

# Thu thập, xử lý và phân tích dữ liệu

Nguyễn Bích Ngọc

[nguyenbichngoc@vnu.edu.vn](mailto:nguyenbichngoc@vnu.edu.vn)



VIETNAM NATIONAL UNIVERSITY, HANOI  
SCHOOL OF INTERDISCIPLINARY SCIENCES AND ARTS

# Thông tin lớp học

- Thời khóa biểu:
  - Lý thuyết: 24/11/2024 (S & C), 01/12/2024 (S)
  - Thực hành: 01/12/2024 (C)
- Phần mềm: MS Excel, (R & RStudio)
- Kiểm tra đánh giá – giữa kỳ 30%
  - Phân tích bộ dữ liệu cho trước sử dụng các công cụ trong khóa học
  - Viết báo cáo về kết quả phân tích. Lưu ý: chỉ đưa ra kết quả tính toán và đồ thị là không đủ, cần thảo luận ý nghĩa của các kết quả đó.

# Mục tiêu lớp học

- Giới thiệu khái niệm cơ bản
- Thảo luận định hướng phương pháp sử dụng (tên phương pháp)
- Thực hành phân tích cơ bản với Excel hoặc R

# Tài liệu tham khảo

- **The practice of social research – Earl Babbie (15<sup>th</sup> 2020)**
- **Understanding research methods – Coursera**  
[\(https://www.coursera.org/learn/research-methods/home/info\)](https://www.coursera.org/learn/research-methods/home/info)
- **Fundamentals of data visualization – Claus O. Wilke**  
[\(https://clauswilke.com/dataviz/index.html\)](https://clauswilke.com/dataviz/index.html)
- **Applied statistics with R – David Dalpiaz** (<https://book.stat420.org/>)
- Từng bước nhập môn nghiên cứu khoa học xã hội – Phạm Hiệp & cộng sự (2022)
- Cẩm nang nghiên cứu khoa học: từ ý tưởng đến công bố – Nguyễn Văn Tuấn (2<sup>nd</sup> edition, 2020)

# Nội dung

Giới thiệu chung

Thiết kế nghiên cứu (Định lượng)

Thu thập dữ liệu

Phân tích dữ liệu

# Nội dung

Giới thiệu chung

Thiết kế nghiên cứu (Định lượng)

Thu thập dữ liệu

Phân tích dữ liệu

# Khoa học

# Khoa học

## Science

---

[Article](#) [Talk](#)

---

From Wikipedia, the free encyclopedia

*For a topical guide, see [Outline of science](#). For other uses, see [Science \(disambiguation\)](#).*

**Science** is a rigorous, systematic endeavor that builds and organizes knowledge in the form of testable explanations and predictions about the universe.<sup>[1][2]</sup>

# Khoa học?

## Science

---

[Article](#) [Talk](#)

---

From Wikipedia, the free encyclopedia

*For a topical guide, see [Outline of science](#). For other uses, see [Science \(disambiguation\)](#).*

**Science** is a rigorous, systematic endeavor that builds and organizes knowledge in the form of testable explanations and predictions about the universe.<sup>[1][2]</sup>

# Khoa học?

## Science

[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

*For a topical guide, see [Outline of science](#). For other uses, see [Science \(disambiguation\)](#).*

**Science** is a rigorous, systematic endeavor that builds and organizes knowledge in the form of testable explanations and predictions about the universe.<sup>[1][2]</sup>

A theory that can't be proved wrong is nonscientific - Karl Popper

# Ví dụ

Một số di sản vật thể có linh khí mà người ta chỉ có thể cảm nhận chứ không thể đo đạc bằng các thiết bị vật lý

# Ví dụ

Một số di sản vật thể có linh khí mà người ta chỉ có thể cảm nhận chứ không thể đo đạc bằng các thiết bị vật lý

- Không thể quan sát, đo lường một cách có hệ thống
- Tính chủ quan
- Không thể lặp lại
- Thuộc phạm trù tâm linh, siêu nhiên, và triết học

# Nghiên cứu



*“Honey, come look! I’ve found some information all the world’s top scientists and doctors missed.”*

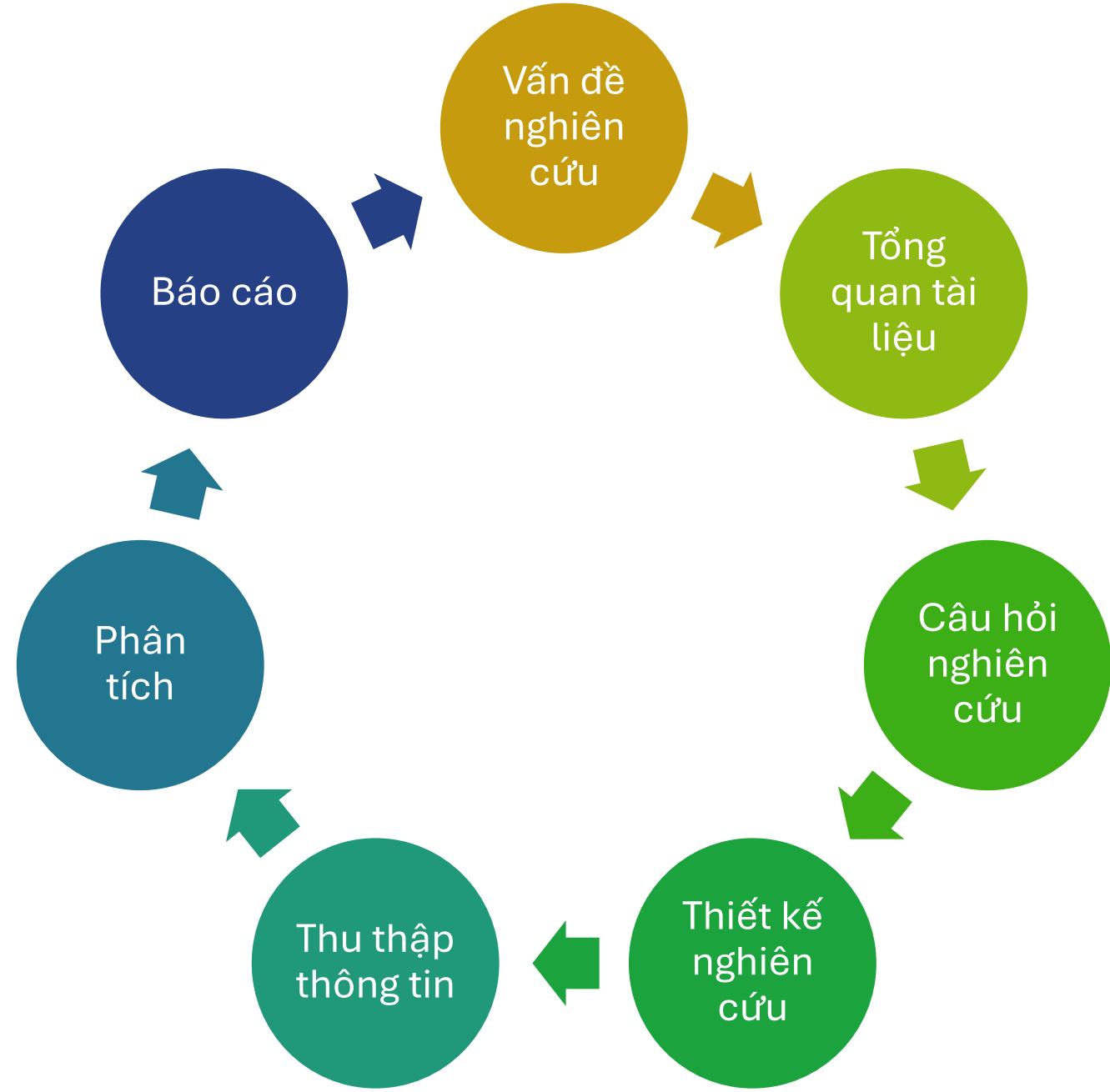
# Nghiên cứu

Research is systematic inquiry that helps to make sense of the world and that helps to make sensible the debates and interpretations that we have of issues of contemporary significance.

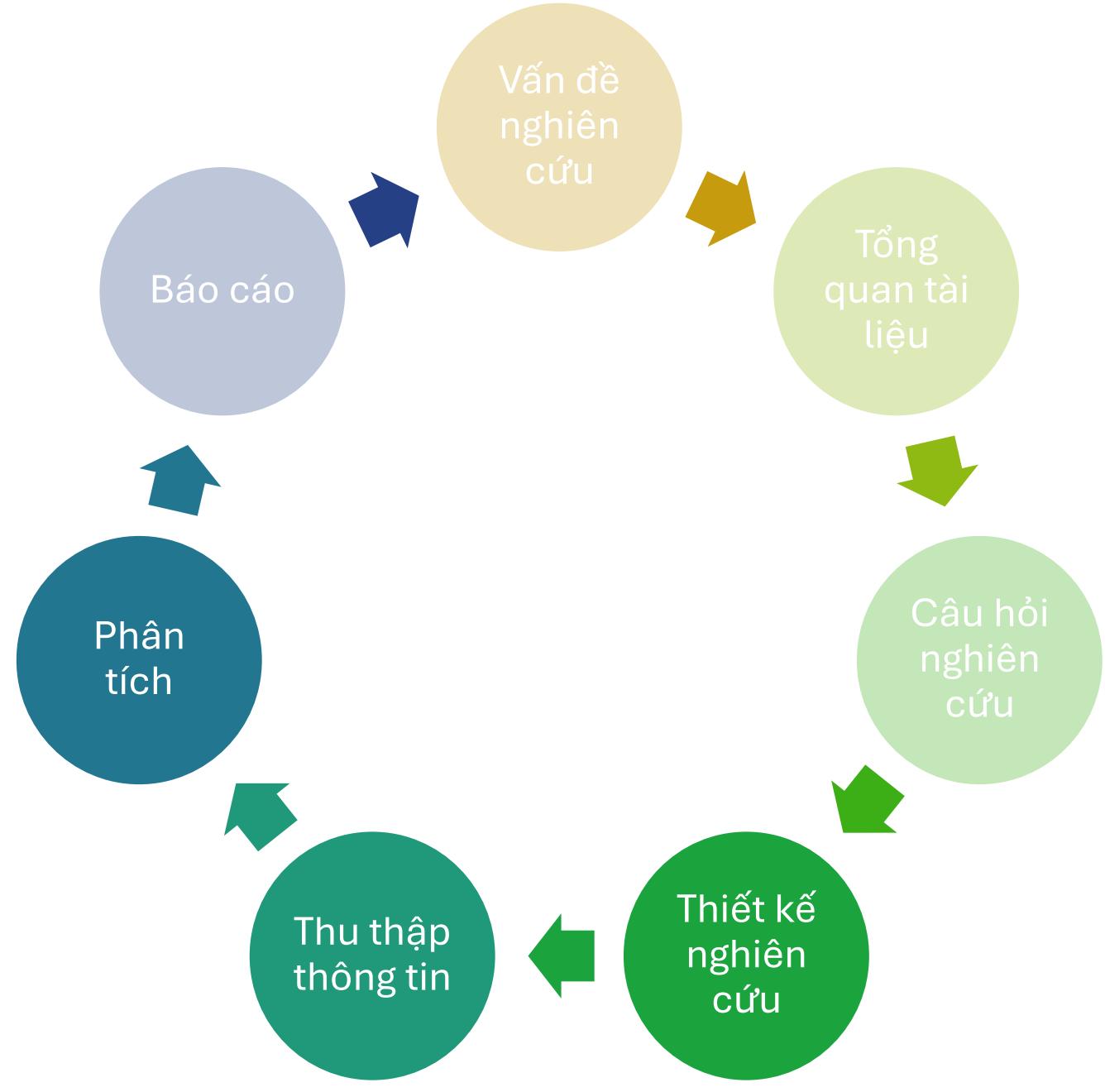
Professor Sandra Halperin

<https://www.coursera.org/learn/research-methods/home/info>

# Quá trình nghiên cứu

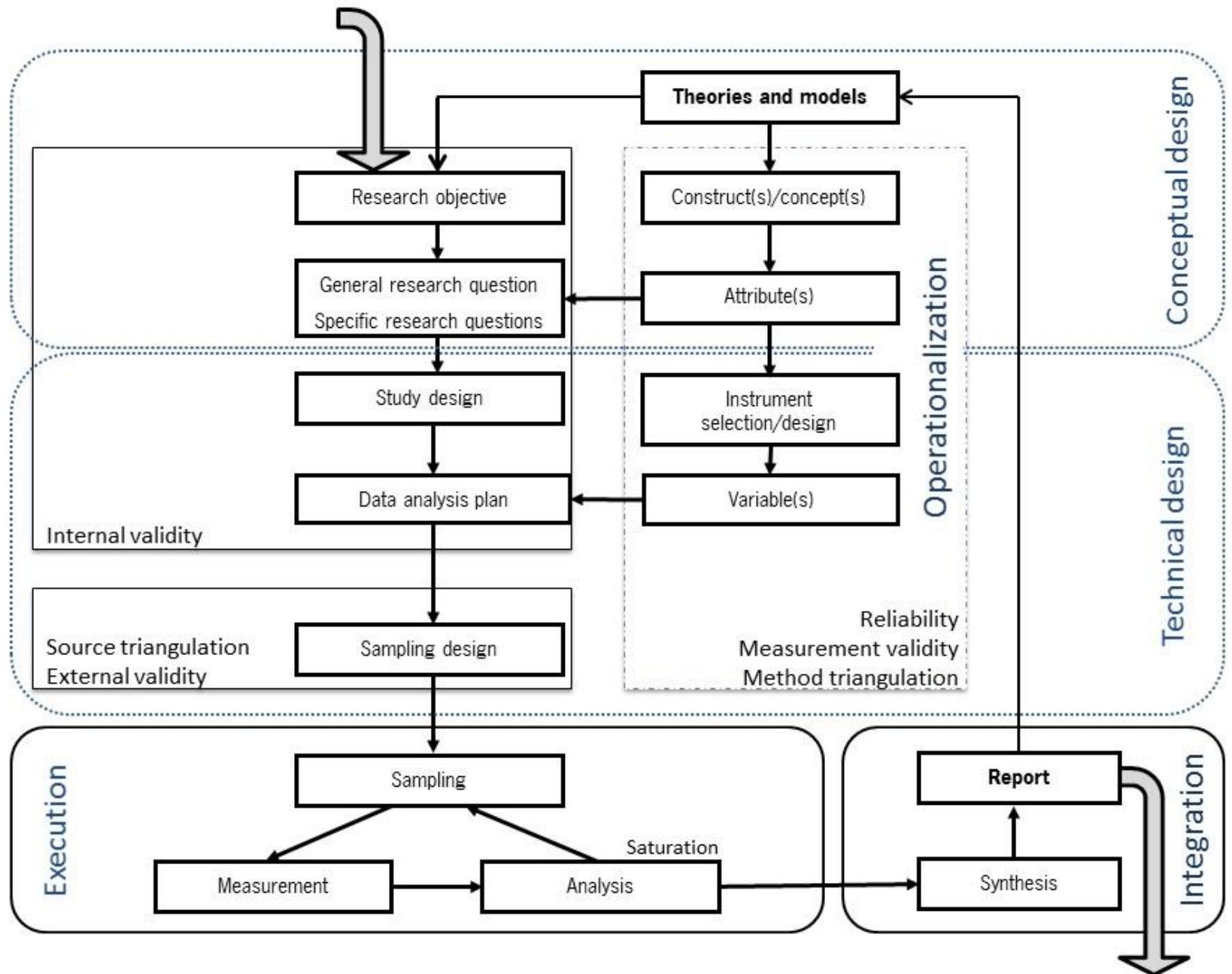


# Quá trình nghiên cứu



# Quá trình nghiên cứu

Tobi, H., & Kampen, J. K. (2018). Research design: The methodology for interdisciplinary research framework. *Quality & Quantity*, 52(3), 1209–1225.  
<https://doi.org/10.1007/s11135-017-0513-8>



# Nội dung

Giới thiệu chung

Thiết kế nghiên cứu (Định lượng)

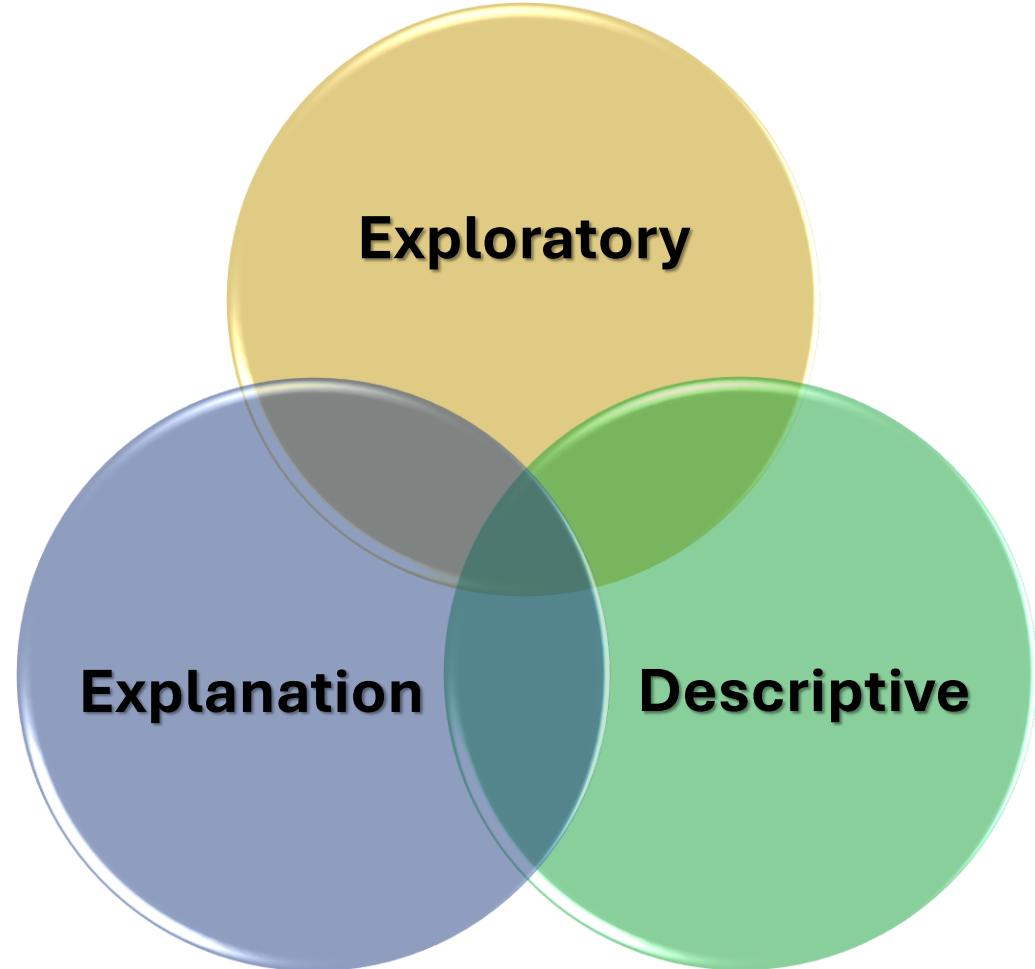
Thu thập dữ liệu

Phân tích dữ liệu

# Thiết kế nghiên cứu

- **What:** Cái gì mà chúng ta muốn quan sát tìm hiểu
- **Why:** Tại sao chúng ta muốn quan sát tìm hiểu về nó
- **How:** Làm sao để có thể quan sát, tìm hiểu, hay đo lường về nó

# Các mục đích của nghiên cứu



# Thiết kế nghiên cứu

## Định lượng

- Tập trung vào các yếu tố có thể đo lường
- Tổng quát hóa những thông tin thu được để diễn giải về thế giới

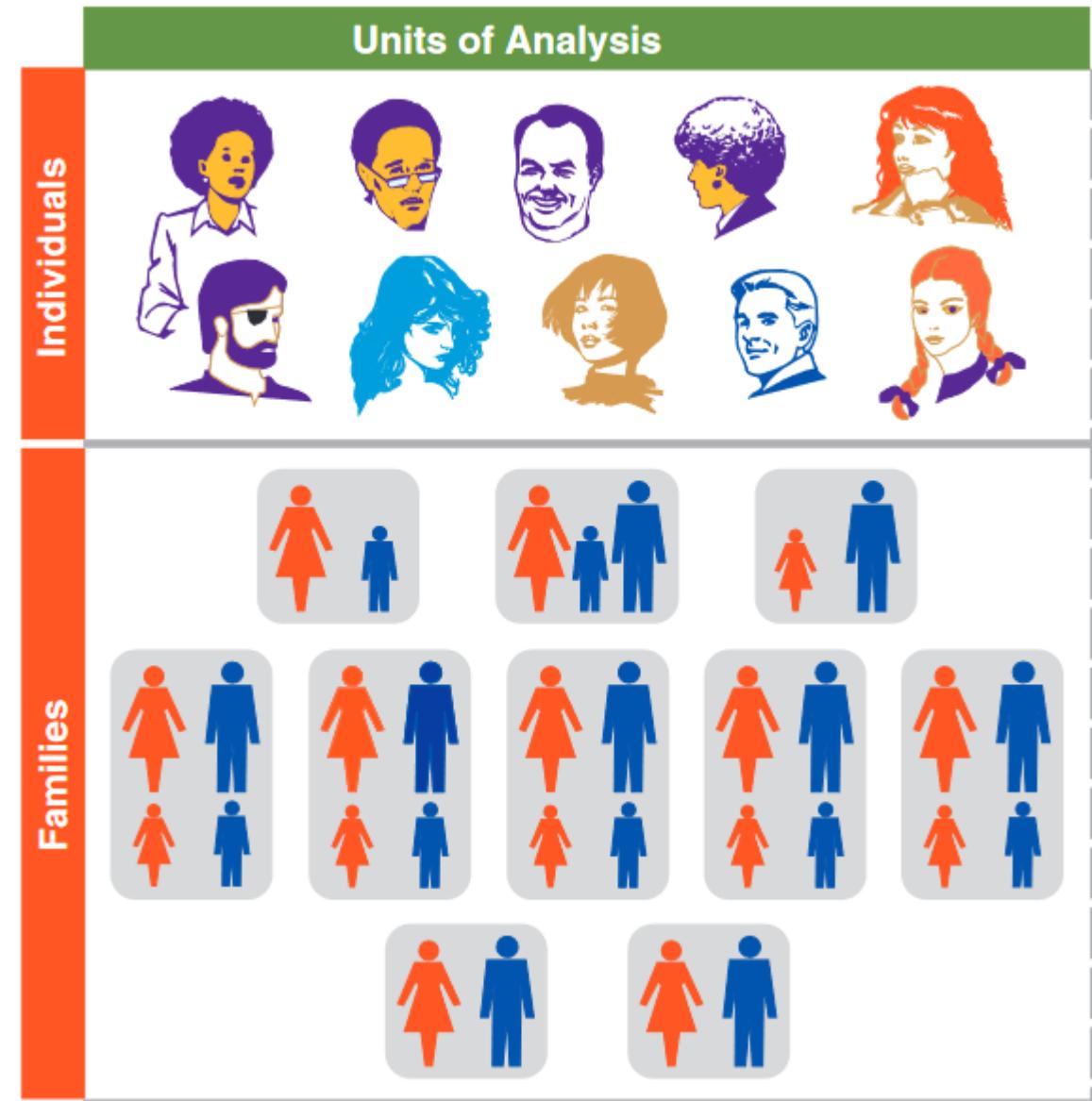
## Định tính

- Các thông tin thu thập thường mang tính diễn giải cao
- Tập trung vào tìm hiểu phân tích sâu đối tượng và bối cảnh

## Hỗn hợp

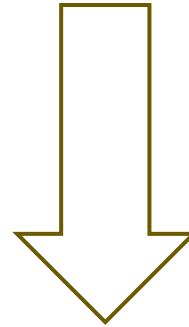
- Kết hợp cả 2 phương pháp
- Cân bằng giữa phân tích sâu về đối tượng quan tâm và tổng quan hóa bối cảnh

# Đối tượng nghiên cứu



# Quần thể - mẫu

Quần thể



Mẫu



# Tính đại diện

- Quan trọng nếu muốn dùng đặc tính của mẫu để nói về đặc tính của quần thể

- Mẫu đại diện trong một số bối cảnh là bất khả thi

[https://youtu.be/rxv\\_sb-wOkY](https://youtu.be/rxv_sb-wOkY)

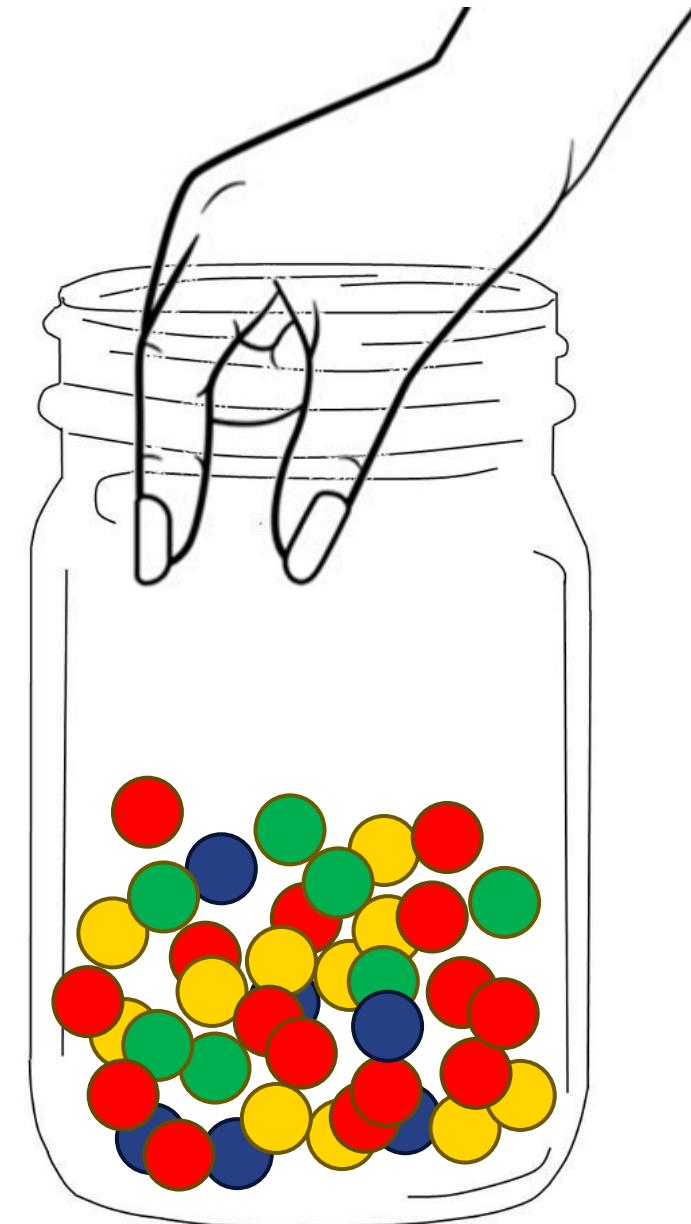
- Cỡ mẫu

[https://youtu.be/Uyd\\_Fk9cDjA?si=1uTujNmKJQmWSCtT](https://youtu.be/Uyd_Fk9cDjA?si=1uTujNmKJQmWSCtT)

[https://nckh.huph.edu.vn/sites/nckh.huph.edu.vn/files/Ph%C6%B0%C6%A1ng%20ph%C3%A1p%20ch%E1%BB%8Dn%20m%E1%BA%ABu%20v%C3%A0%20t%C3%ADnh%20t%C3%ADnh%20m%E1%BA%ABu\\_revised%20l%E1%BA%A7n%201\\_5.8.2020\\_0.pdf](https://nckh.huph.edu.vn/sites/nckh.huph.edu.vn/files/Ph%C6%B0%C6%A1ng%20ph%C3%A1p%20ch%E1%BB%8Dn%20m%E1%BA%ABu%20v%C3%A0%20t%C3%ADnh%20t%C3%ADnh%20m%E1%BA%ABu_revised%20l%E1%BA%A7n%201_5.8.2020_0.pdf)

# Phương pháp lấy mẫu

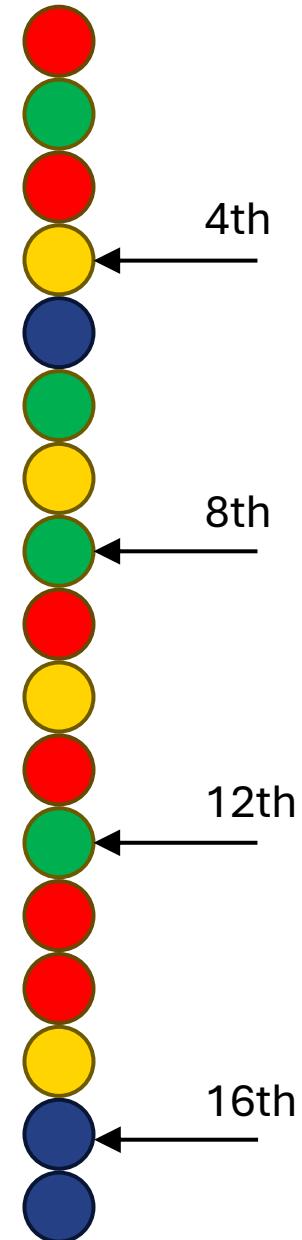
- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)



# Phương pháp lấy mẫu

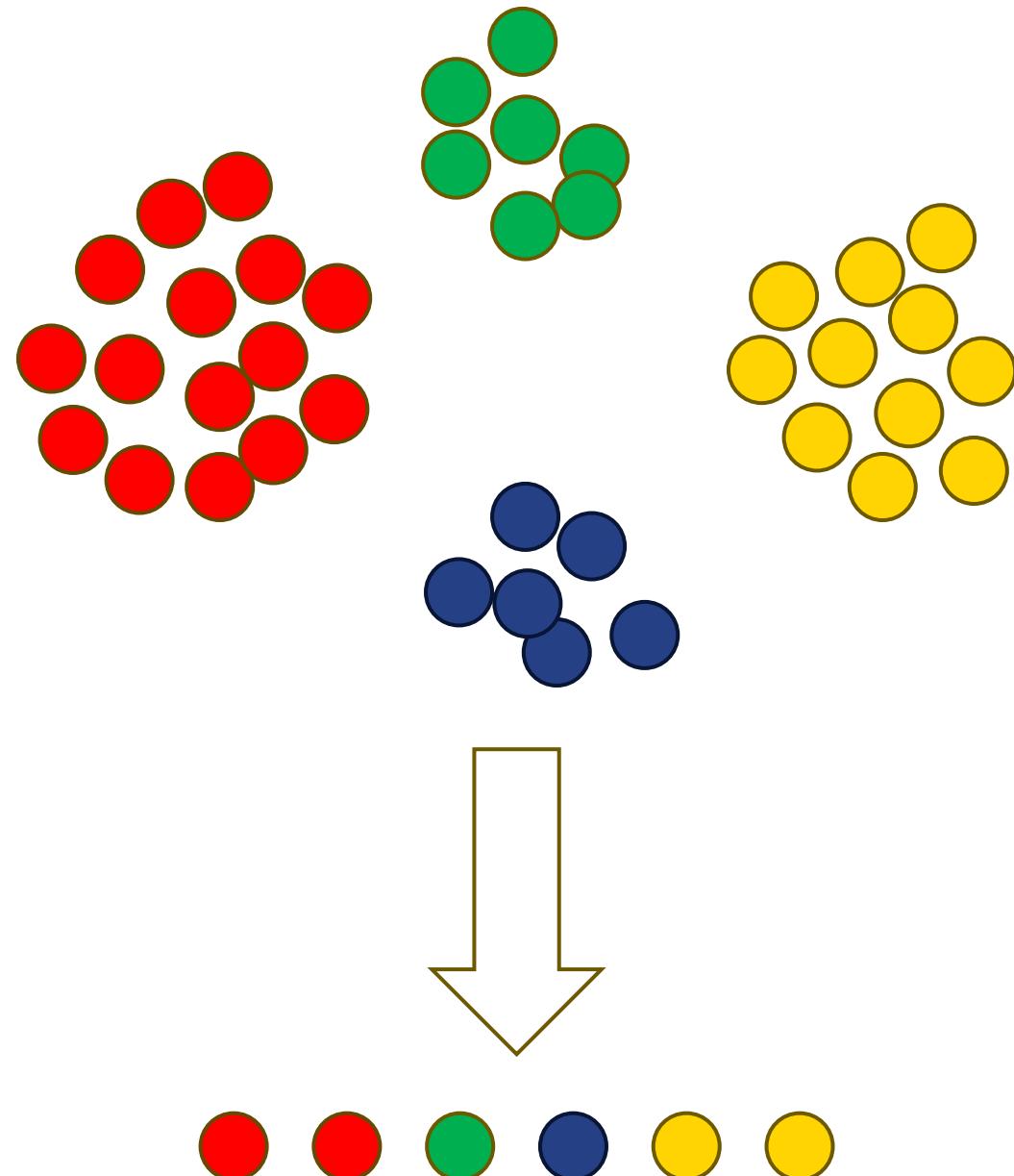
- Mẫu ngẫu nhiên  
(Probability/Random sample)

- Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
- Mẫu ngẫu nhiên hệ thống  
(Systematic sample)



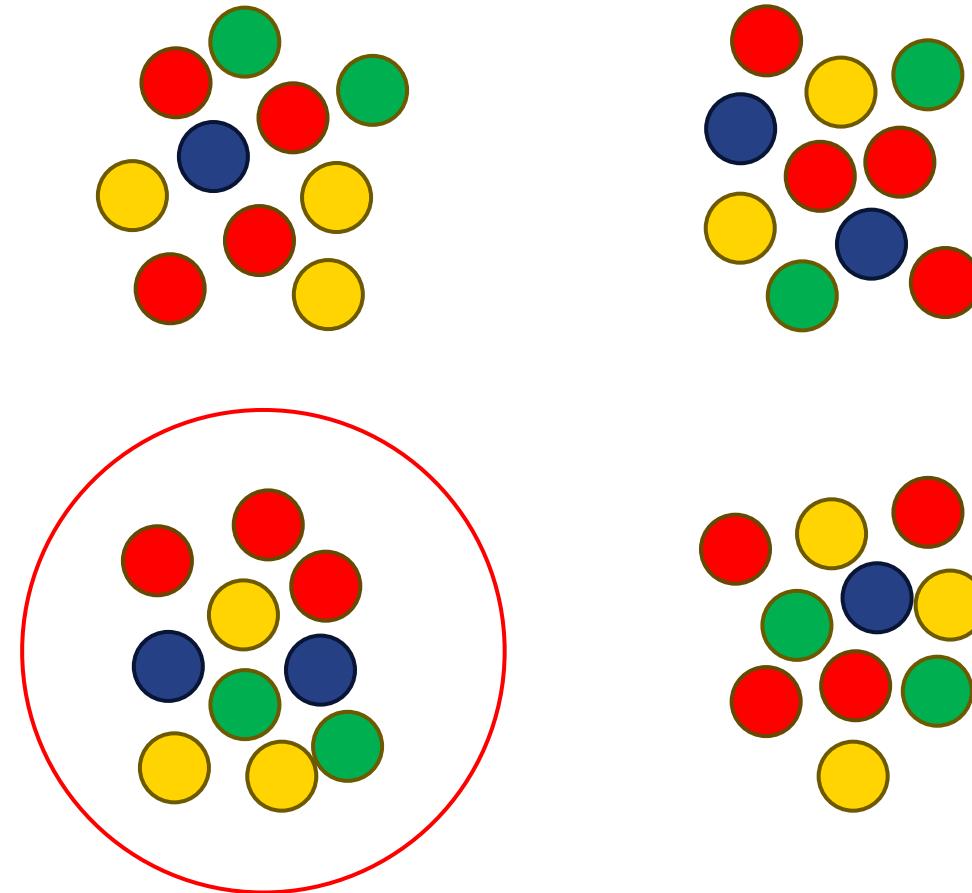
# Phương pháp lấy mẫu

- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)



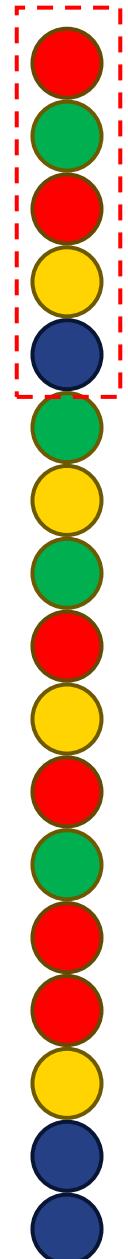
# Phương pháp lấy mẫu

- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)
  - Mẫu ngẫu nhiên cụm (Cluster sample)



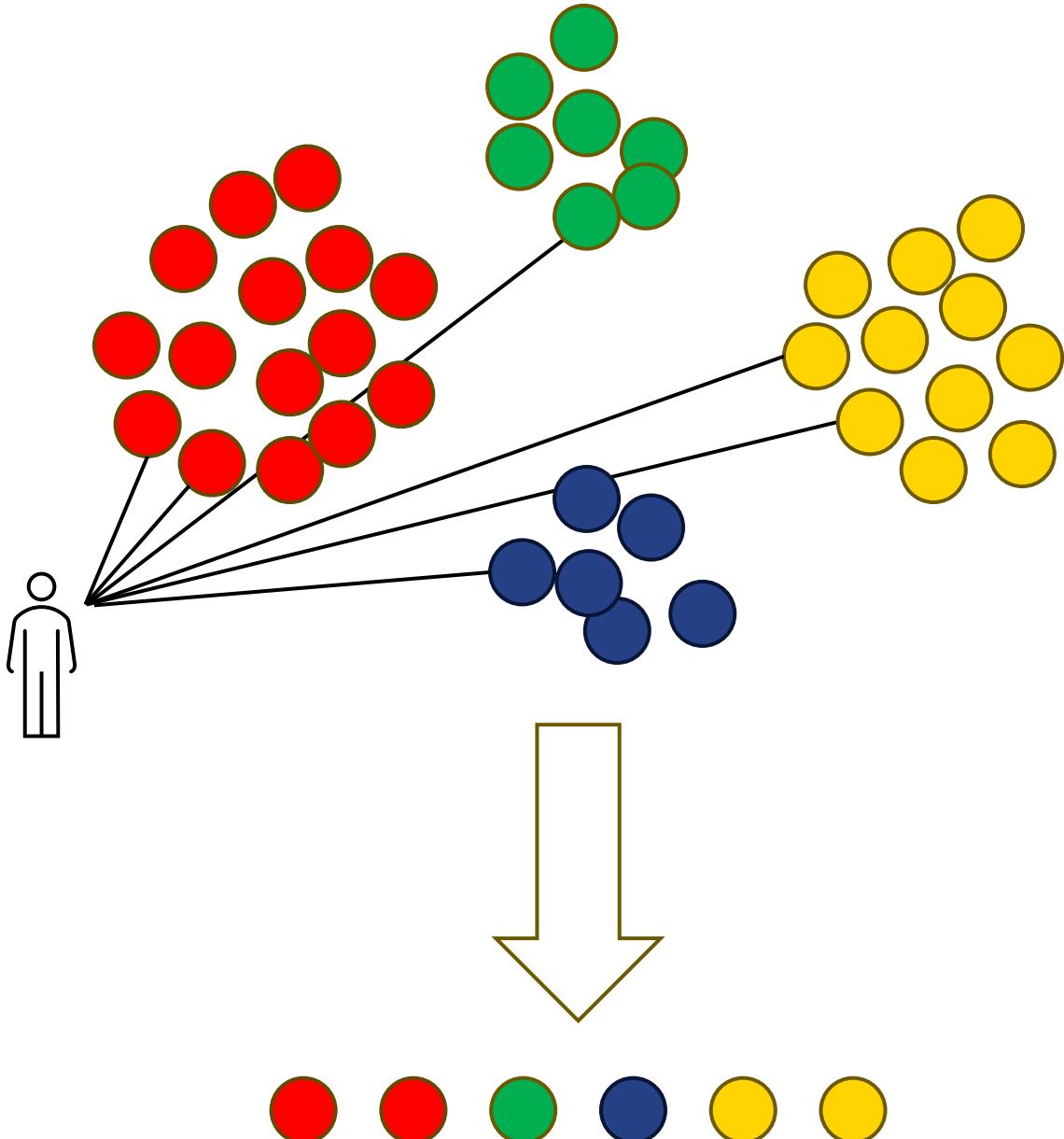
# Phương pháp lấy mẫu

- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)
  - Mẫu ngẫu nhiên cụm (Cluster sample)
- Mẫu không ngẫu nhiên  
(Nonprobability sample)
  - Mẫu thuận tiện (Convenience sample)



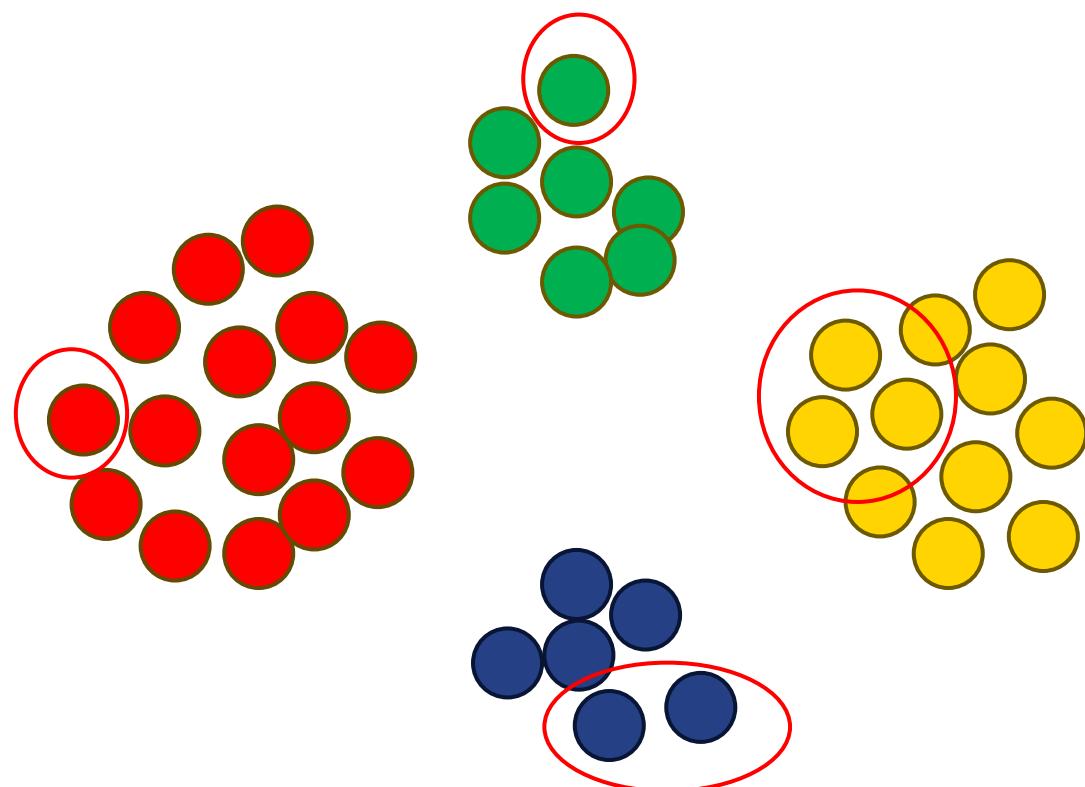
# Phương pháp lấy mẫu

- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)
  - Mẫu ngẫu nhiên cụm (Cluster sample)
- Mẫu không ngẫu nhiên  
(Nonprobability sample)
  - Mẫu thuận tiện (Convenience sample)
  - Mẫu hạn ngạnh (Quota sample)



# Phương pháp lấy mẫu

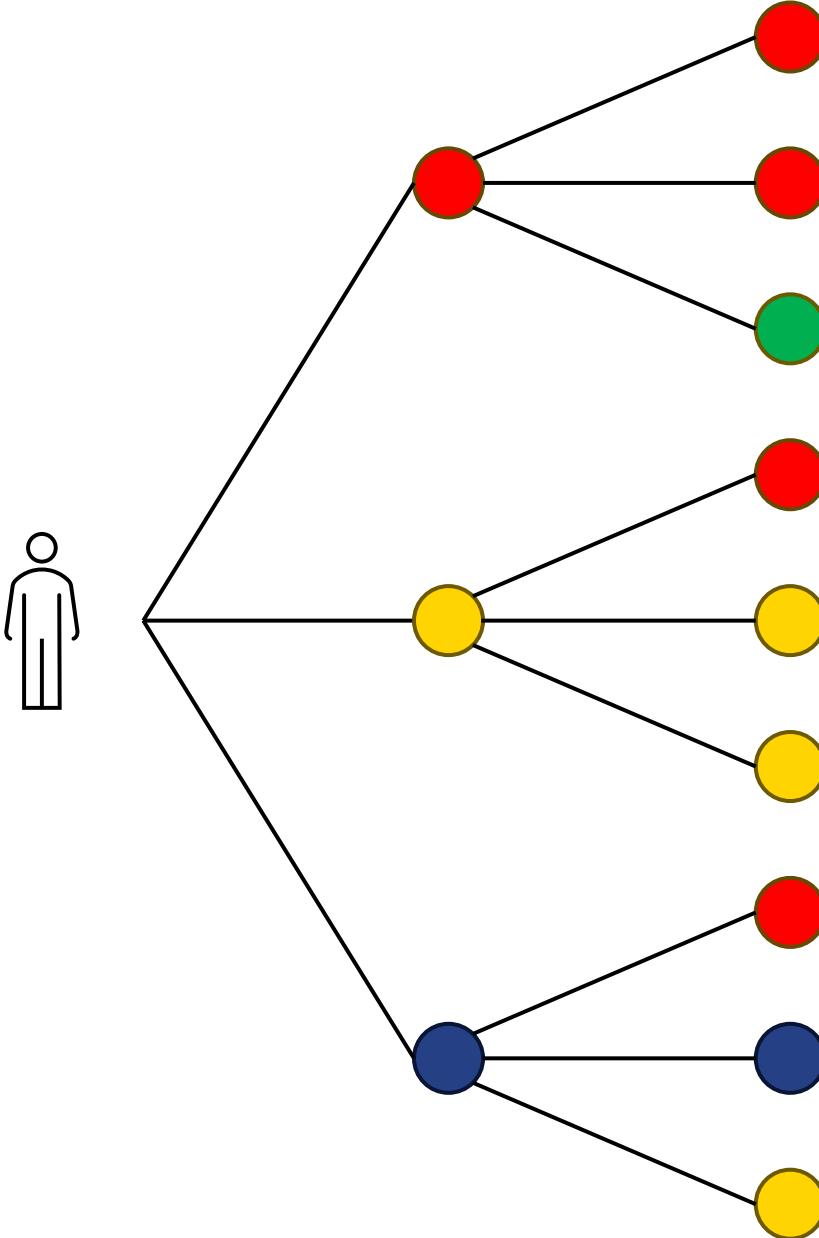
- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)
  - Mẫu ngẫu nhiên cụm (Cluster sample)
- Mẫu không ngẫu nhiên  
(Nonprobability sample)
  - Mẫu thuận tiện (Convenience sample)
  - Mẫu hạn ngạnh (Quota sample)
  - Mẫu có mục đích  
(Judgement (or purposive) sample)



- ✓ Sống ở thành phố này hơn 1 năm
- ✓ Vào công viên tập thể dục nhiều hơn 2 lần/tuần

# Phương pháp lấy mẫu

- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)
  - Mẫu ngẫu nhiên cụm (Cluster sample)
- Mẫu không ngẫu nhiên  
(Nonprobability sample)
  - Mẫu thuận tiện (Convenience sample)
  - Mẫu hạn ngạnh (Quota sample)
  - Mẫu có mục đích  
(Judgement (or purposive) sample)
  - Mẫu bóng tuyết (Snowball sample)

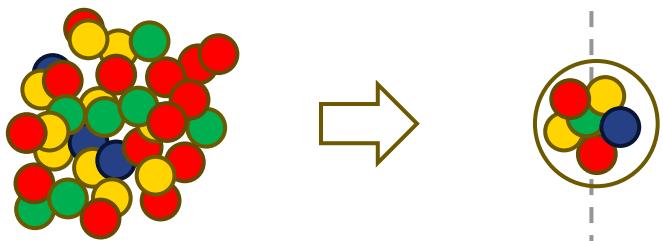


# Phương pháp lấy mẫu

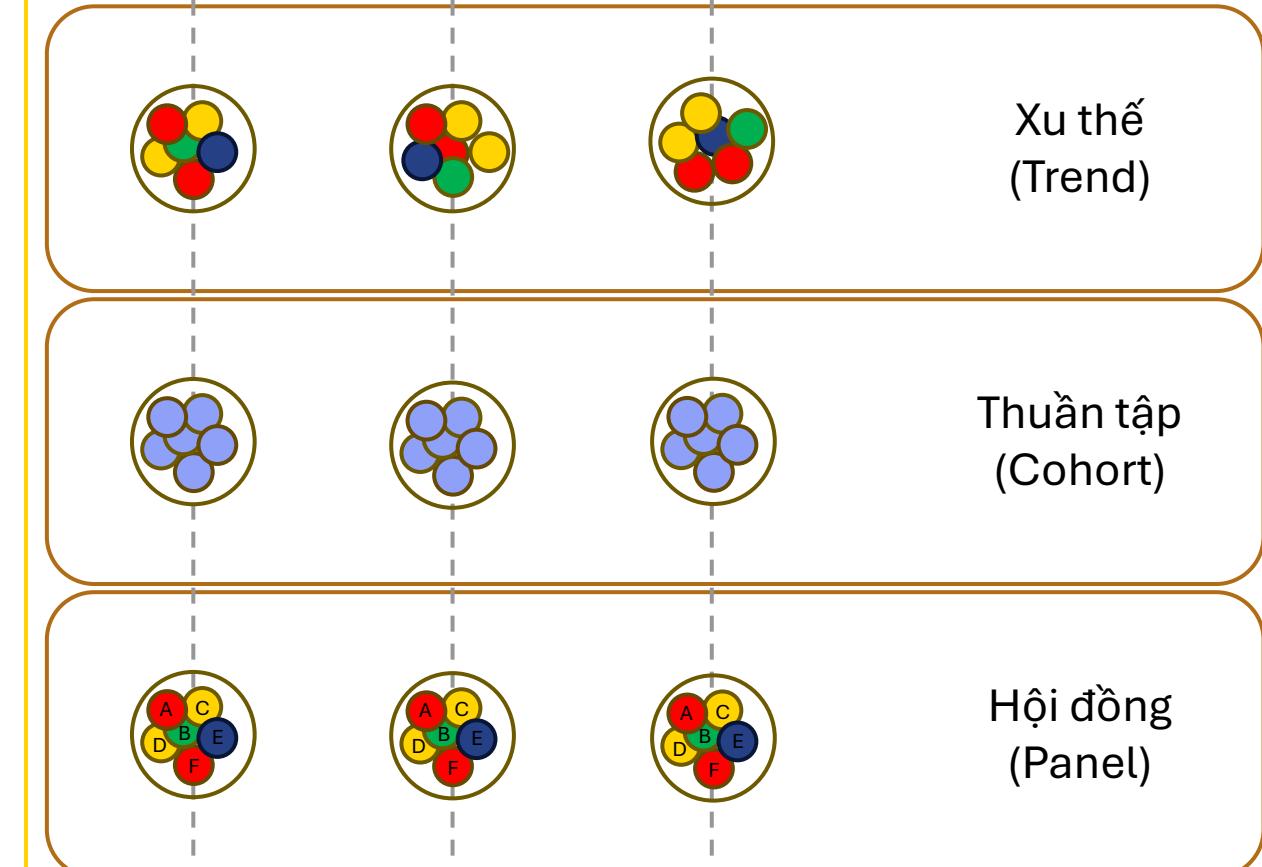
- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)
  - Mẫu ngẫu nhiên cụm (Cluster sample)
- Mẫu không ngẫu nhiên  
(Nonprobability sample)
  - Mẫu thuận tiện (Convenience sample)
  - Mẫu hạn ngạnh (Quota sample)
  - Mẫu có mục đích  
(Judgement (or purposive) sample)
  - Mẫu bóng tuyết (Snowball sample)
  - Khi lấy mẫu ngẫu nhiên là không khả thi
  - Đơn giản và tiết kiệm hơn
  - Cần có biện luận chặt chẽ về sự lựa chọn phương pháp chọn mẫu
  - Sử dụng trọng số để khắc phục tính không đại diện trong quá trình xử lý số liệu về sau

# Yếu tố thời gian của nghiên cứu

Cắt ngang/Cross-sectional



Cắt dọc/Longitudinal



# Yếu tố thời gian của nghiên cứu

	Cross-Sectional	Longitudinal		
		Trend	Cohort	Panel
Snapshot in time	X			
Measurements across time		X	X	X
Follow age group across time			X	
Study same people over time				X

# Nội dung

Giới thiệu chung

Thiết kế nghiên cứu (Định lượng)

Thu thập dữ liệu

Phân tích dữ liệu

# Dữ liệu

Dữ liệu là các sự kiện, con số, quan sát hoặc ghi chép có thể ở dạng hình ảnh, âm thanh, văn bản hoặc các phép đo vật lý.

