nghiên cứu trên internet hoặc nghiên cứu trực tuyến vì các lý do sau:.
  - Tiết kiệm chi phí
  - Tiết kiệm thời gian.
  - Phạm vi tiếp cận rộng.
  - Tính linh hoạt của khảo sát vì người trả lời có thể tự do tham gia cuộc khảo sát vào thời gian rảnh bằng máy tính để bàn, máy tính xách tay, máy tính bảng hoặc thiết bị di động.

4.3.3. Khảo sát/Bảng câu hỏi (Surveys/Questionnaires)

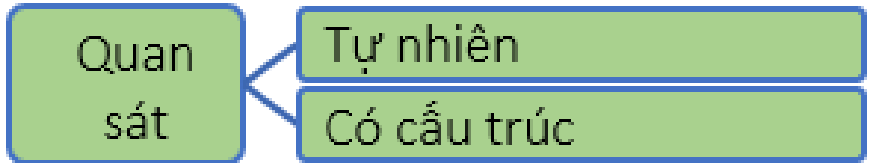
- Hai loại bảng câu hỏi khảo sát được sử dụng để thu thập dữ liệu trực tuyến cho dữ liệu định lượng.

(ii).- Bảng câu hỏi qua email (Mail Questionnaire)

- Trong bảng câu hỏi qua email, khảo sát được gửi đến một loạt đối tượng mẫu, cho phép phân tích viên kết nối với nhiều đối tượng.
- Một trong những lợi ích chính của bảng câu hỏi qua thư là tất cả các câu trả lời đều ẩn danh và người trả lời được phép dành bao nhiêu thời gian tùy thích để hoàn thành bản khảo sát và hoàn toàn trung thực về câu trả lời mà không sợ bị thành kiến.
- Bảng câu hỏi này thường bao gồm một gói chứa trang bìa giới thiệu với người tham gia về loại nghiên cứu và lý do tại sao nó được thực hiện để thu thập dữ liệu trực tuyến.
- Do bảng câu hỏi qua mail có tỷ lệ rời bỏ cao hơn so với các hình thức bảng câu hỏi khác, do đó hình thức này thường được đi kèm với:
  - Có hình thức khuyến khích/khuyến mãi đi kèm nếu người nhận phản hồi khảo sát.
  - Hoặc sau khi khảo sát được gởi qua mail 1 thời gian nhất định, thường có thêm lời nhắc người tham gia hoàn thành khảo sát.

4.3.4. Quan sát (Observations)

- Nhà nghiên cứu thu thập dữ liệu định lượng thông qua quan sát có hệ thống bằng cách sử dụng các kỹ thuật như đếm số người có mặt tại sự kiện cụ thể tại một thời điểm cụ thể và địa điểm cụ thể hoặc số người tham dự sự kiện ở một địa điểm được chỉ định.
- Phân loại:



4.3.4. Quan sát (Observations)

- Phân loại:

  - *Quan sát tự nhiên (Naturalistic observation)*
    - Được sử dụng để thu thập cả hai loại dữ liệu; định tính và định lượng.
    - Phân tích viên cần những kỹ năng quan sát nhạy bén và các giác quan để có được dữ liệu số về "cái gì" chứ không phải về "tại sao" và "như thế nào".

  - *Quan sát có cấu trúc (Structured observation)*
    - Nhà nghiên cứu, thay vì quan sát mọi thứ, chỉ tập trung vào những hành vi quan tâm rất cụ thể. Nó cho phép họ định lượng các hành vi mà họ đang quan sát.
    - Khi các quan sát định tính yêu cầu sự đánh giá từ phía người quan sát – nó thường được mô tả là mã hóa, đòi hỏi phải xác định rõ ràng một tập hợp các hành vi mục tiêu.

    - Quan sát có cấu trúc được sử dụng để thu thập dữ liệu định lượng hơn là định tính.

4.3. Các phương pháp được sử dụng để thu thập dữ liệu định lượng

4.3.5. Thu thập dựa trên tài liệu đã lưu trữ trước đó

- Đây là một cách hiệu quả để thu thập dữ liệu thiết thực, có chất lượng từ quá khứ.
- Ba loại tài liệu chính để thu thập dữ liệu nghiên cứu định lượng hỗ trợ.
  - *Hồ sơ công khai (Public Records)*: là các hồ sơ chính thức, được lưu trữ liên tục của một tổ chức. Ví dụ: báo cáo thường niên, sổ tay chính sách, hoạt động sinh viên, hoạt động trò chơi trong trường đại học, v.v.
  - *Tài liệu cá nhân (Personal Documents)*: Ngược lại với tài liệu công khai, loại tài liệu này xem xét tài liệu liên quan đến tài khoản cá nhân của cá nhân về hành động, hành vi, sức khỏe, vóc dáng, v.v. Ví dụ: chiều cao và cân nặng của học sinh, khoảng cách và phương tiện học sinh đang di chuyển để đến trường học, v.v.
  - *Bằng chứng vật chất (Physical Evidence)*: Bằng chứng vật chất hoặc tài liệu vật chất đề cập đến những thành tựu trước đây của một cá nhân hoặc của một tổ chức về mặt điểm số, doanh thu, lợi nhuận, tỷ lệ tăng trưởng, ...

NỘI DUNG CHƯƠNG 2

1. Những thách thức đối với các nhà khoa học dữ liệu
2. Chiến lược lấy mẫu để đảm bảo kết quả phù hợp
3. Thu thập dữ liệu định tính (*Qualitative Data*)
4. Thu thập dữ liệu định lượng (*Quantitative Data*)
5. Bài thực hành

5. BÀI THỰC HÀNH

i. Giả sử có nhu cầu phân tích về sức khỏe tâm thần (Metal_Health) như mô tả trong các file được gởi kèm theo bài giảng.

SV tự tìm hiểu và thiết kế phiếu khảo sát sao cho dữ liệu thu được phù hợp với các file mô tả được gởi kèm.

ii. Dựa vào bài thực hành trên, SV tự thiết kế phiếu khảo sát cho đề tài cá nhân ở cuối môn học.


Q&A