Nguồn: <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch1/definitions/5214853-eng.htm>

# Dữ liệu phi cấu trúc và có cấu trúc

# Dữ liệu phi cấu trúc

## Dữ liệu có cấu trúc

Name	era	count	occup	HPI
Aristotle	ancient	Greece	Philosopher	31.99
Mozart	modern	Austria	Composer	30.51
Shakespeare	renaissance	UK	Writer	30.44
Michelangelo	renaissance	Italy	Painter	30.44
Einstein	contemporary	Germany	Physicist	30.21
Beethoven	modern	Germany	Composer	30.11
Van Gogh	modern	Netherland	Painter	29.74
Frida Kahlo	contemporary	Mexico	Painter	27.05

# Dữ liệu phi cấu trúc và có cấu trúc

## Dữ liệu phi cấu trúc

### Act 1, Scene 1

[Enter Sampson and Gregory, two high-ranking servants of the Capulet household, carrying swords and shields. Gregory is making fun of Sampson, who sees himself as a fearsome fighter]

**Sampson**

Gregory, on my word, we'll not carry coals.



**Gregory**

No, for then we should be colliers.  
coal workers

**Sampson**

I mean, an we be in choler we'll draw.  
if angered (our swords)

**Gregory**

Ay, while you live, draw your neck out of collar.

**Sampson**

I strike quickly, being moved.  
provoked

1

2

3

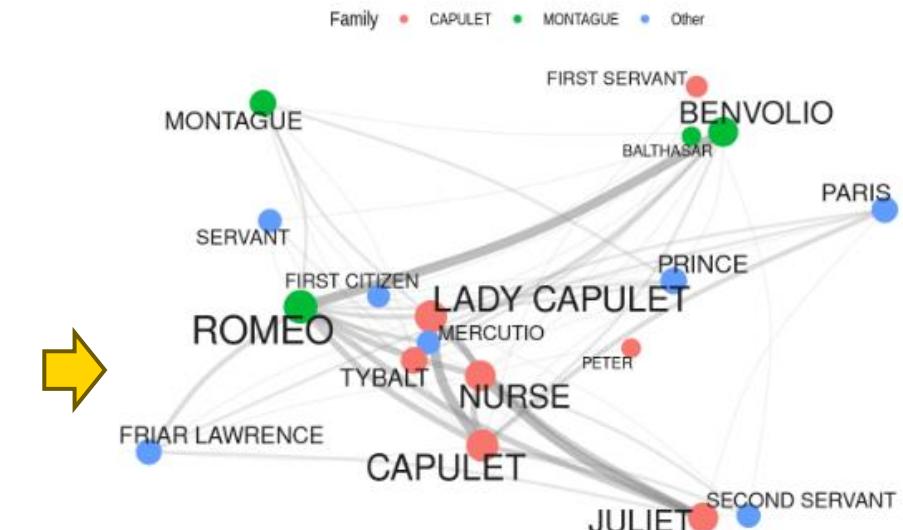
4

5

actscene	person	contrib	occurrences
ACT I_SCENE I	BENVOLIO	24	7
ACT I_SCENE I	CAPULET	2	9
ACT I_SCENE I	FIRST CITIZEN	1	2
ACT I_SCENE I	LADY CAPULET	1	10
ACT I_SCENE I	MONTAGUE	6	3
ACT I_SCENE I	PRINCE	1	3
ACT I_SCENE I	ROMEO	16	14
ACT I_SCENE I	TYBALT	2	3
ACT I_SCENE II	BENVOLIO	5	7
ACT I_SCENE II	CAPULET	3	9
ACT I_SCENE II	PARIS	2	5
ACT I_SCENE II	ROMEO	11	14
ACT I_SCENE II	SERVANT	8	3
ACT I_SCENE III	JULIET	5	11
ACT I_SCENE III	LADY CAPULET	11	10

## Dữ liệu có cấu trúc

Persona	BALTHASAR	BENVOLIO	CAPULET	FIRST CITIZEN	FIRST SERVANT
BALTHASAR	0	0	1	0	0
BENVOLIO	0	0	3	2	1
CAPULET	1	3	0	1	2
FIRST CITIZEN	0	2	1	0	0
FIRST SERVANT	0	1	2	0	0



# Cấu trúc bảng dữ liệu

## 10 mét súng ngắn hơi nam – Olympic 2016

Tên	Quốc tịch	Xếp hạng	Chiều cao	Cân nặng
Hoàng Xuân Vinh	Việt Nam	1 	1.75	75
Felipe Almeida Wu	Brazil	2 	1.69	69
Pang Wei	Trung Quốc	3 	1.79	73
Juraj Tužinský	Slovakia	4	1.84	78
Jin Jong-oh	Hàn Quốc	5	1.75	78
Giuseppe Giordano	Ý	6	1.70	74
Vladimir Gontcharov	Nga	7	1.67	59
Jitu Rai	Ấn Độ	8	1.63	64

# Cấu trúc bảng dữ liệu

## 10 mét súng ngắn hơi nam – Olympic 2016

### Đối tượng quan sát

Tên	Quốc tịch	Xếp hạng	Chiều cao	Cân nặng
Hàng Xuân Vinh	Việt Nam	1	1.75	75
Fábio Almeida Wu	Brazil	2	1.60	60
Peng Wei	Trung Quốc	3	1.73	73
Jaraj Tuzinský	Slovakia	4	1.84	76
Jin Jong-ch	Hàn Quốc	5	1.75	70
Giacomo Giordano	Ý	6	1.70	74
Vladimir Gontcharov	Nga	7	1.67	59
Jitu Rai	Ấn Độ	8	1.63	64

# Cấu trúc bảng dữ liệu

## 10 mét súng ngắn hơi nam – Olympic 2016

Biến số

Tên	Quốc tịch	Xếp hạng	Chiều cao	Cân nặng
Hoàng Xuân Vinh	Việt Nam	1 	1.75	75
Felipe Almeida Wu	Brazil	2 	1.69	69
Pang Wei	Trung Quốc	3 	1.79	73
Juraj Tužinský	Slovakia	4	1.84	78
Jin Jong-oh	Hàn Quốc	5	1.75	78
Giuseppe Giordano	Ý	6	1.70	74
Vladimir Gontcharov	Nga	7	1.67	59
Jitu Rai	Ấn Độ	8	1.65	64

# Cấu trúc bảng dữ liệu

## 10 mét súng ngắn hơi nam – Olympic 2016

### Giá trị

Tên	Quốc tịch	Xếp hạng	Chiều cao	Cân nặng
Hoàng Xuân Vinh	Việt Nam	1	1.75	75
Felipe Almeida Wu	Brazil	2	1.69	69
Pang Wei	Trung Quốc	3	1.79	73
Juraj Tužinský	Slovakia	4	1.84	78
Jin Jong-oh	Hàn Quốc	5	1.75	78
Giuseppe Giordano	Ý	6	1.70	74
Vladimir Gontcharov	Nga	7	1.67	59
Jitu Rai	Ấn Độ	8	1.63	64

# Cấu trúc bảng dữ liệu

## 10 mét súng ngắn hơi nam – Olympic 2016

Biến định danh		Biến thứ bậc		
Tên	Quốc tịch	Xếp hạng	Chiều cao	Cân nặng
Hoàng Xuân Vinh	Việt Nam	1 	1.75	75
Felipe Almeida Wu	Brazil	2 	1.69	69
Pang Wei	Trung Quốc	3 	1.79	73
Juraj Tužinský	Slovakia	4	1.84	78
Jin Jong-oh	Hàn Quốc	5	1.75	78
Giuseppe Giordano	Ý	6	1.70	74
Vladimir Gontcharov	Nga	7	1.67	59
Jitu Rai	Ấn Độ	8	1.63	64

# Biến định lượng/liên tục

- Ví dụ:
  - GDP
  - Lượng khách tham quan các công trình văn hóa hàng năm
  - Kinh phí đầu tư tu sửa
  - Số nghệ nhân hiện tại
- Đặc điểm
  - Được biểu diễn bằng số, cho phép thực hiện các phép toán
  - Có thể đo lường được
  - Liên tục (GDP) hoặc rời rạc (số nghệ nhân)

# Biến định tính – Biến định danh

- Ví dụ:
  - Giới tính: Nam, Nữ, X
  - Loại hình di tích: Đền, đền, chùa, lăng, thành cổ
  - Nguồn kinh phí bảo tồn: Nhà nước, Tư nhân, Xã hội hóa
- Đặc điểm
  - Được biểu diễn bởi tên (ký tự chữ)
  - Thường mang tính phân loại và không có tính thứ bậc

# Biến định tính/rời rạc – Biến thứ bậc

- Ví dụ:
  - Thang Likert: hoàn toàn không đồng ý – hoàn toàn đồng ý
  - Mức độ bảo tồn: Được bảo tồn tốt, Đang được phục dựng, Nguy cấp
  - Cấp độ di tích: Tỉnh, Quốc gia, Quốc gia đặc biệt
  - Đánh giá/chấm điểm: 1 – 5 ★
- Đặc điểm
  - Có tính thứ bậc tự nhiên
  - Không thể khẳng định khoảng cách bằng nhau giữa các giá trị

# Bài tập

Name	era	count	occup	HPI
Aristotle	ancient	Greece	Philosopher	31.99
Mozart	modern	Austria	Composer	30.51
Shakespeare	renaissance	UK	Writer	30.44
Michelangelo	renaissance	Italy	Painter	30.44
Einstein	contemporary	Germany	Physicist	30.21
Beethoven	modern	Germany	Composer	30.11
Van Gogh	modern	Netherland	Painter	29.74
Frida Kahlo	contemporary	Mexico	Painter	27.05

# Nguồn dữ liệu



## Sơ cấp

Tự mình thu thập

- Thực nghiệm
- Khảo sát/Bảng hỏi
- Điền dã/Đo đạc thực địa
- Phỏng vấn sâu/Thảo luận nhóm



## Thứ cấp

Đã được thu thập từ trước

- Khai thác tài liệu
- Cơ sở dữ liệu mở
- Khảo sát/Bảng hỏi của nghiên cứu khác

# Thách thức và cơ hội – Dữ liệu sơ cấp

## Thách thức

- Chi phí và thời gian
- Độ tin cậy phụ thuộc nhiều vào thiết kế nghiên cứu
- Khó khăn trong tiếp cận đối tượng quan sát
- Cần đảm bảo quyền riêng tư và bảo mật
- Khó có thể làm các nghiên cứu theo thời gian

## Cơ hội

- Phù hợp với mục tiêu cụ thể
- Cập nhật
- Kiểm soát chất lượng

# Thách thức và cơ hội – Dữ liệu thứ cấp

## Thách thức

- Không phù hợp với mục tiêu nghiên cứu
- Độ tin cậy phụ thuộc vào nguồn số liệu
- Quyền truy cập
- Định dạng không đồng nhất
- Kém cập nhật

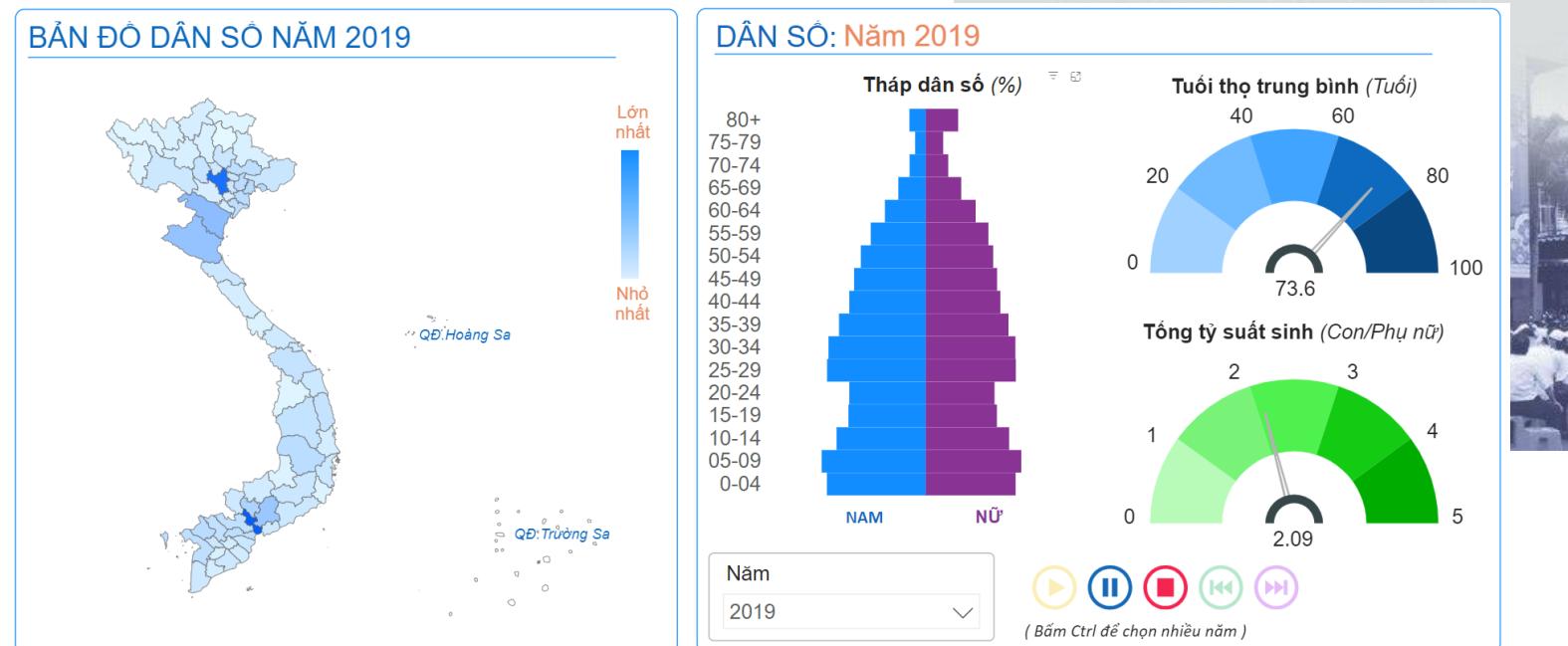
## Cơ hội

- Tiết kiệm thời gian và kinh phí
- Đa dạng
- Dữ liệu thường đã được làm sạch
- Độ phủ rộng

# Nguồn dữ liệu thứ cấp

- **Tổng điều tra dân số (Census)**

- ~100% dân số
- Cơ quan nhà nước
- Tiêu chuẩn vàng
- Tần suất thưa
- Không dễ tiếp cận



# Nguồn dữ liệu thứ cấp

- **Cổng dữ liệu mở của chính phủ và các thành phố**
  - <https://data.gov.vn/>
  - <https://data.hochiminhcity.gov.vn/>
  - <https://data.hanoi.gov.vn/>
- **Của các bộ ban ngành**
  - <https://opendata.monre.gov.vn/>
  - <https://data.mpi.gov.vn/Pages/default.aspx>

The screenshot shows the homepage of data.gov.vn. At the top, there's a search bar with placeholder text 'Bạn cần tìm dữ liệu gì?' and a magnifying glass icon. Below the search bar is the website's logo 'data.gov.vn' and a sub-header 'Điểm đầu mối công bố dữ liệu mở, cung cấp thông tin về chia sẻ dữ liệu của cơ quan nhà nước'. A green button labeled 'DỮ LIỆU MỞ' is visible. To the right, there are several icons representing different sectors: 'Địa phương', 'Kinh tế, thương mại', 'Xã hội', 'Tài nguyên', 'Biển đảo', and 'Công nghệ thông tin'. On the right side of the header, there are links for 'Lĩnh vực', 'Dịch vụ công trực tuyến', 'Tiện ích', and 'Đăng nhập'. Below the header, there's a large section titled 'Kho dữ liệu chuyển đổi số và Cổng dữ liệu mở thành phố Hà Nội'. This section includes a sub-section titled '(Data warehouse)' with a definition: 'Data warehouse (DW) hay kho dữ liệu là một hệ thống lưu trữ dữ liệu từ nhiều nguồn, nhiều môi trường khác nhau.' There are also illustrations of people interacting with data on screens and clouds.

The screenshot shows two parts of the website for the Ministry of Planning and Investment. The top part is the homepage with a green background featuring a globe and the text 'CỘNG DỮ LIỆU MỞ NGÀNH TÀI NGUYÊN MÔI TRƯỜNG'. It has a search bar and a sub-header 'Điểm đầu mối công bố dữ liệu mở, cung cấp thông tin về chia sẻ dữ liệu của cơ quan nhà nước'. The bottom part is a detailed view of the data catalog for the ministry, titled 'BỘ KẾ HOẠCH VÀ ĐẦU TƯ CỘNG DỮ LIỆU CỦA BỘ KẾ HOẠCH VÀ ĐẦU TƯ'. It features a red header with the ministry's logo and a navigation menu with links like 'Dữ liệu Bộ Kế hoạch và Đầu tư', 'Danh mục cơ sở dữ liệu', 'Danh mục dữ liệu mở', 'Văn bản - Hướng dẫn', 'Ứng dụng', 'Cung cấp thông tin và dịch vụ công trực tuyến', and 'Văn bản QPPL'. Below the menu, there's a section titled 'DANH MỤC DỮ LIỆU MỞ CỦA BỘ KẾ HOẠCH VÀ ĐẦU TƯ' with a grid of nine items, each with a thumbnail and a brief description:

Số liệu thống kê tổng hợp về đăng ký doanh nghiệp	Số liệu thống kê tổng hợp về đầu tư trực tiếp nước ngoài	Số liệu thống kê tổng hợp về đầu tư trên hệ thống mạng đầu tư quốc gia
Dữ liệu về dự án đầu công	Thông tin kinh tế - xã hội	Số liệu kinh tế - xã hội
Chỉ tiêu kinh tế - xã hội	Số liệu về doanh nghiệp nhà nước	Dữ liệu về thủ tục hành chính của bộ
Dữ liệu công khai dự toán ngân sách nhà nước	Dữ liệu về văn bản pháp luật của Bộ Kế hoạch và Đầu tư	

# Nguồn dữ liệu thứ cấp

- Khảo sát/nghiên cứu được thực hiện bởi các NGOs, trường đại học
  - <https://mics.unicef.org/country-profiles/viet-nam/4316#survey-dissemination>



Mẫu điều tra			
Hộ		Kiểm tra chất lượng nước	
• Được chọn	14.000	• Được chọn <sup>1</sup>	3.500
• Tìm thấy	13.511	• Tìm thấy	3.373
• Đã phỏng vấn	13.359	• Tỷ lệ trả lời (%)	98,2
• Tỷ lệ trả lời (%)	98,9	◦ Hộ	
		◦ Nguồn nước	98,1
Phụ nữ (từ 15-49 tuổi)		Trẻ em dưới 5 tuổi	
• Đủ điều kiện phỏng vấn	11.294	• Đủ điều kiện phỏng vấn	4.404
• Đã phỏng vấn	10.770	• Mẹ/người chăm sóc được phỏng vấn	4.329
• Tỷ lệ trả lời (%)	95,4	• Tỷ lệ trả lời (%)	98,3
Nam giới (từ 15-49 tuổi)		Trẻ em từ 5-17 tuổi	
• Số lượng trong các hộ đã phỏng vấn	11.009	• Số lượng trong các hộ đã phỏng vấn	10.869
• Đủ điều kiện phỏng vấn <sup>2</sup>	5.429	• Đủ điều kiện phỏng vấn <sup>3</sup>	7.003
• Đã phỏng vấn	4.923	• Mẹ/người chăm sóc đã phỏng vấn	6.894
• Tỷ lệ trả lời (%)	90,7	• Tỷ lệ trả lời (%)	98,4

## Dữ liệu vùng về các dịch vụ cơ bản

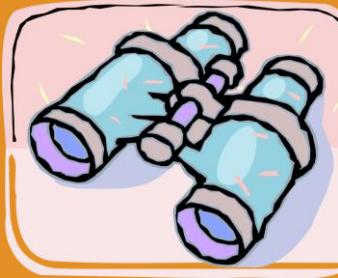
Phần trăm dân số sử dụng dịch vụ nước uống, công trình vệ sinh, và vệ sinh cơ bản, chia theo vùng/thành phố

Vùng/thành phố	Nước uống cơ bản	Công trình vệ sinh cơ bản	Chỗ rửa tay cơ bản
Cả nước	97,8	89,9	90,3
Đồng bằng sông Hồng	99,6	97,0	91,6
Hà Nội	99,4	95,9	96,4
Trung du và miền núi phía Bắc	93,8	85	84,9
Bắc Trung Bộ và Duyên hải miền Trung	97,3	93,3	92
Tây Nguyên	94,2	79,4	77,8
Đông Nam Bộ	99,3	96,3	93,5
TP Hồ Chí Minh	99,6	95,7	93,1
Đồng bằng sông Cửu Long	98,5	76,6	91,2

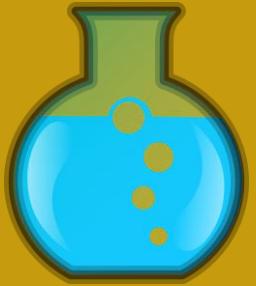
# Thu thập dữ liệu sơ cấp



Bảng hỏi  
(Questionnaire)



Hiện trường  
(Field data collection)



Thực nghiệm  
(Experiments)



Nguồn lực  
cộng đồng  
(Crowd-sourcing)

# Bảng hỏi

- Tập trung vào khía cạnh con người
- Sự tiếp nhận, phản ứng, thái độ của cộng đồng
- Sử dụng kết hợp với các nguồn dữ liệu khác



La Rotonde, Dmitry Spiros

# Bảng hỏi

## **Người tham gia tự điền      Phỏng vấn**

---

- |   |  |
|---|--|
| <ul style="list-style-type: none"><li>• Trực tuyến</li><li>• Gửi qua bưu điện</li></ul> | <ul style="list-style-type: none"><li>• Gặp mặt trực tiếp</li><li>• Qua điện thoại</li></ul> |
|---|--|

# Bảng hỏi

	<b>Người tham gia tự điền</b>	<b>Phỏng vấn</b>
	<ul style="list-style-type: none"><li>• Tiết kiệm thời gian, chi phí</li><li>• Thuận tiện, chủ động cho người tham gia</li><li>• Đảm bảo tính riêng tư</li><li>• Phạm vi tiếp cận rộng</li></ul>	<ul style="list-style-type: none"><li>• Tỷ lệ trả lời cao hơn</li><li>• Có thể thu được các thông tin chi tiết và phong phú</li></ul>
	<ul style="list-style-type: none"><li>• Tỷ lệ trả lời thấp/không tính được</li><li>• Hiểu sai câu hỏi</li><li>• Giới hạn lượng và loại câu hỏi</li><li>• Khó kiểm tra tính xác thực của người tham gia</li></ul>	<ul style="list-style-type: none"><li>• Tốn chi phí và thời gian</li><li>• Tác động của phỏng vấn viên</li></ul>

# Bảng hỏi

- Phạm trù (Construct) cần quan tâm
  - Là gì?
  - **Làm sao để đo lường?**



# Bảng hỏi – loại câu hỏi

- Câu hỏi đóng
  - Đơn giản hơn cho việc trả lời của người tham gia và việc xử lý, phân tích của người thu thập
  - Chịu ảnh hưởng góc nhìn của người phát triển bảng hỏi và đôi khi không đảm bảo tính bao trùm
  - Yes/No, câu hỏi nhiều lựa chọn, Thang Likert
- Câu hỏi mở
  - Giúp thu thập thông tin chi tiết và phong phú hơn
  - Thường đòi hỏi nhiều thời gian để xử lý và phân tích

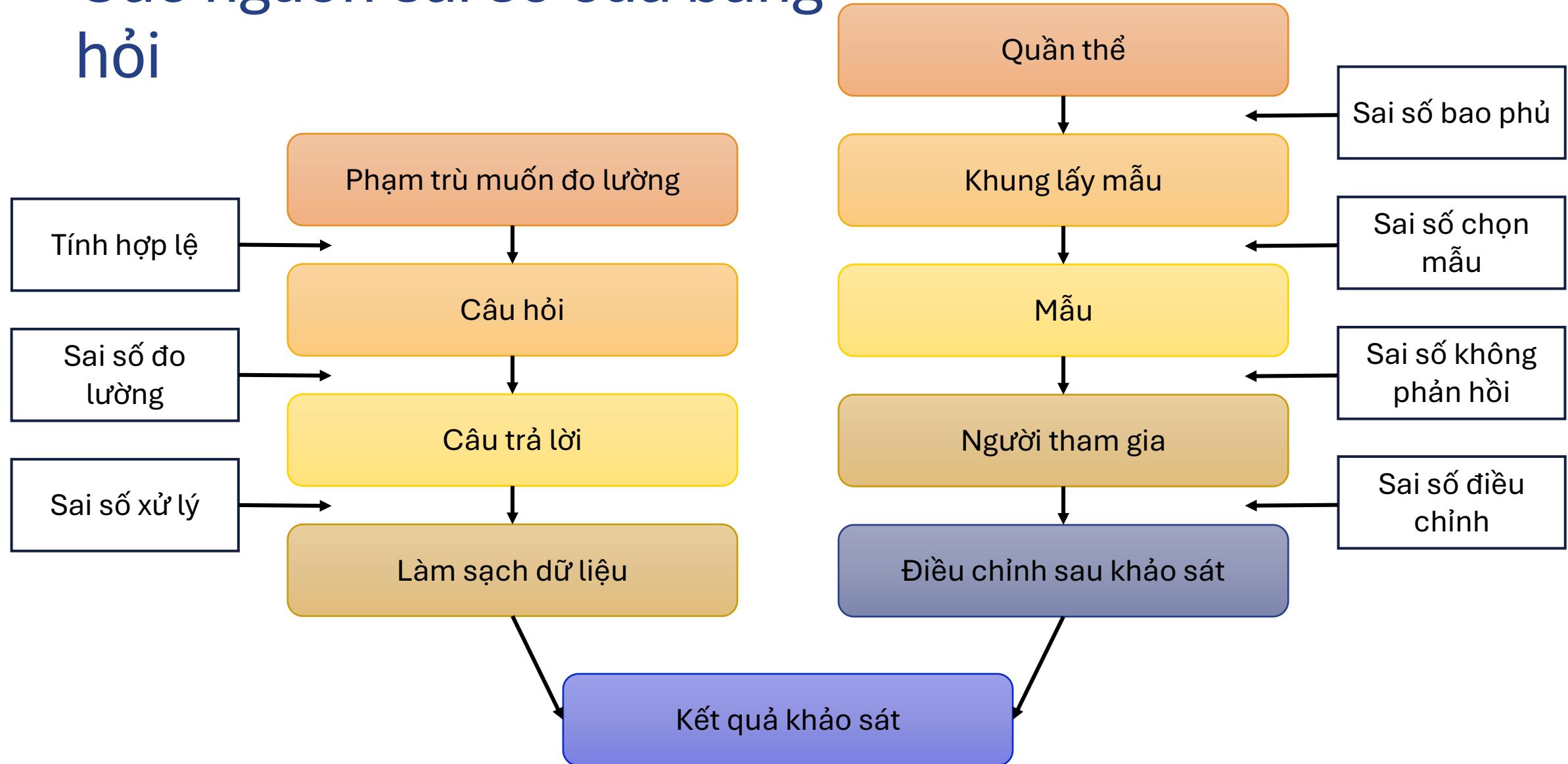
# Lưu ý khi thiết kế bảng hỏi

- Luôn **tập trung** vào câu hỏi hoặc vấn đề quan tâm
- Đưa hướng dẫn trả lời rõ ràng
- Giữ bảng hỏi và các câu hỏi ngắn gọn, đơn giản, dễ hiểu (! Dễ hiểu với người lập bảng hỏi chưa chắc đã dễ hiểu với người tham gia)
- Không yêu cầu trả lời tất cả các câu hỏi
- Sử dụng “Chuyển đến” để chuyển đến cụm câu hỏi liên quan
- Kết thúc bằng các câu hỏi nhân khẩu học như “Tuổi, giới tính, v.v...”

# Lưu ý khi thiết kế bảng hỏi

- Đối với phạm trù tiềm ẩn: dùng nhiều hơn 1 câu hỏi
- Tránh câu hỏi kép
- Không đặt câu hỏi mang tính dẫn dắt câu trả lời
- Đưa câu hỏi cụ thể tránh đòi hỏi người trả lời phải gợi nhớ quá nhiều
- **Thử nghiệm, thử nghiệm, thử nghiệm**

# Các nguồn sai số của bảng hỏi



# Hiện trường



# Hiện trường

- Có cấu trúc
  - Có lịch trình cụ thể
  - Có yếu tố quan sát cụ thể
  - Giúp giảm thiên kiến
  - Có thể lặp lại và kiểm chứng
- Phi cấu trúc
  - Thu thập dữ liệu và thông tin nhiều nhất có thể
  - Dựa vào trực giác của người quan sát
- Vấn đề đạo đức và quyền riêng tư (chụp ảnh, đồng ý của người tham gia)

# Thực nghiệm

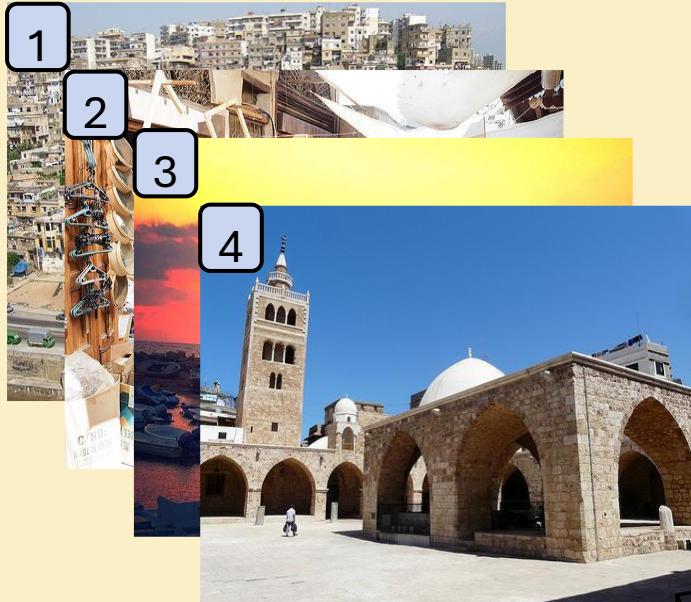


Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, 322(5908), 1681–1685.  
<https://doi.org/10.1126/science.1161405>

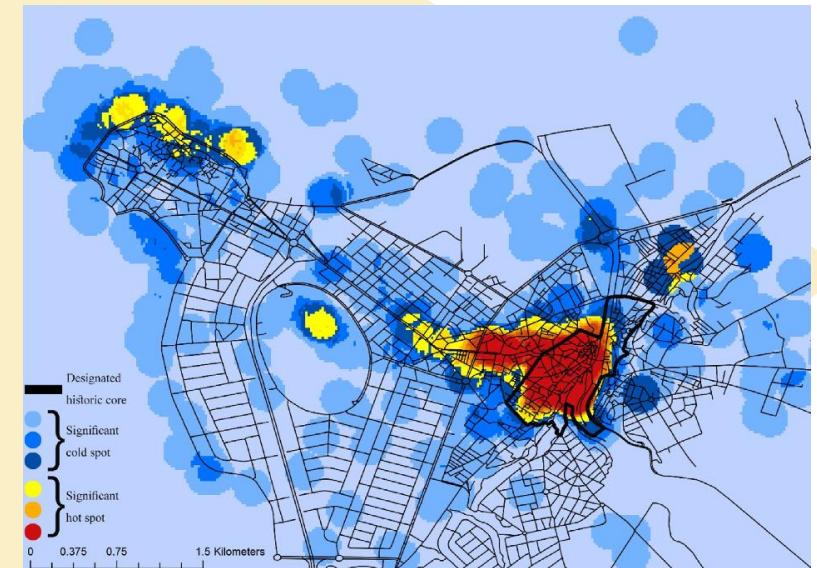
# Nguồn lực đám đông

- Crowd-sourcing/Citizen science (Khoa học công dân)
- Vận dụng nguồn lực từ cộng đồng
- Ưu điểm
  - Tăng cường nguồn dữ liệu
  - Giảm thiểu chi phí
  - Tăng cường sự tham gia/quan tâm của cộng đồng
- Hình thức
  - Khảo sát trực tuyến
  - Ứng dụng điện thoại di động
  - Phân tích dữ liệu từ mạng xã hội

# Nguồn lực đám đông



Mã	Phân loại	Số lượt xem	Số lượt tương tác
1	Cảnh thành phố	5262	213
2	Chợ	2654	102
3	Cảnh thiên nhiên	18265	574
4	Công trình văn hóa	7512	356
...	...	...	...



Ginzarly, M., Pereira Roders, A., & Teller, J. (2019). Mapping historic urban landscape values through social media. *Journal of Cultural Heritage*, 36, 1–11. <https://doi.org/10.1016/j.culher.2018.10.002>

# Nội dung

Giới thiệu chung

Thiết kế nghiên cứu (Định lượng)

Thu thập dữ liệu

Phân tích dữ liệu

# Các bước phân tích

- Tiền xử lý dữ liệu
- Tìm hiểu dữ liệu/Biểu diễn dữ liệu
- Thống kê mô tả
- Thống kê suy luận

# Các bước phân tích

- Tiền xử lý dữ liệu
- Tìm hiểu dữ liệu/Biểu diễn dữ liệu
- Thống kê mô tả
- Thống kê suy luận

# Tiền xử lý dữ liệu

## Dữ liệu thực tế thường

- Không đầy đủ: người trả lời bỏ qua câu hỏi, thiết bị đo gặp sự cố dẫn đến gián đoạn trong dữ liệu
- Chứa nhiều dữ liệu nhiễu (noise), dữ liệu ngoại lai (outliers)
- Không thống nhất

# Đo lường chất lượng dữ liệu

- Độ chính xác
  - Sai lệch do thiết bị đo hoặc quá trình ghi nhận dữ liệu
- Tính đầy đủ
  - Thiết bị đo gấp sự cố, gián đoạn trong kết nối dữ liệu, người tham gia không cung cấp câu trả lời
- Tính nhất quán
  - Viết hoa viết thường, Hà Nội và Hanoi, độ phân giải không gian và thời gian, 7 và bảy
- Tính cập nhật
  - Dữ liệu có cập nhật so với thực tế
- Tính hợp lệ
  - Vd: email không có @
- Tính duy nhất
  - Trùng lắp dữ liệu

# Tiền xử lý dữ liệu

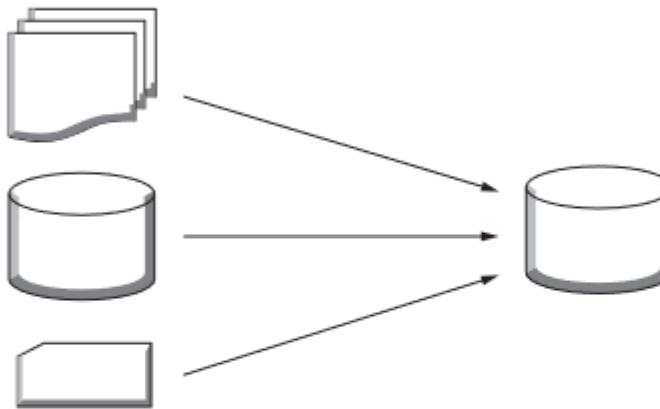
... là bước quan trọng để chuyển từ dữ liệu thô sang dạng dữ liệu  
**sẵn sàng sử dụng** cho bước phân tích

# Tiền xử lý dữ liệu

- Làm sạch dữ liệu



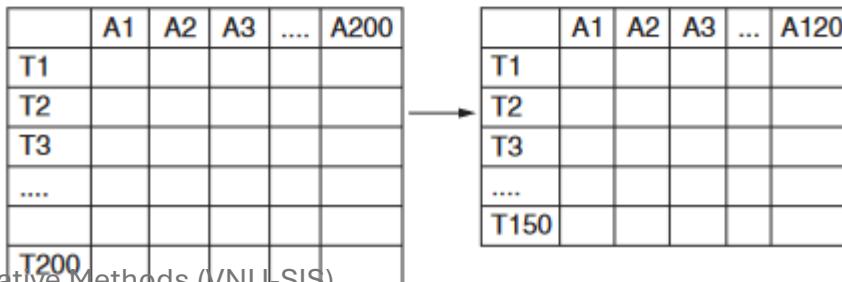
- Tích hợp dữ liệu



- Biến đổi dữ liệu

-17, 25, 39, 128, -39 → 0.17, 0.25, 0.39, 1.28, -0.39

- Tinh giản dữ liệu



# Các bước phân tích

- Tiền xử lý dữ liệu
- Tìm hiểu dữ liệu/Biểu diễn dữ liệu
- Thống kê mô tả
- Thống kê suy luận

# Tìm hiểu dữ liệu/Biểu diễn dữ liệu

- Là bước không thể bỏ qua
- Giúp phát hiện những vấn đề trong dữ liệu
- Giúp có hình dung chung về dữ liệu và các mối tương quan giữa các dữ liệu
- Phát triển giả thuyết mới

# Biểu diễn dữ liệu

- BMI và số bước chân
  - Nam
  - Nữ
- 2 nhóm sinh viên
  - Giả thuyết: Tương quan giữa số bước chân và BMI
  - Có thể rút ra được gì từ tập dữ liệu này

a

The image shows two side-by-side windows of the 'Lister' application. Both windows have a title bar 'Lister - [c:\Users\YAN...]' and a menu bar with 'File', 'Edit', 'Options', 'Encoding', and 'Help'. The left window displays data from row 3 to 4, with columns 'ID', 'steps', and 'bmi'. The right window displays data from row 12 to 42, also with columns 'ID', 'steps', and 'bmi'. The data in both windows is identical.

ID	steps	bmi
3	15000	17.0
4	14861	17.2
5		
9		
12		
14		
15	15000	16.9
16	15000	16.9
21	6	16.8
23	14861	16.8
26	7	16.8
28	14699	17.3
31	10	16.9
33	11	16.9
34	13	16.8
35	14	16.8
36	15	16.8
38	16	16.8
39	17	16.8
41	18	16.8
44	19	16.8
45	20	16.8
27	21	16.8
29	22	16.8
30	23	16.8
32	24	16.8
37	25	16.8
40	26	16.8
42	27	16.8
43	28	16.8
44	29	16.8
45	30	16.8
46	31	16.8
47	32	16.8
48	33	16.8
49	34	16.8
50	35	16.8
51	36	16.8
52	37	16.8
53	38	16.8
54	39	16.8
55	40	16.8
56	41	16.8
57	42	16.8
58	43	16.8
59	44	16.8
60	45	16.8
61	46	16.8
62	47	16.8
63	48	16.8
64	49	16.8
65	50	16.8
66	51	16.8
67	52	16.8
68	53	16.8
69	54	16.8
70	55	16.8
71	56	16.8
72	57	16.8
73	58	16.8
74	59	16.8
75	60	16.8
76	61	16.8
77	62	16.8
78	63	16.8
79	64	16.8
80	65	16.8
81	66	16.8
82	67	16.8
83	68	16.8
84	69	16.8
85	70	16.8
86	71	16.8
87	72	16.8
88	73	16.8
89	74	16.8
90	75	16.8
91	76	16.8
92	77	16.8
93	78	16.8
94	79	16.8
95	80	16.8
96	81	16.8
97	82	16.8
98	83	16.8
99	84	16.8
100	85	16.8
101	86	16.8
102	87	16.8
103	88	16.8
104	89	16.8
105	90	16.8
106	91	16.8
107	92	16.8
108	93	16.8
109	94	16.8
110	95	16.8
111	96	16.8
112	97	16.8
113	98	16.8
114	99	16.8
115	100	16.8
116	101	16.8
117	102	16.8
118	103	16.8
119	104	16.8
120	105	16.8
121	106	16.8
122	107	16.8
123	108	16.8
124	109	16.8
125	110	16.8
126	111	16.8
127	112	16.8
128	113	16.8
129	114	16.8
130	115	16.8
131	116	16.8
132	117	16.8
133	118	16.8
134	119	16.8
135	120	16.8
136	121	16.8
137	122	16.8
138	123	16.8
139	124	16.8
140	125	16.8
141	126	16.8
142	127	16.8
143	128	16.8
144	129	16.8
145	130	16.8
146	131	16.8
147	132	16.8
148	133	16.8
149	134	16.8
150	135	16.8
151	136	16.8
152	137	16.8
153	138	16.8
154	139	16.8
155	140	16.8
156	141	16.8
157	142	16.8
158	143	16.8
159	144	16.8
160	145	16.8
161	146	16.8
162	147	16.8
163	148	16.8
164	149	16.8
165	150	16.8
166	151	16.8
167	152	16.8
168	153	16.8
169	154	16.8
170	155	16.8
171	156	16.8
172	157	16.8
173	158	16.8
174	159	16.8
175	160	16.8
176	161	16.8
177	162	16.8
178	163	16.8
179	164	16.8
180	165	16.8
181	166	16.8
182	167	16.8
183	168	16.8
184	169	16.8
185	170	16.8
186	171	16.8
187	172	16.8
188	173	16.8
189	174	16.8
190	175	16.8
191	176	16.8
192	177	16.8
193	178	16.8
194	179	16.8
195	180	16.8
196	181	16.8
197	182	16.8
198	183	16.8
199	184	16.8
200	185	16.8
201	186	16.8
202	187	16.8
203	188	16.8
204	189	16.8
205	190	16.8
206	191	16.8
207	192	16.8
208	193	16.8
209	194	16.8
210	195	16.8
211	196	16.8
212	197	16.8
213	198	16.8
214	199	16.8
215	200	16.8
216	201	16.8
217	202	16.8
218	203	16.8
219	204	16.8
220	205	16.8
221	206	16.8
222	207	16.8
223	208	16.8
224	209	16.8
225	210	16.8
226	211	16.8
227	212	16.8
228	213	16.8
229	214	16.8
230	215	16.8
231	216	16.8
232	217	16.8
233	218	16.8
234	219	16.8
235	220	16.8
236	221	16.8
237	222	16.8
238	223	16.8
239	224	16.8
240	225	16.8
241	226	16.8
242	227	16.8
243	228	16.8
244	229	16.8
245	230	16.8
246	231	16.8
247	232	16.8
248	233	16.8
249	234	16.8
250	235	16.8
251	236	16.8
252	237	16.8
253	238	16.8
254	239	16.8
255	240	16.8
256	241	16.8
257	242	16.8
258	243	16.8
259	244	16.8
260	245	16.8
261	246	16.8
262	247	16.8
263	248	16.8
264	249	16.8
265	250	16.8
266	251	16.8
267	252	16.8
268	253	16.8
269	254	16.8
270	255	16.8
271	256	16.8
272	257	16.8
273	258	16.8
274	259	16.8
275	260	16.8
276	261	16.8
277	262	16.8
278	263	16.8
279	264	16.8
280	265	16.8
281	266	16.8
282	267	16.8
283	268	16.8
284	269	16.8
285	270	16.8
286	271	16.8
287	272	16.8
288	273	16.8
289	274	16.8
290	275	16.8
291	276	16.8
292	277	16.8
293	278	16.8
294	279	16.8
295	280	16.8
296	281	16.8
297	282	16.8
298	283	16.8
299	284	16.8
300	285	16.8
301	286	16.8
302	287	16.8
303	288	16.8
304	289	16.8
305	290	16.8
306	291	16.8
307	292	16.8
308	293	16.8
309	294	16.8
310	295	16.8
311	296	16.8
312	297	16.8
313	298	16.8
314	299	16.8
315	300	16.8
316	301	16.8
317	302	16.8
318	303	16.8
319	304	16.8
320	305	16.8
321	306	16.8
322	307	16.8
323	308	16.8
324	309	16.8
325	310	16.8
326	311	16.8
327	312	16.8
328	313	16.8
329	314	16.8
330	315	16.8
331	316	16.8
332	317	16.8
333	318	16.8
334	319	16.8
335	320	16.8
336	321	16.8
337	322	16.8
338	323	16.8
339	324	16.8
340	325	16.8
341	326	16.8
342	327	16.8
343	328	16.8
344	329	16.8
345	330	16.8
346	331	16.8
347	332	16.8
348	333	16.8
349	334	16.8
350	335	16.8
351	336	16.8
352	337	16.8
353	338	16.8
354	339	16.8
355	340	16.8
356	341	16.8
357	342	16.8
358	343	16.8
359	344	16.8
360	345	16.8
361	346	16.8
362	347	16.8
363	348	16.8
364	349	16.8
365	350	16.8
366	351	16.8
367	352	16.8
368	353	16.8
369	354	16.8
370	355	16.8
371	356	16.8
372	357	16.8
373	358	16.8
374	359	16.8
375	360	16.8
376	361	16.8
377	362	16.8
378	363	16.8
379	364	16.8
380	365	16.8
381	366	16.8
382	367	16.8
383	368	16.8
384	369	

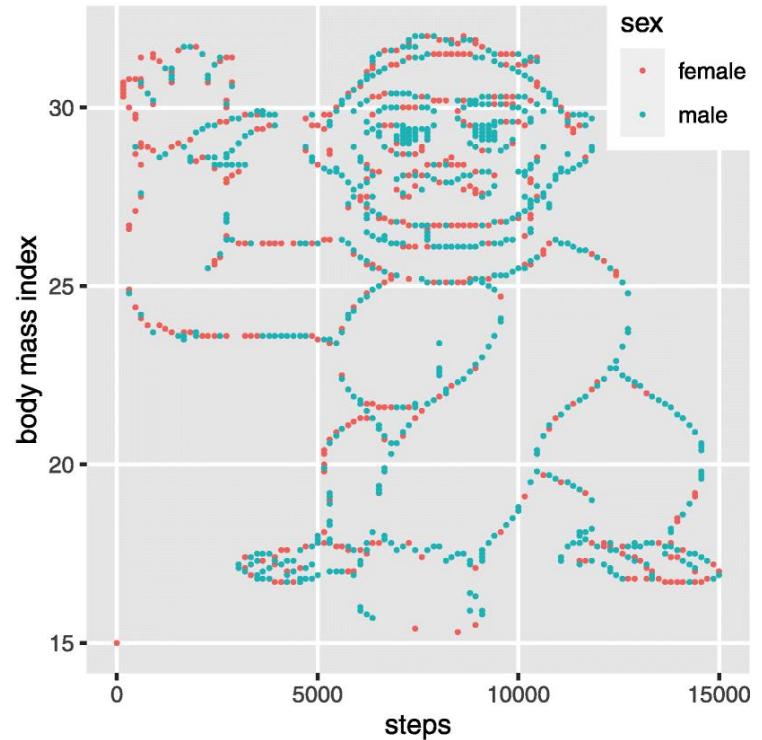
# Biểu diễn dữ liệu

- BMI và số bước chân
  - Nam
  - Nữ
- 2 nhóm sinh viên
  - Giả thuyết: Tương quan giữa số bước chân và BMI
  - Có thể rút ra được gì từ tập dữ liệu này

a

ID	steps	bmi	
3	15000	17.0	
4	14861	17.2	
5			
9			
12			
14			
15	15000	16.9	
16	15000	16.9	
21	6	14861	16.8
23	7	14861	16.8
26	8	14699	17.3
28	10	14560	20.5
31	11	14560	20.6
33	13	14560	20.5
34	17	14560	20.4
35	18	14560	20.4
36	19	14560	19.8
38	20	14560	19.7
39	22	14560	19.7
41	24	14560	19.6
44	25	14560	19.6
45	27	14560	19.6
29	29	14560	17.4
30	30	14560	17.4
32	32	14398	20.9
37	37	14398	17.5
40	40	14398	17.1
42	42	14259	21.1
43	43	14259	21.1
44	44	14259	20.0

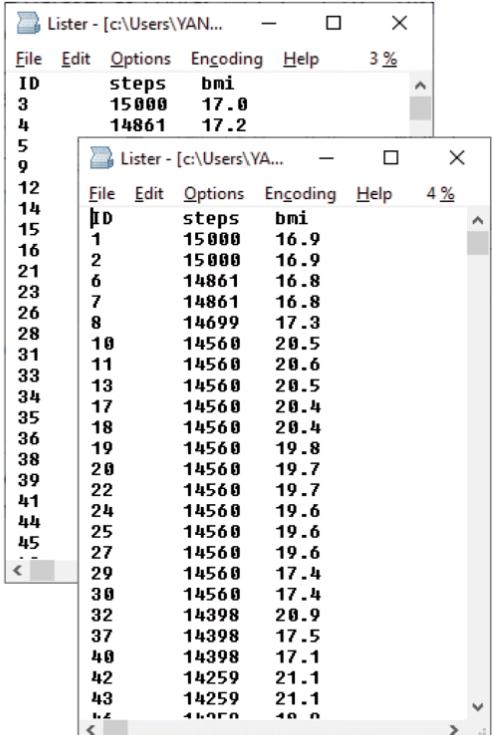
b



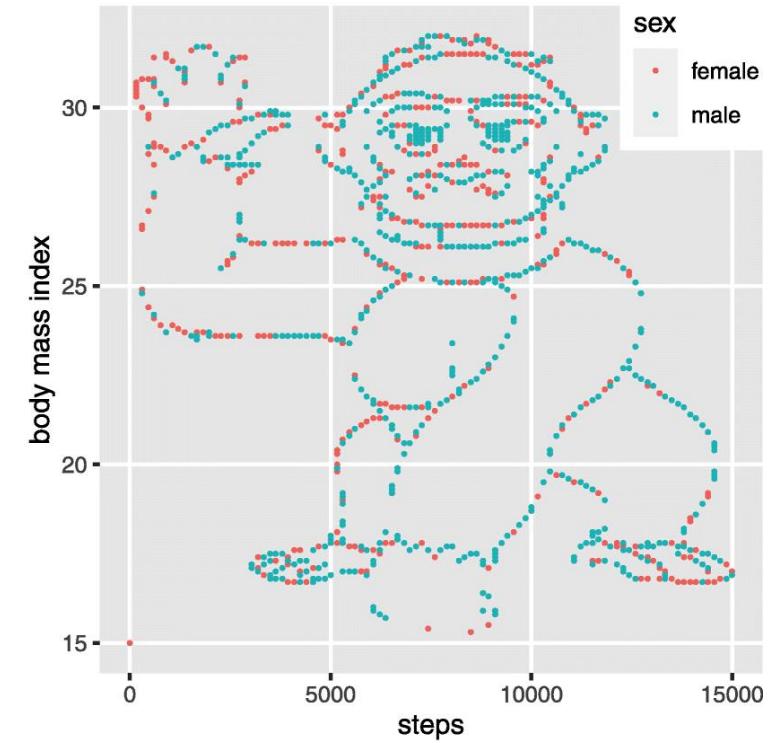
# Biểu diễn dữ liệu

- BMI và số bước chân
  - Nam
  - Nữ
- 2 nhóm sinh viên
  - Giả thuyết: Tương quan giữa số bước chân và BMI
  - Có thể rút ra được gì từ tập dữ liệu này

a



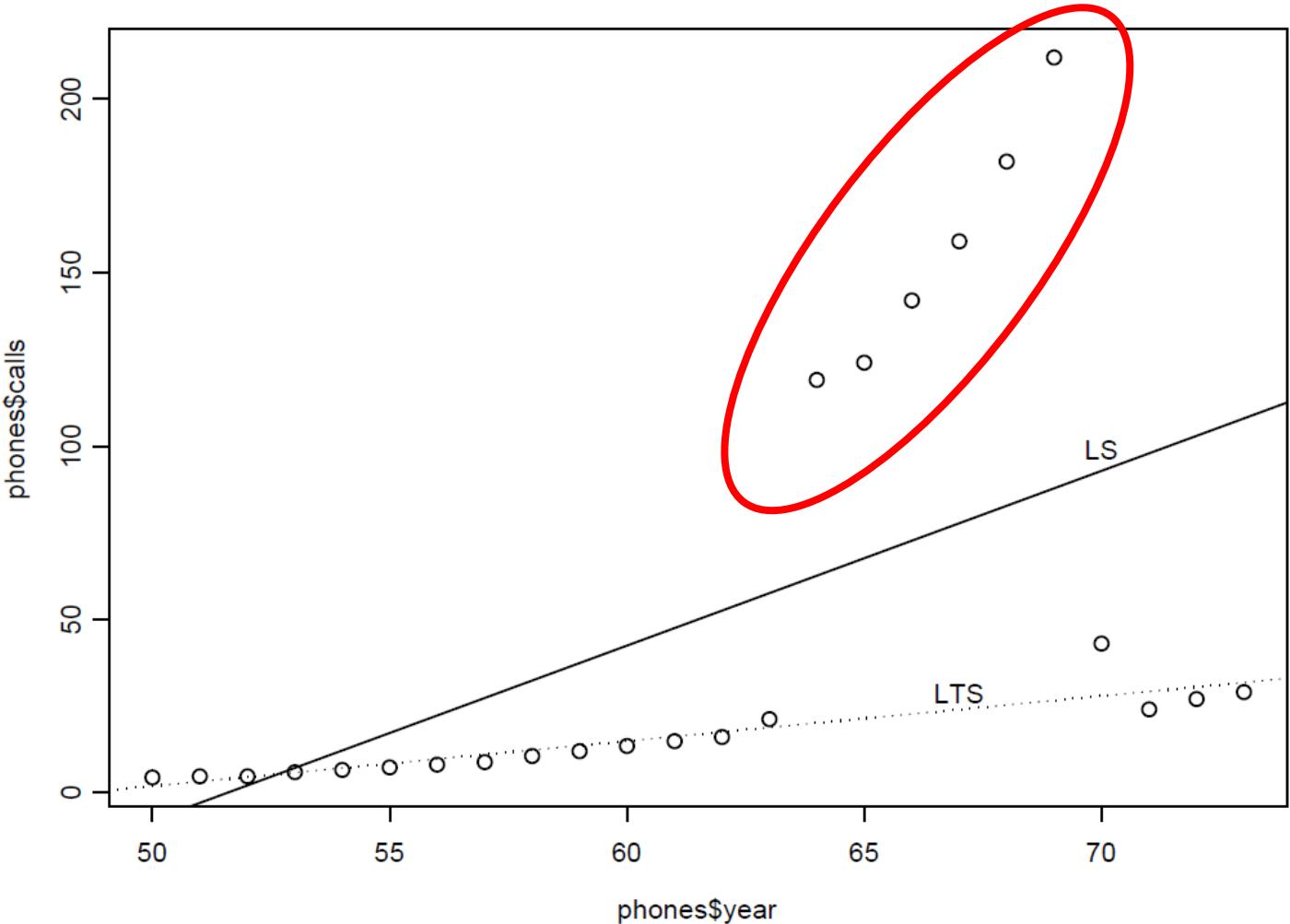
b



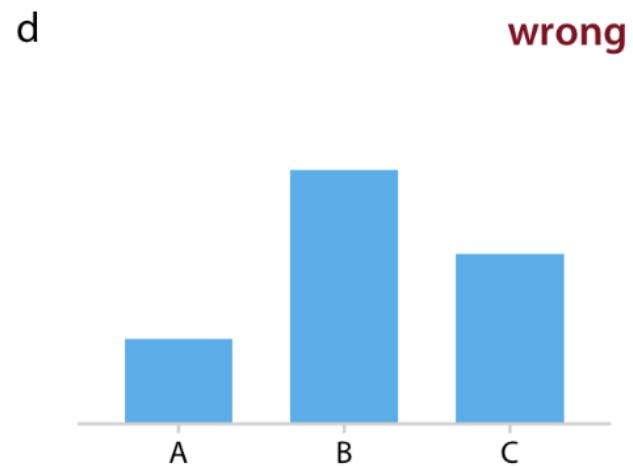
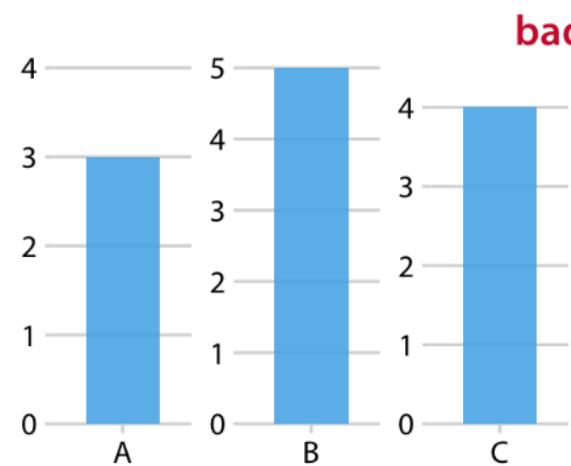
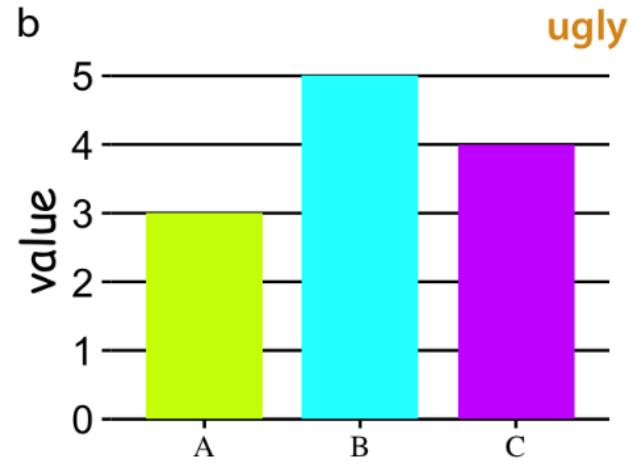
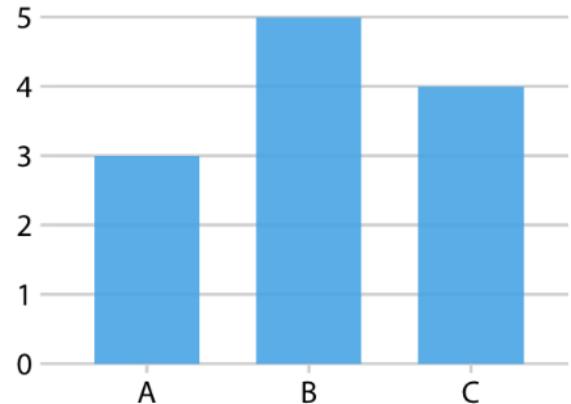
	Gorilla <u>not</u> discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

# Biểu diễn dữ liệu

- Dữ liệu điện thoại
- Cuộc gọi (triệu) ra nước ngoài từ Bỉ từ 1950-1973.
- Dữ liệu bất thường từ 1964-1970



# Biểu diễn dữ liệu

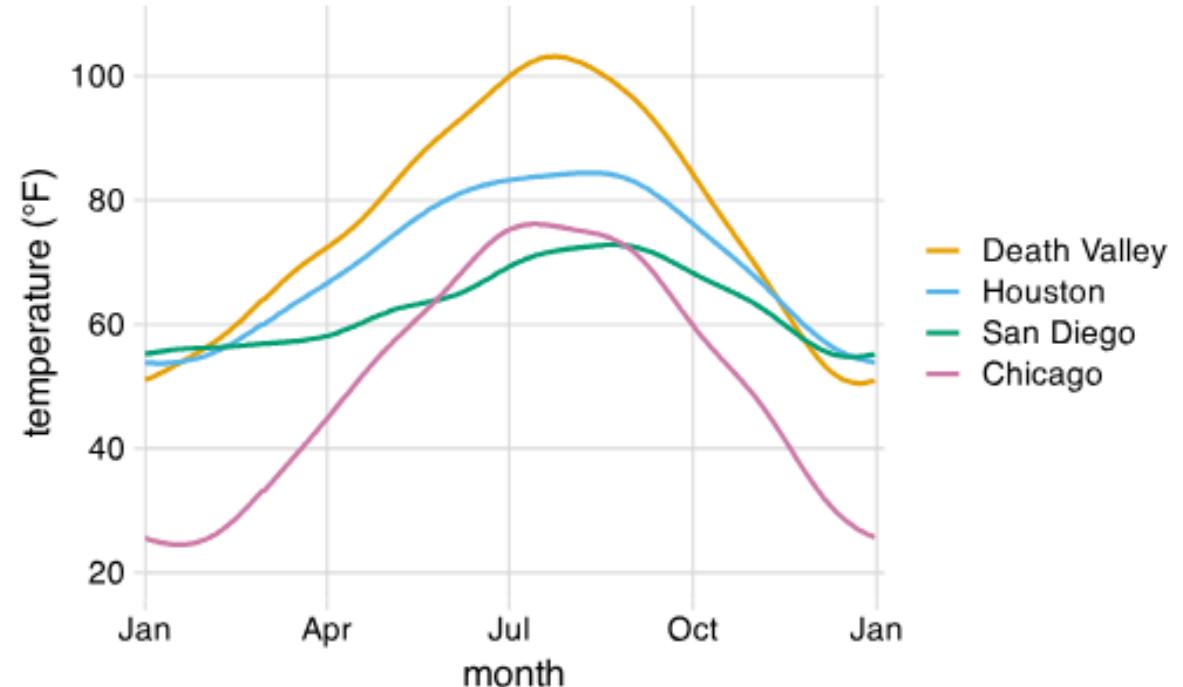


- Rõ ràng
- Chính xác, không gây hiểu nhầm
- Nêu bật thông điệp chính
- [https://www.ted.com/talks/hans\\_rosling\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen](https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen)

# Biểu diễn dữ liệu

- Sử dụng **vị trí** để thể hiện nhiệt độ

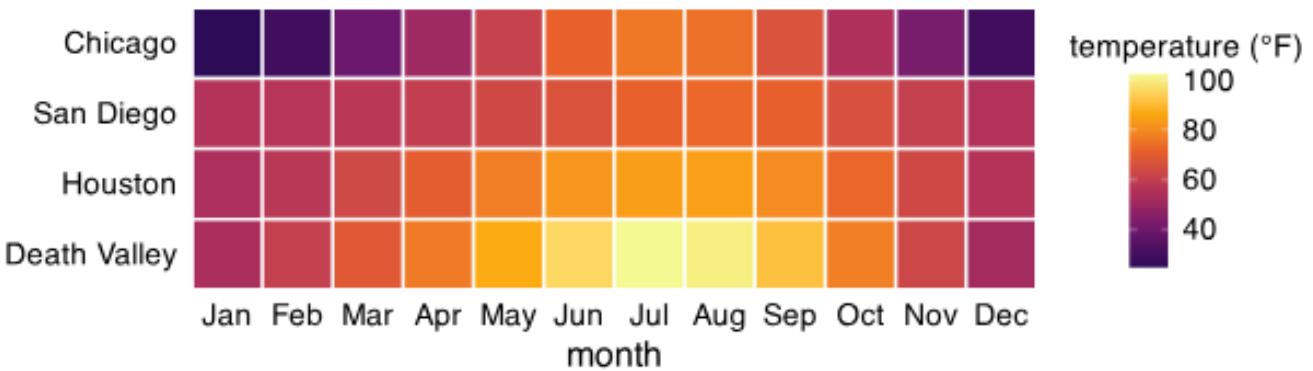
location	day_of_year	month	temperature
Death Valley	1	01	51.0
Death Valley	2	01	51.2
Death Valley	3	01	51.3
Death Valley	4	01	51.4
Death Valley	5	01	51.6
Death Valley	6	01	51.7
Death Valley	7	01	51.9
Death Valley	8	01	52.0
Death Valley	9	01	52.2
Death Valley	10	01	52.3
Death Valley	11	01	52.5



# Biểu diễn dữ liệu

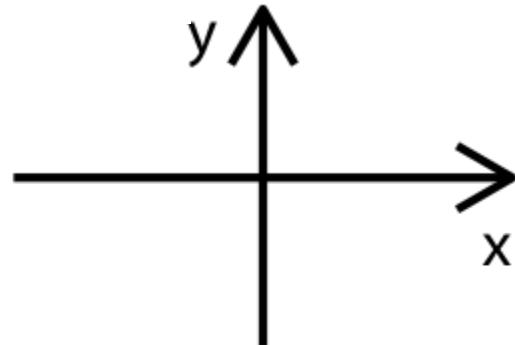
- Sử dụng **màu sắc** để thể hiện nhiệt độ

location	day_of_year	month	temperature
Death Valley	1	01	51.0
Death Valley	2	01	51.2
Death Valley	3	01	51.3
Death Valley	4	01	51.4
Death Valley	5	01	51.6
Death Valley	6	01	51.7
Death Valley	7	01	51.9
Death Valley	8	01	52.0
Death Valley	9	01	52.2
Death Valley	10	01	52.3
Death Valley	11	01	52.5

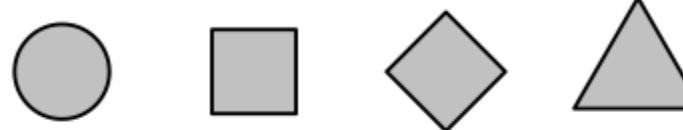


# Biểu diễn dữ liệu

vị trí



hình dáng



kích thước



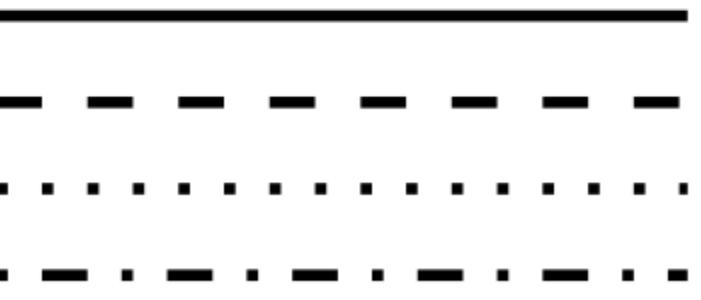
màu sắc



độ dày đường

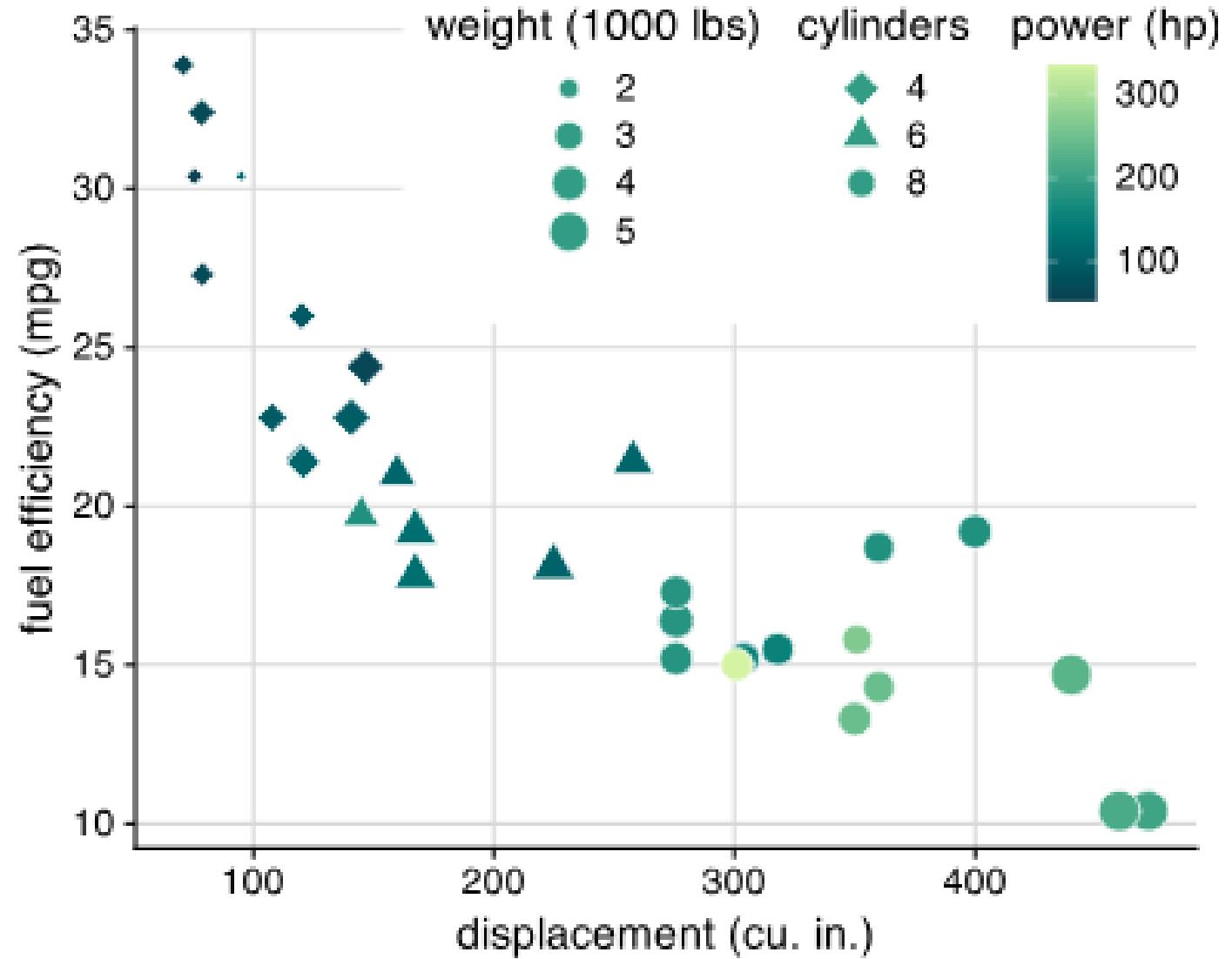


loại đường

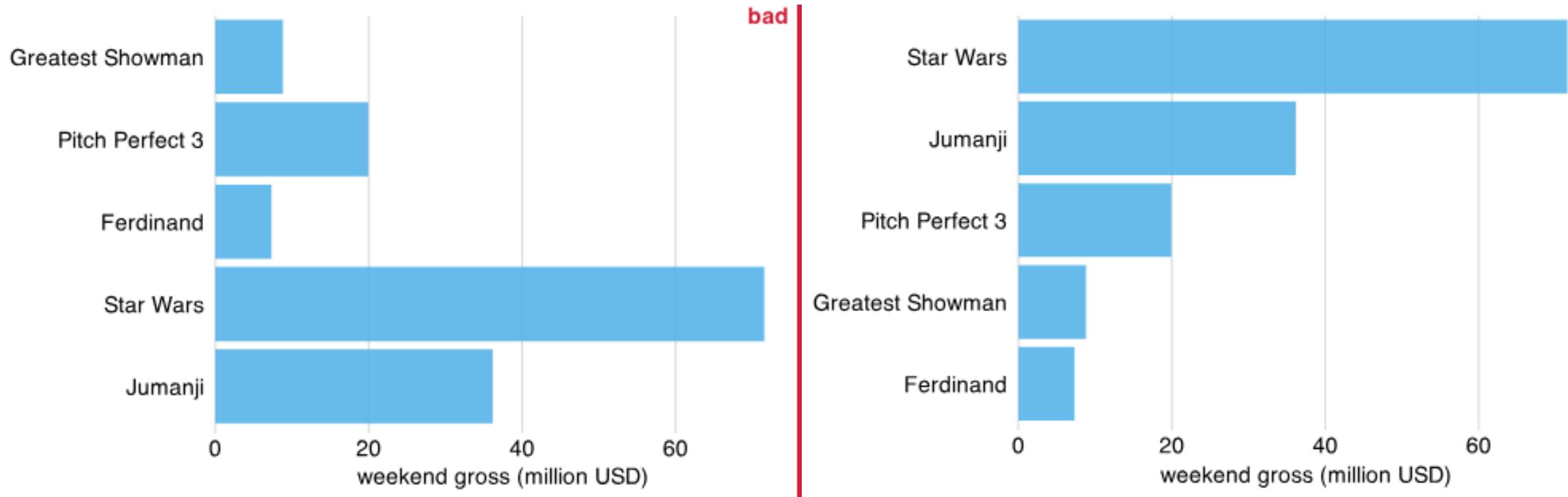


# Biểu diễn dữ liệu

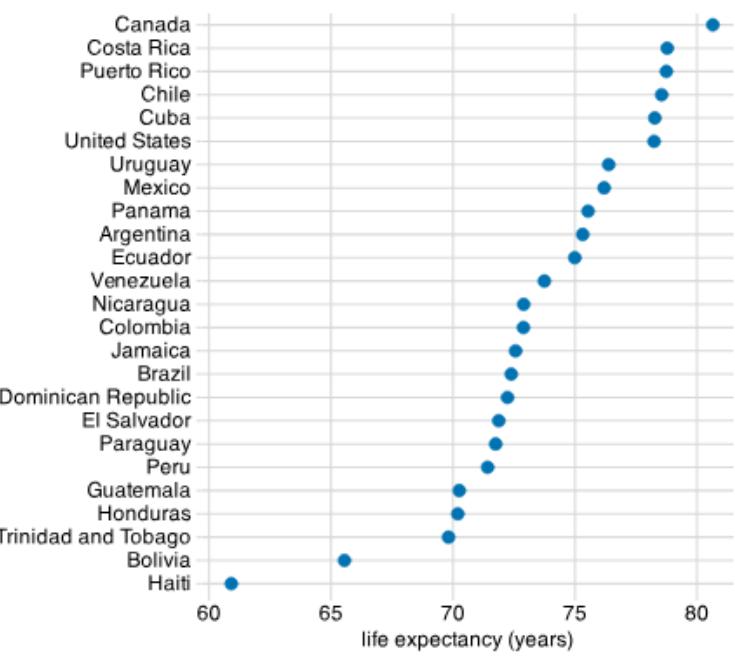
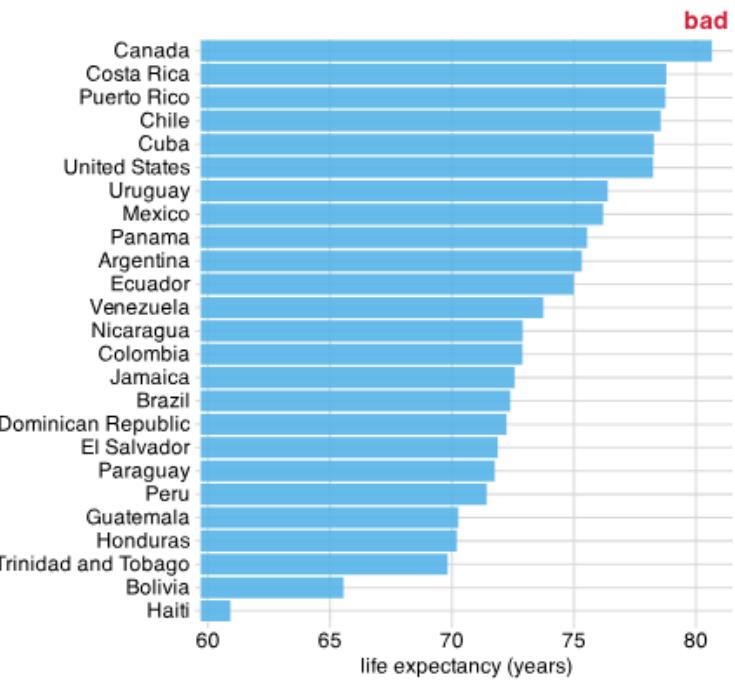
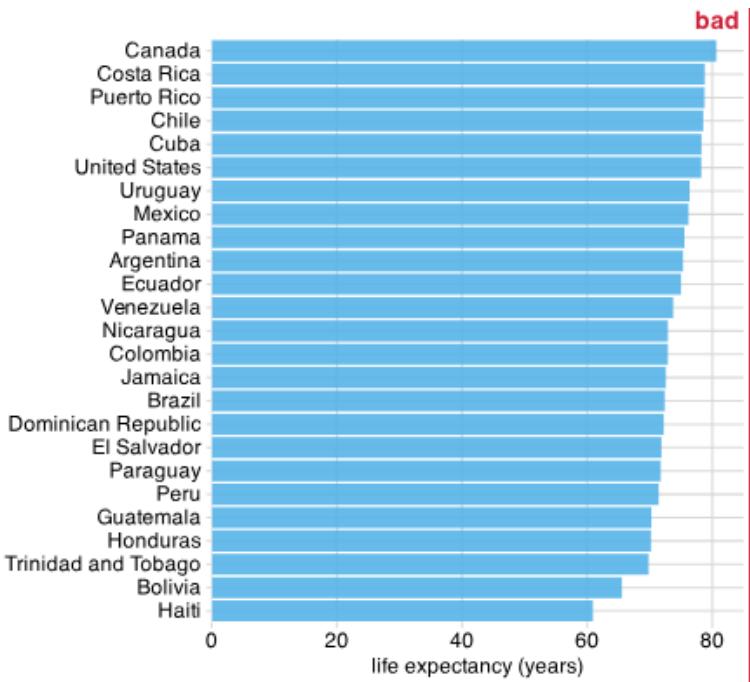
- Có thể sử dụng nhiều yếu tố hội họa trong cùng một đồ thị



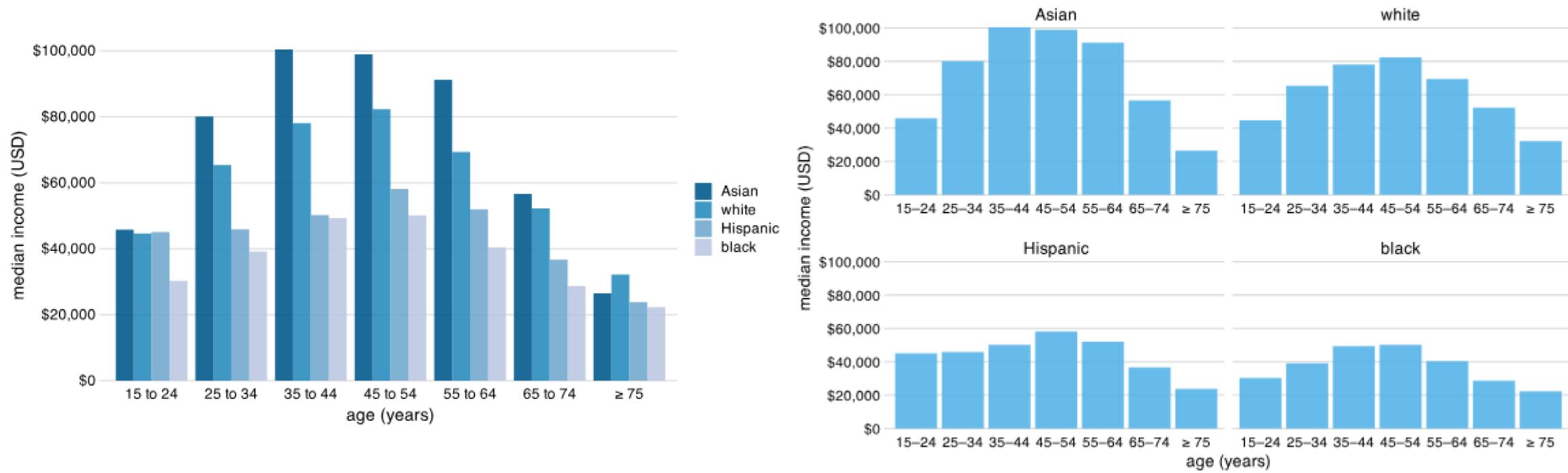
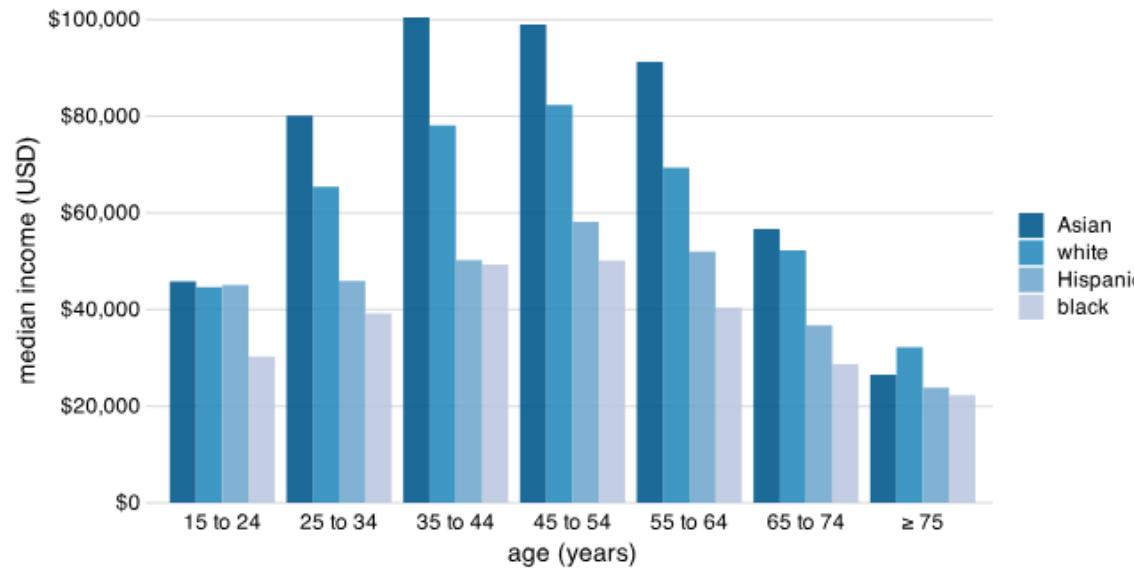
# Biểu diễn dữ liệu - lượng



# Biểu diễn dữ liệu - lượng



# Biểu diễn dữ liệu - lượng



# Biểu diễn dữ liệu – phân phối

- Xác định các khoảng
- Đếm số trường hợp

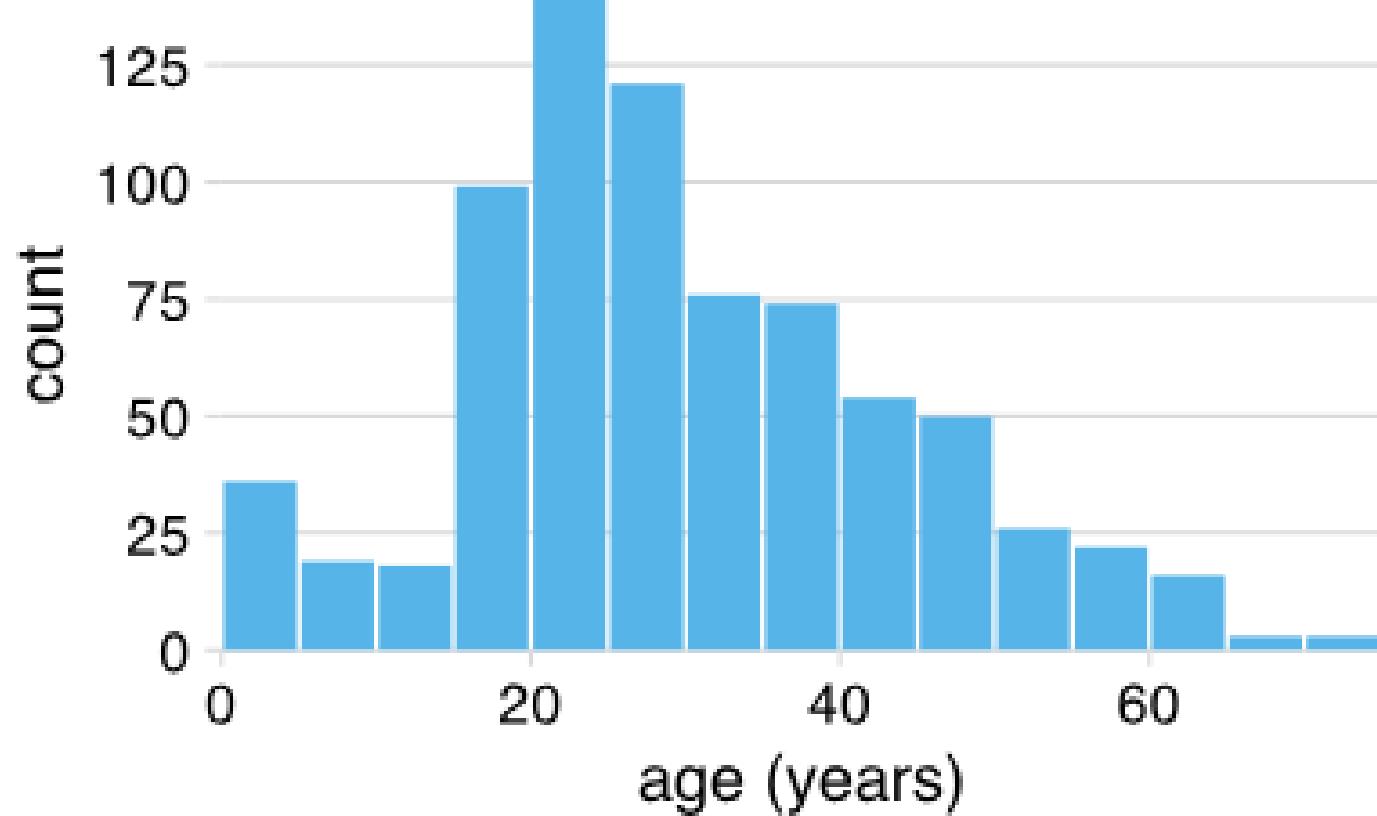
age range	count	age range	count
0-5	36	41-45	54
6-10	19	46-50	50
11-15	18	51-55	26
16-20	99	56-60	22
21-25	139	61-65	16
26-30	121	66-70	3
31-35	76	71-75	3
36-40	74	76-80	0

# Biểu diễn dữ liệu – phân phối

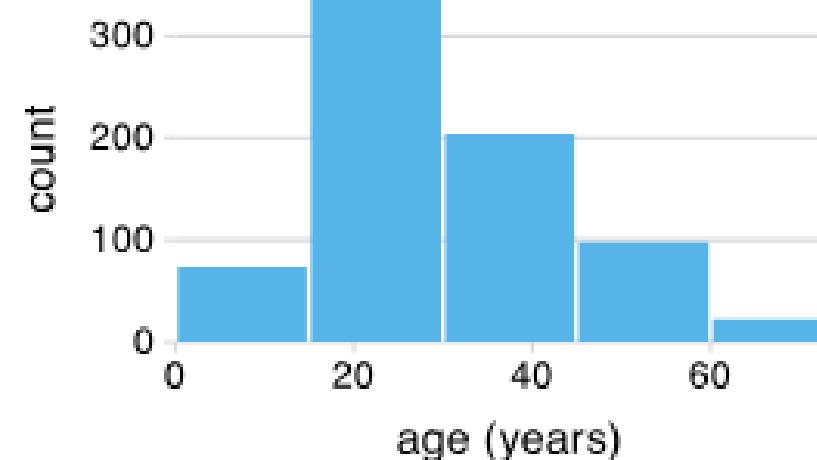
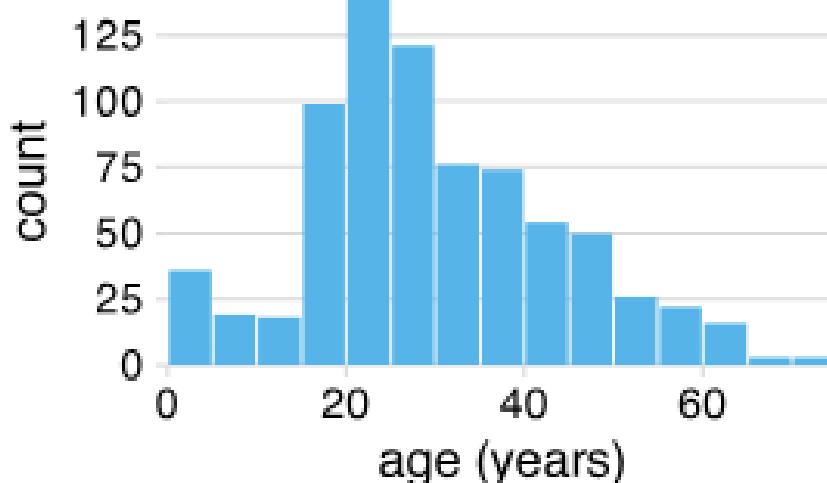
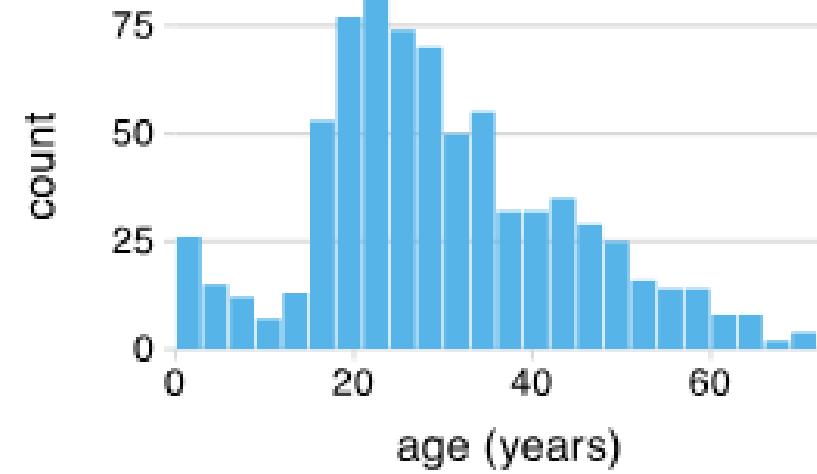
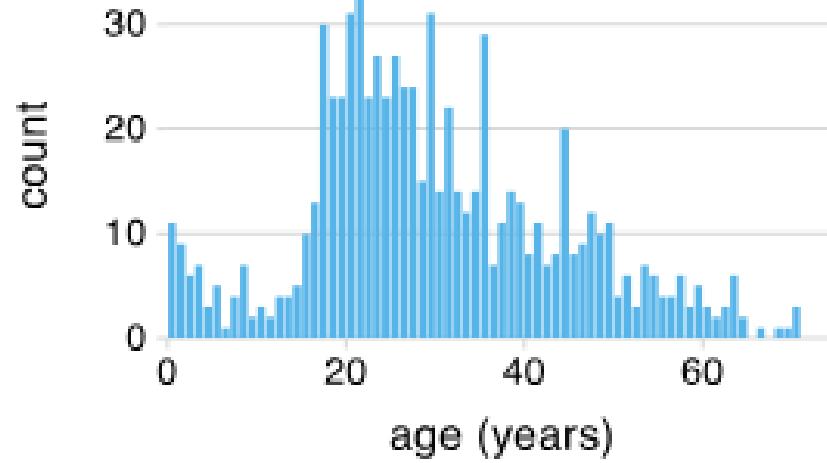
- Xác định các khoảng
- Đếm số trường hợp

age range	count
0-5	36
6-10	19
11-15	18
16-20	99
21-25	139
26-30	121
31-35	76
36-40	74

age range	count
41-45	54
46-50	50
51-55	26
56-60	22
61-65	16
66-70	3
71-75	3
76-80	0

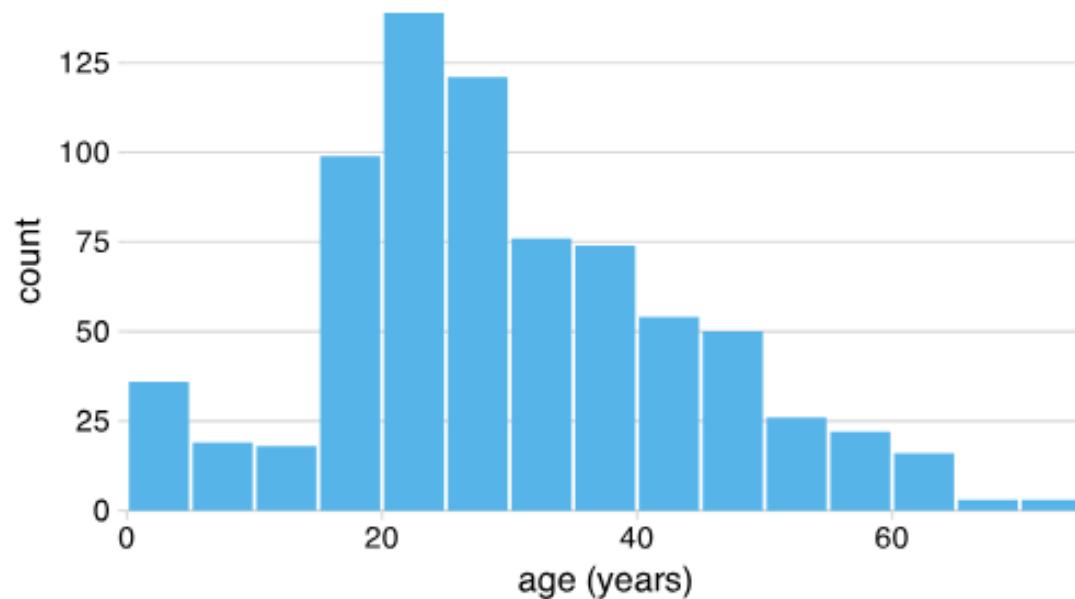


# Biểu diễn dữ liệu – phân phối

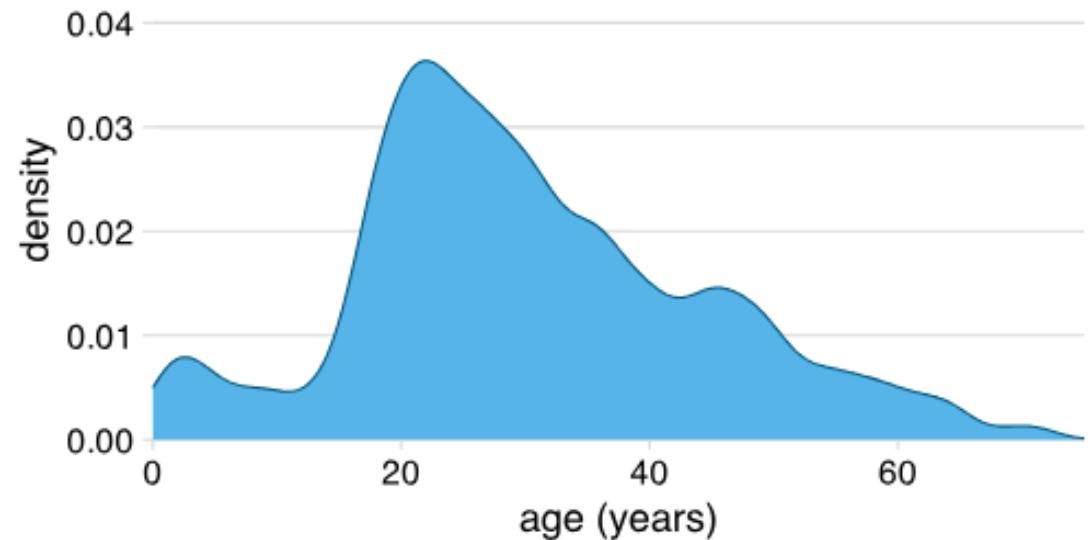


# Biểu diễn dữ liệu – phân phối

Histogram



Kernel density estimate



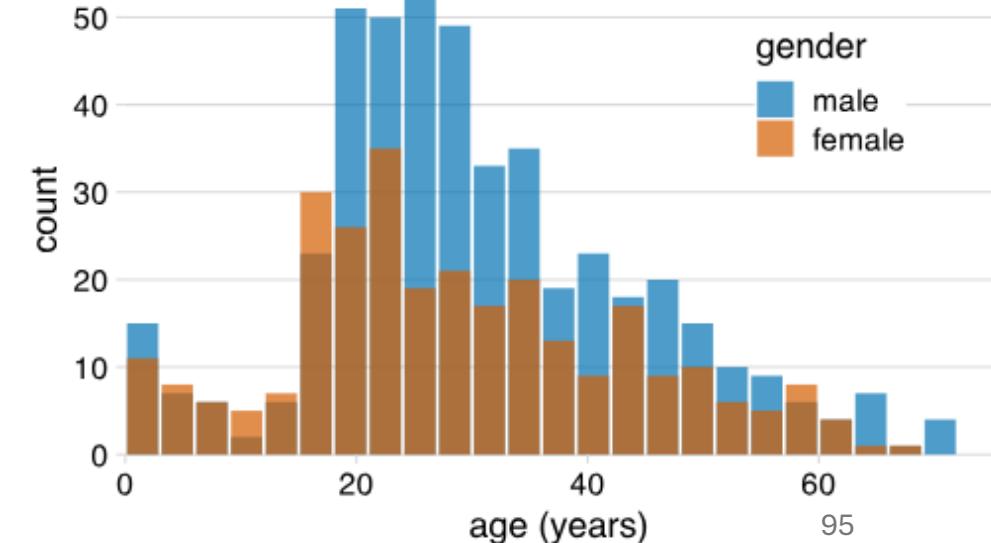
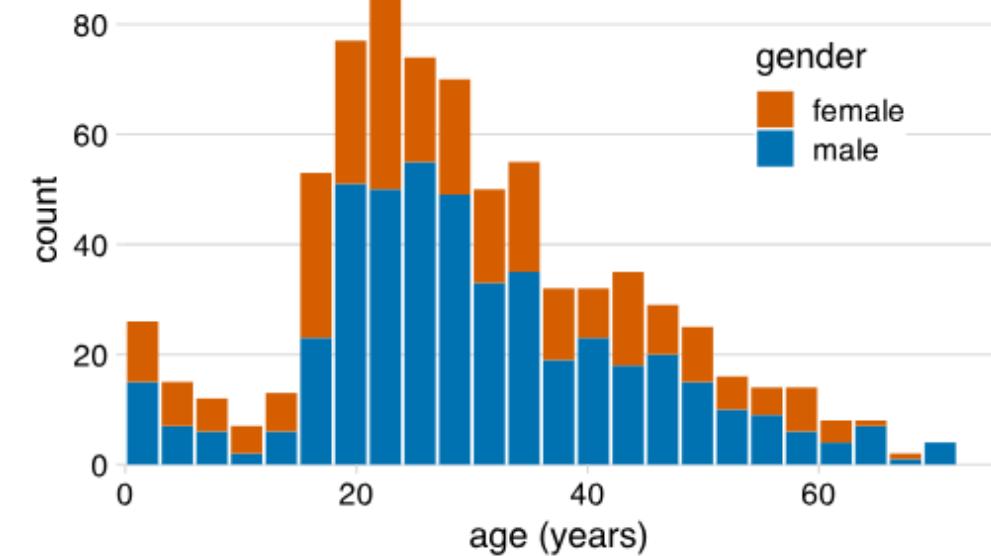
# Biểu diễn dữ liệu – phân phối

**female**

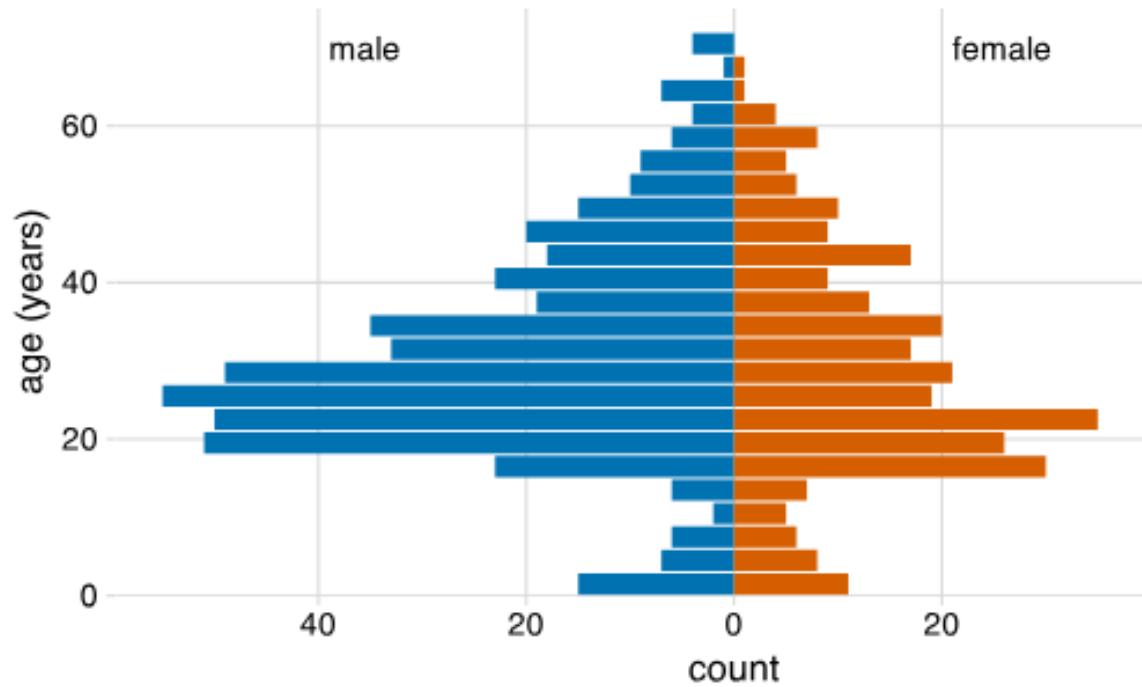
age_c	count
[0,3]	11
(3,6]	8
(6,9]	6
(9,12]	5
(12,15]	7
(15,18]	30
(18,21]	26
(21,24]	35
(24,27]	19
(27,30]	21

**male**

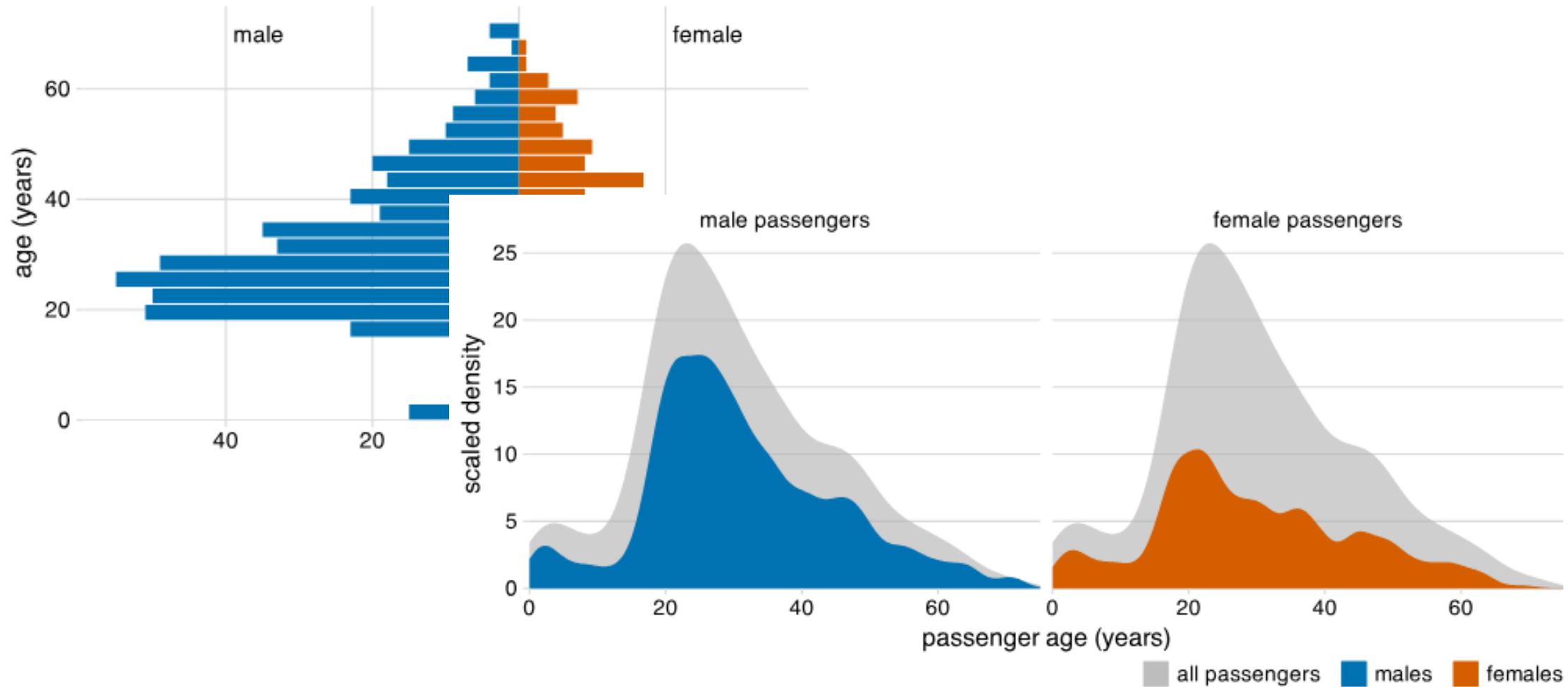
age_c	count
[0,3]	15
(3,6]	7
(6,9]	6
(9,12]	2
(12,15]	6
(15,18]	23
(18,21]	51
(21,24]	50
(24,27]	55
(27,30]	49



# Biểu diễn dữ liệu – phân phối

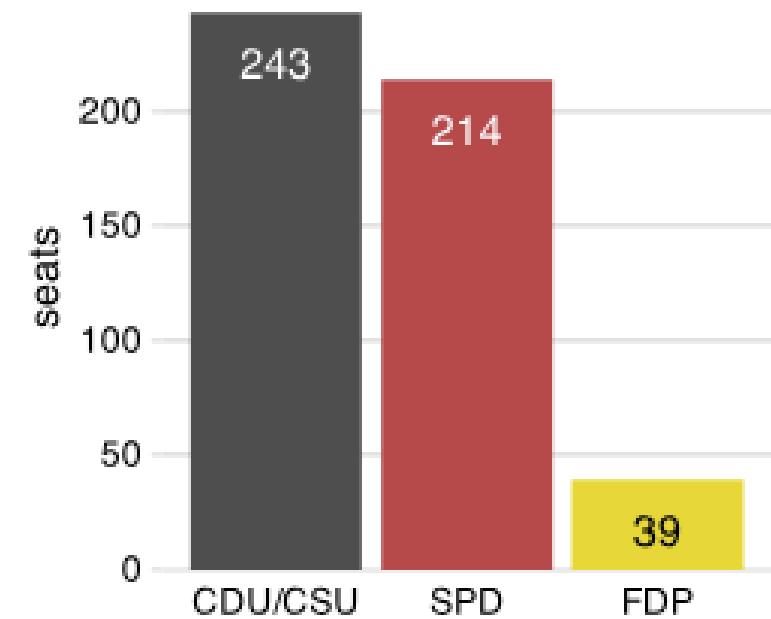
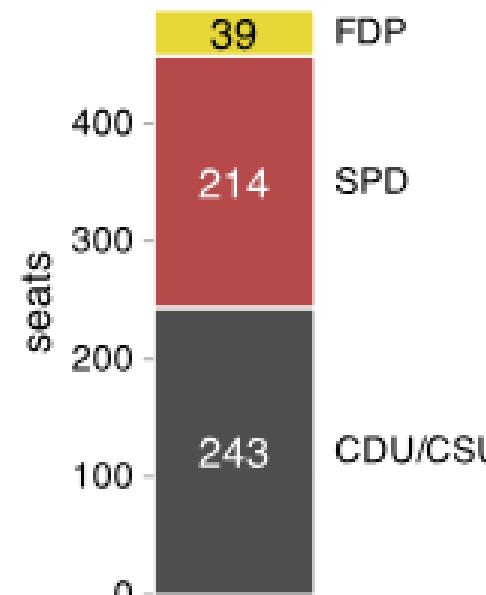
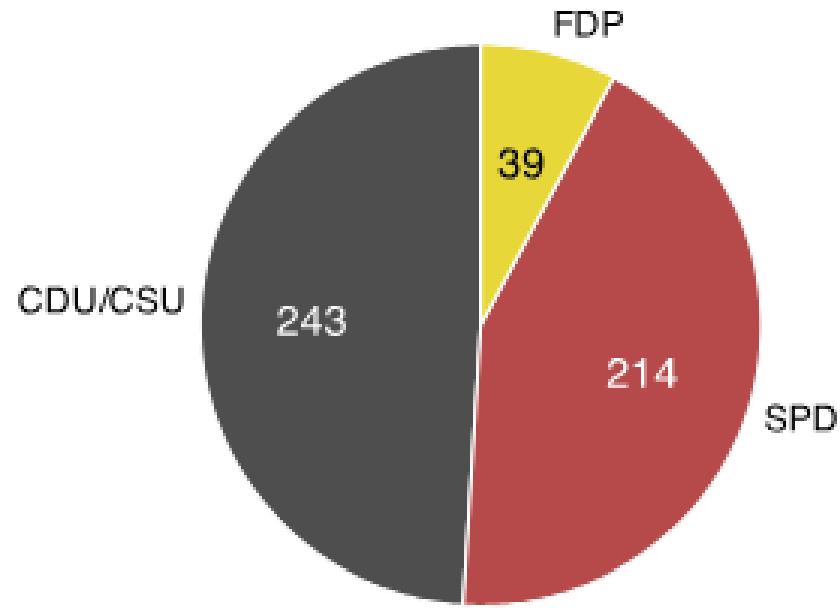


# Biểu diễn dữ liệu – phân phối

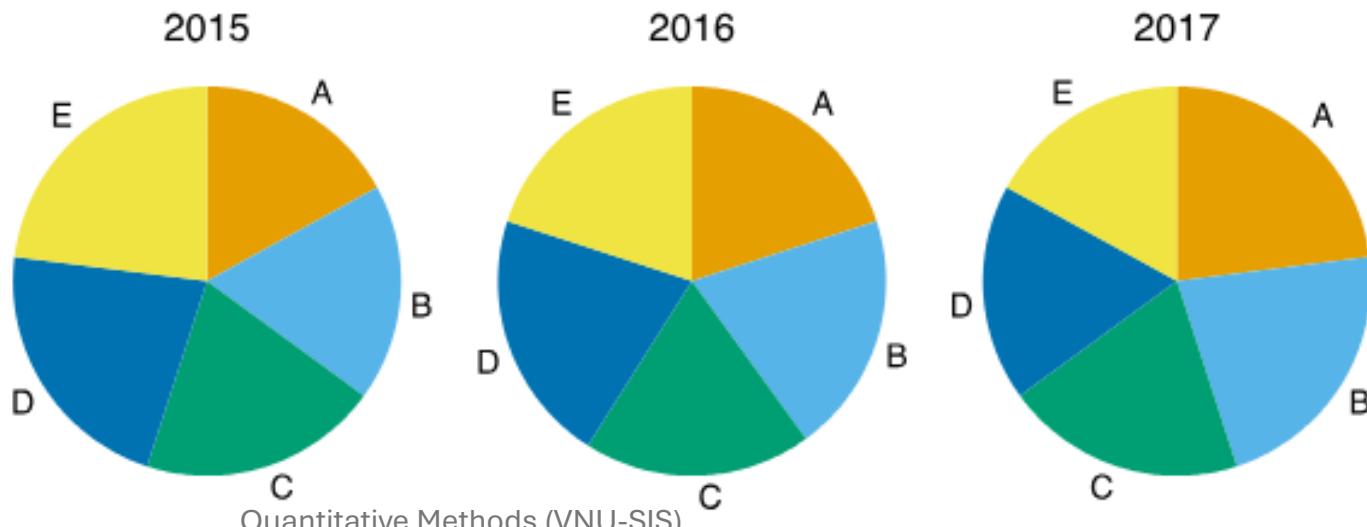
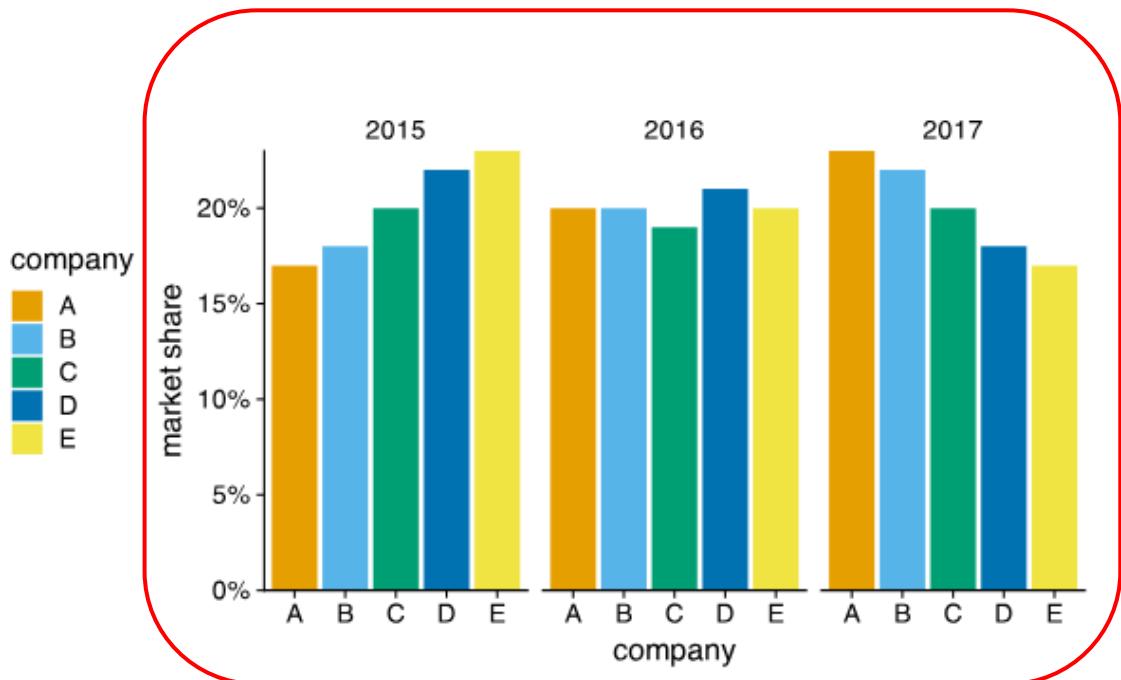
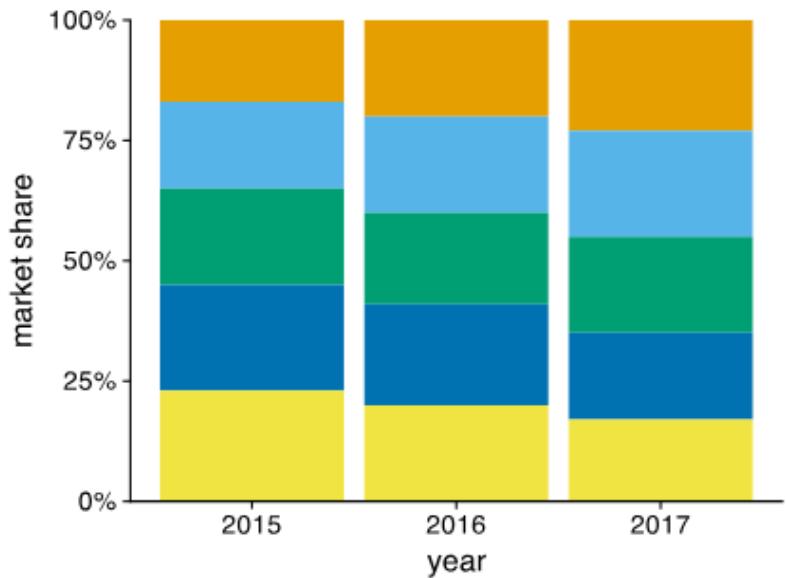


# Biểu diễn dữ liệu – tỷ lệ

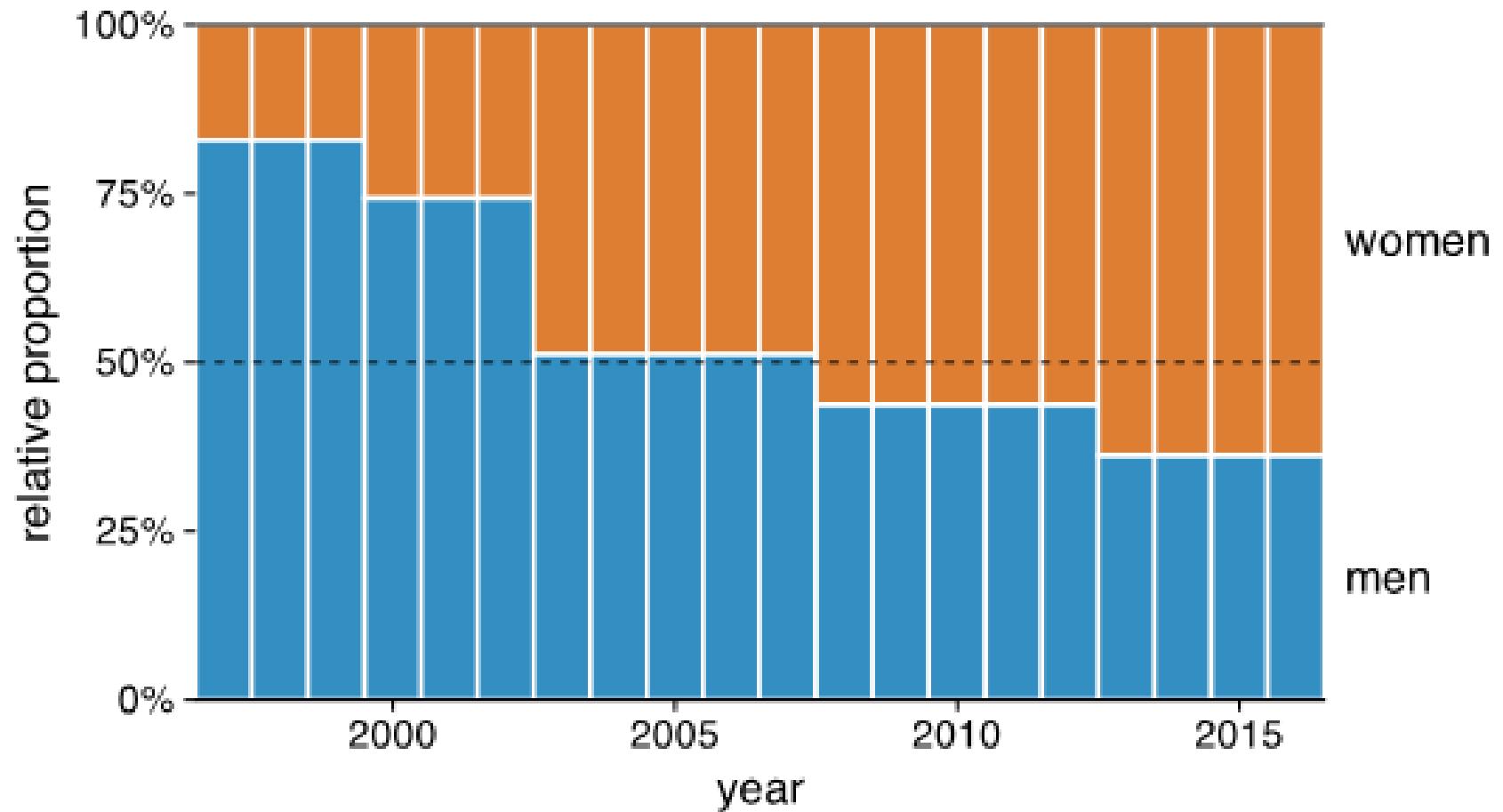
Thành phần đảng phái của Quốc hội Đức khóa 8,  
giai đoạn 1976–1980.



# Biểu diễn dữ liệu – tỷ lệ



# Biểu diễn dữ liệu – tỷ lệ



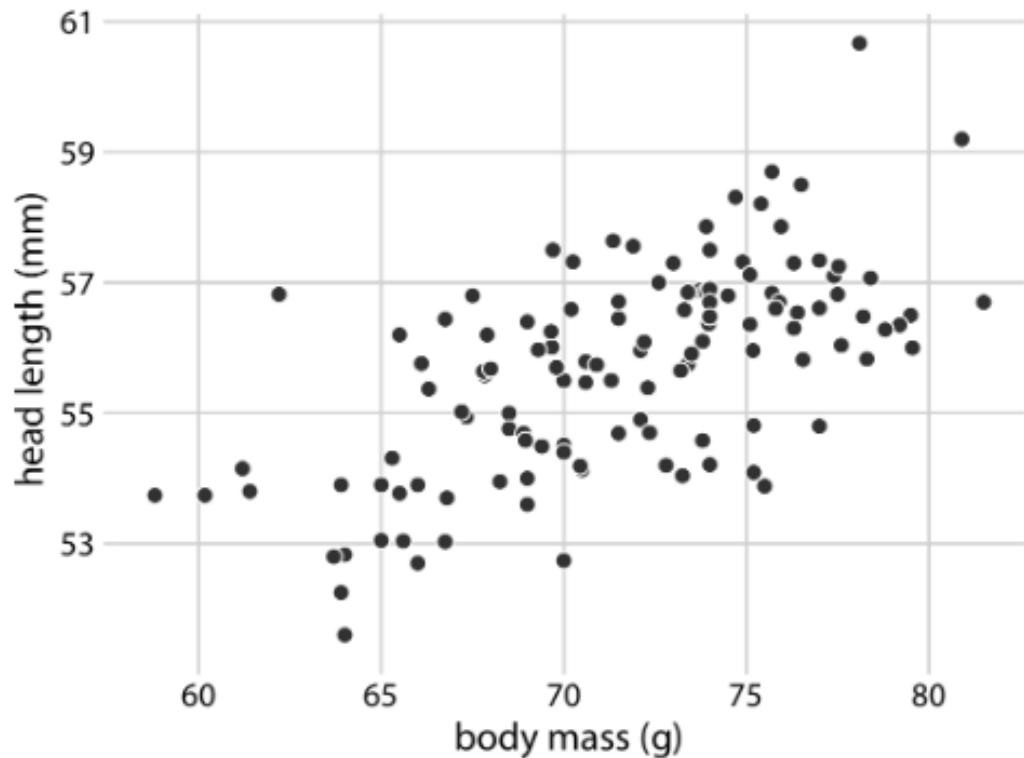
Thay đổi trong thành phần giới tính của quốc hội Rwanda  
từ năm 1997 đến năm 2016

# Biểu diễn dữ liệu – tỷ lệ

	Biểu đồ quạt	Biểu đồ cột xếp chồng	Biểu đồ cột xếp cạnh
Cho phép so sánh dễ dàng tỷ lệ tương đối	✗	✗	✓
Hiển thị dữ liệu dưới dạng tỷ lệ của tổng thể	✓	✓	✗
Nhấn mạnh các phân số đơn giản ( $1/2$ , $1/3$ , ...)	✓	✗	✗
Ưa nhìn cho các tập dữ liệu nhỏ	✓	✗	✓
Hiệu quả thể hiện với một số lượng lớn các tập con	✗	✗	✓
Hiệu quả thể hiện với chuỗi thời gian	✗	✓	✗

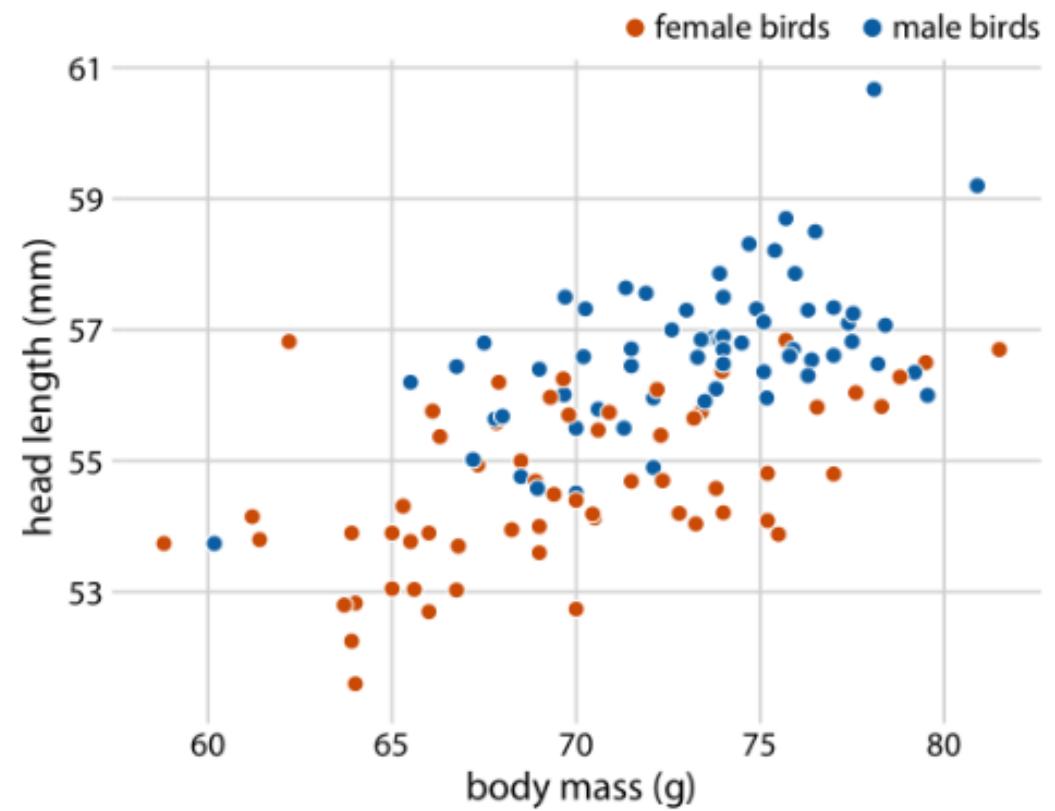
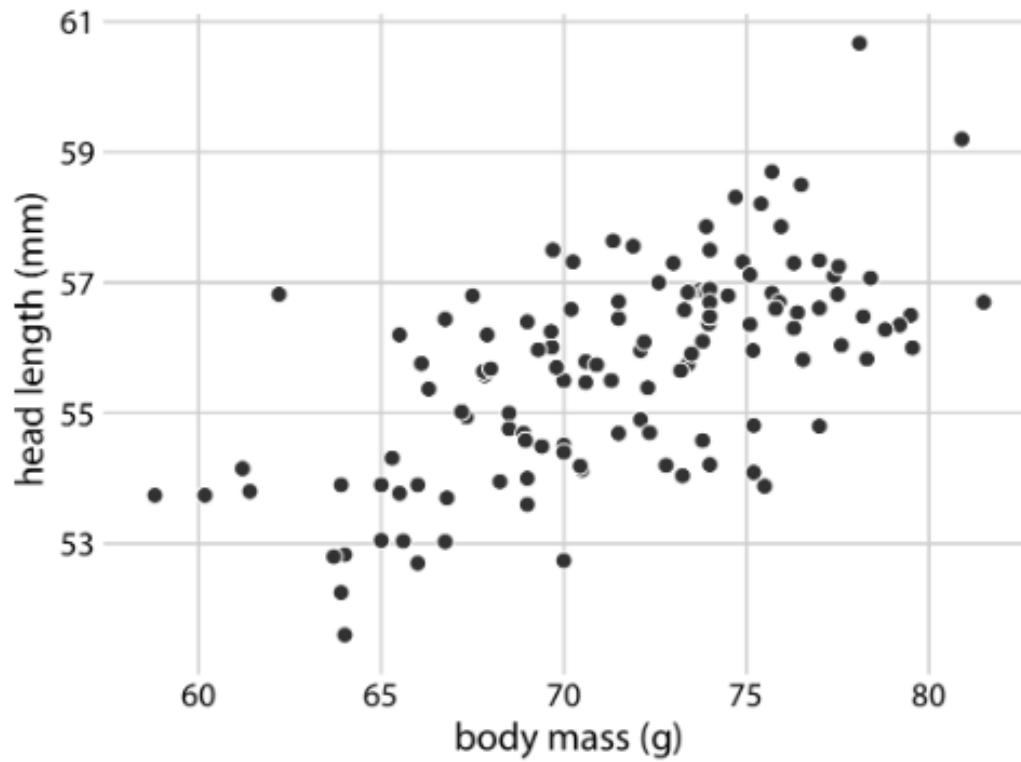
# Biểu diễn dữ liệu – quan hệ

123 chim giẻ cùi lam (blue jay)



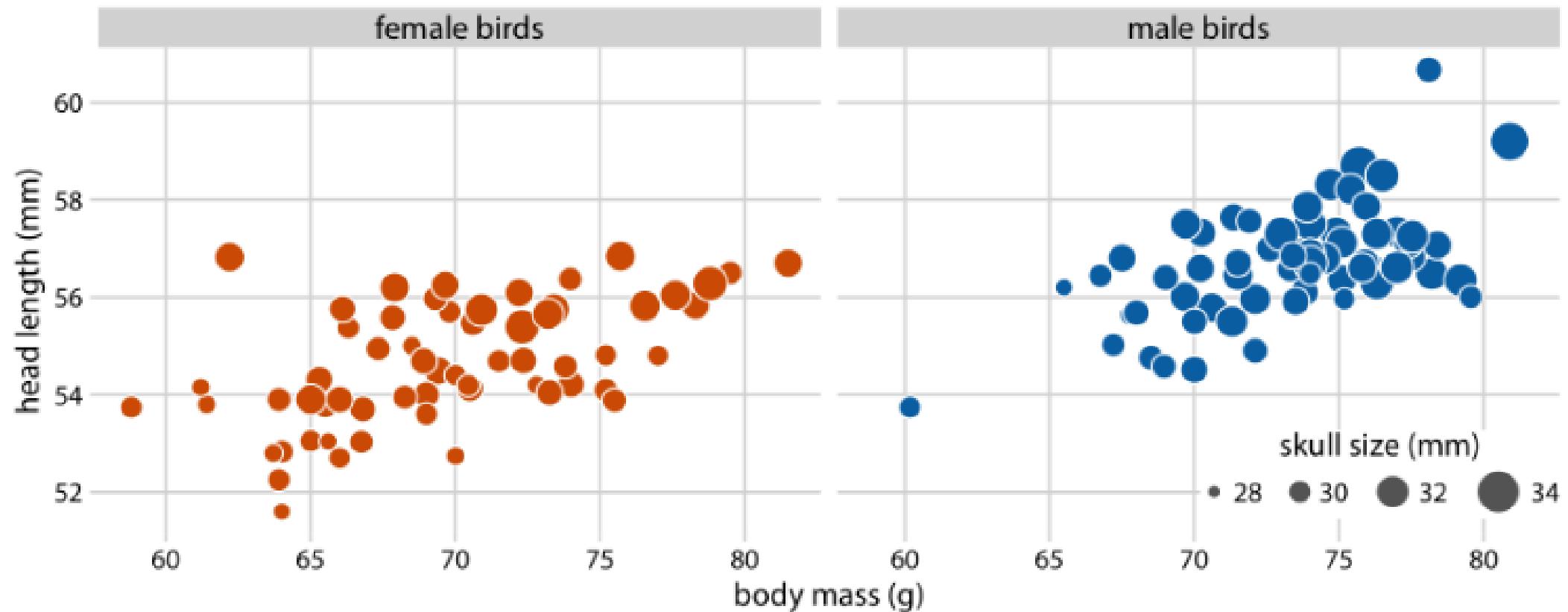
# Biểu diễn dữ liệu – quan hệ

123 chim giẻ cùi lam (blue jay)

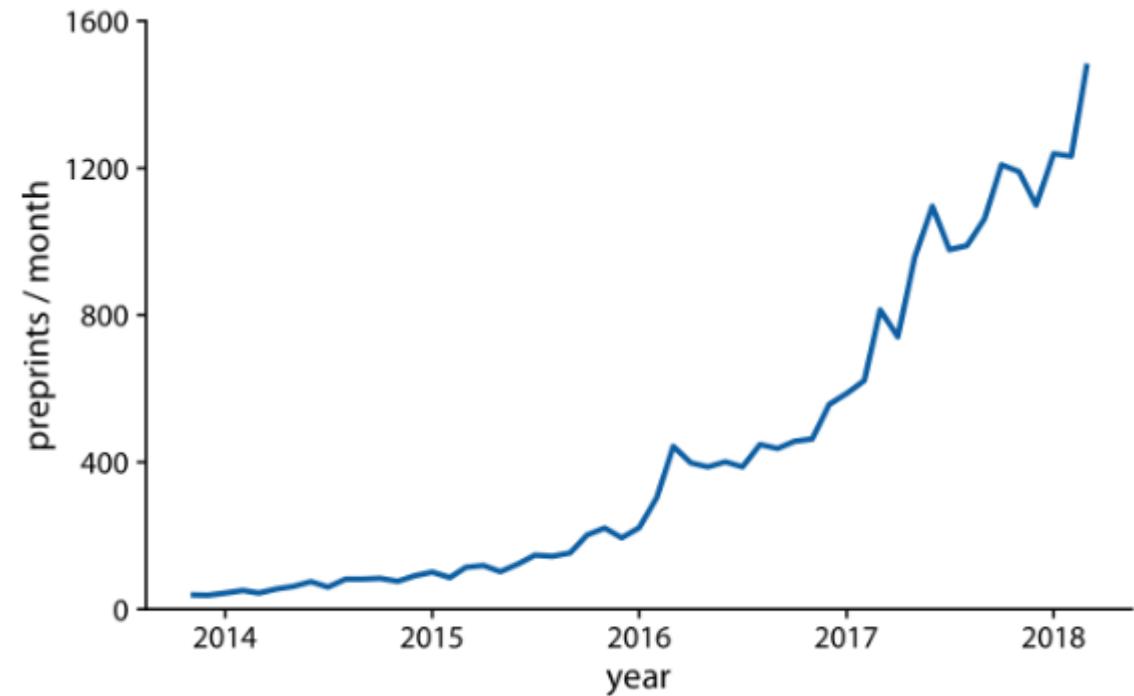
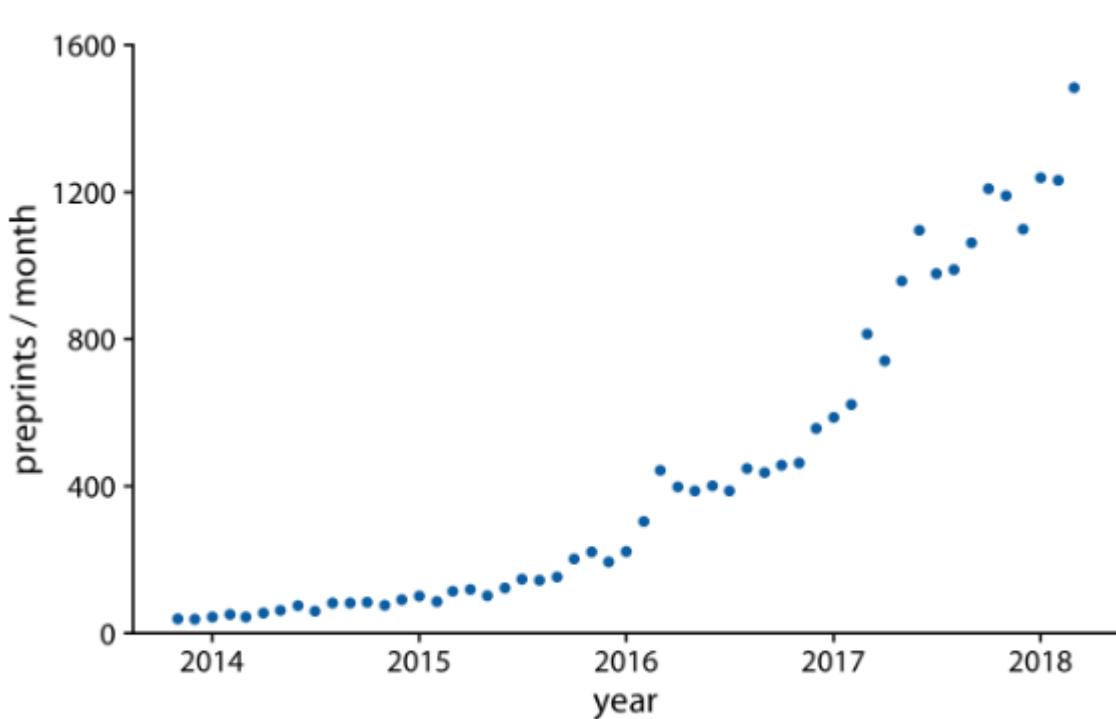


# Biểu diễn dữ liệu – quan hệ

123 chim giẻ cùi lam (blue jay)

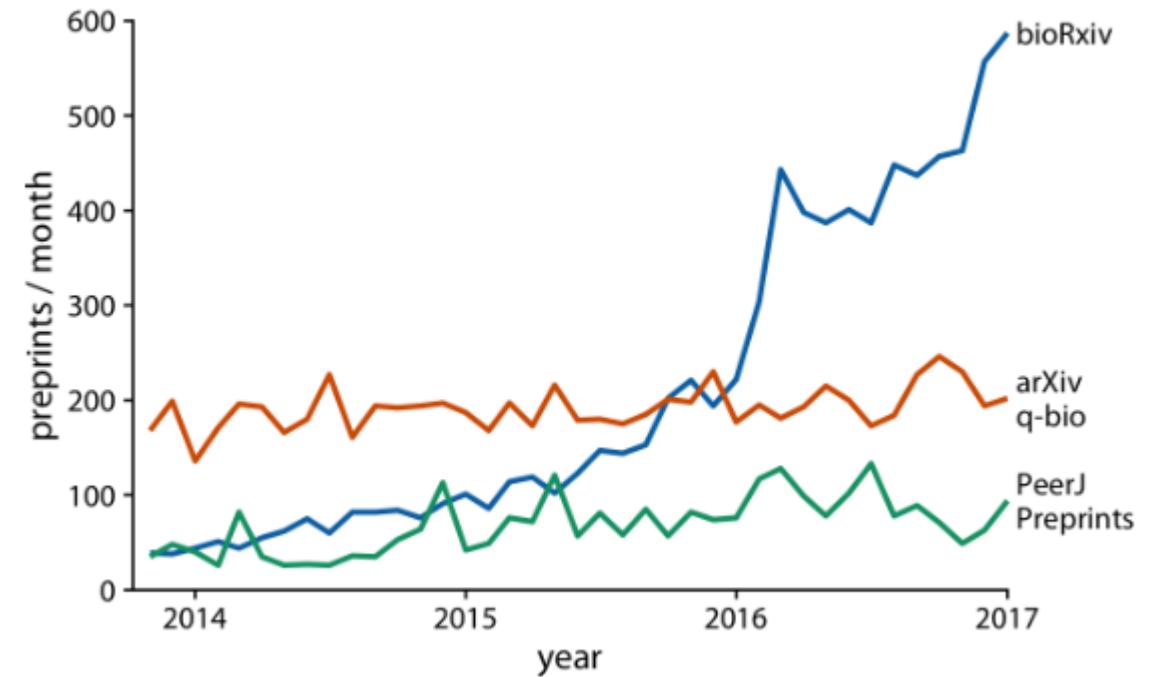
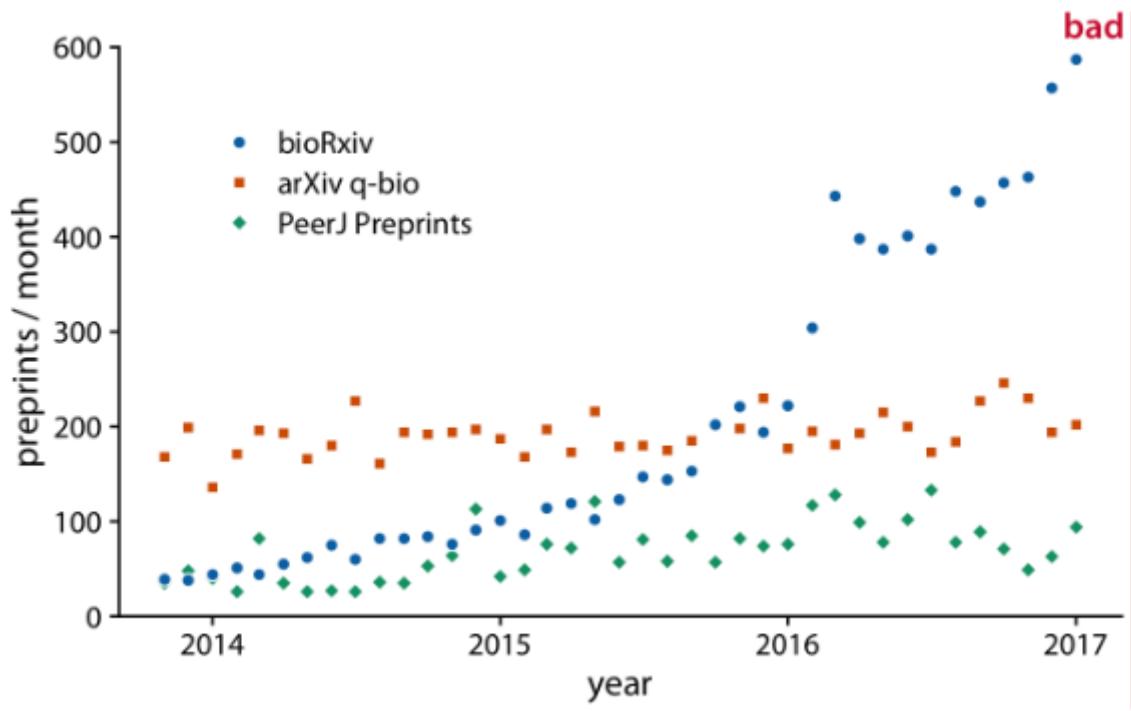


# Biểu diễn dữ liệu – xu hướng



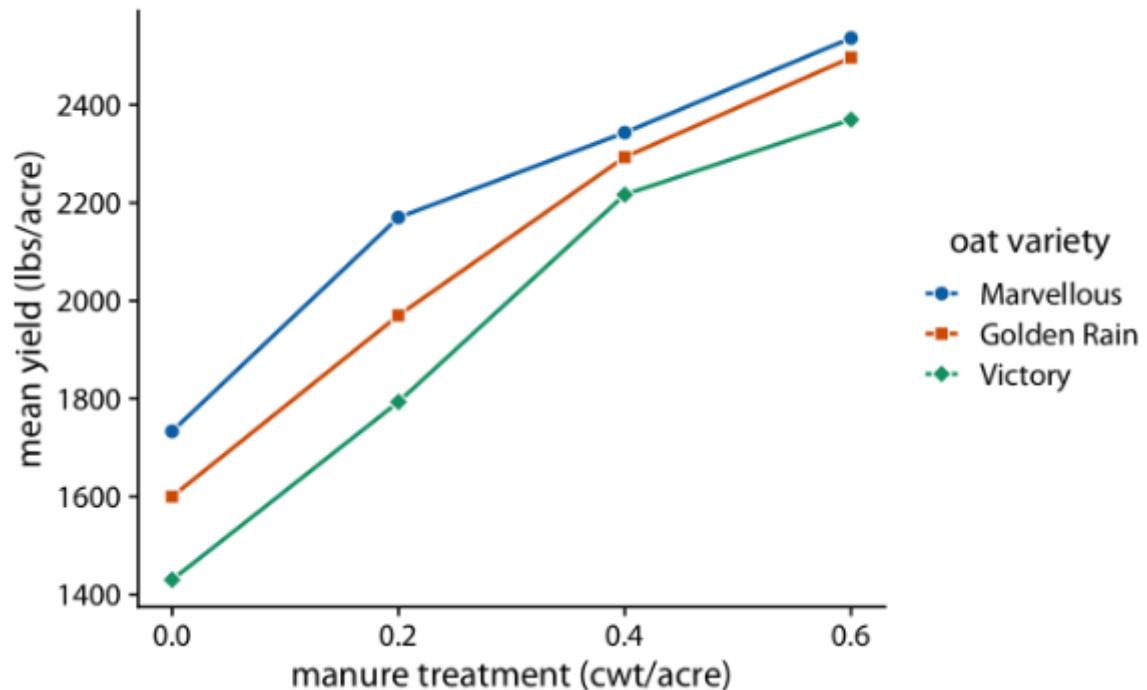
Số lượng bài báo nộp tính theo tháng lên bioRxiv

# Biểu diễn dữ liệu – xu hướng



Số lượng bài báo nộp tính theo tháng lên bioRxiv

# Biểu diễn dữ liệu – xu hướng



Sản lượng lúa mì trên hàm lượng phân bón sử dụng

# Các bước phân tích

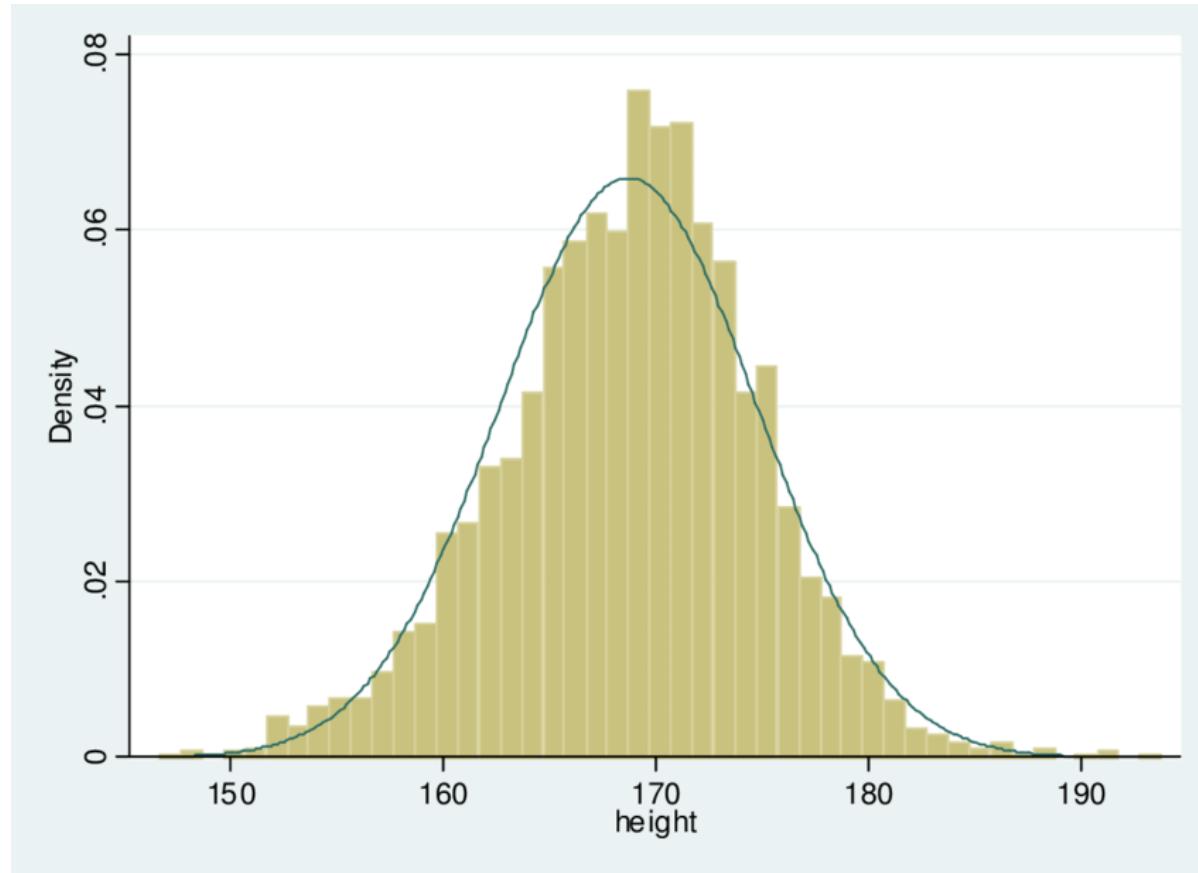
- Tiền xử lý dữ liệu
- Tìm hiểu dữ liệu/Biểu diễn dữ liệu
- Thống kê mô tả
- Thống kê suy luận

# Thống kê mô tả

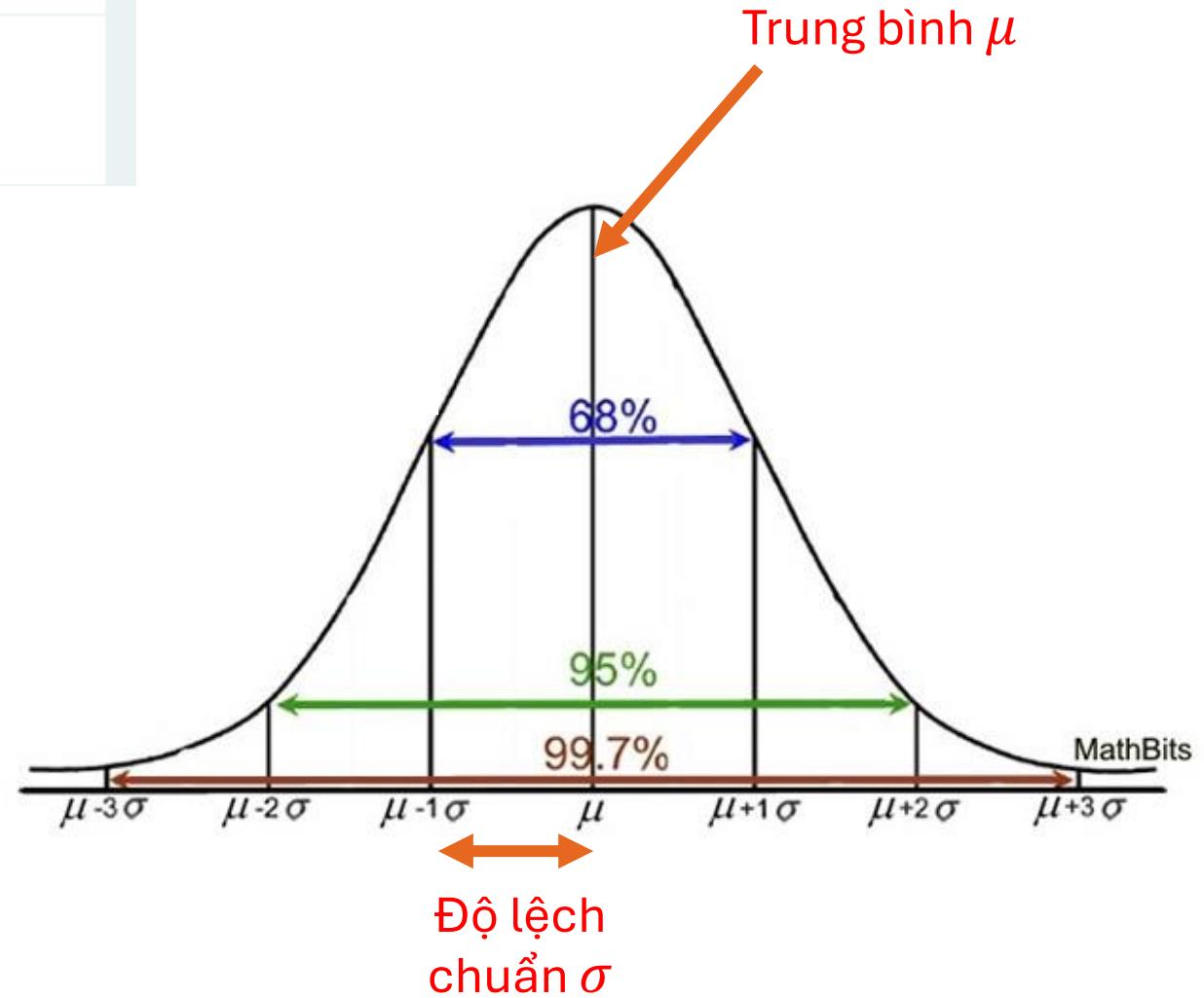
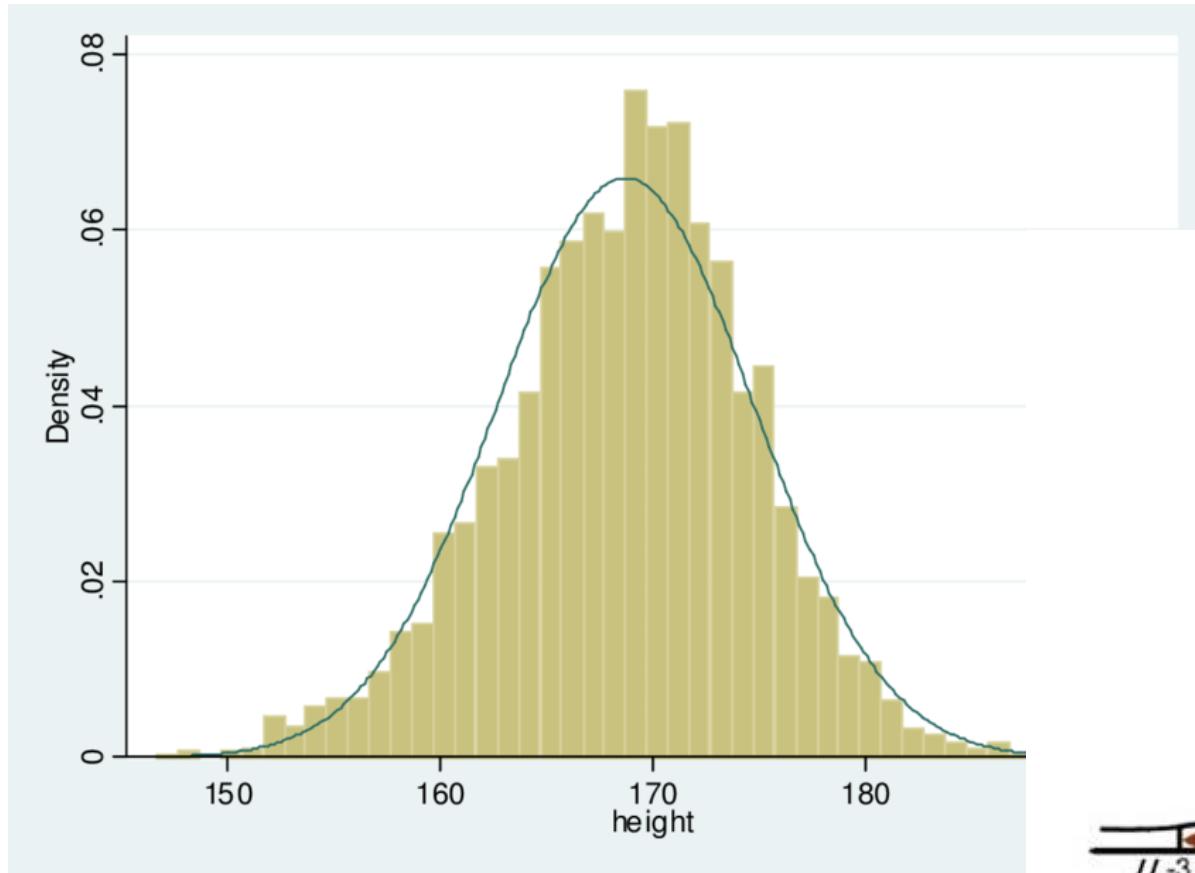
- visitors: lượng khách tham quan (k người)
- budget: kinh phí vận hành & trùng tu (kUSD)
- events: số sự kiện được tổ chức
- level: cấp độ di tích
- category: loại hình di tích

site	visitors	budget	events	level	category
1	298	166	5	National-Special	Museum
	78	107	6	Provincial	Historical
	252	134	7	National	Monument
2	262	228	5	National	Archaeological
	92	177	0	Provincial	Site
	236	213	3	National	Museum
3	120	148	6	Provincial	Museum
	101	99	3	Provincial	Museum
	55	111	6	Provincial	Museum
4	100	163	4	Provincial	Cultural Center
	115	162	8	Provincial	Historical
	280	223	8	National-Special	Monument
5					Museum

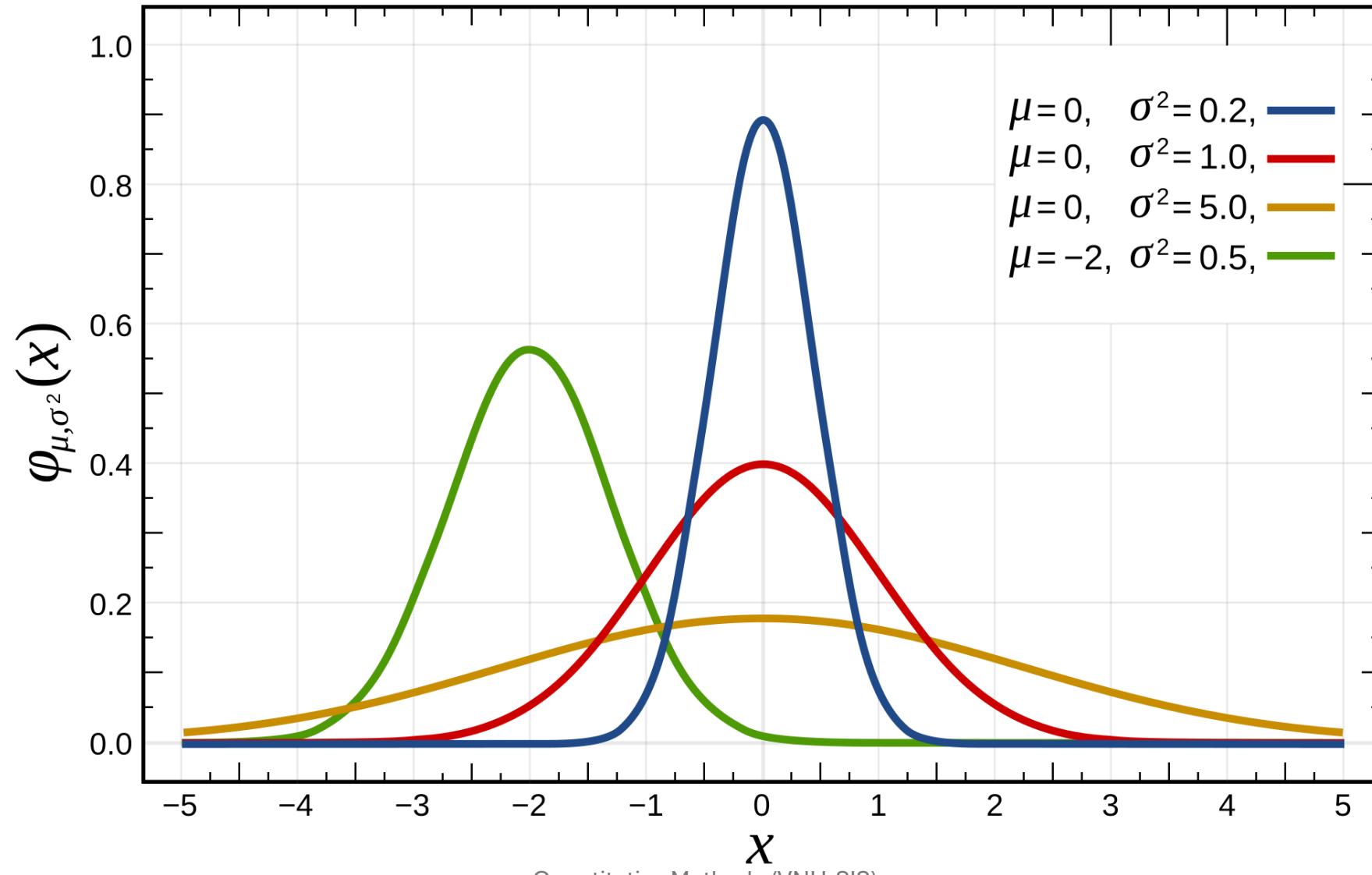
# Thống kê mô tả - Biến liên tục



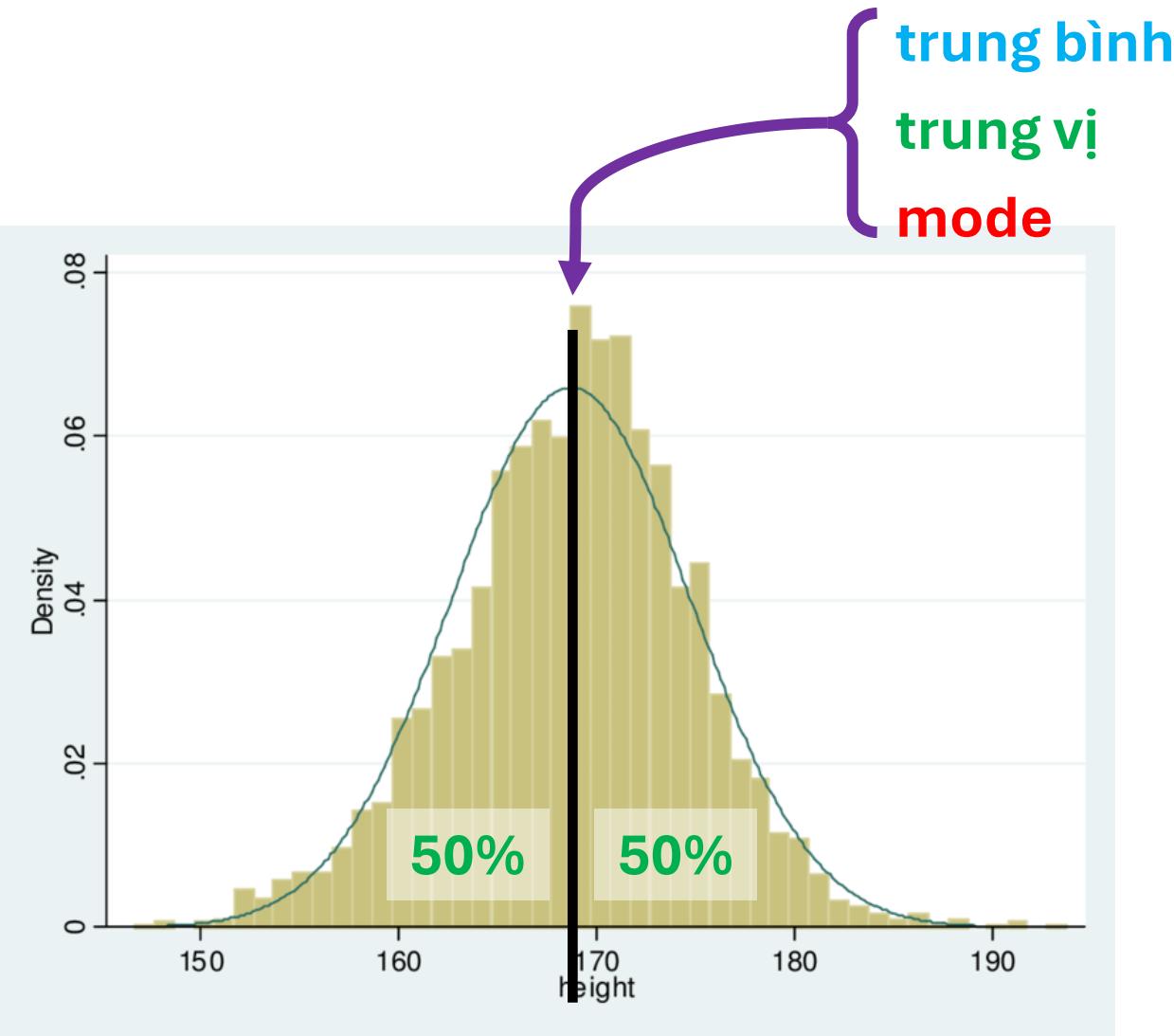
# Thống kê mô tả - Biến liên tục



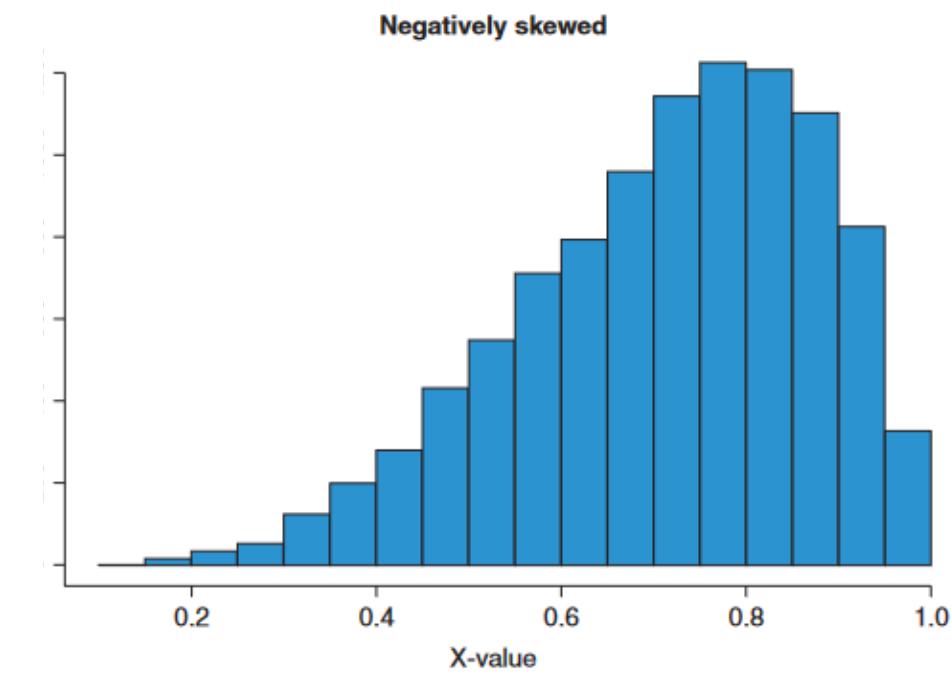
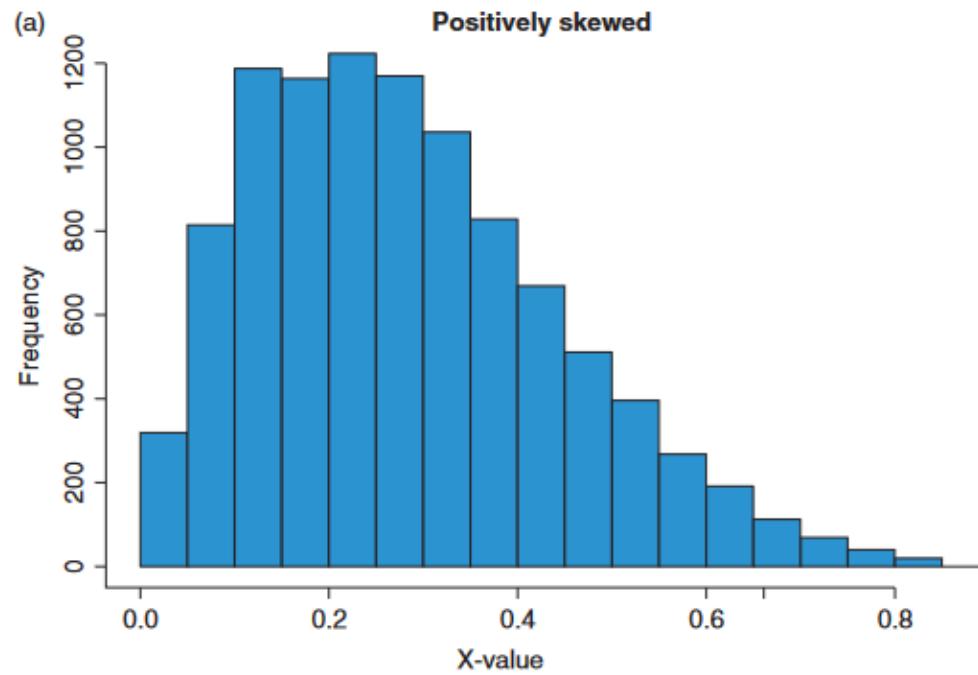
# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục



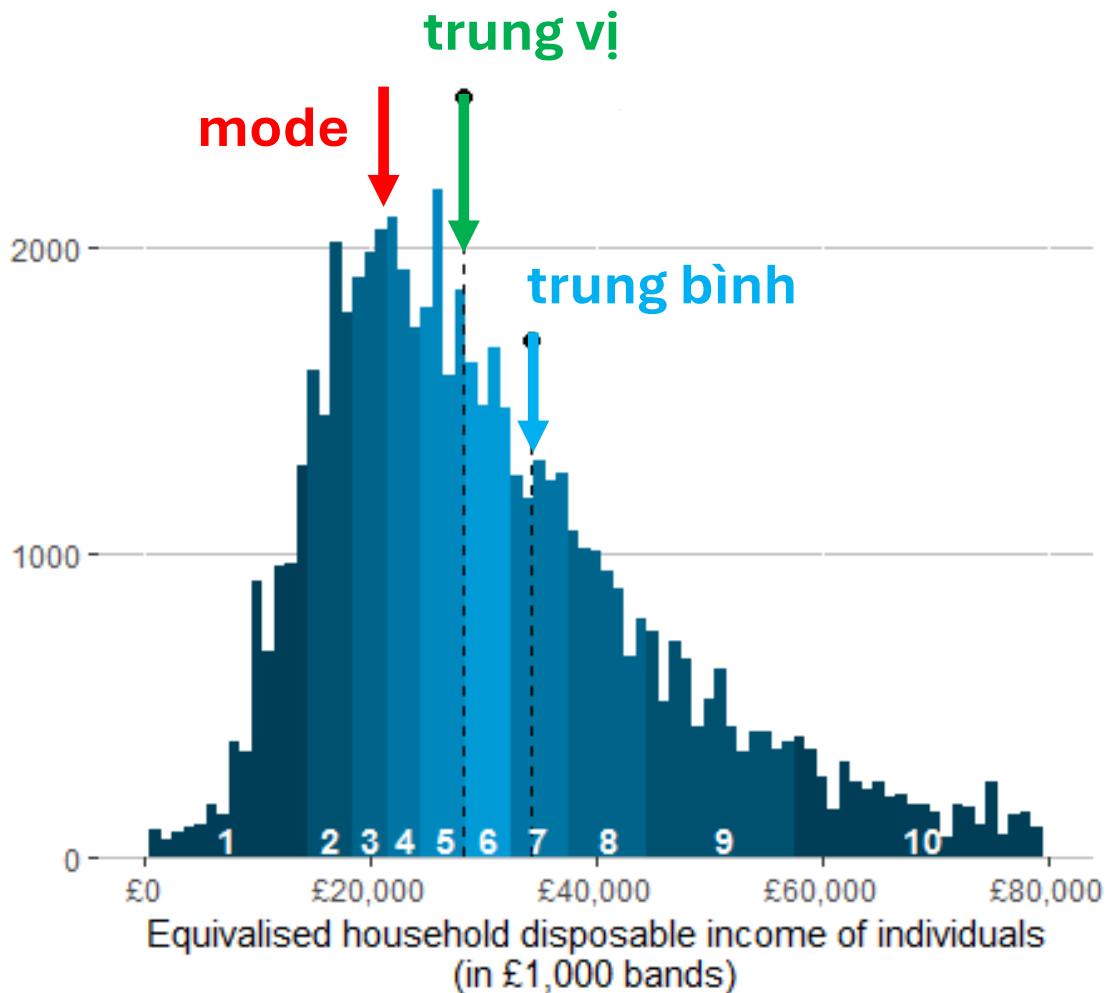
# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

Number of individuals (in thousands)

3000

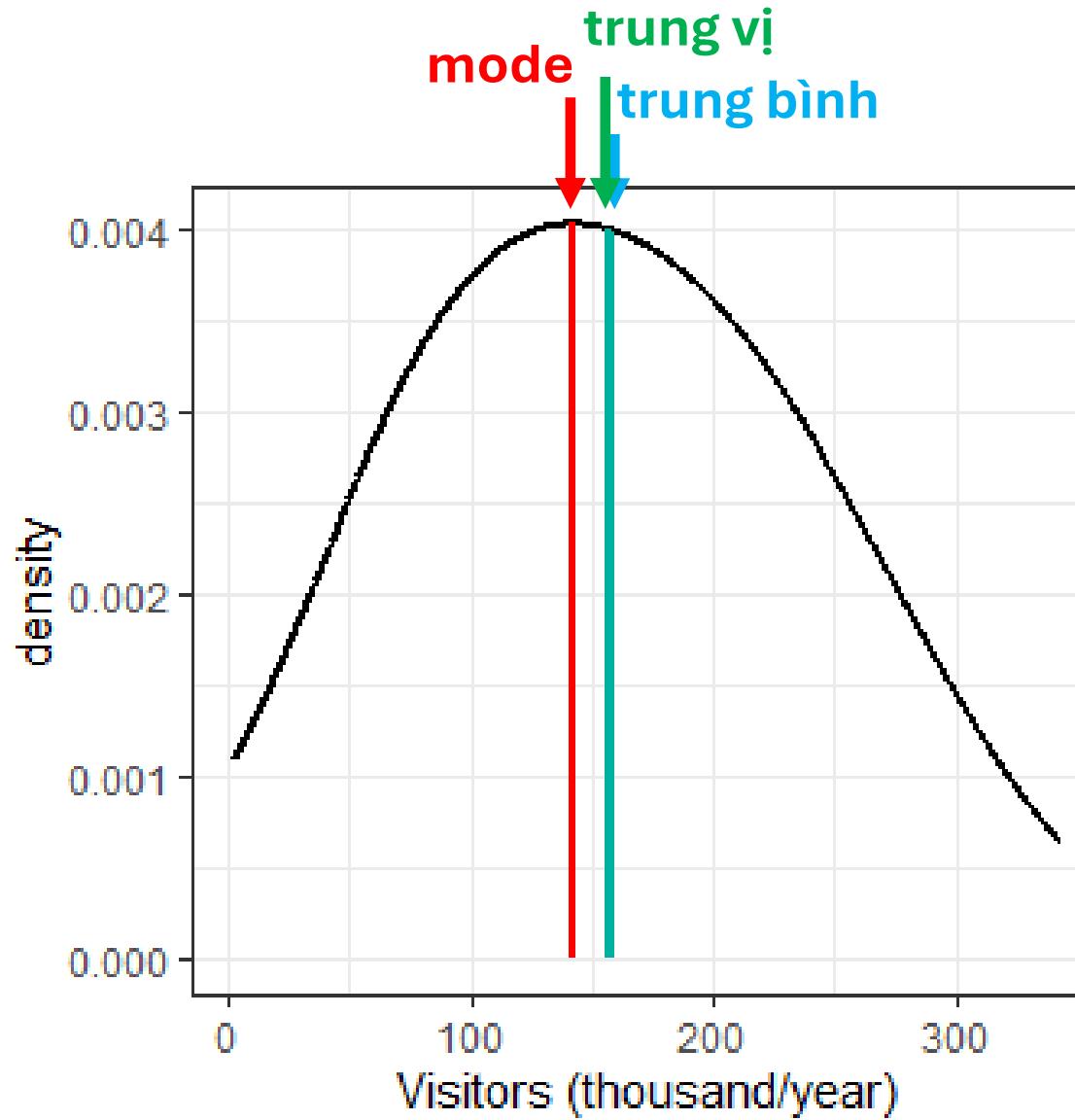


# Thống kê mô tả - Biến liên tục

- visitors: lượng khách tham quan (k người)
- budget: kinh phí vận hành & trùng tu (kUSD)
- events: số sự kiện được tổ chức
- level: cấp độ di tích
- category: loại hình di tích

site	visitors	budget	events	level	category
1	298	166	5	National-Special	Museum
2	78	107	6	Provincial	Historical
3	252	134	7	National	Monument
4	262	228	5	National	Archaeological Site
5	92	177	0	Provincial	Museum
6	236	213	3	National	Archaeological Site
7	120	148	6	Provincial	Museum
8	101	99	3	Provincial	Museum
9	55	111	6	Provincial	Museum
10	100	163	4	Provincial	Cultural Center
11	115	162	8	Provincial	Historical
12	280	223	8	National-Special	Monument
					Museum

# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

trung bình = 158

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

site	visitors	budget	events	level	category
1	298	166	5	National-Special	Museum
2	78	107	6	Provincial	Historical Monument
3	252	134	7	National	Archaeological Site
4	262	228	5	National	Museum
5	92	177	0	Provincial	Archaeological Site
6	236	213	3	National	Museum
7	120	148	6	Provincial	Museum
8	101	99	3	Provincial	Museum
9	55	111	6	Provincial	Museum
10	100	163	4	Provincial	Cultural Center
11	115	162	8	Provincial	Historical Monument
12	280	223	8	National-Special	Museum

# Thống kê mô tả - Biến liên tục

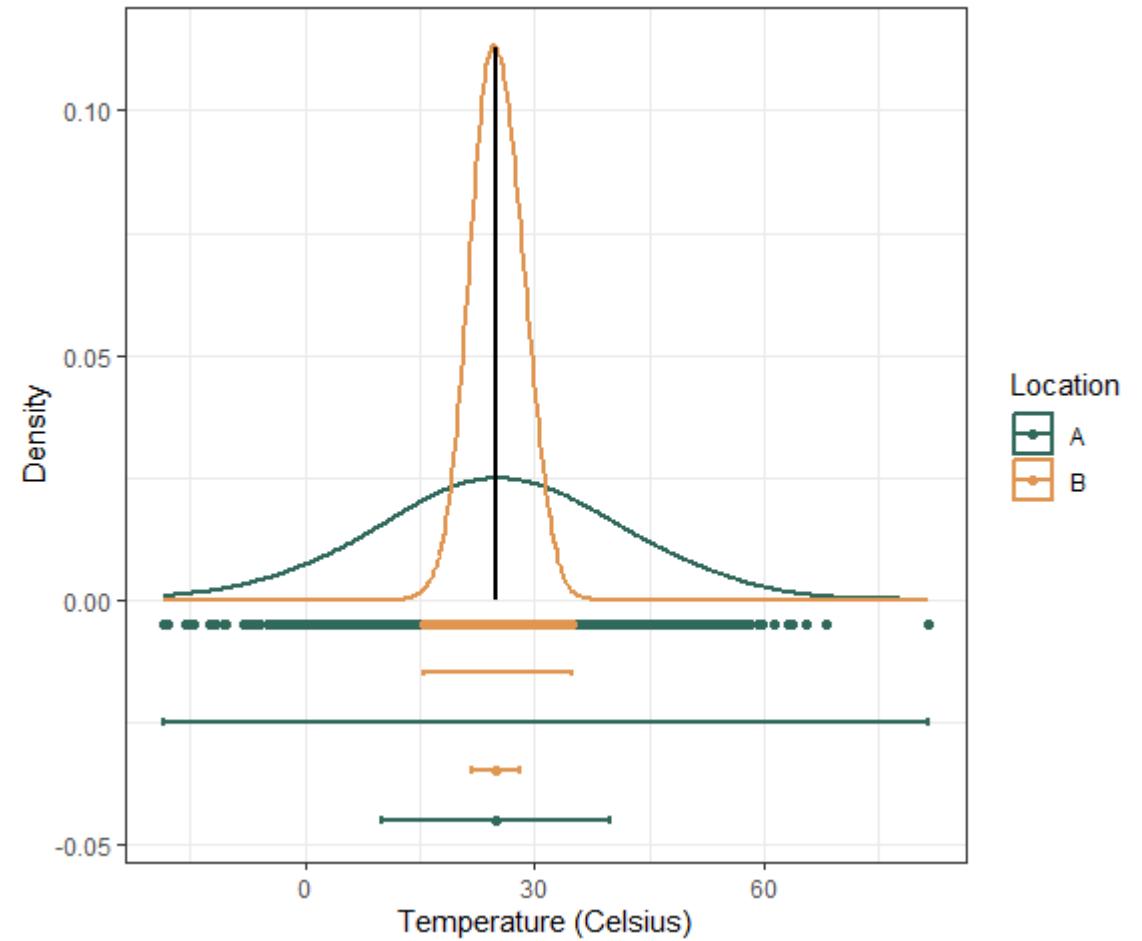
site	visitors	budget	events	level	category
40	2	84	5	Provincial	Archaeological Site
...	...	...	...	...	...
105	96	107	4	Provincial	Historical Monument
82	97	159	7	Provincial	Historical Monument
...	...	...	...	...	...
58	157	118	5	Provincial	Museum
68	157	188	2	Provincial	Cultural Center
...	...	...	...	...	...
101	213	203	1	National	Historical Monument
41	214	95	9	National	Museum
...	...	...	...	...	...
140	343	301	5	National-Special	Cultural Center

# Thống kê mô tả - Biến liên tục

trung vị = 157

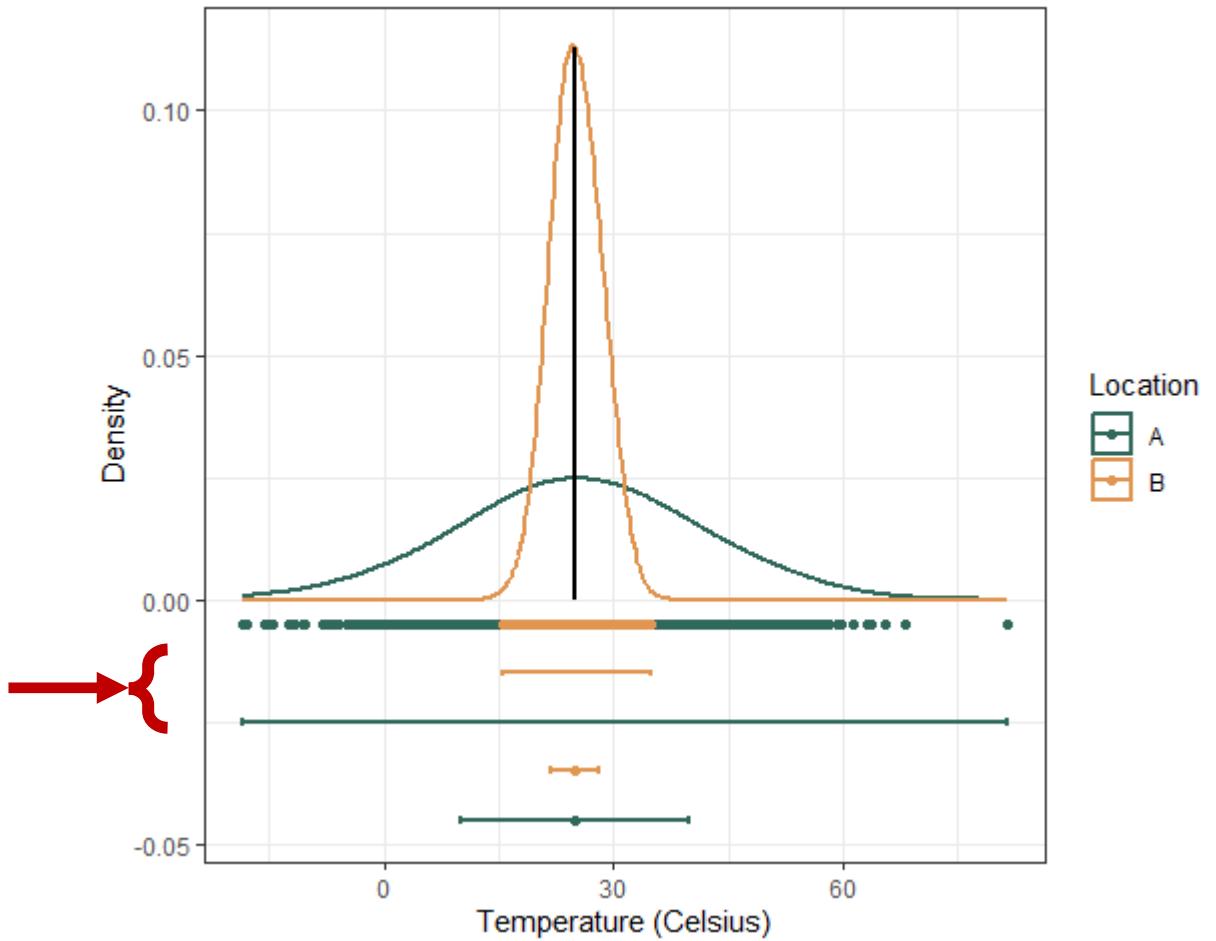
site	visitors	budget	events	level	category
40	2	84	5	Provincial	Archaeological Site
105	96	107	4	Provincial	Historical Monument
82	97	159	7	Provincial	Historical Monument
...	...	...	...	...	...
58	157	118	5	Provincial	Museum
68	157	188	2	Provincial	Cultural Center
...	...	...	...	...	...
101	213	203	1	National	Historical Monument
41	214	95	9	National	Museum
...	...	...	...	...	...
140	343	301	5	National-Special	Cultural Center

# Thống kê mô tả - Biến liên tục

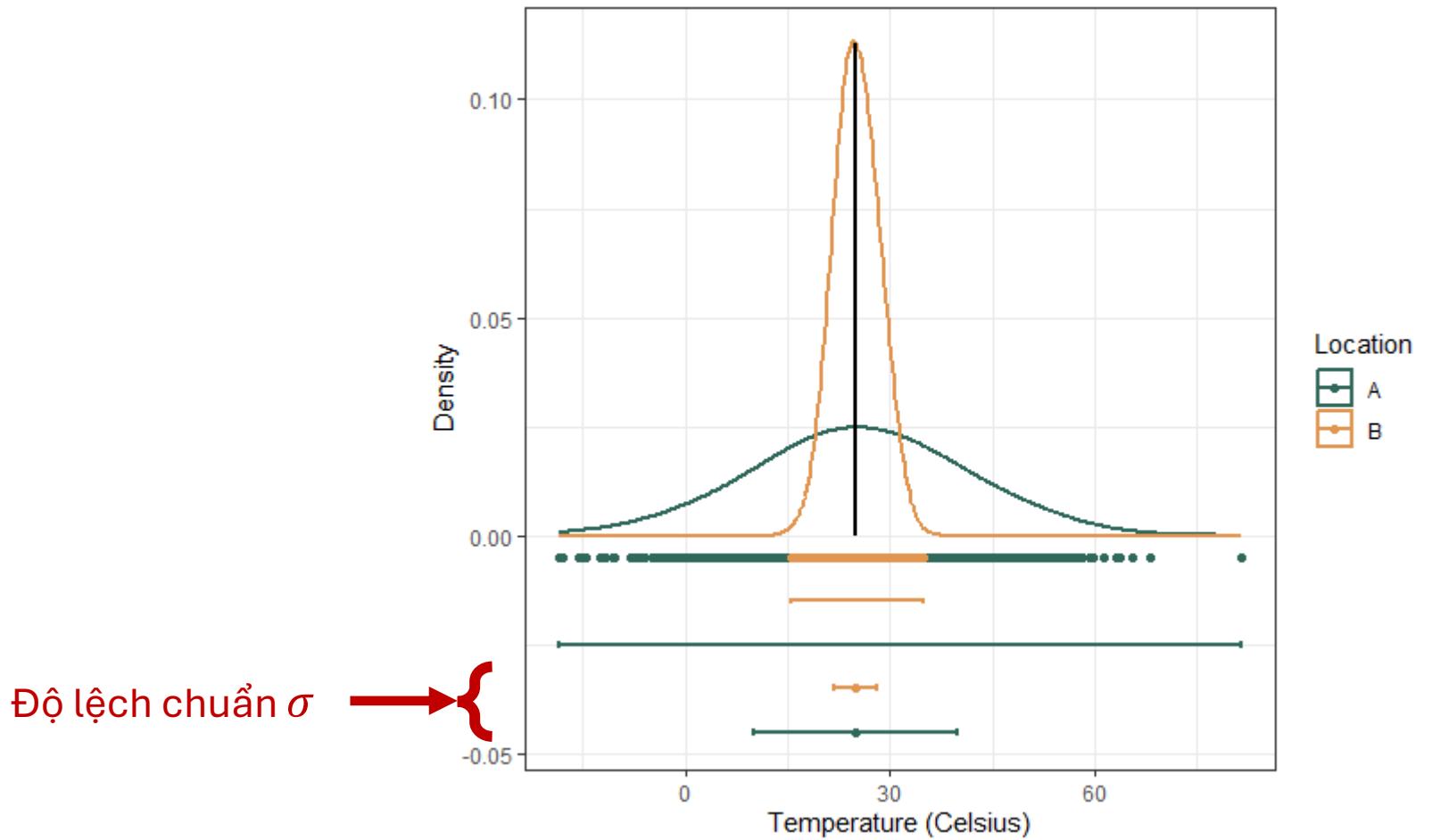


# Thống kê mô tả - Biến liên tục

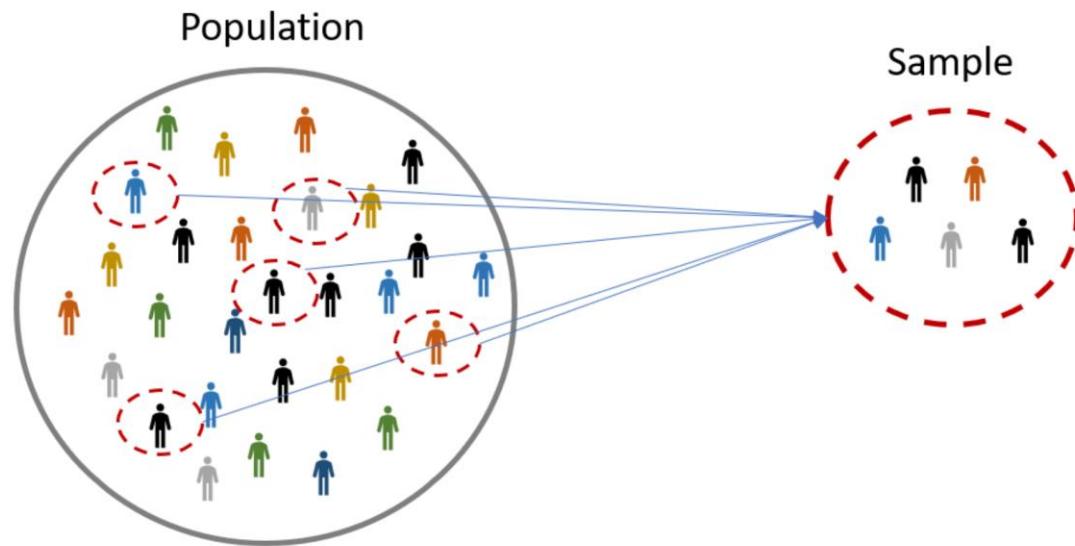
Khoảng biến thiên  
 $\min(x) \rightarrow \max(x)$



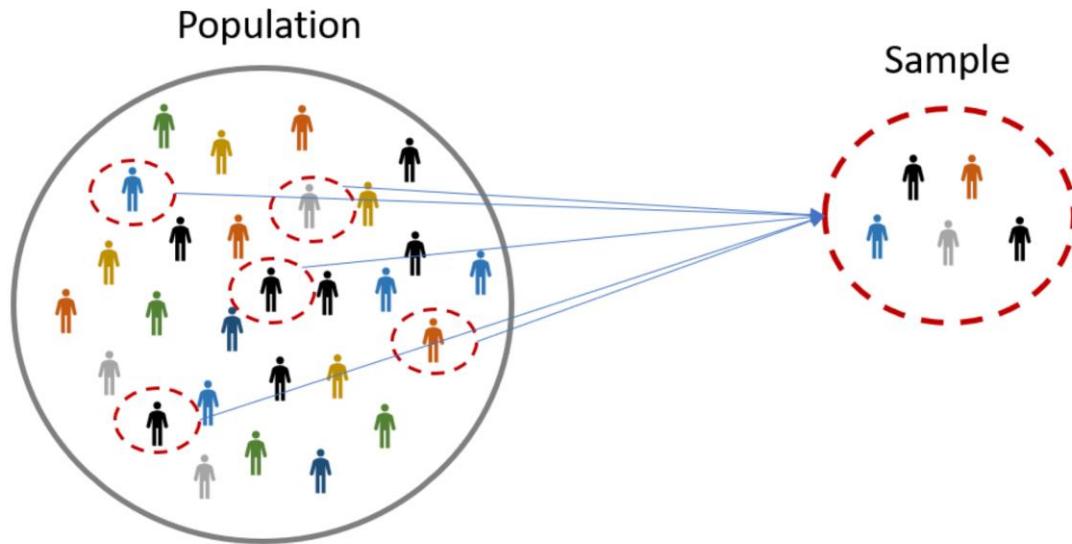
# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

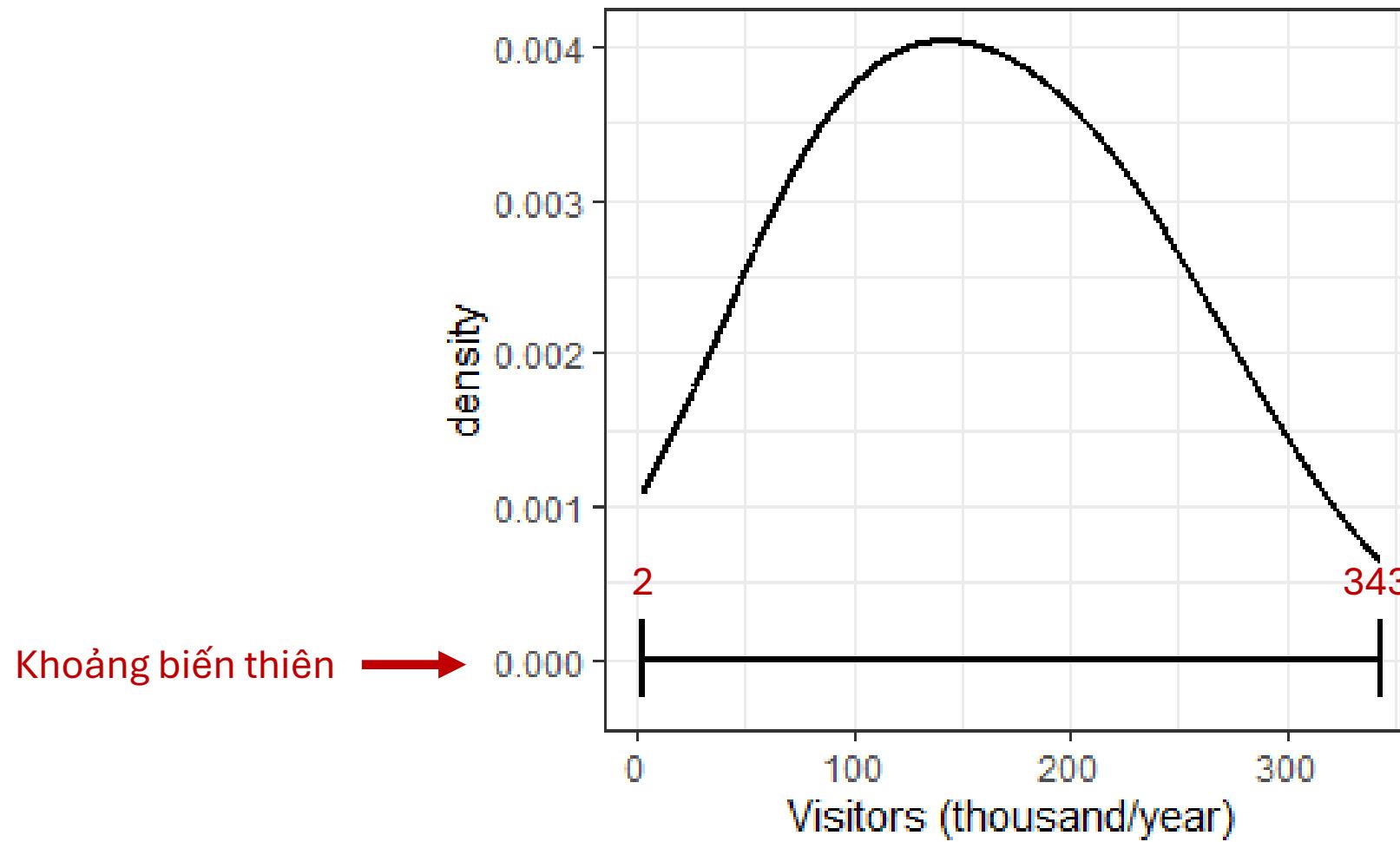


$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{N}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)}$$

$\sigma$	Độ lệch chuẩn của quần thể
$\sigma^2$	Phương sai của quần thể
$X_i$	Giá trị của đối tượng $i$
$\bar{X}$	Trung bình của quần thể
$N$	Số lượng đối tượng trong quần thể
$s$	Độ lệch chuẩn của mẫu
$s^2$	Phương sai của mẫu
$x_i$	Giá trị của đối tượng $i$
$\bar{x}$	Trung bình của mẫu
$n$	Số lượng đối tượng trong mẫu

# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)}$$

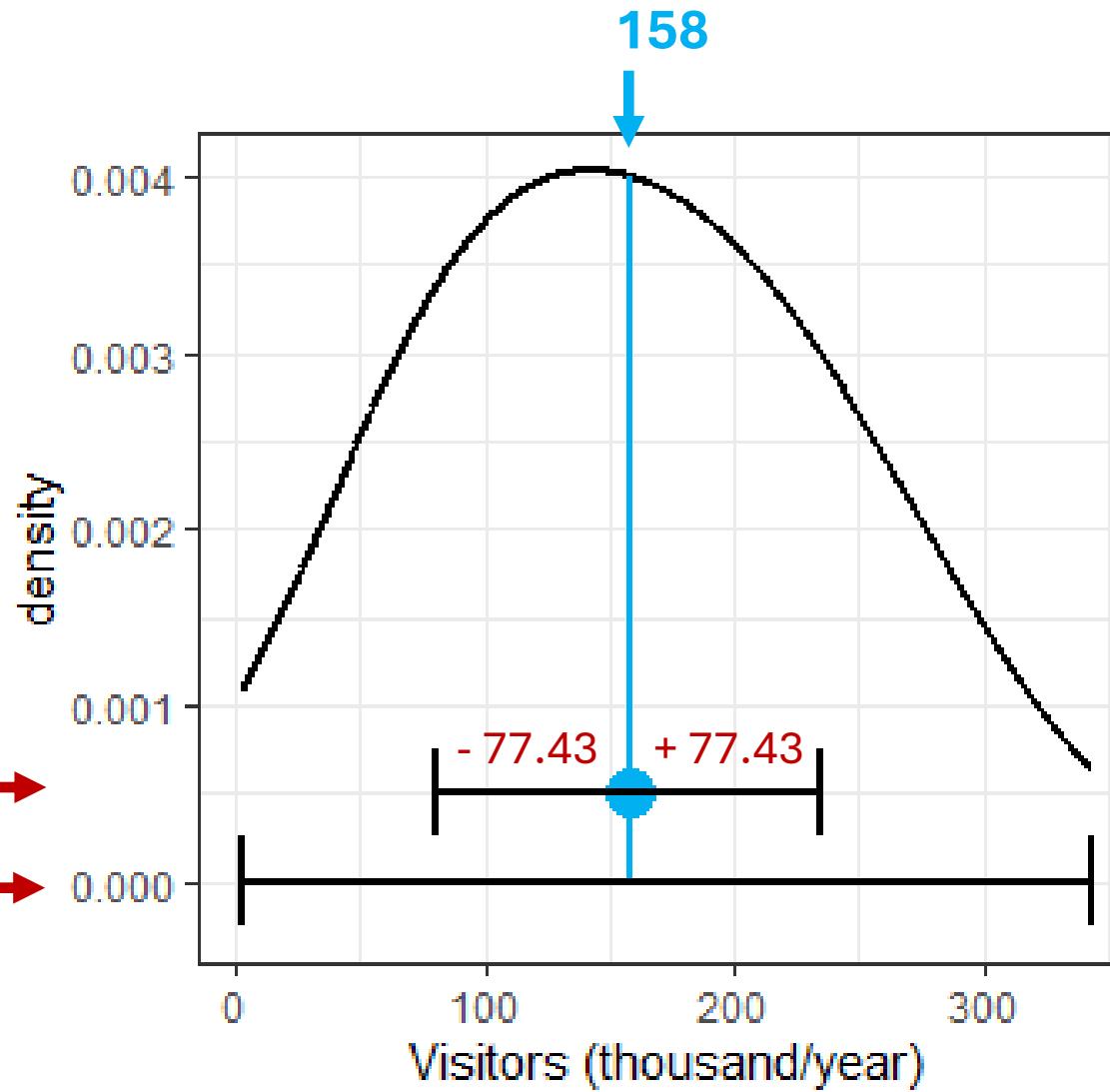
$$s = 77.43$$

$$s^2 = 5995.95$$

Độ lệch chuẩn



Khoảng biến thiên



# Thống kê mô tả - Biến liên tục

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)}$$

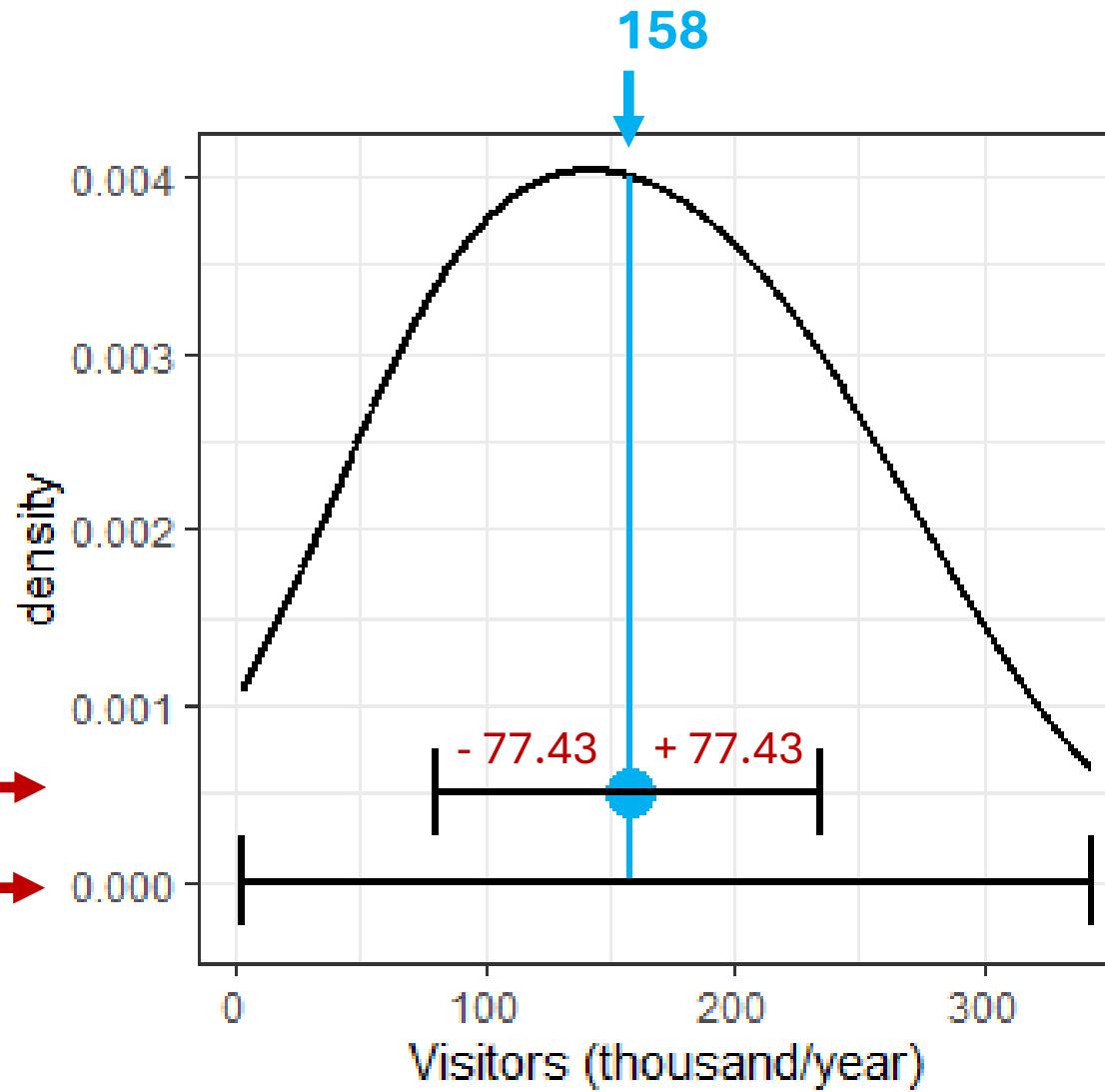
$$s = 77.43$$

$$s^2 = 5995.95$$

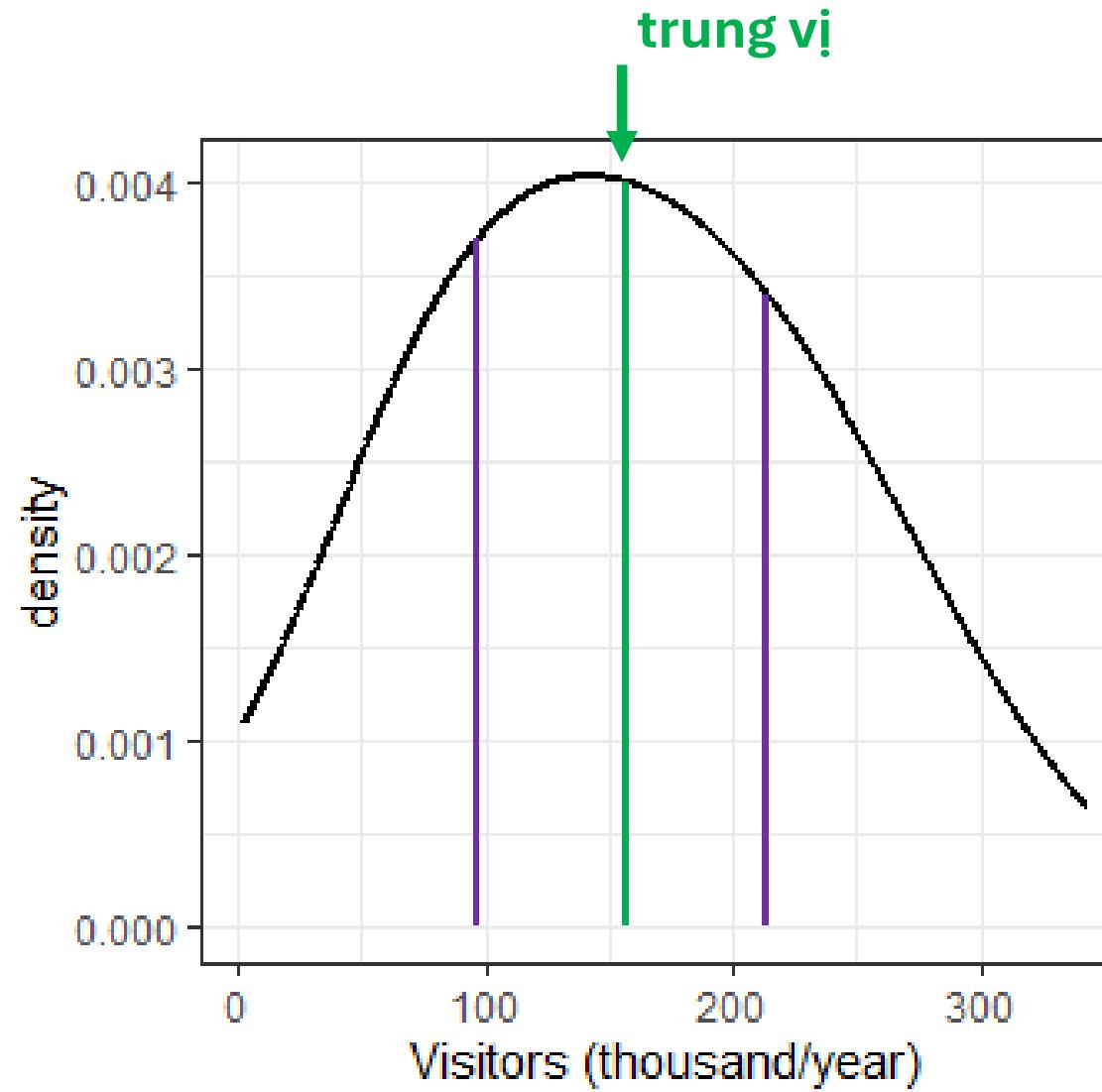
Độ lệch chuẩn



Khoảng biến thiên

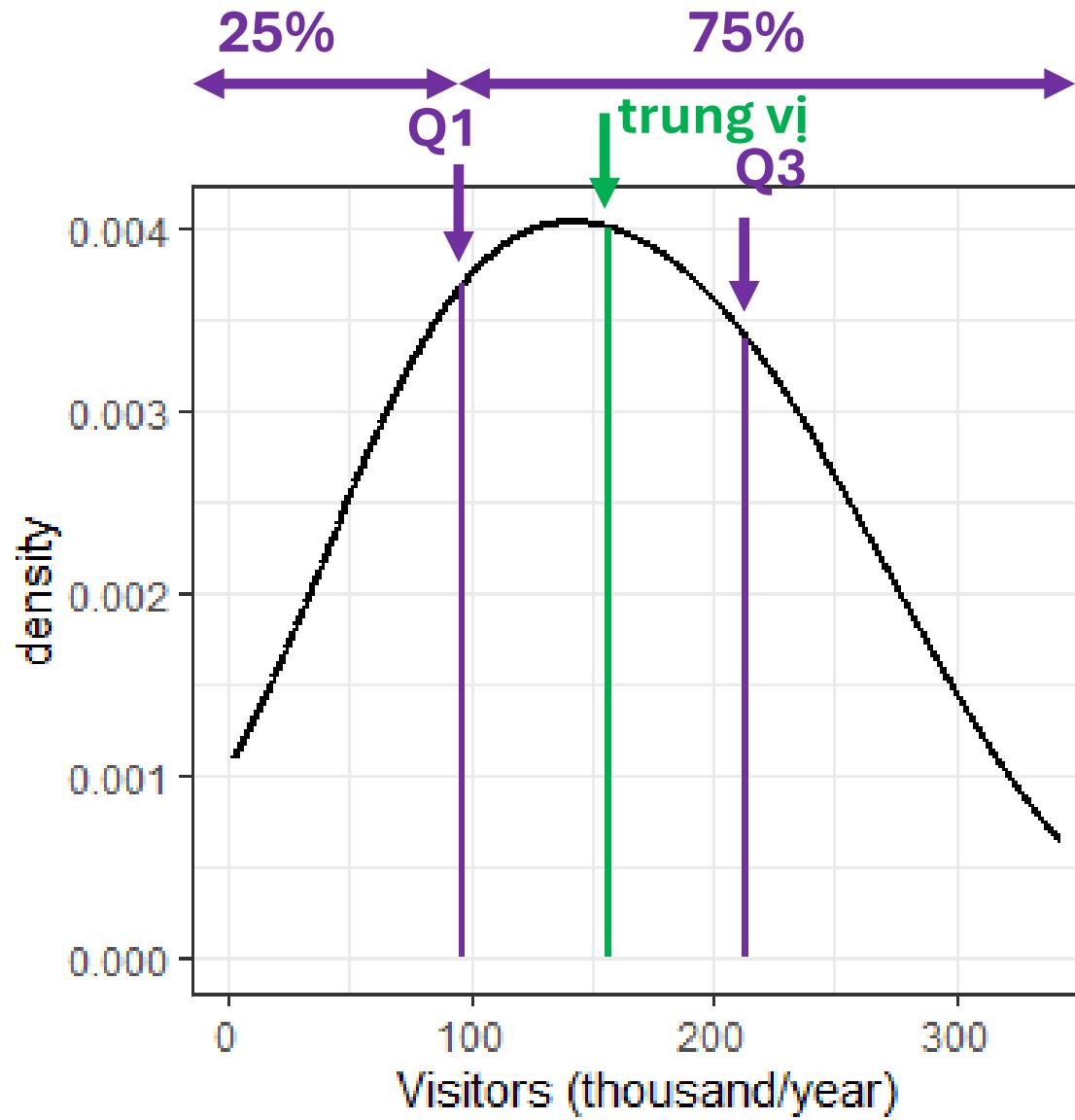


# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

Tứ phân vị



# Thống kê mô tả - Biến liên tục

tứ phân vị 1 = Q1 = 96.5

trung vị = Q2 = 157

tứ phân vị 3 = Q3 = 213.5

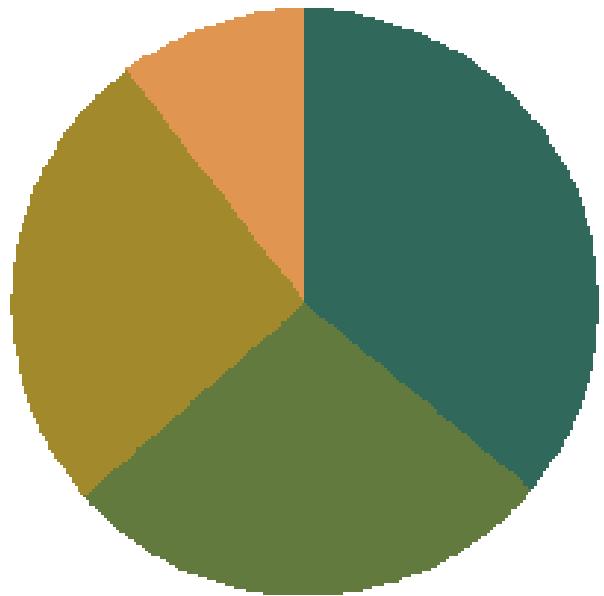
site	visitors	budget	events	level	category
40	2	84	5	Provincial	Archaeological Site
105	96	107	4	Provincial	Historical Monument
82	97	159	7	Provincial	Historical Monument
...	...	...	...	...	...
58	157	118	5	Provincial	Museum
68	157	188	2	Provincial	Cultural Center
...	...	...	...	...	...
101	213	203	1	National	Historical Monument
41	214	95	9	National	Museum
...	...	...	...	...	...
140	343	301	5	National-Special	Cultural Center

# Thống kê mô tả - Biến định tính

- visitors: lượng khách tham quan (k người)
- budget: kinh phí vận hành & trùng tu (kUSD)
- events: số sự kiện được tổ chức
- level: cấp độ di tích
- category: loại hình di tích

site	visitors	budget	events	level	category
1	298	166	5	National-Special Provincial	Museum
	78	107	6		Historical
	252	134	7	National	Monument
3	262	228	5	National	Archaeological
	92	177	0	Provincial	Site
4	236	213	3	National	Museum
	120	148	6	Provincial	Museum
6	101	99	3	Provincial	Museum
	55	111	6	Provincial	Museum
8	100	163	4	Provincial	Cultural Center
	115	162	8	Provincial	Historical
12	280	223	8	National-Special	Monument
					Museum

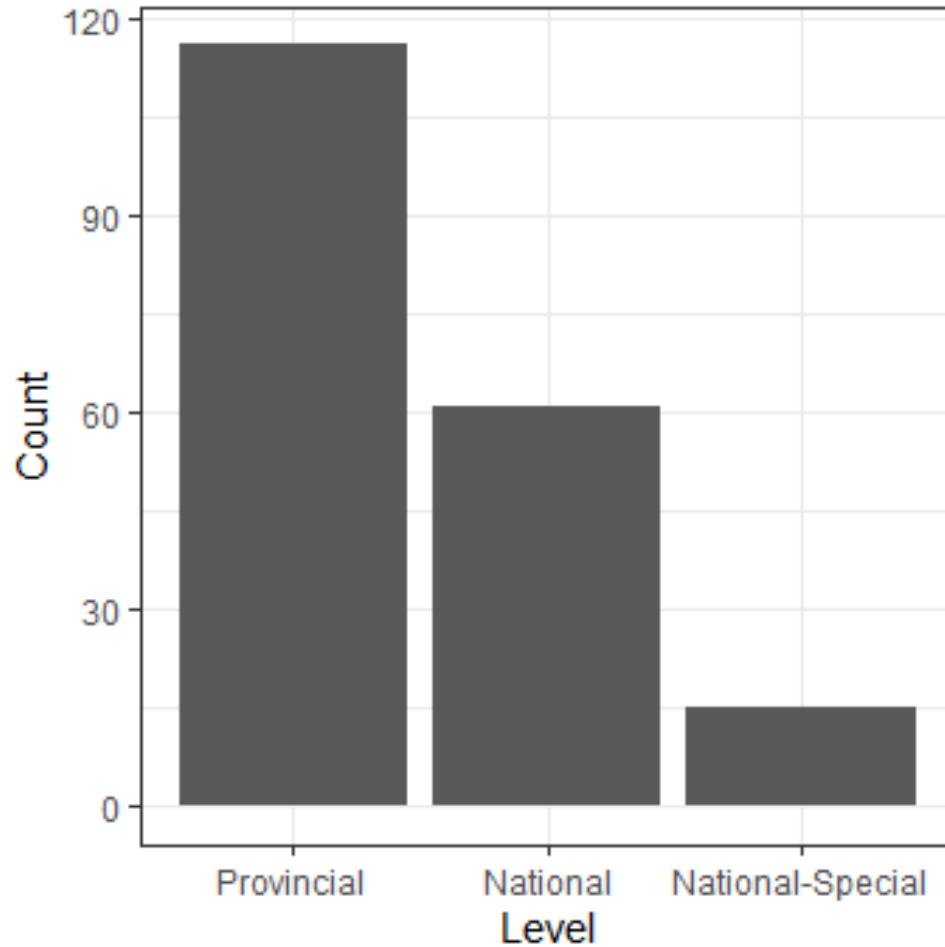
# Thống kê mô tả - Biến định tính



Category	Count	%
Historical Monument	69	35.94
Museum	53	27.60
Archaeological Site	50	26.04
Cultural Center	20	10.42
<b>Total</b>	<b>192</b>	<b>100.00</b>

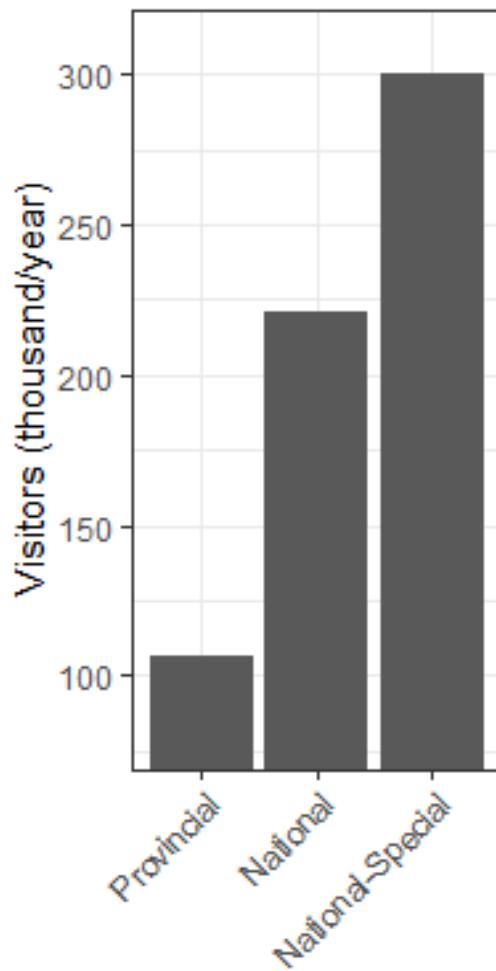
Category	Count	%
Historical Monument	69	35.94
Museum	53	27.60
Archaeological Site	50	26.04
Cultural Center	20	10.42
<b>Total</b>	<b>192</b>	<b>100.00</b>

# Thống kê mô tả - Biến định tính

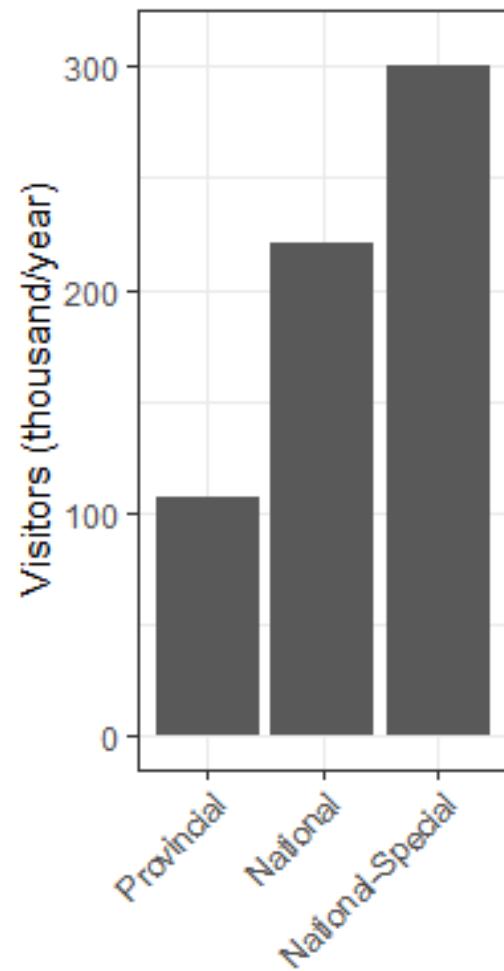
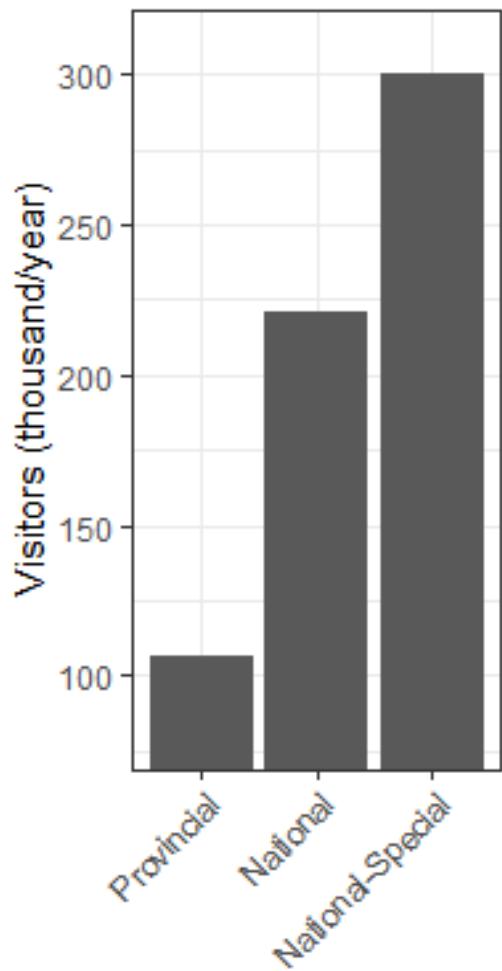


Income	Count	%
Provincial	116	60.42
National	61	21.77
National-Special	15	7.81
<b>Total</b>	<b>192</b>	<b>100.00</b>

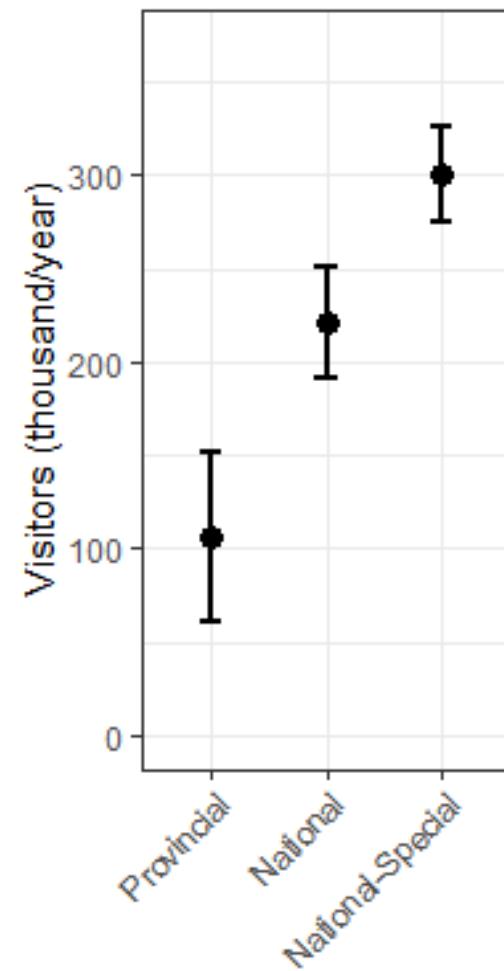
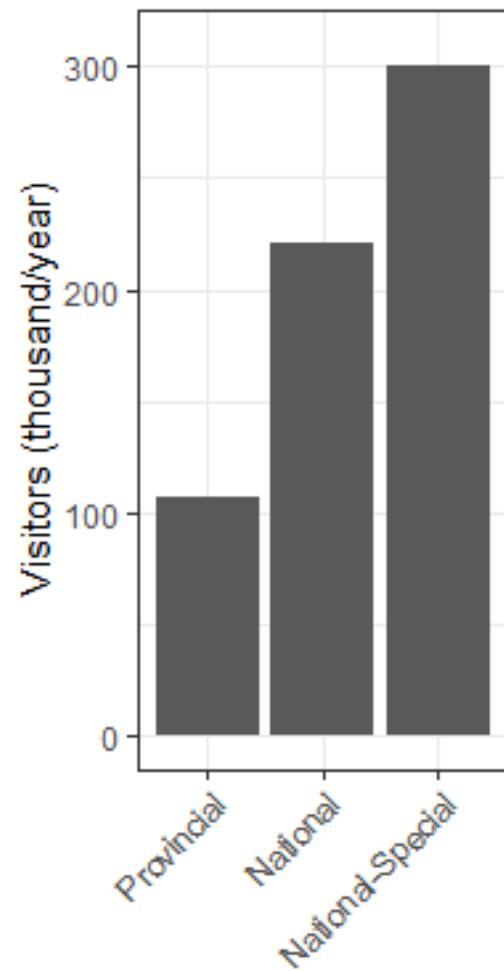
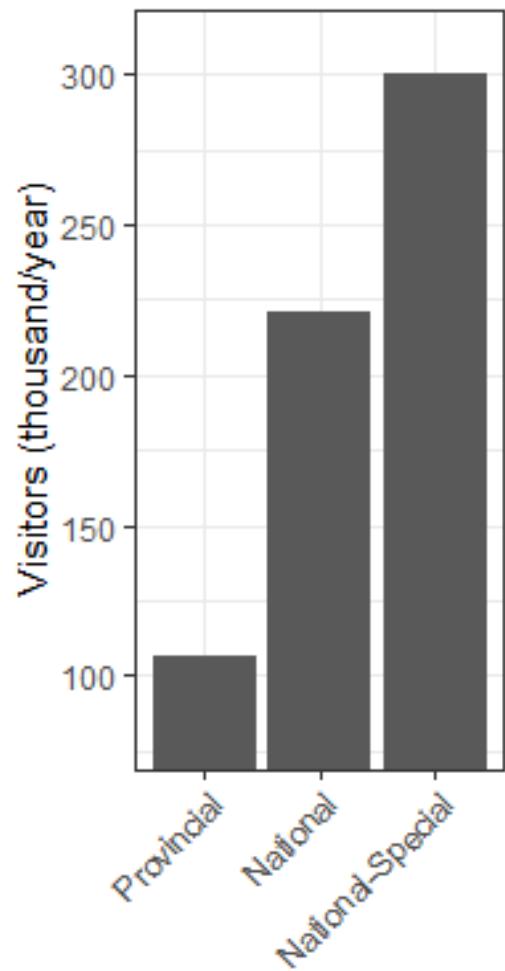
# Thống kê mô tả - Cho từng nhóm



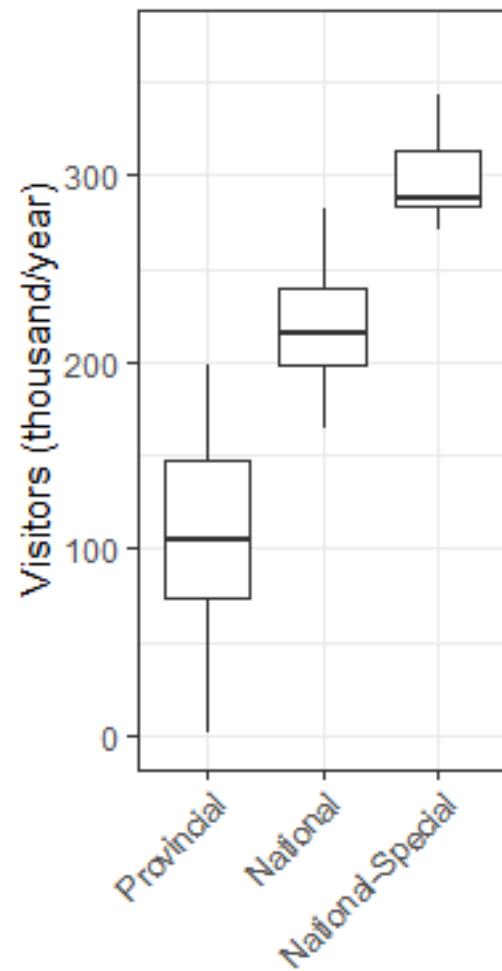
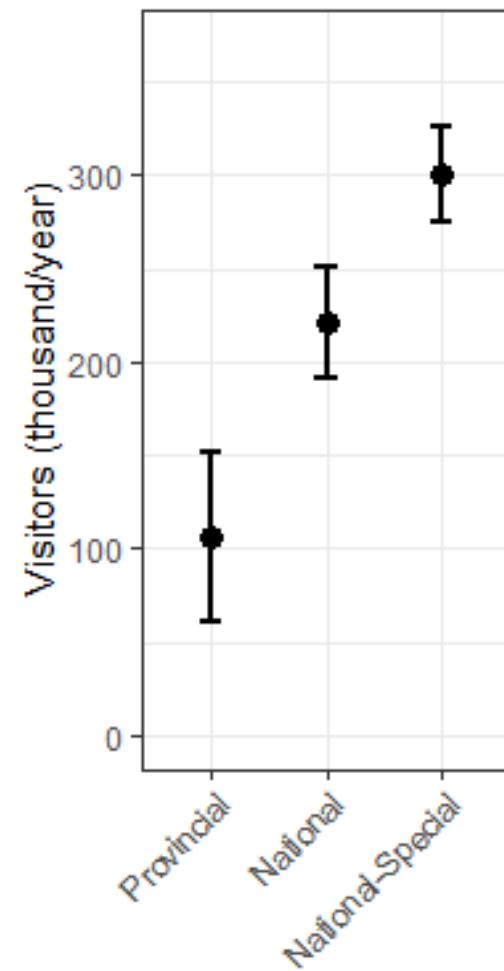
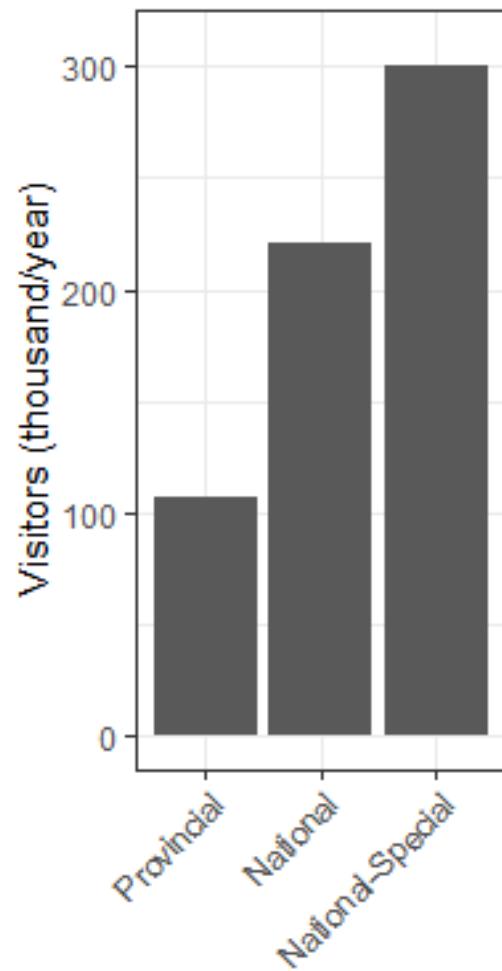
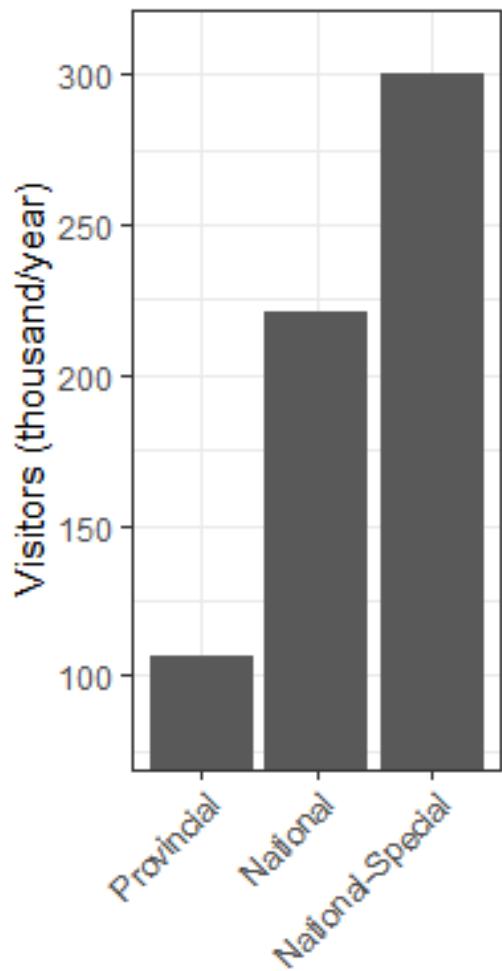
# Thống kê mô tả - Cho từng nhóm



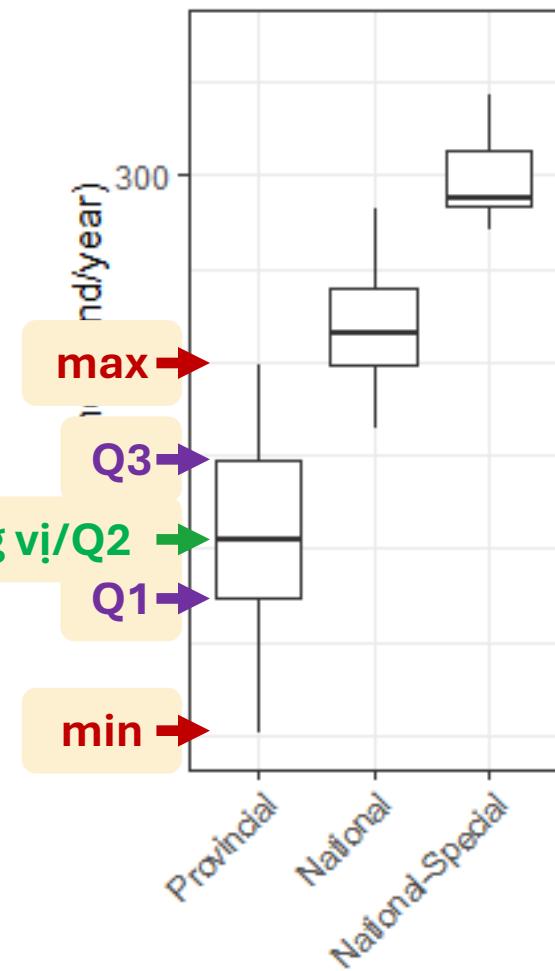
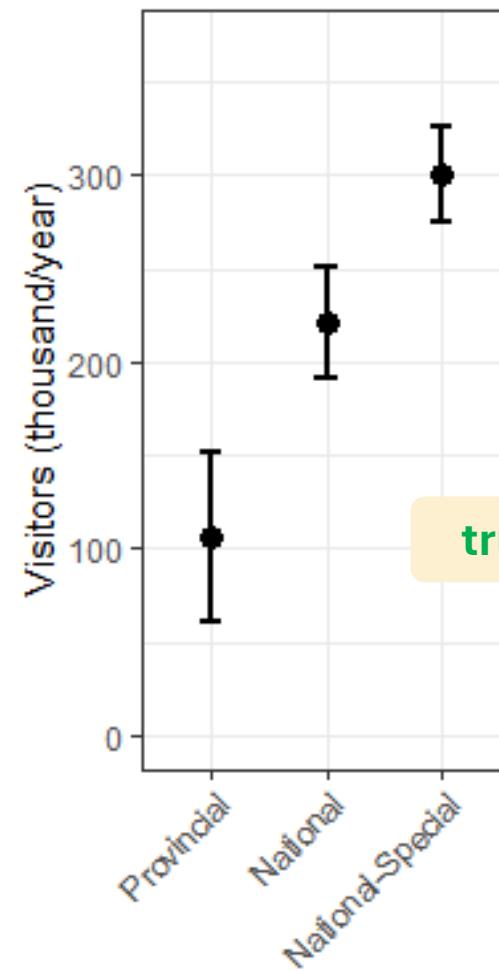
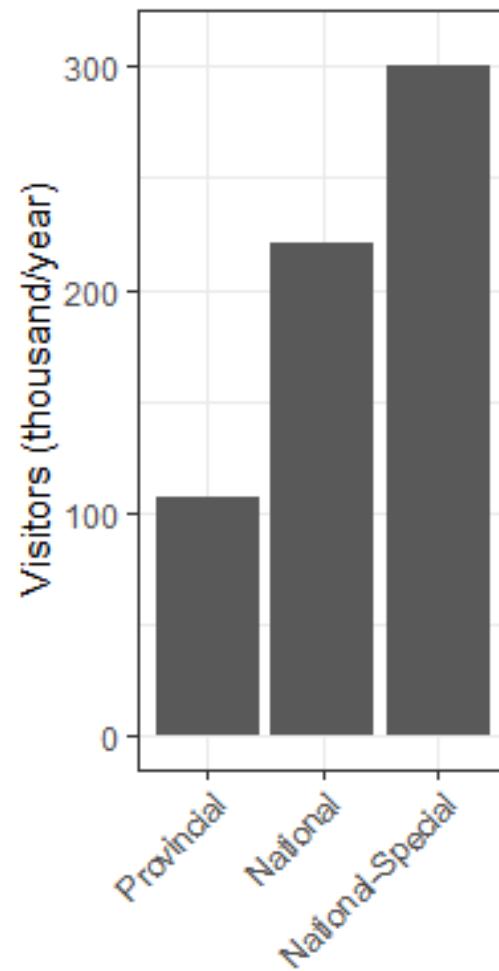
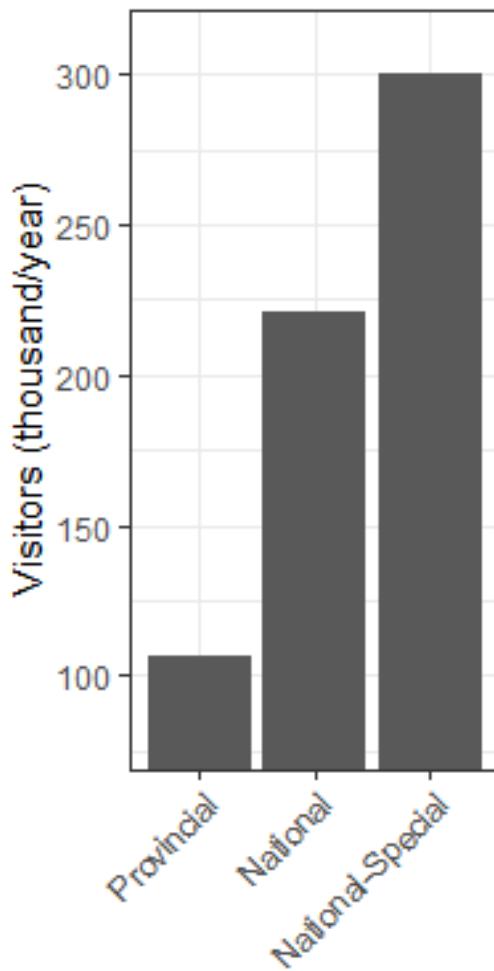
# Thống kê mô tả - Cho từng nhóm



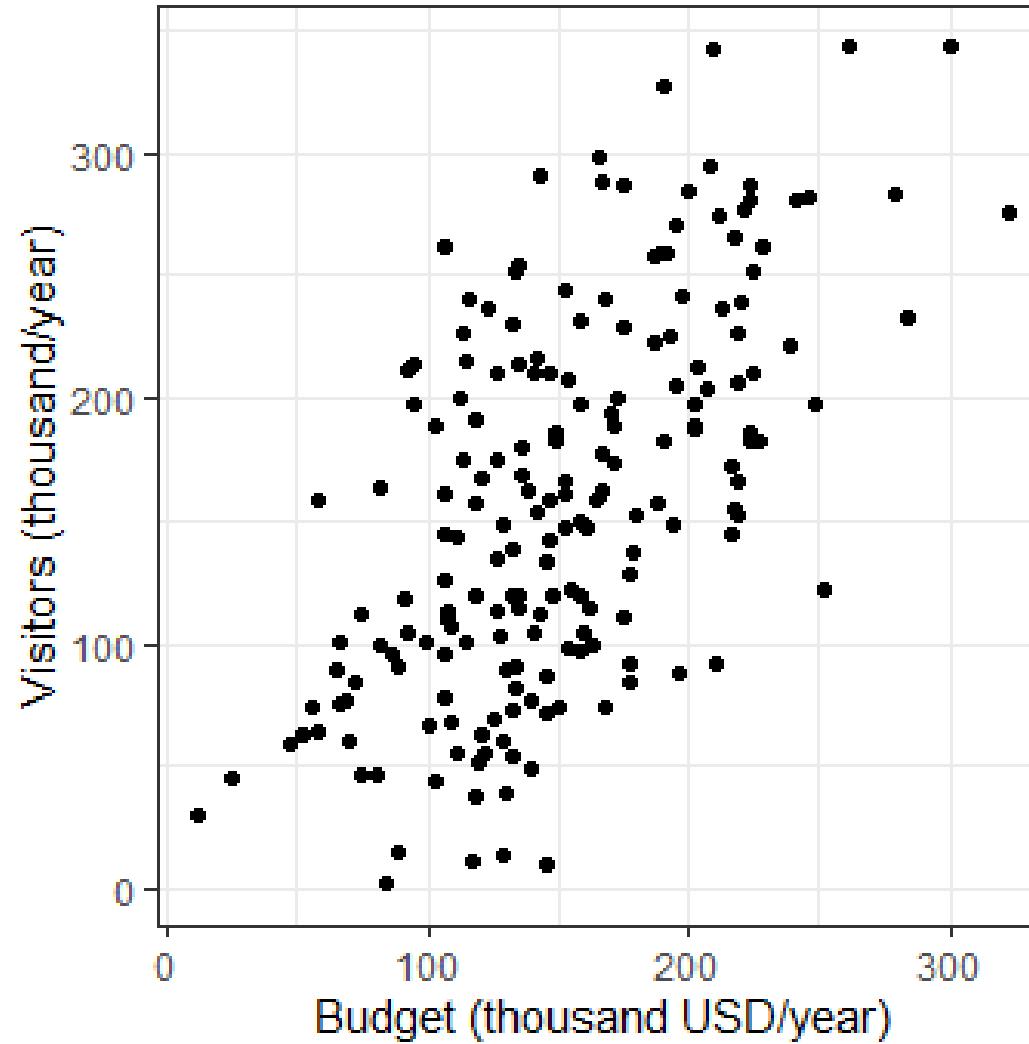
# Thống kê mô tả - Cho từng nhóm



# Thống kê mô tả - Cho từng nhóm



# Thống kê mô tả - Tương quan



# Thống kê mô tả - Tương quan

- Pearson's  $r$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$r$  Hệ số tương quan Pearson

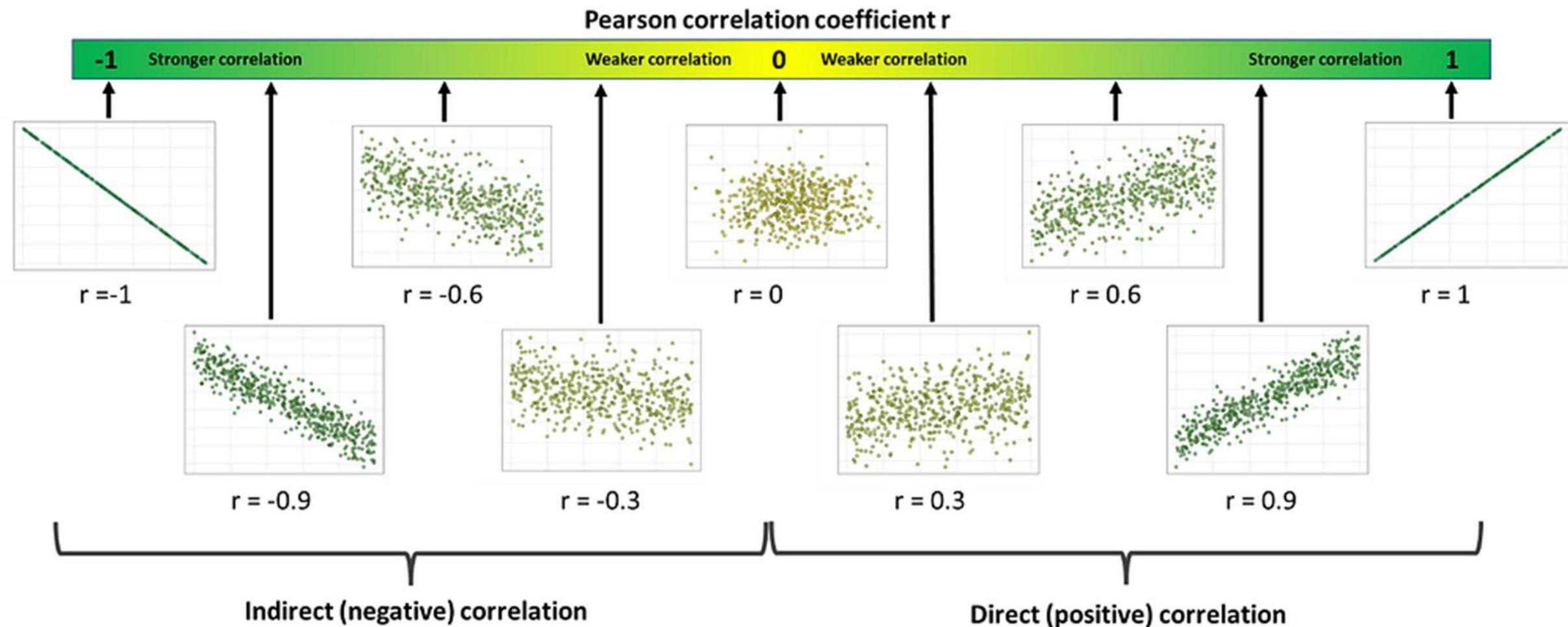
$x_i$  Giá trị biến  $x$  của đối tượng quan sát  $i$

$y_i$  Giá trị biến  $y$  của đối tượng quan sát  $i$

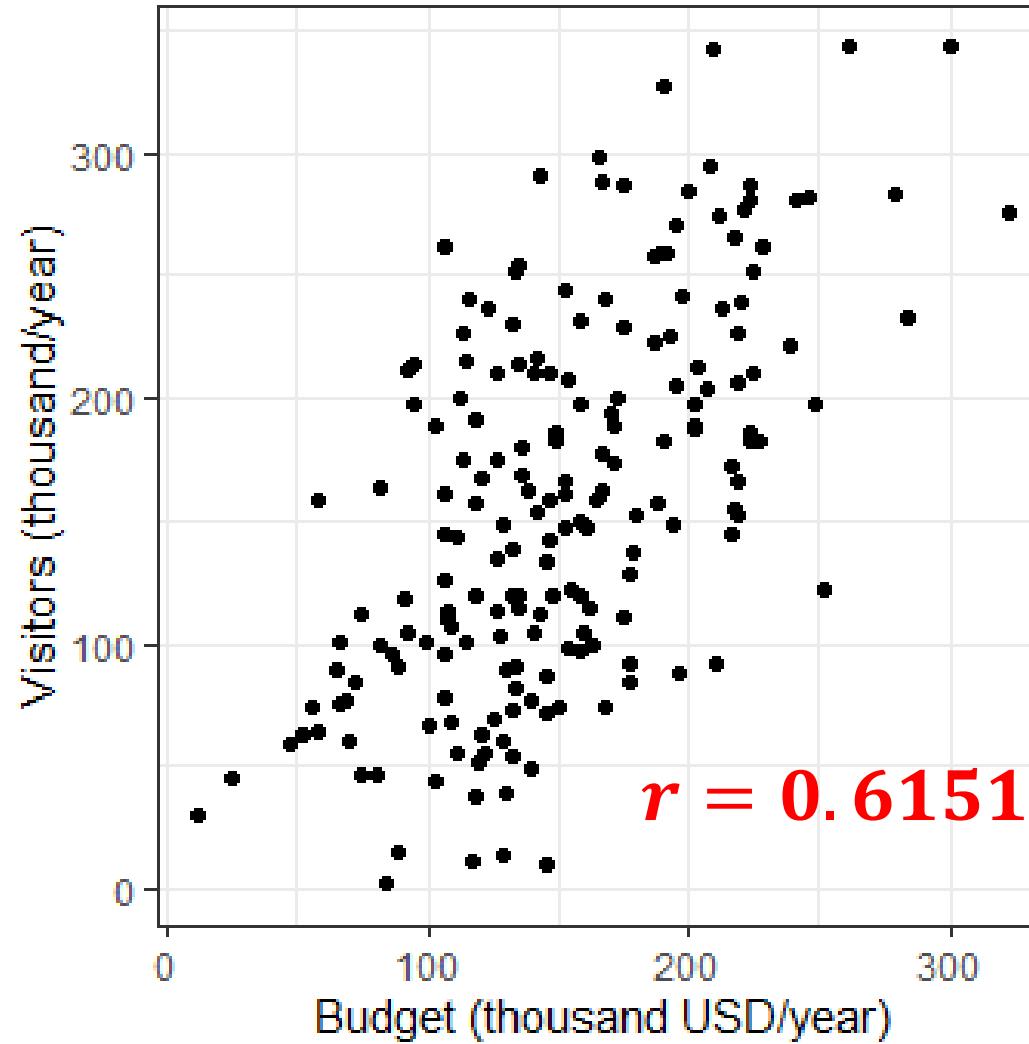
$\bar{x}$  Trung bình biến  $x$

$\bar{y}$  Trung bình biến  $y$

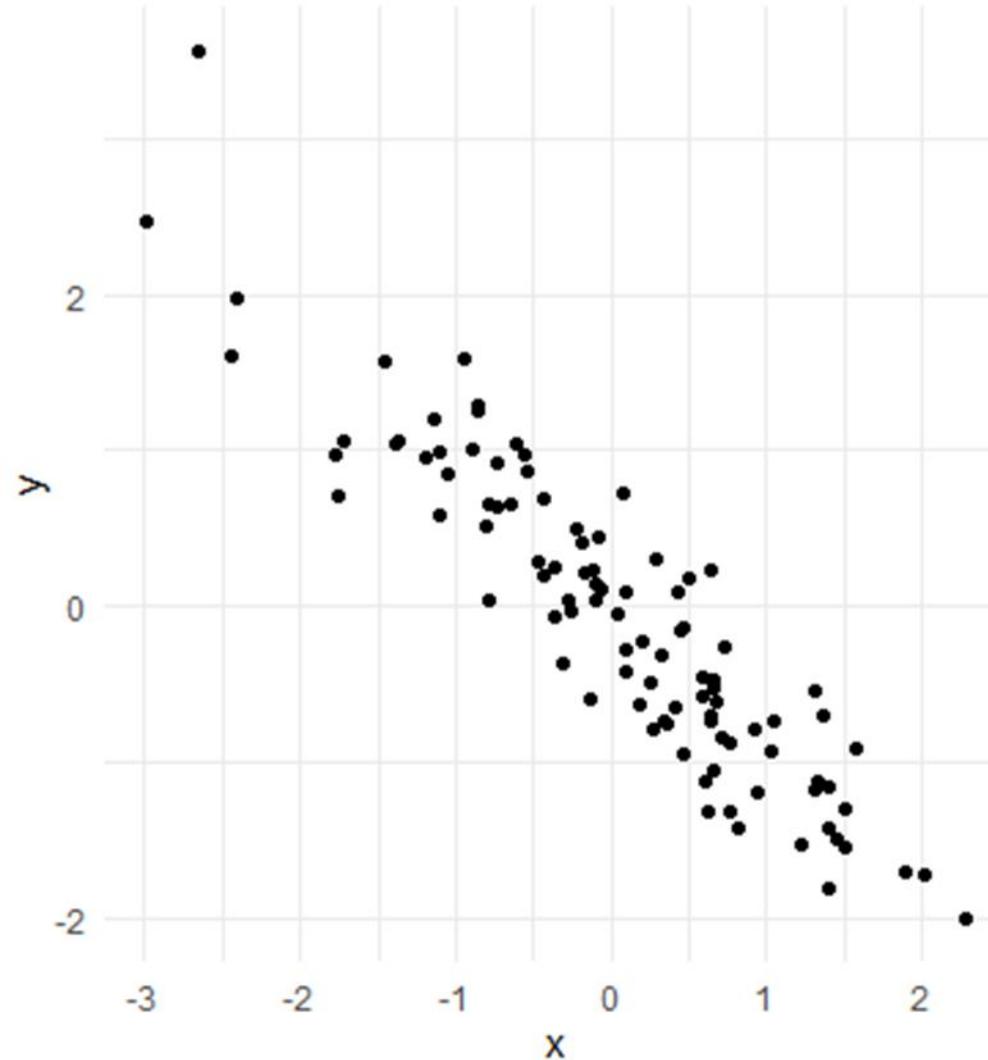
# Thống kê mô tả - Tương quan



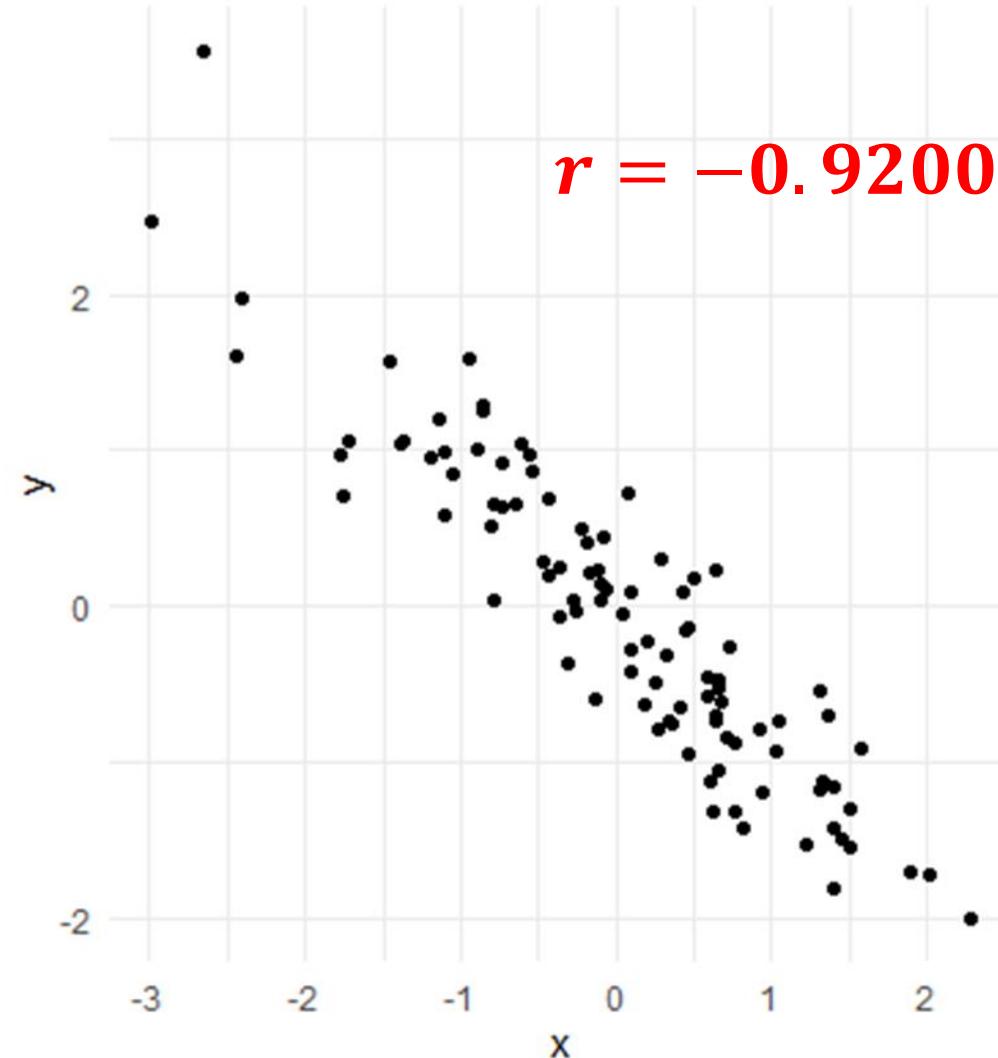
# Thống kê mô tả - Tương quan



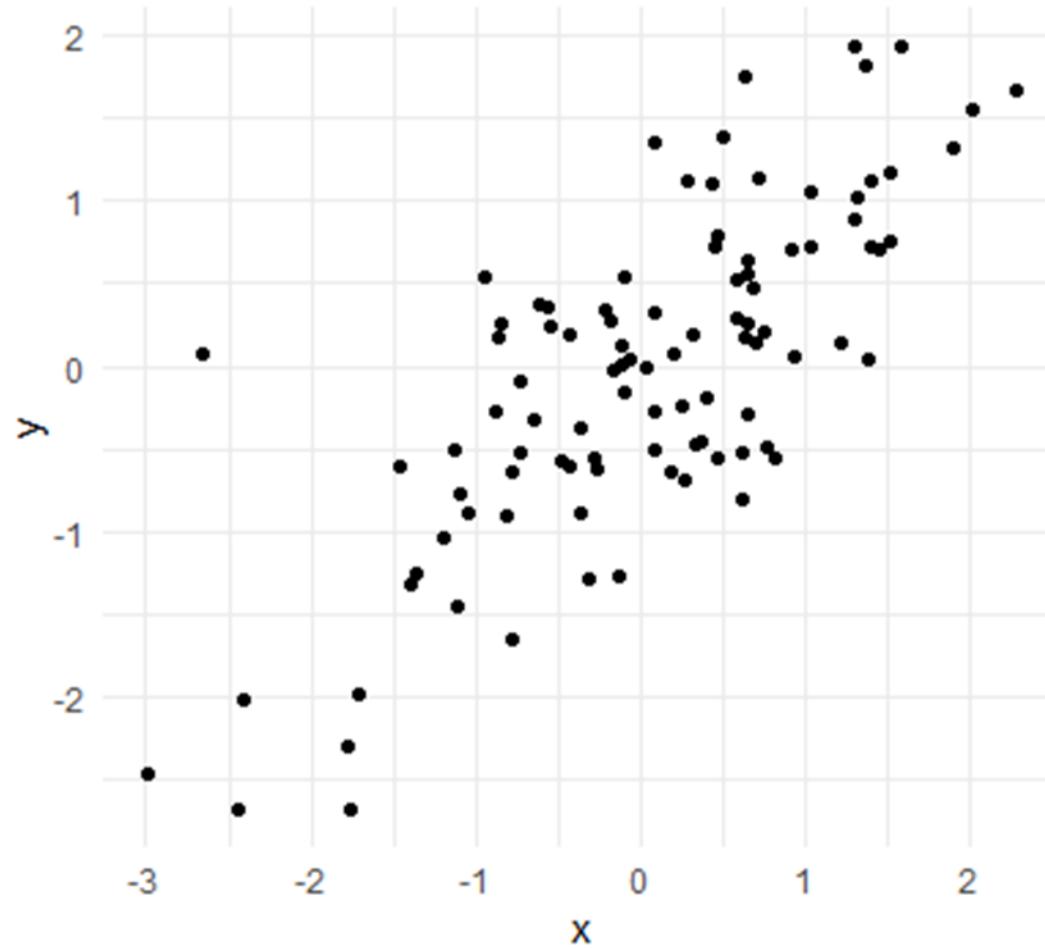
# Thống kê mô tả - Tương quan



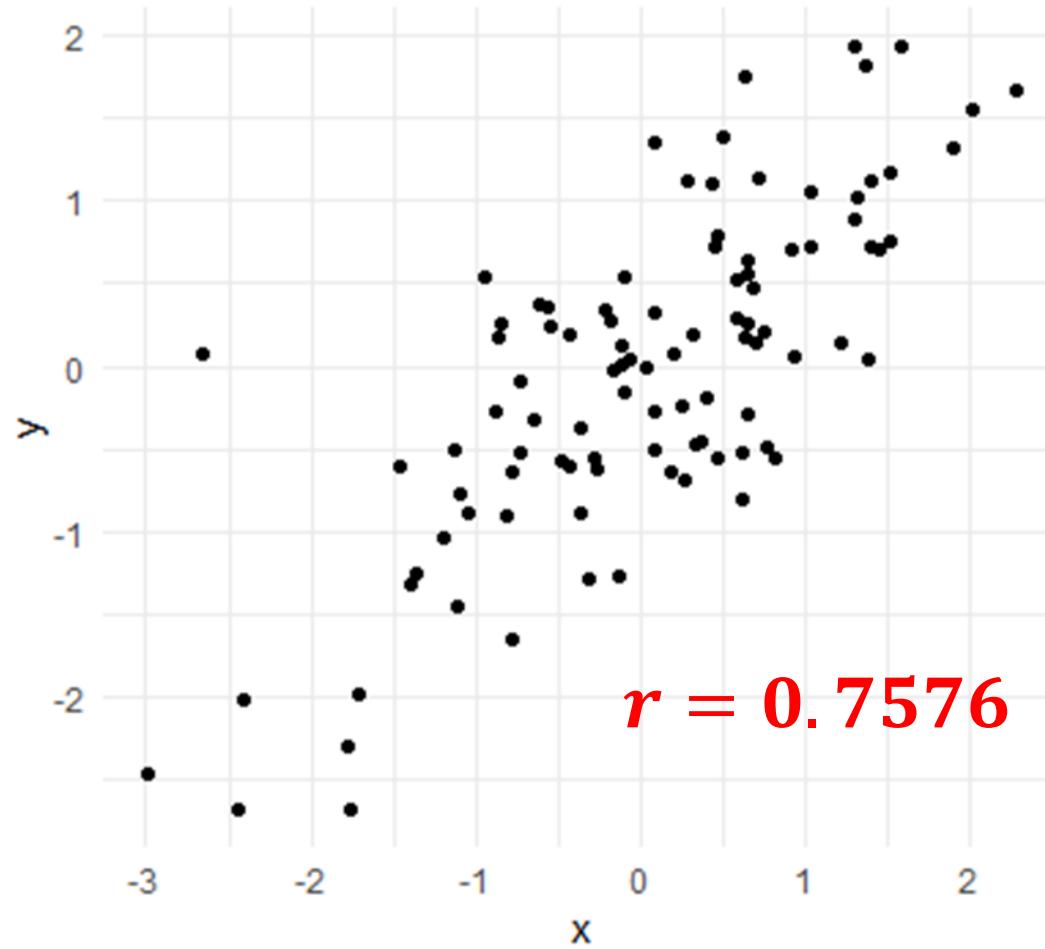
# Thống kê mô tả - Tương quan



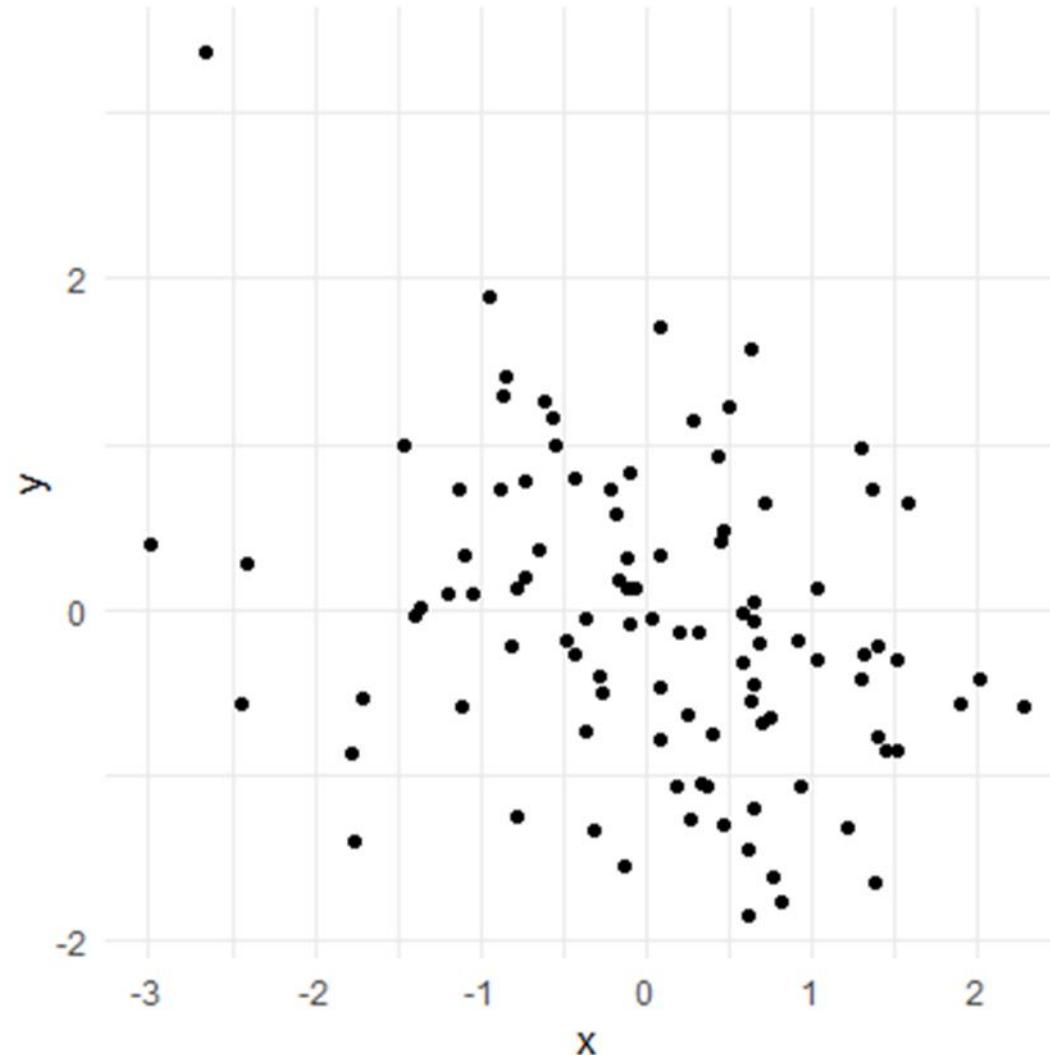
# Thống kê mô tả - Tương quan



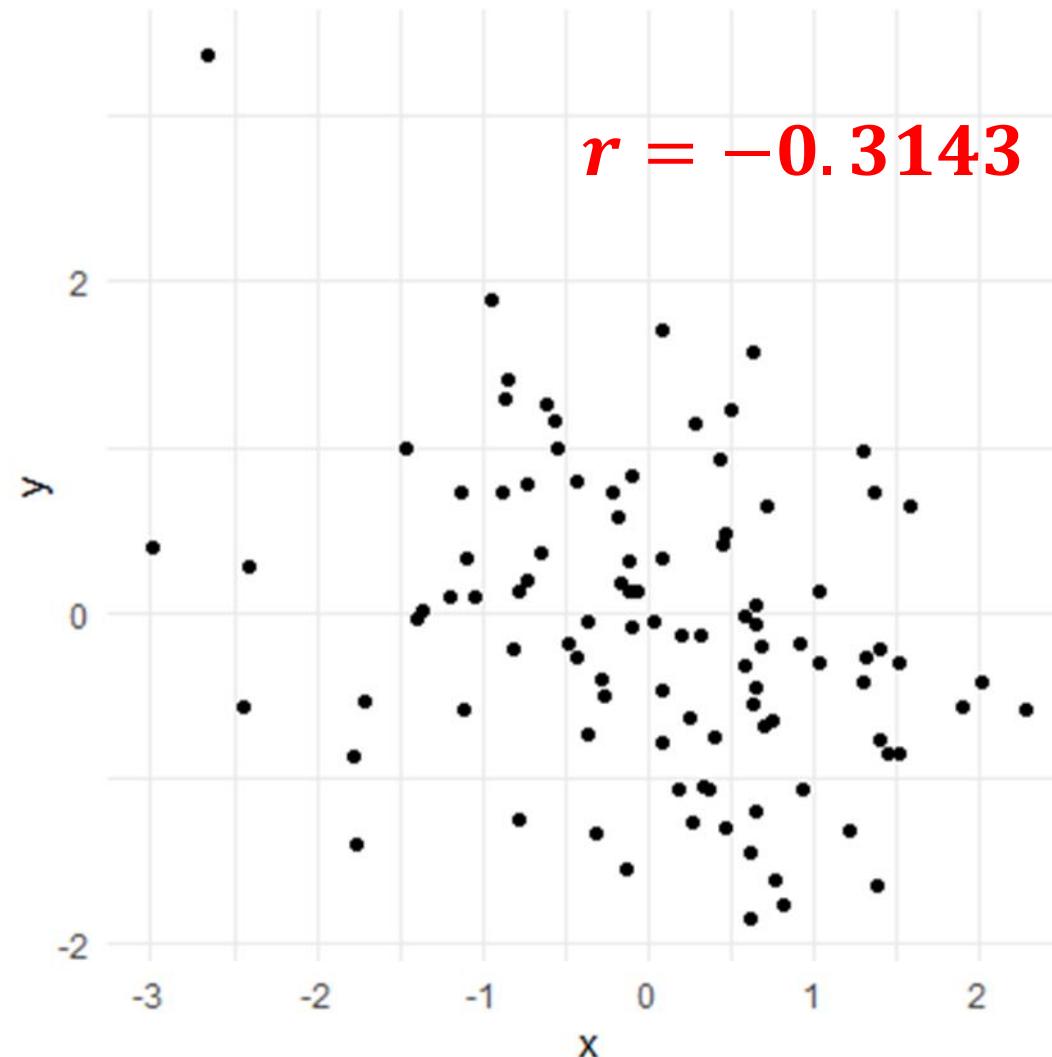
# Thống kê mô tả - Tương quan



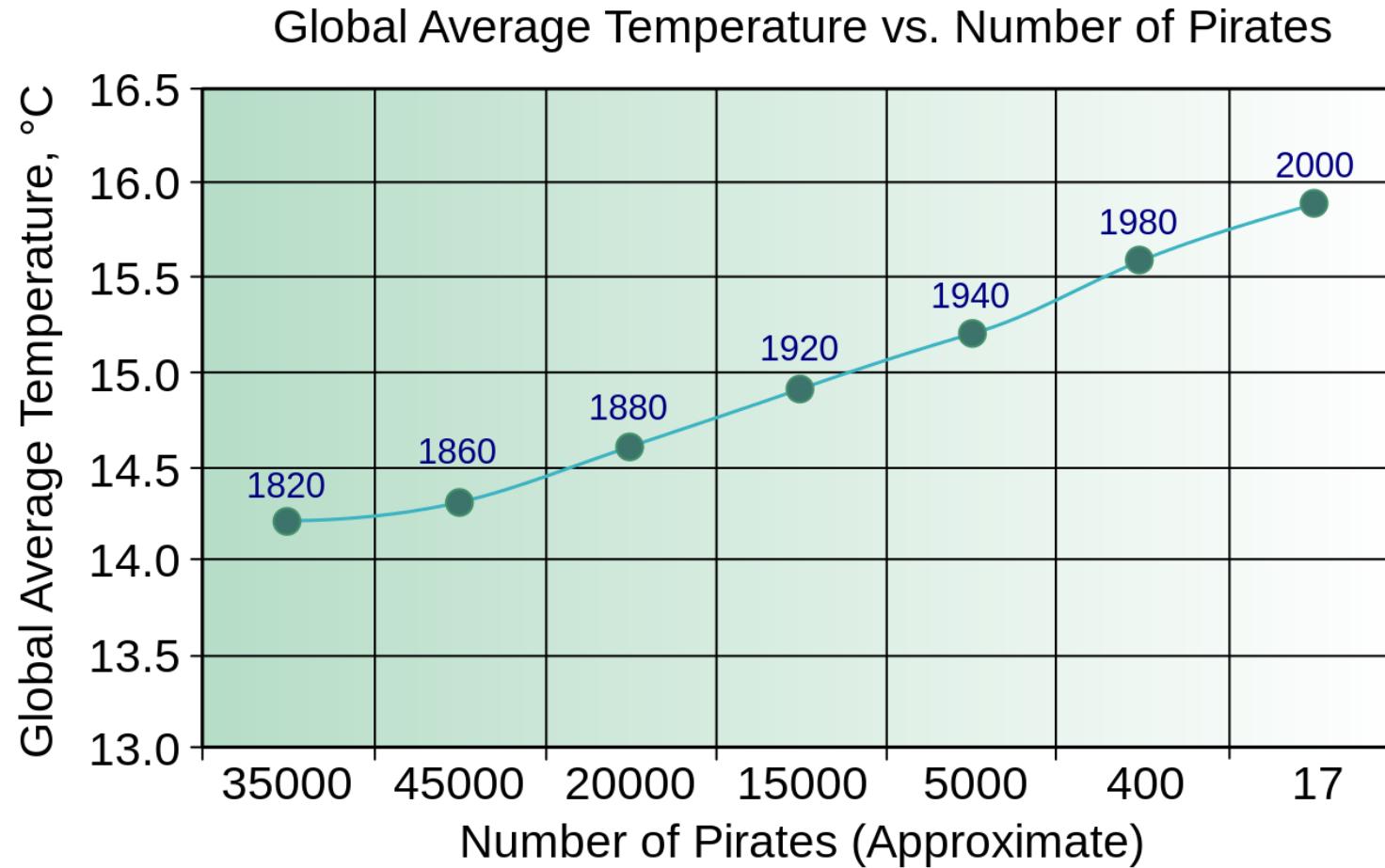
# Thống kê mô tả - Tương quan



# Thống kê mô tả - Tương quan



# Thống kê mô tả - Tương quan vs. nhân quả



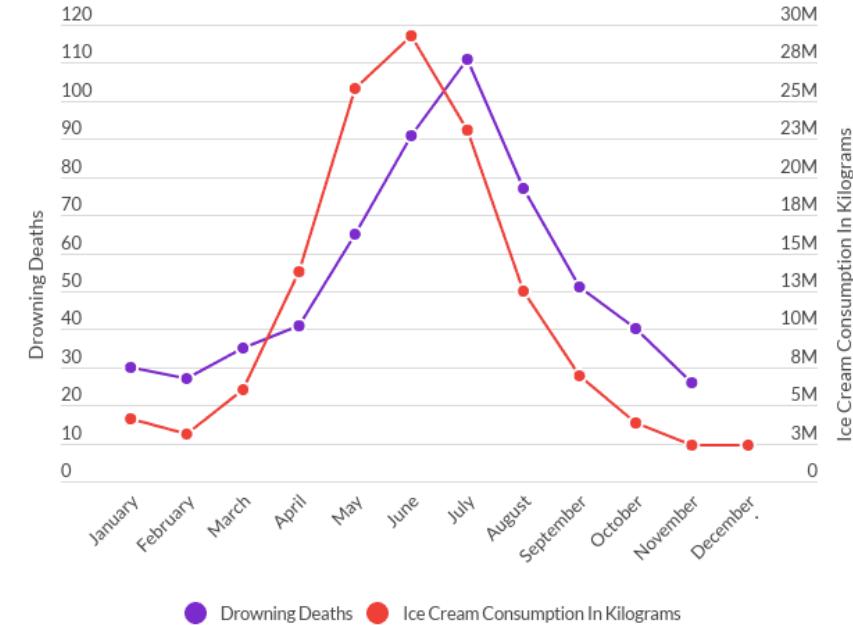
# Thống kê mô tả - Tương quan vs. nhân quả



# Thống kê mô tả - Tương quan vs. nhân quả

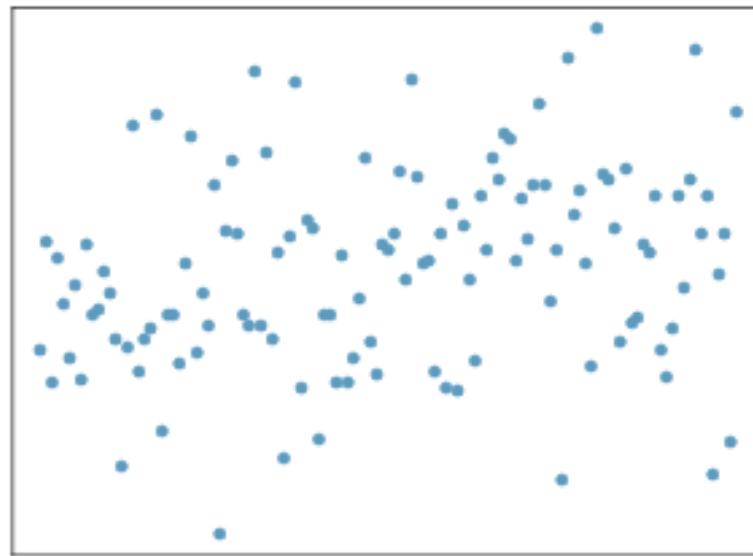


Drowning Deaths and Ice Cream Consumption by Month in Spain (2018)

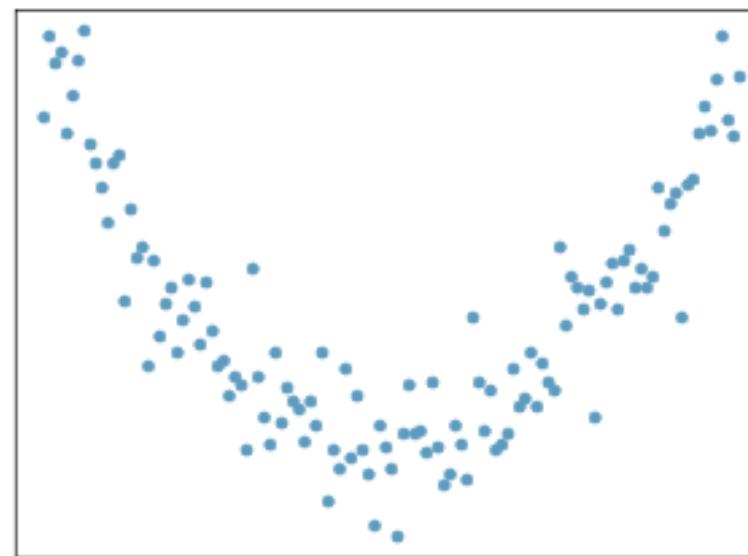


Statista (2020)

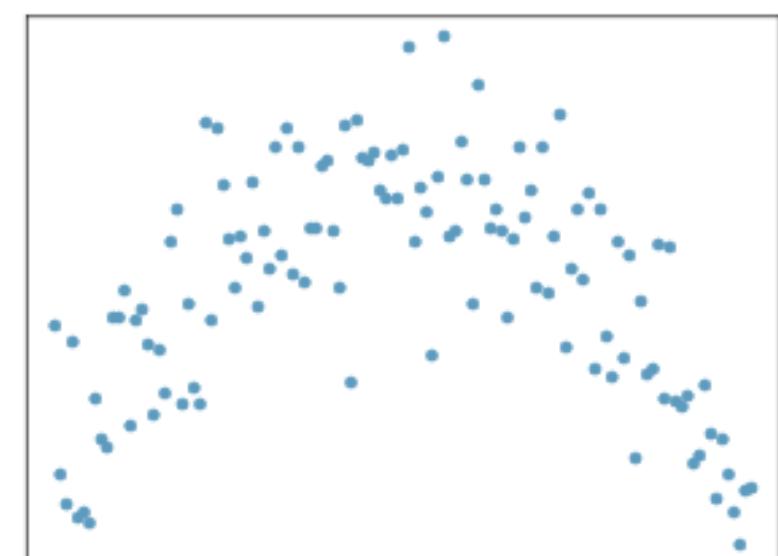
# Thống kê mô tả - Tương quan vs. nhân quả



Pearson's  $r \approx 0.8$



$r \approx -0.8$



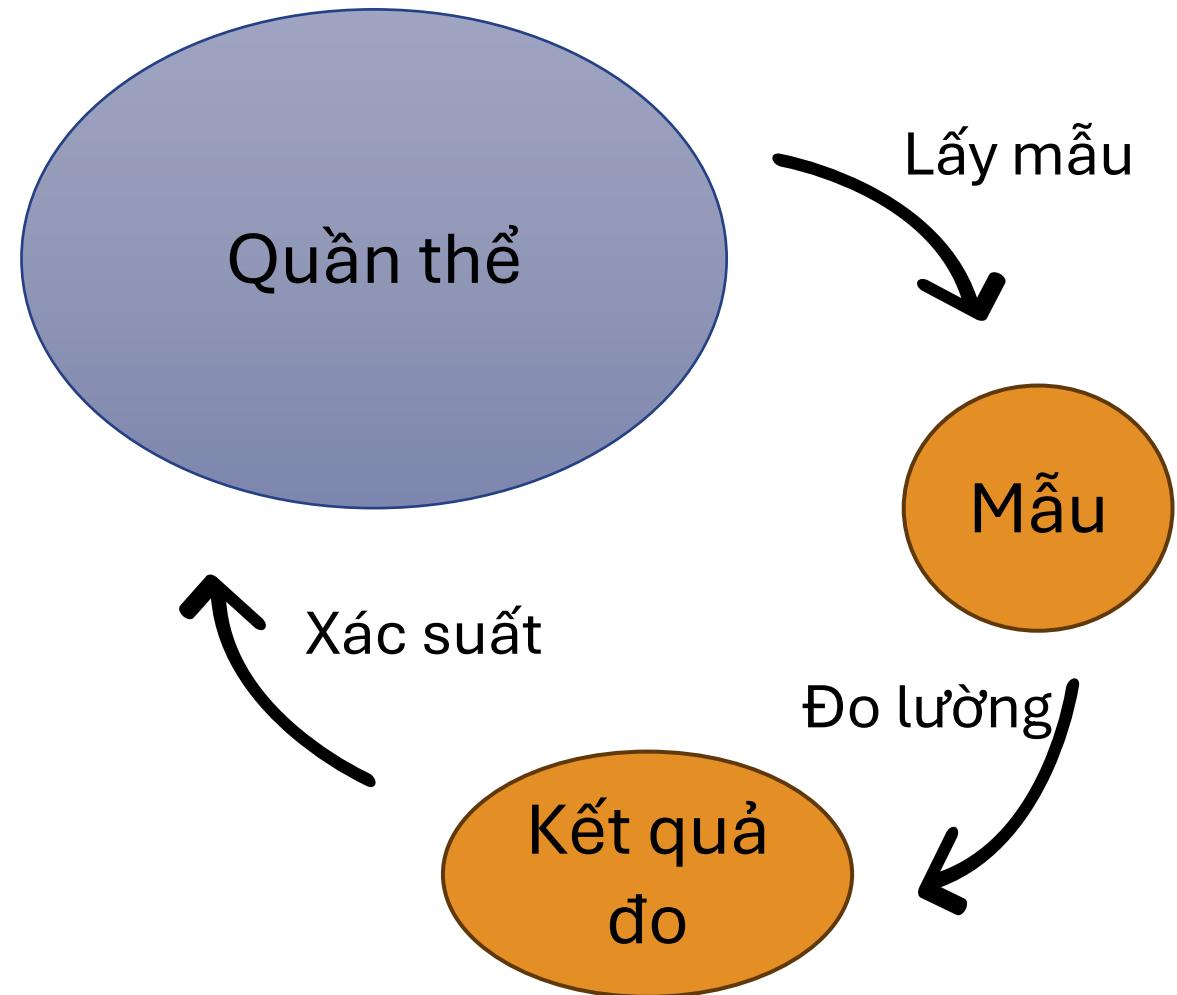
$r \approx 0$

# Các bước phân tích

- Tiền xử lý dữ liệu
- Tìm hiểu dữ liệu/Biểu diễn dữ liệu
- Thống kê mô tả
- Thống kê suy luận

# Thống kê suy luận

- Dùng thông tin về mẫu để suy luận về quần thể
- Kiểm định các giả thuyết (hypothesis) nghiên cứu
- Đưa ra kết luận về mối quan hệ giữa các biến



# Thống kê suy luận – Kiểm định thống kê

## Lượng khách

Trung bình      Độ lệch chuẩn



**Museum**

164

83



**Cultural  
Center**

147

70

# Thống kê suy luận – Kiểm định thống kê

## Giả thuyết không $H_0$

- Không có sự khác biệt lượng khách giữa hai nhóm

$$\mu_1 = \mu_2$$

## Giả thuyết thay thế $H_a$

- Trung bình lượng khách tham quan đến bảo tàng đông hơn

$$\mu_1 > \mu_2$$

\* Thông thường, giả thuyết thú vị với nhà nghiên cứu là giả thuyết thay thế

# Thống kê suy luận – Giá trị p

- Xác suất để thu được kết quả tương tự hoặc cực đoan hơn khi giả thiết rằng giả thuyết không là đúng
- Giá trị  $p < 0.05 \Rightarrow$  có ý nghĩa về mặt thống kê

	$H_0$ đúng	$H_0$ sai
Bắc bỏ $H_0$	Lỗi loại I	✓
Không bắc bỏ $H_0$	✓	Lỗi loại II

- Ý nghĩa về mặt thống kê vs. ý nghĩa thực tế

# Thống kê suy luận – Kiểm định thống kê

## Lượng khách

Trung bình      Độ lệch chuẩn



**Museum**

164

83

p-value = 0.1854



**Cultural  
Center**

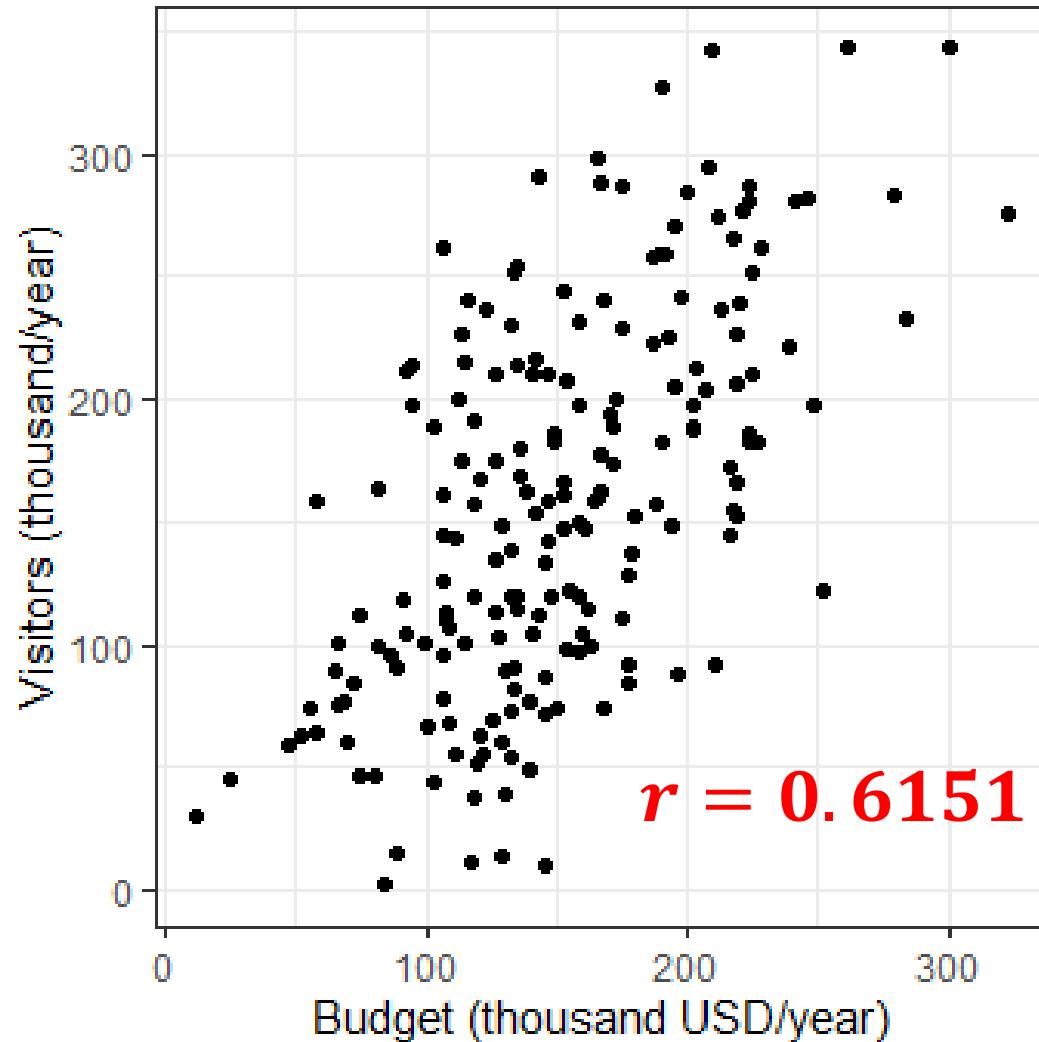
147

70

# Thống kê suy luận – Một số kỹ thuật

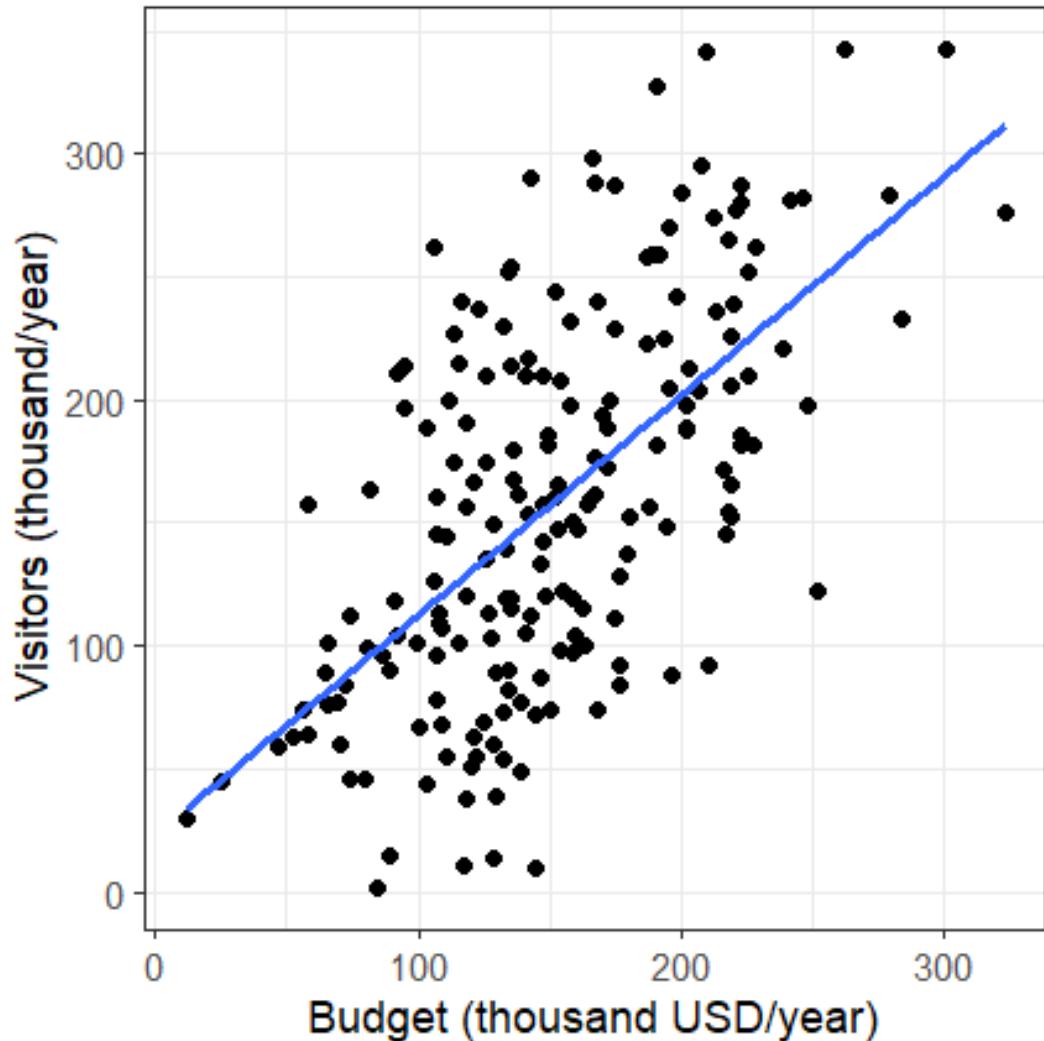
- Hồi quy tuyến tính đơn giản (Simple linear regression)
- Hồi quy tuyến tính bội (Multiple linear regression)
- Hồi quy tuyến tính suy rộng (Generalized linear regression)
- Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp (Multilevel model/Mixed effects model)
- Phân tích nhân tố (Factor analysis)
- Phân tích thành phần chính (PCA - Principal component analysis)
- Mô hình phương trình cấu trúc (SEM – Structural equation model)
- Thống kê không gian (Spatial statistics)
- ...

# Thống kê suy luận – Hồi quy tuyến tính



$$\begin{aligned}x &\sim y \\y &\sim x\end{aligned}$$

# Thống kê suy luận – Hồi quy tuyến tính



Lượng  
khách

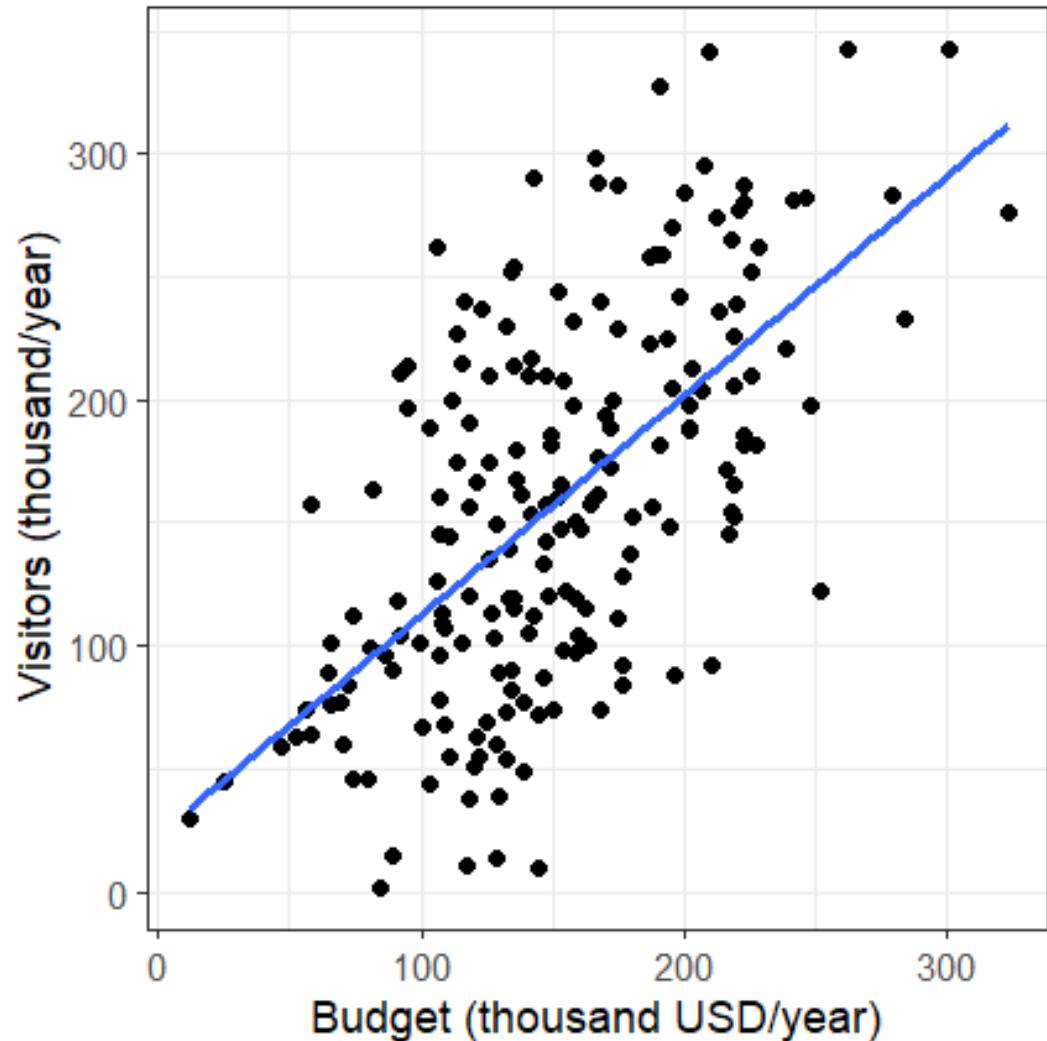


Kinh phí

$$y = \alpha + \beta * x$$

- Dùng  $x$  để **giải thích** và **dự đoán** sự biến thiên trong  $y$
- $y$ : Biến phụ thuộc, biến kết quả (Dependent/Outcome variable)
- $x$ : Biến độc lập, biến giải thích (Independent/Explanatory variable)

# Thống kê suy luận – Hồi quy tuyến tính



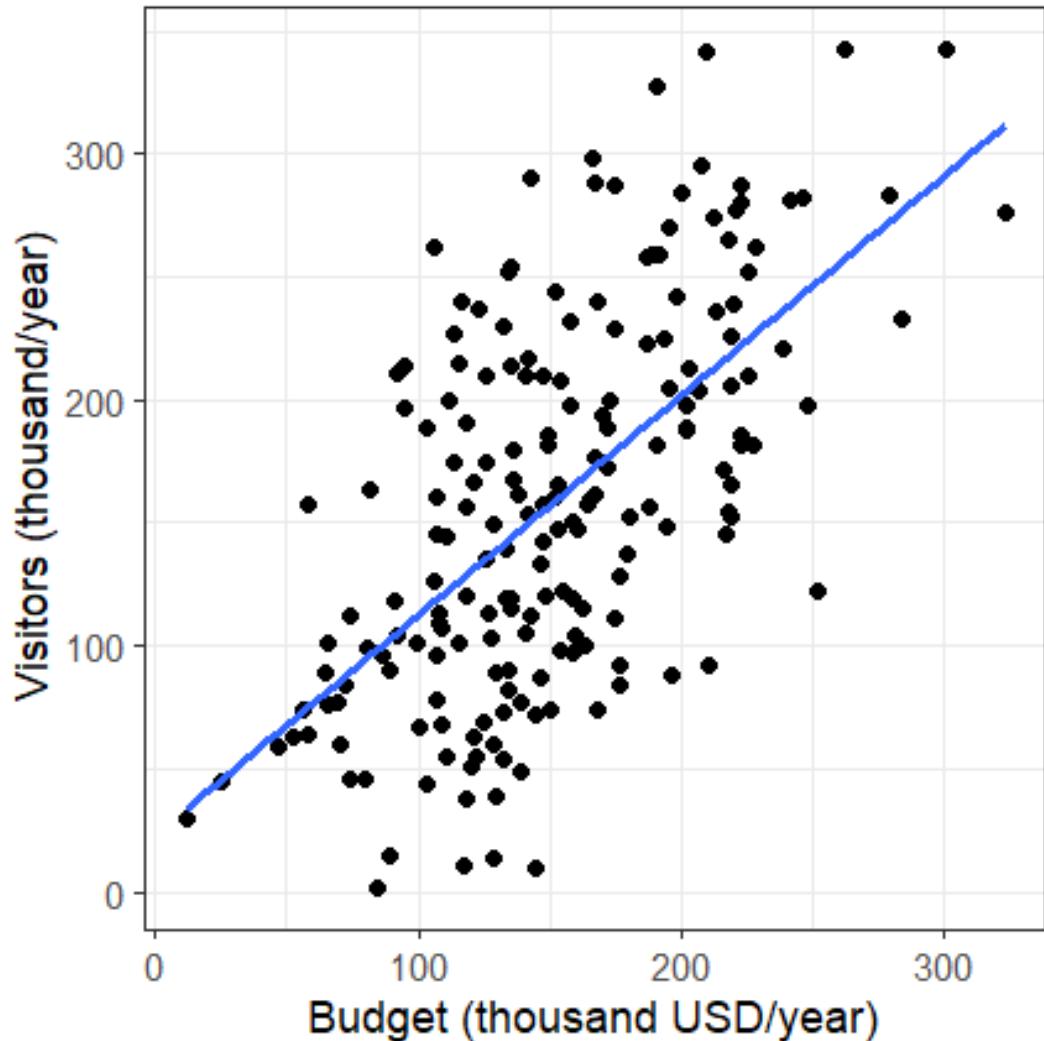
Lượng  
khách



Kinh phí

$$y = 22.74 + 0.89 * x$$

# Thống kê suy luận – Hồi quy tuyến tính



Lượng  
khách

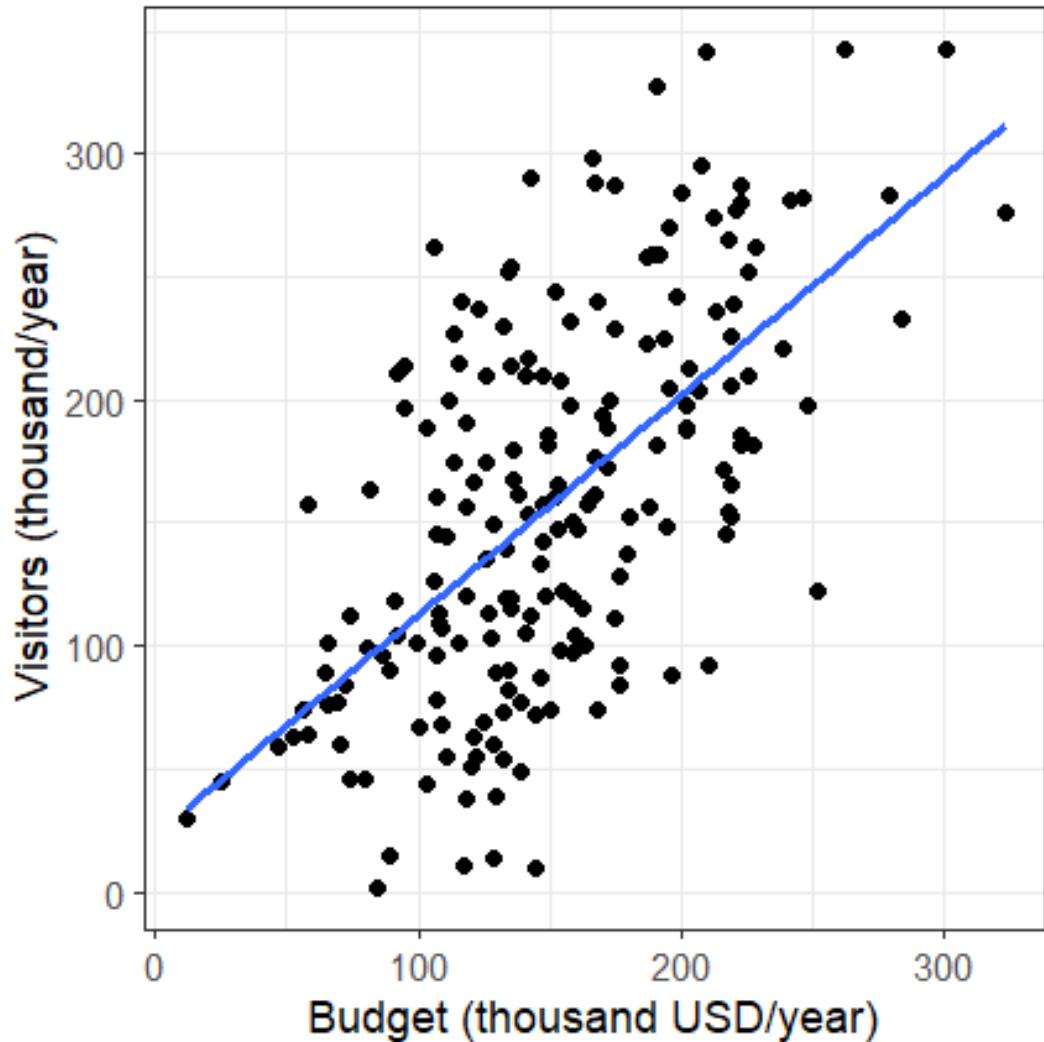
Kinh phí



$$y = 22.74 + 0.89 * x$$

- $\beta = 0.89$  – mức độ thay đổi của giá trị trung bình của y khi x thay đổi 1 đơn vị
- Khi kinh phí thay đổi 1 đơn vị (1k USD), lượng khách trung bình thay đổi 1 lượng bằng 890 người

# Thống kê suy luận – Hồi quy tuyến tính



Lượng  
khách

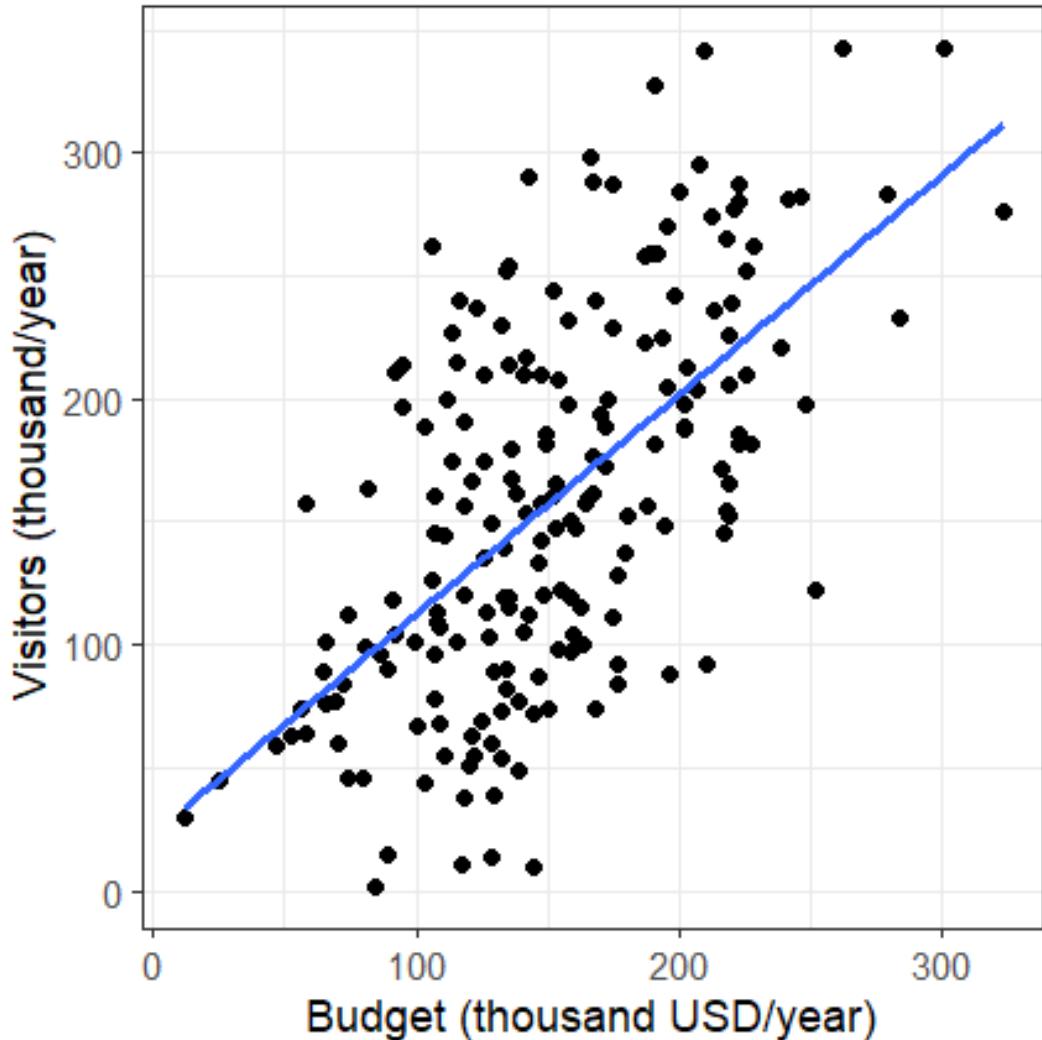


Kinh phí

$$y = 22.74 + 0.89 * x$$

- $\beta > 0$  – kinh phí tăng, lượng khách tăng
- $\beta < 0$  – kinh phí tăng, lượng khách giảm
- $\beta = 0$  – kinh phí không ảnh hưởng đến lượng khách

# Thống kê suy luận – Hồi quy tuyến tính



Lượng  
khách

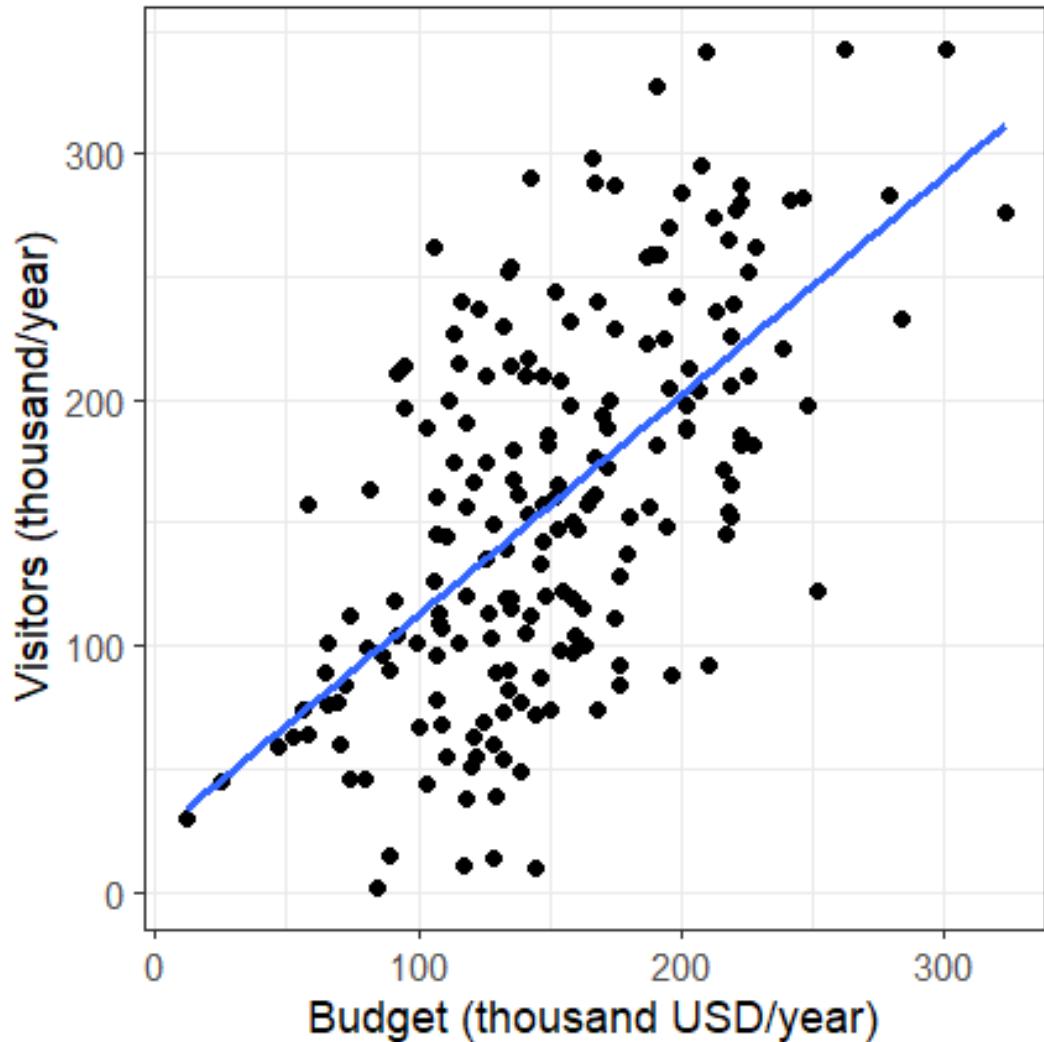
Kinh phí



$$y = 22.74 + 0.89 * x$$

- $H_a: \beta \neq 0$
- $H_0: \beta = 0$
- p-value < 0.05 ( $< 2.2 \times 10^{-16}$ )

# Thống kê suy luận – Hồi quy tuyến tính



Lượng  
khách

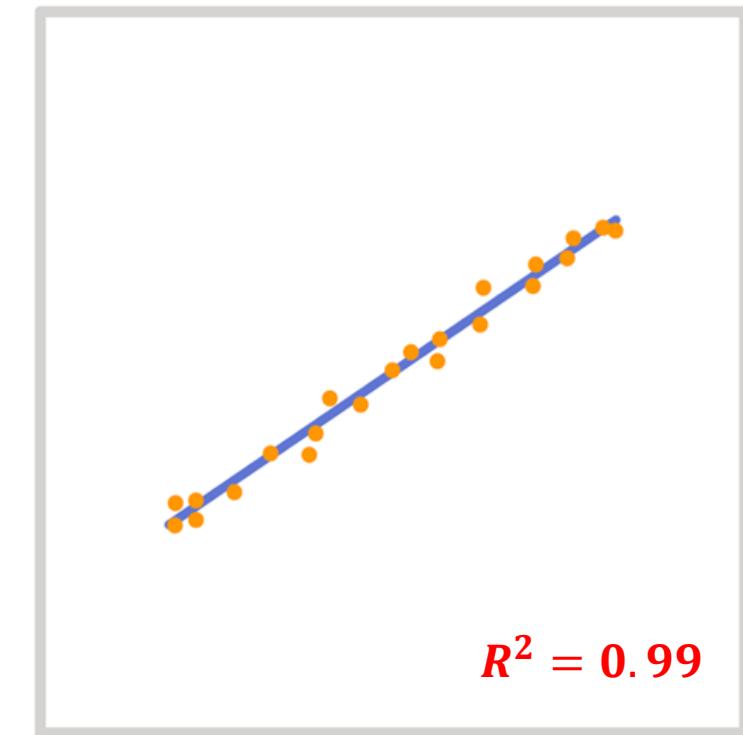
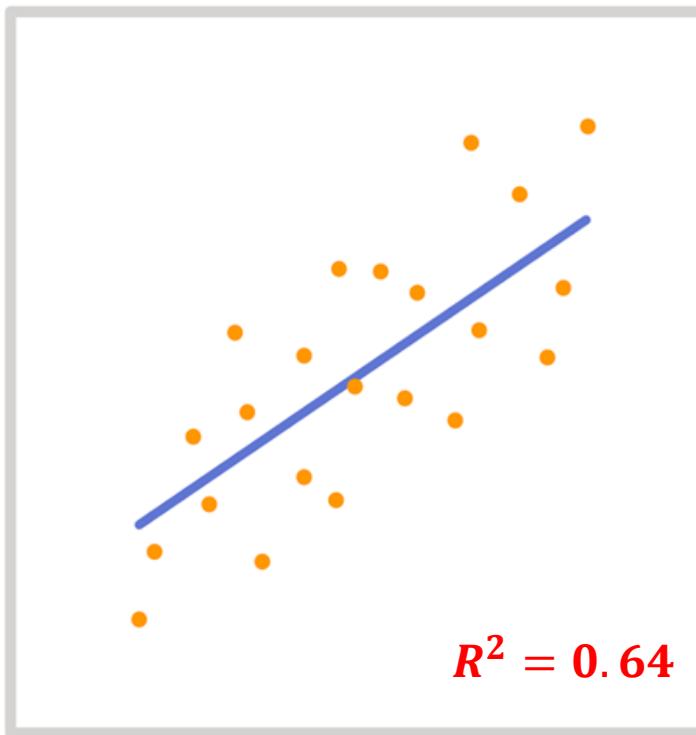
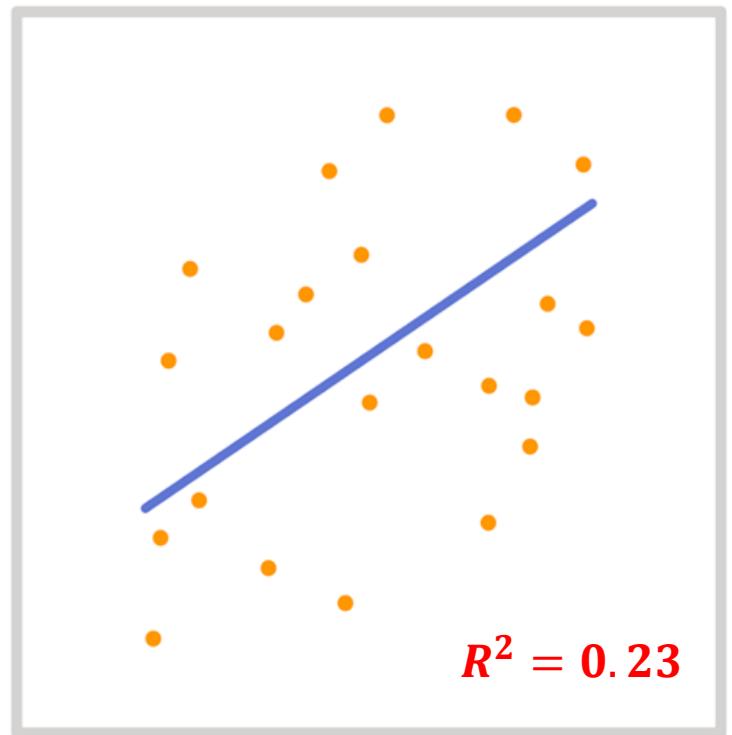
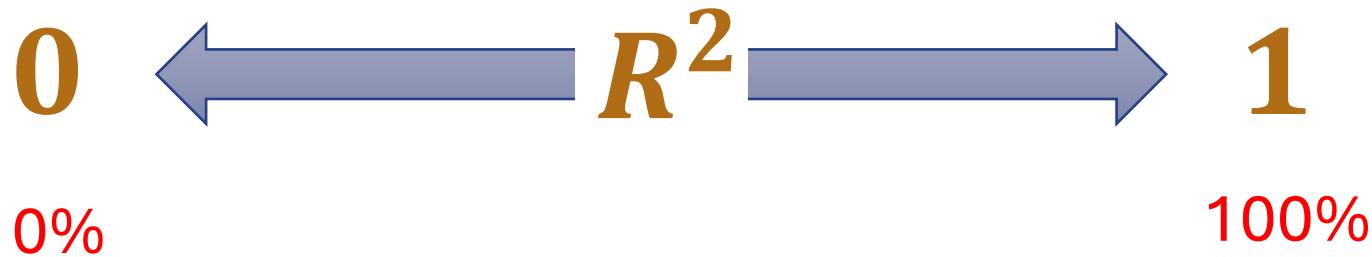


Kinh phí

$$y = 22.74 + 0.89 * x$$

- $H_a: \beta \neq 0$
- $H_0: \beta = 0$
- p-value < 0.05 ( $< 2.2 \times 10^{-16}$ )
- $R^2 = 0.3784$

# Thống kê suy luận – Hồi quy tuyến tính



# Thống kê suy luận – Hồi quy tuyến tính

$y$ : là biến định lượng  
 $x$ : là biến định lượng

# Thống kê suy luận – Hồi quy tuyến tính

$y$ : là biến định lượng  
 $x$ : là biến **định tính?**

Lượng  
khách

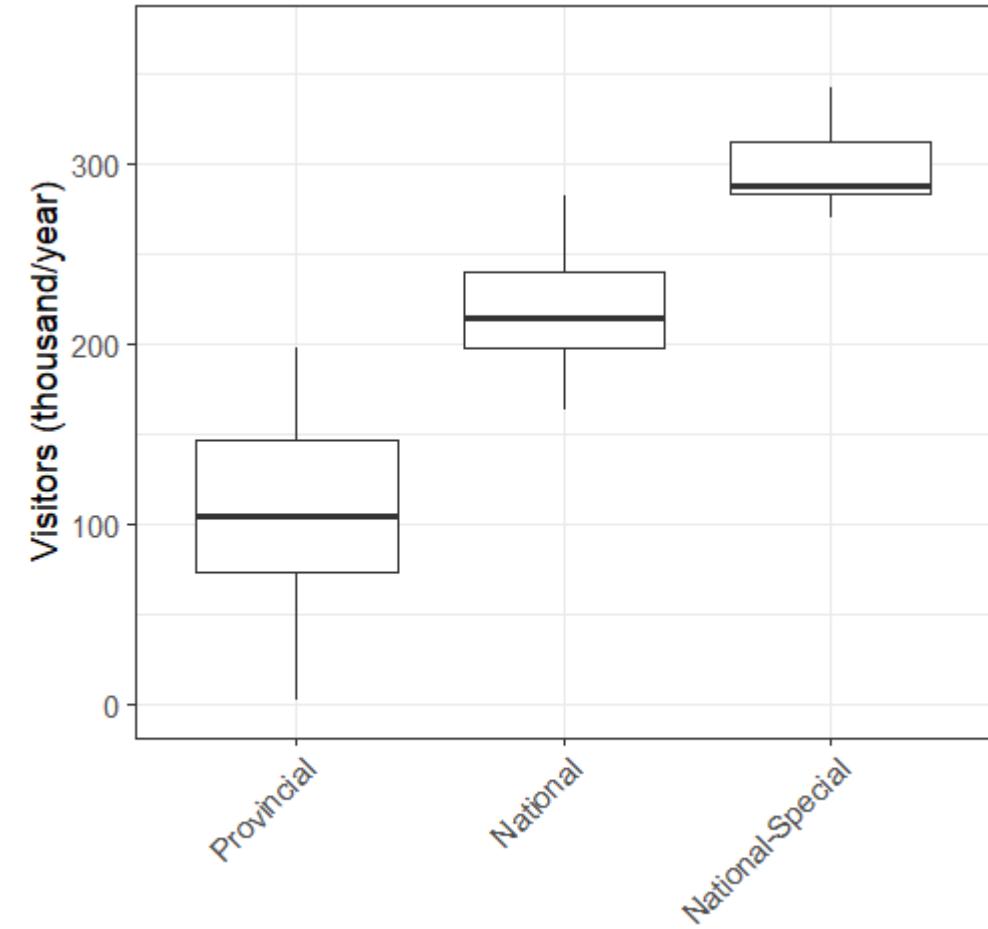
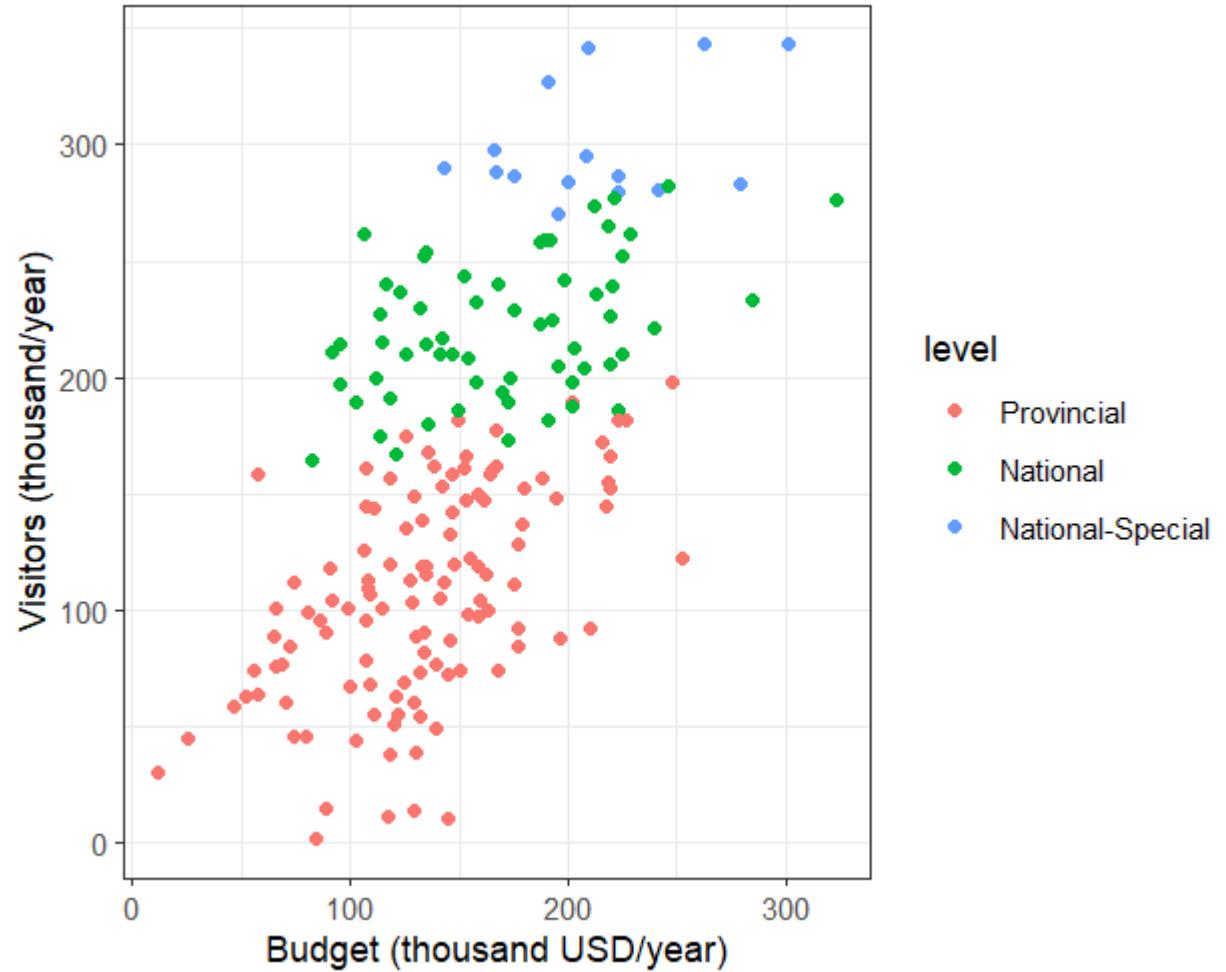


Cấp độ  
di sản

# Thống kê suy luận – Hồi quy tuyến tính

site	visitors	budget	events	level	category
1	298	166	5	National-Special	Museum
2	78	107	6	Provincial	Historical Monument
3	252	134	7	National	Archaeological Site
4	262	228	5	National	Museum
5	92	177	0	Provincial	Archaeological Site
6	236	213	3	National	Museum
7	120	148	6	Provincial	Museum
8	101	99	3	Provincial	Museum
9	55	111	6	Provincial	Museum
10	100	163	4	Provincial	Cultural Center
11	115	162	8	Provincial	Historical Monument
12	280	223	8	National-Special	Museum

# Thống kê suy luận – Hồi quy tuyến tính



# Thống kê suy luận – Hồi quy tuyến tính

$$y = \alpha + \beta_1 * x_1 + \beta_2 * x_2$$

$$x_1 \begin{cases} = 1 & \text{National} \\ = 0 & \text{Others} \end{cases} \quad x_2 \begin{cases} = 1 & \text{National-Special} \\ = 0 & \text{Others} \end{cases}$$

# Thống kê suy luận – Hồi quy tuyến tính

$$y = \alpha + \beta_1 * x_1 + \beta_2 * x_2$$

$$x_1 \begin{cases} = 1 & \text{National} \\ = 0 & \text{Others} \end{cases} \quad x_2 \begin{cases} = 1 & \text{National-Special} \\ = 0 & \text{Others} \end{cases}$$

Provincial	$x_1 = 0$	&	$x_2 = 0$	$\rightarrow$	$y = \alpha$	3 giá trị trung bình
National	$x_1 = 1$	&	$x_2 = 0$	$\rightarrow$	$y = \alpha + \beta_1$	
National-Special	$x_1 = 0$	&	$x_2 = 1$	$\rightarrow$	$y = \alpha + \beta_2$	

# Thống kê suy luận – Hồi quy tuyến tính

$$y = \alpha + \beta_1 * x_1 + \beta_2 * x_2$$

$$x_1 \begin{cases} = 1 & \text{National} \\ = 0 & \text{Others} \end{cases} \quad x_2 \begin{cases} = 1 & \text{National-Special} \\ = 0 & \text{Others} \end{cases}$$

National-Special vs. Provincial:  $\beta_1 = 114$

$$H_a: \beta_1 \neq 0 \quad \& \quad H_0: \beta_1 = 0$$

National-Special vs. Provincial:  $\beta_2 = 193$

$$H_a: \beta_2 \neq 0 \quad \& \quad H_0: \beta_2 = 0$$

# Thống kê suy luận – Hồi quy tuyến tính

$$y = \alpha + \beta_1 * x_1 + \beta_2 * x_2$$

$$x_1 \begin{cases} = 1 & \text{National} \\ = 0 & \text{Others} \end{cases} \quad x_2 \begin{cases} = 1 & \text{National-Special} \\ = 0 & \text{Others} \end{cases}$$

National-Special vs. Provincial:  $\beta_1 = 114$

$$H_a: \beta_1 \neq 0 \quad \& \quad H_0: \beta_1 = 0$$

*p-value* < 0.05

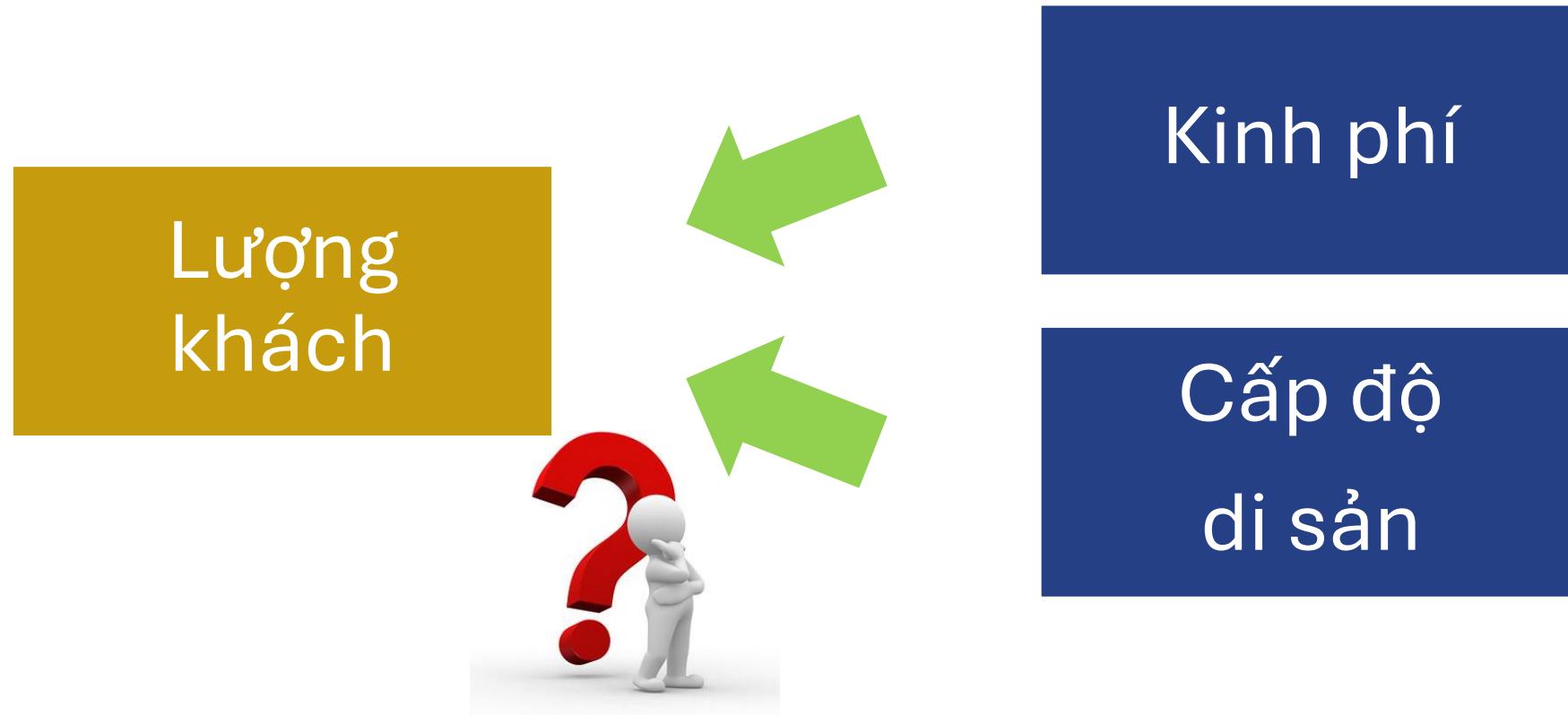
$$R^2 = 0.7401$$

National-Special vs. Provincial:  $\beta_2 = 193$

$$H_a: \beta_2 \neq 0 \quad \& \quad H_0: \beta_2 = 0$$

*p-value* < 0.05

# Thống kê suy luận – Hồi quy tuyến tính



# Thống kê suy luận – Hồi quy tuyến tính

$$y = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3$$

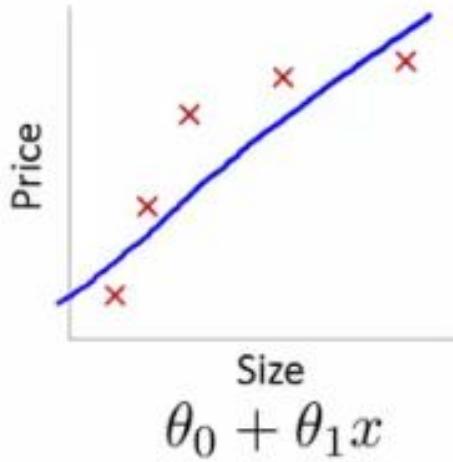
Kinh phí	Cấp độ di sản	Cả 2
$R^2$	0.3784	0.7401
$R^2_{adj}$	0.3751	0.7374

# Hồi quy tuyến tính bội

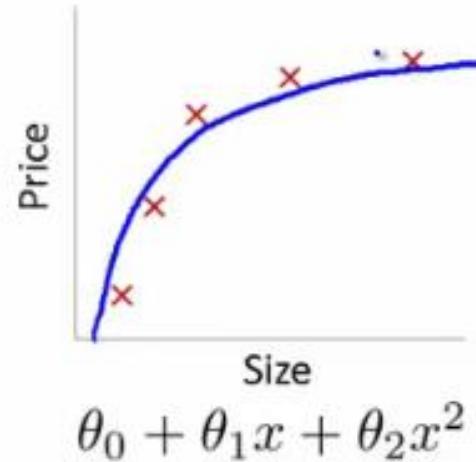
Phương trình càng phức tạp càng tốt???????

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3^2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \dots$$

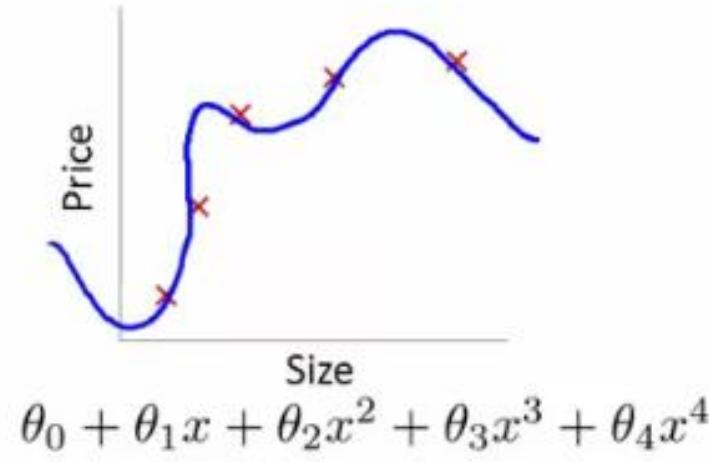
# Quá khớp hoặc thiếu khớp với số liệu



High bias  
(underfit)



"Just right"

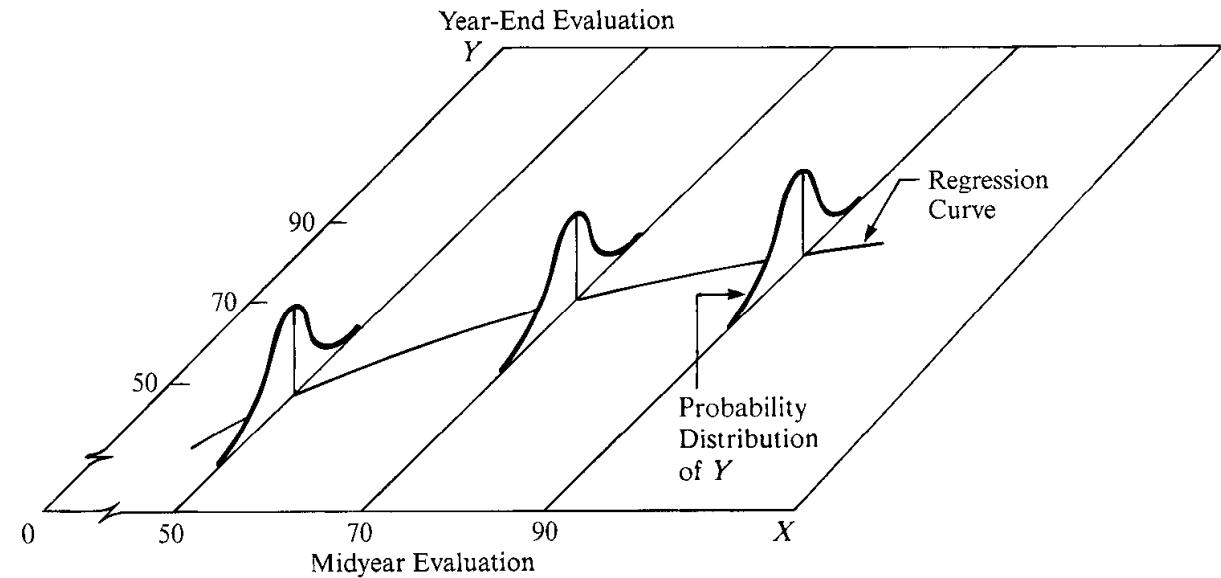


High variance  
(overfit)

- $R^2_{adj}$
- AIC
- BIC
- Predictive power

# Hồi quy tuyến tính – Các giả định

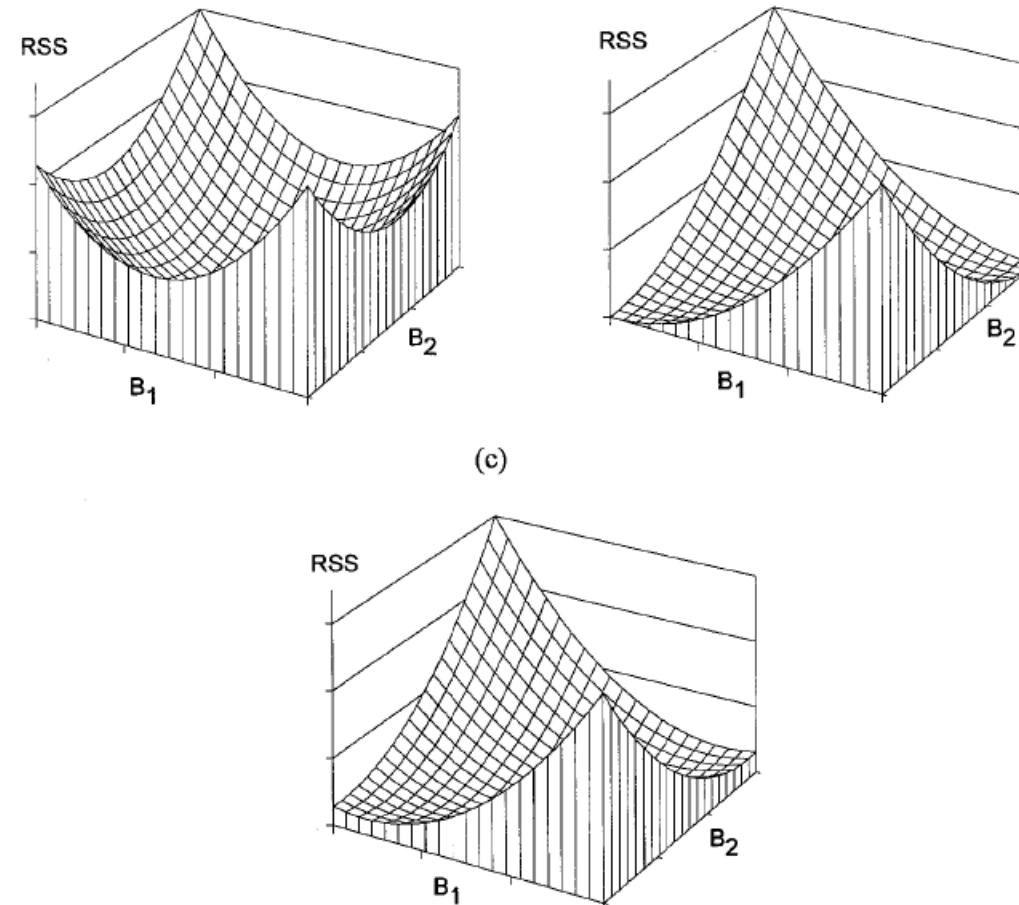
- $Y$  là biến liên tục
- Mối liên hệ tuyến tính giữa  $Y$  với các tham số khảo sát
- Các giá trị  $Y$  độc lập với nhau
- Các sai số ngẫu nhiên tuân theo phân phối chuẩn có cùng phương sai và trung bình = 0



# Tương quan giữa các biến độc lập

## Multicollinearity

Variance Inflation Factor  
(Yếu tố lạm phát phương sai)



# Hồi quy tuyến tính suy rộng (Generalized linear regression)

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \cdots + \beta_p X_{ip}$$

Hồi quy tuyến tính bội

$Y$ : liên tục/định lượng

$X$ : liên tục/định lượng hoặc rời rạc

Khi  $Y$  là biến rời rạc?

# Hồi quy tuyến tính suy rộng (Generalized linear regression)

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2}^2 + \beta_3 X_{i3} + \dots + \beta_p X_{ip}$$

Hồi quy tuyến tính bội

$Y$ : liên tục/định lượng

$X$ : liên tục/định lượng hoặc rời rạc

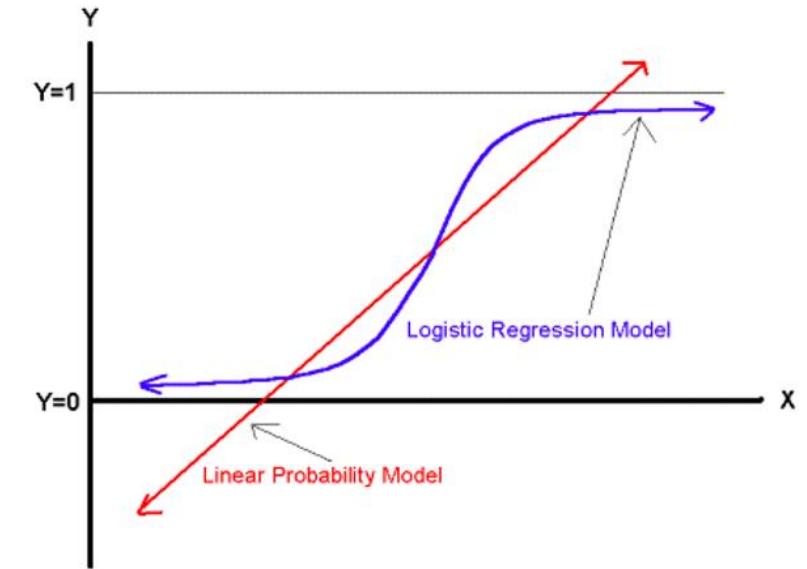
Khi  $Y$  là biến rời rạc?

Nhị phân (Yes/No): Hồi quy Logistic (Logistic regression)

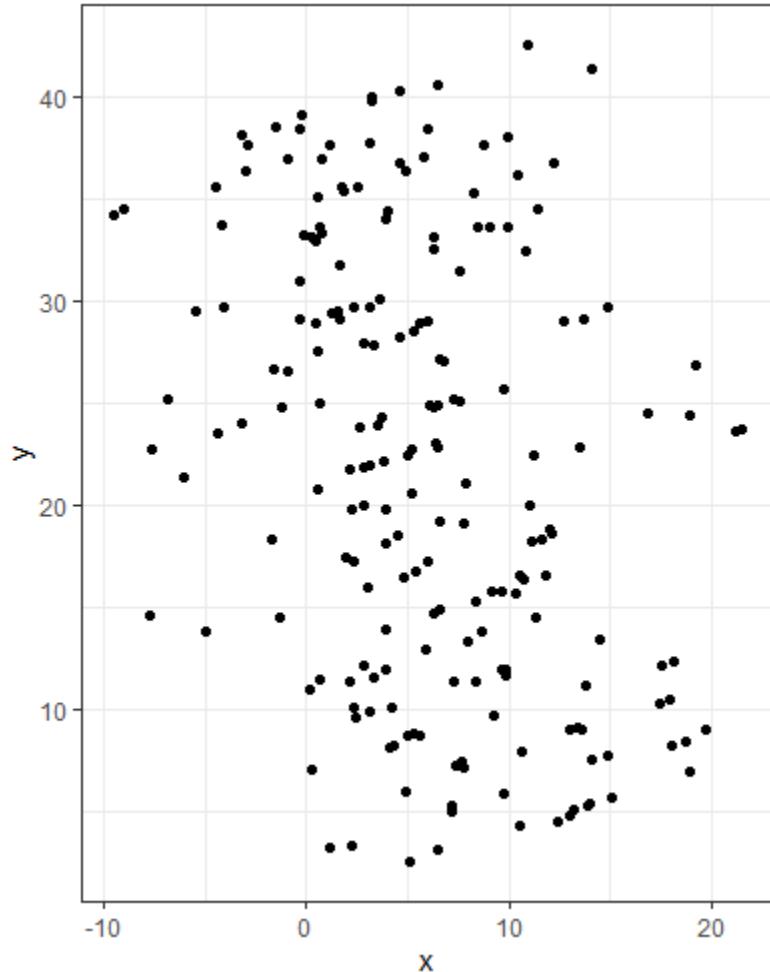
Định danh: Multinomial logistic regression

Thứ bậc: Cumulative logistic regression

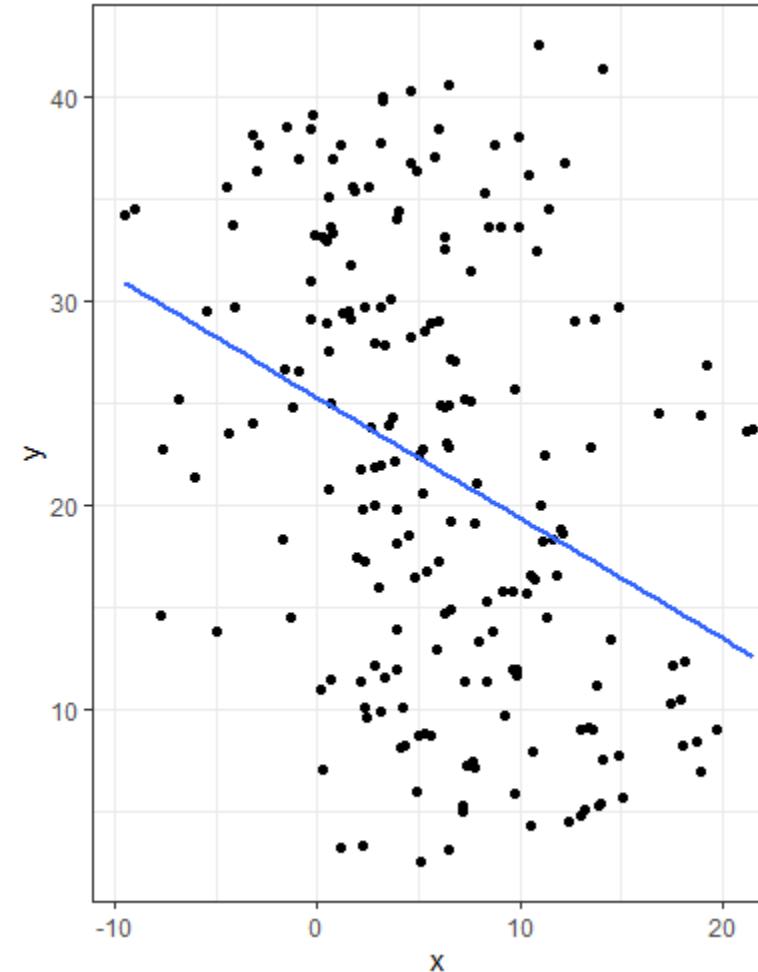
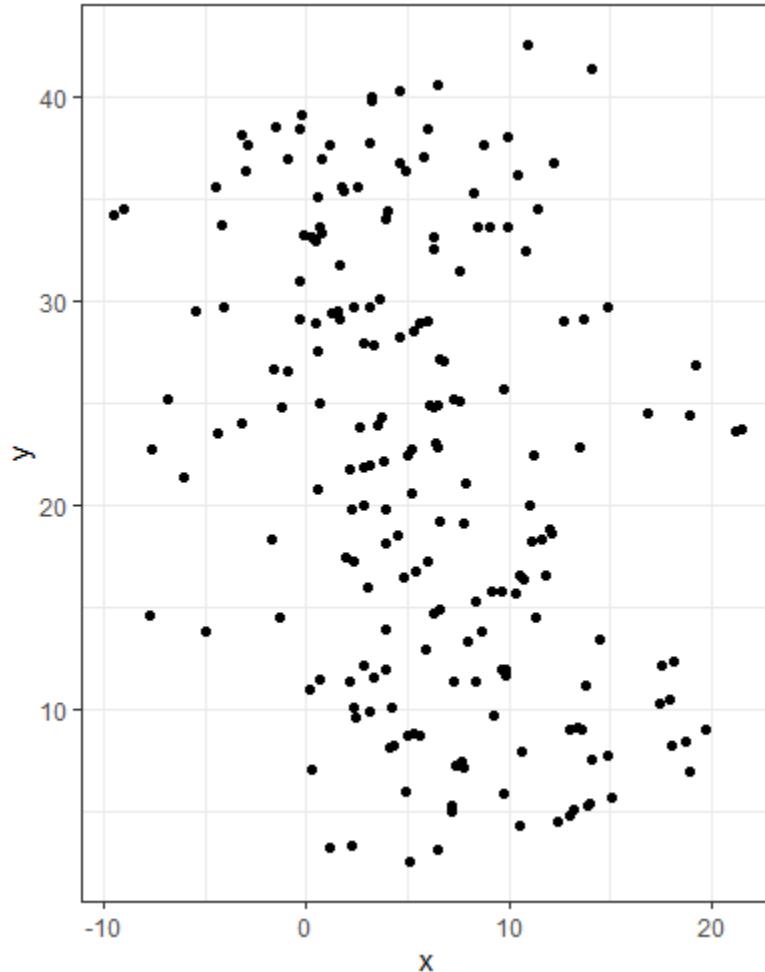
Biến đếm: Poisson regression



# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp Multilevel model/Mixed effect model

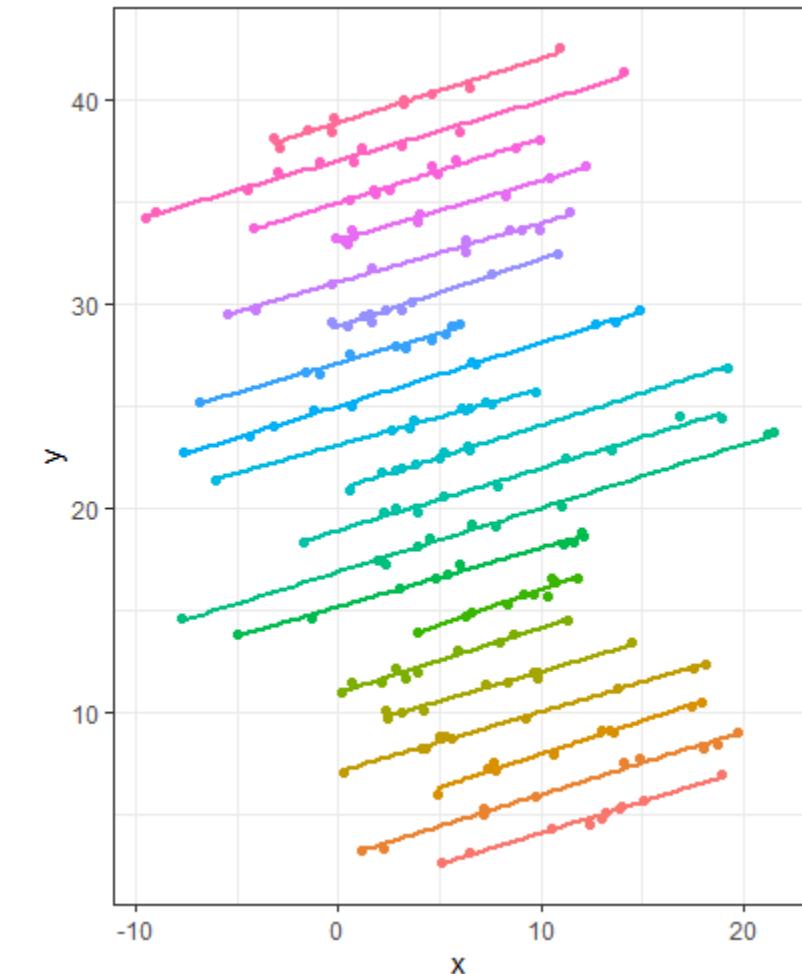
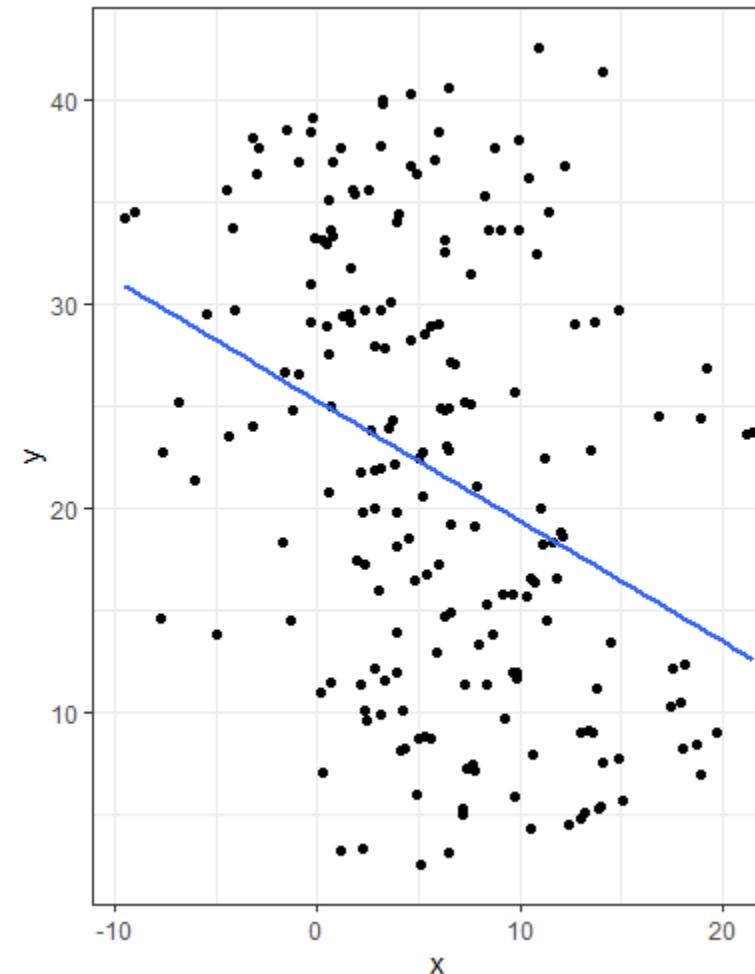
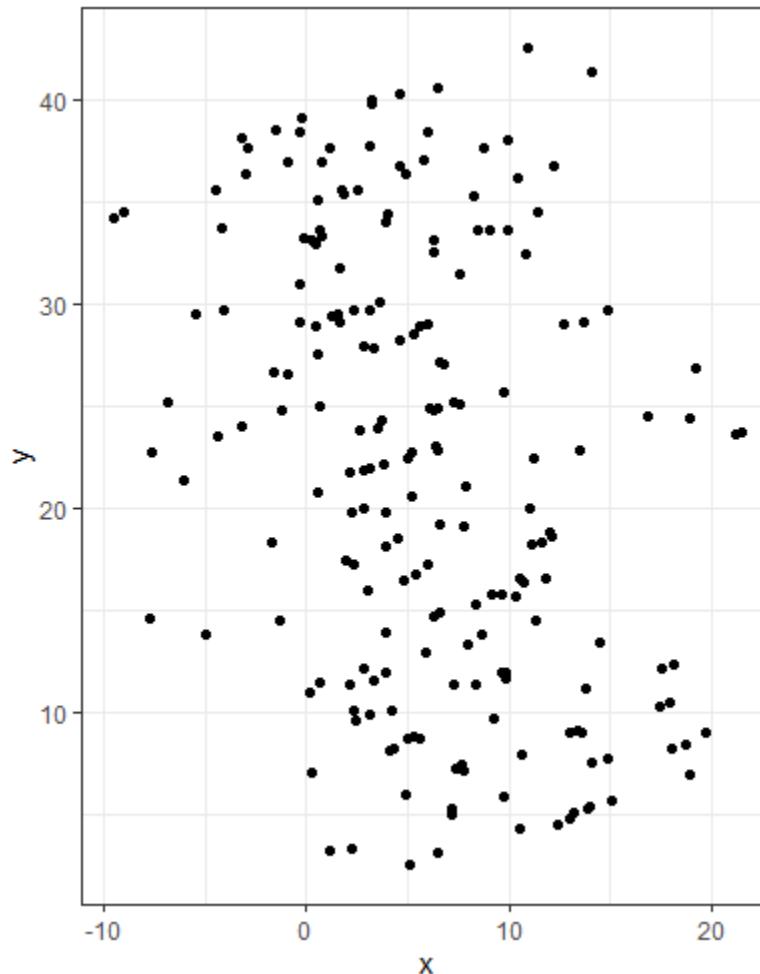


# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp Multilevel model/Mixed effect model

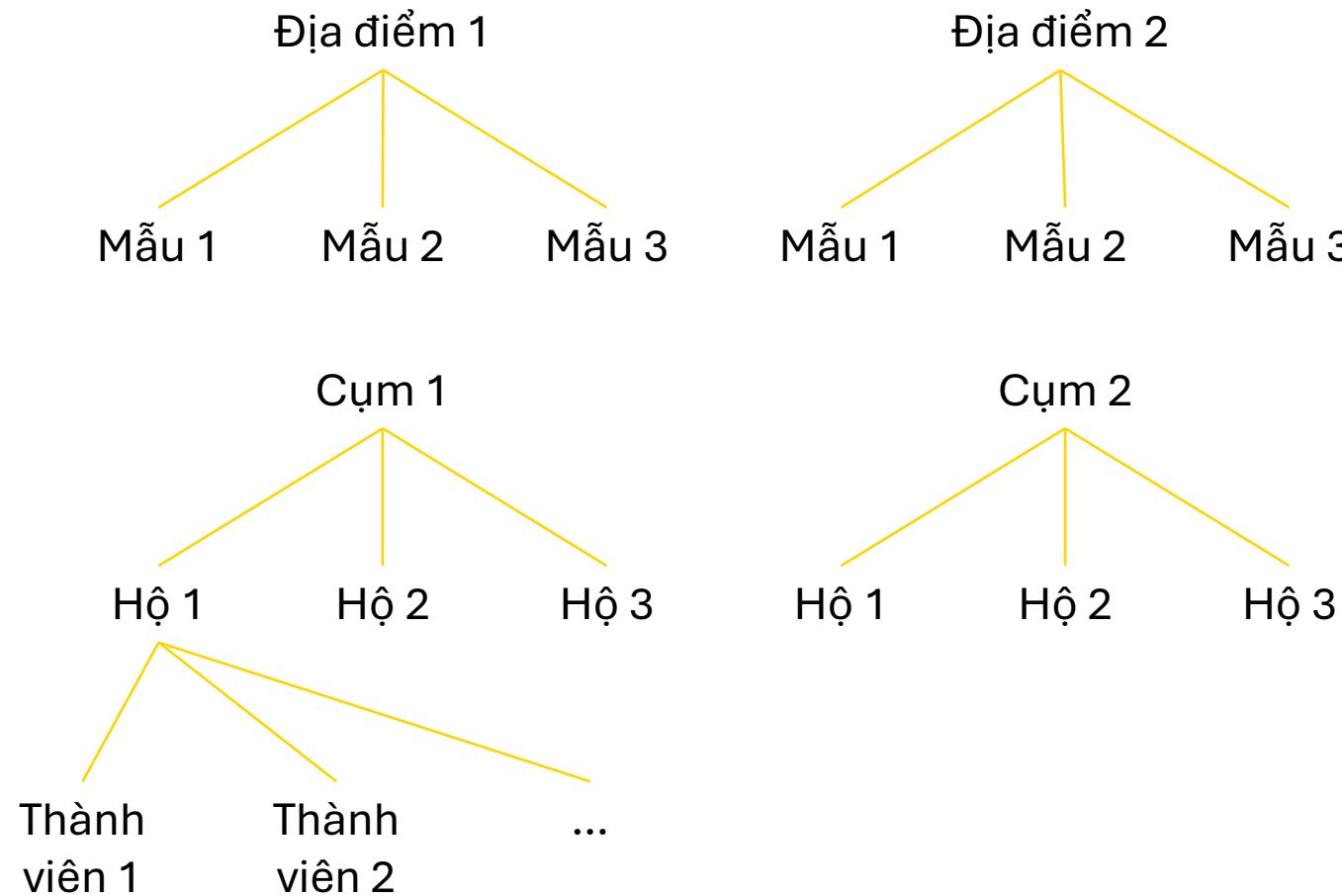


# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

## Multilevel model/Mixed effect model



# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp Multilevel model/Mixed effect model



# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

## Multilevel model/Mixed effect model

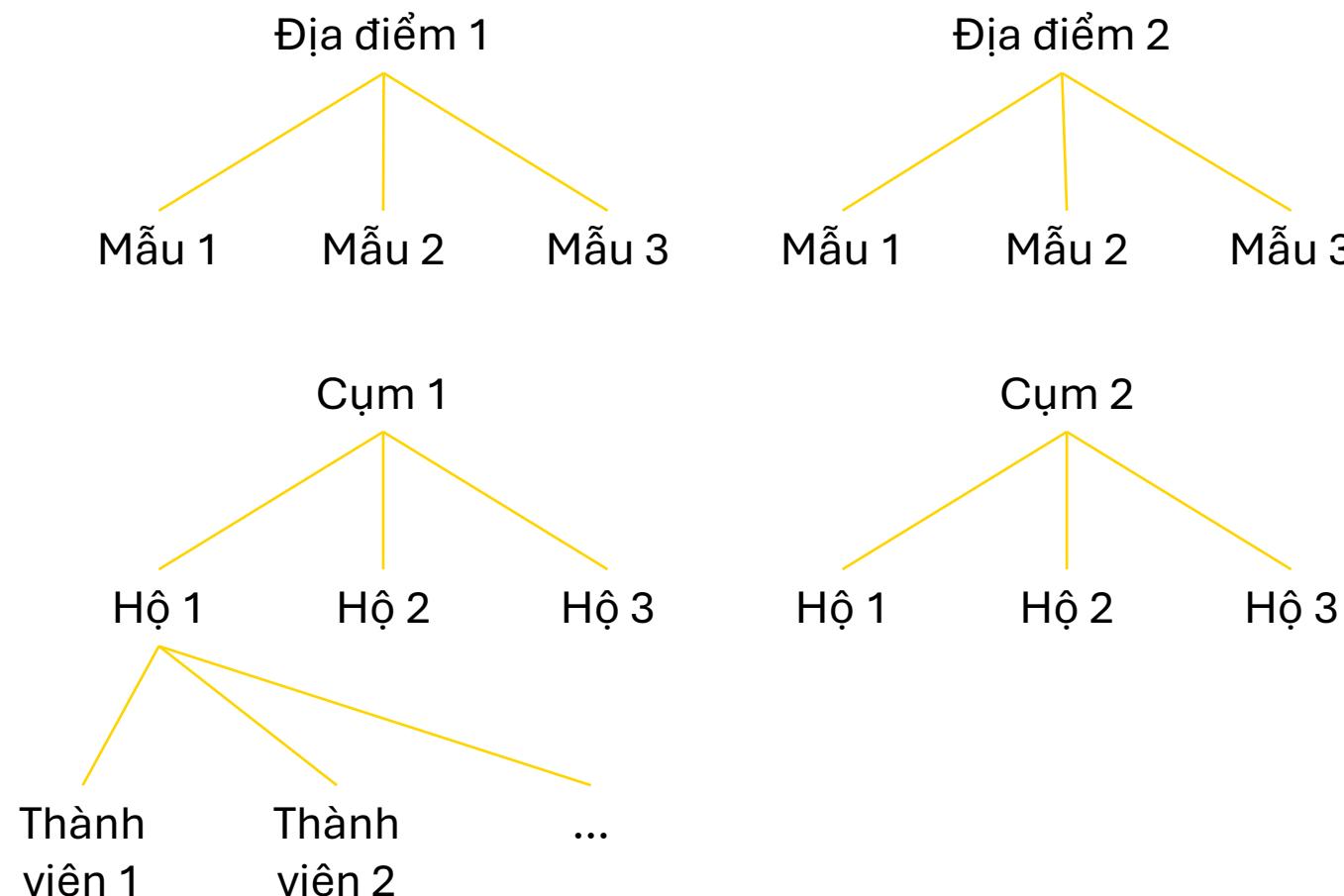
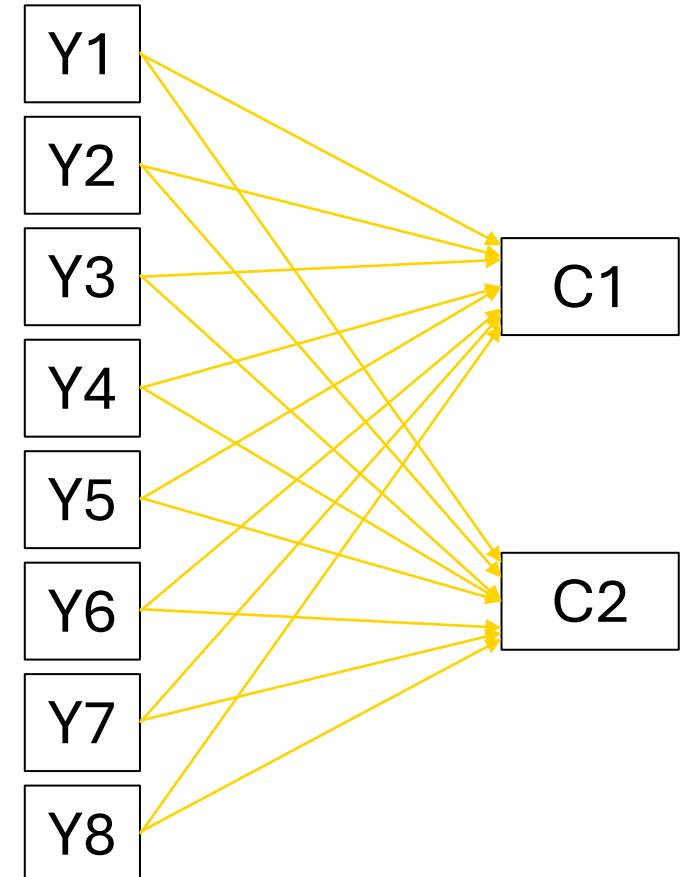


Image by [Chelsea Parlett-Pelleriti](#)

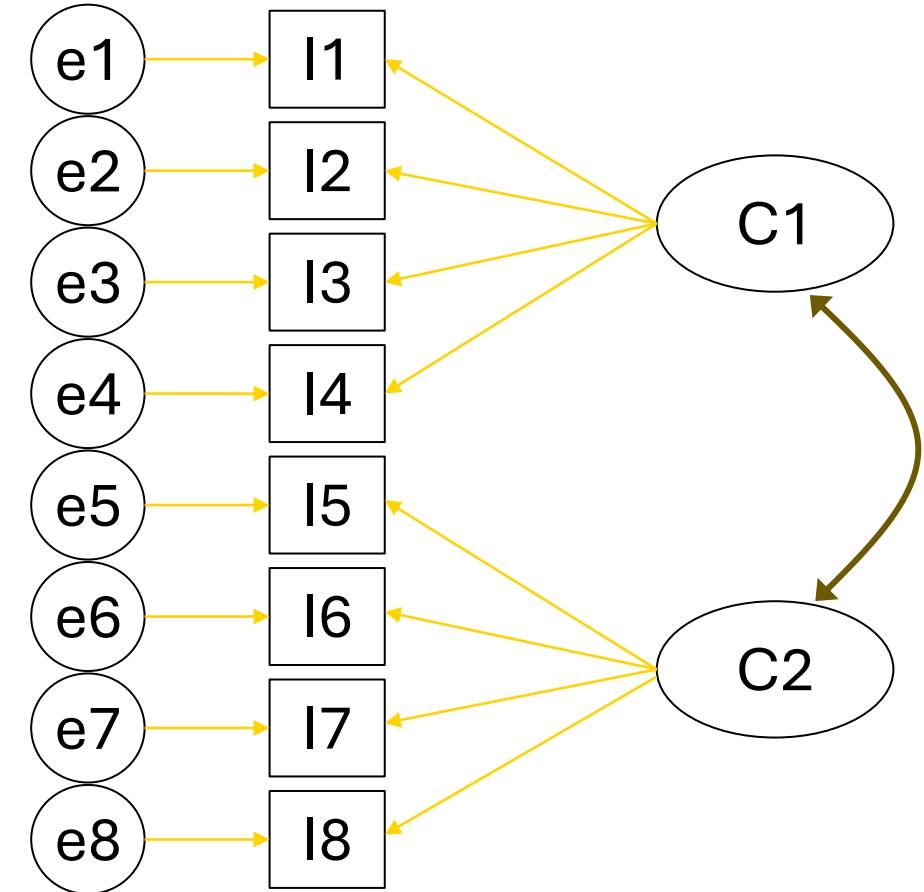
# Phân tích thành phần chính (PCA - Principal component analysis)

- Phương pháp giảm chiều dữ liệu
- Không phân biệt biến độc lập hay phụ thuộc
- Phương pháp khảo sát (không phải phương pháp suy luận)
- Bước trước cho hồi quy tuyến tính để giảm đa cộng tuyến (multicollinearity)



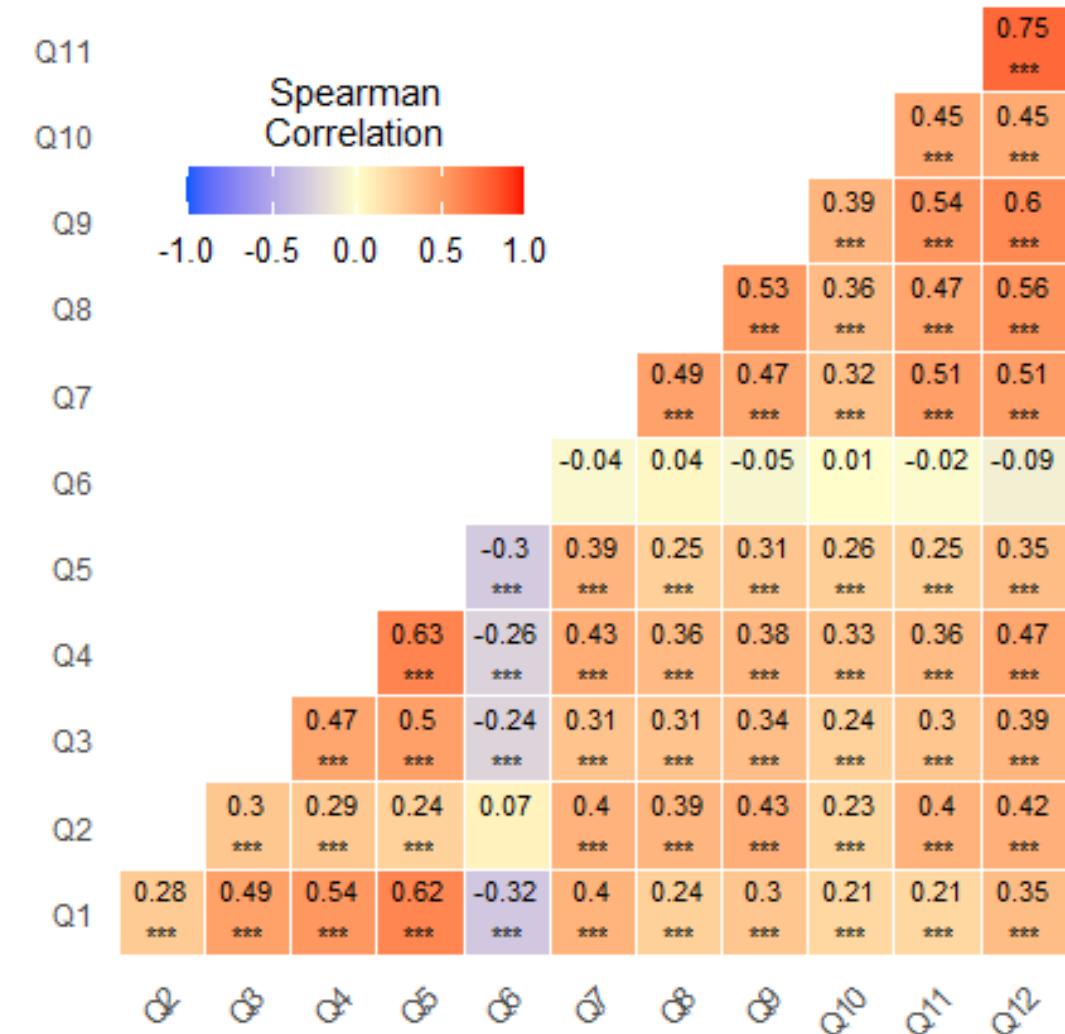
# Phân tích nhân tố (Factor analysis)

- Phân tích nhân tố khám phá/khẳng định (Exploratory/Confirmatory Factor Analysis)
- CFA: Thường áp dụng cho dữ liệu bảng hỏi
- CFA: Đo phạm trù tiềm ẩn (Latent construct)



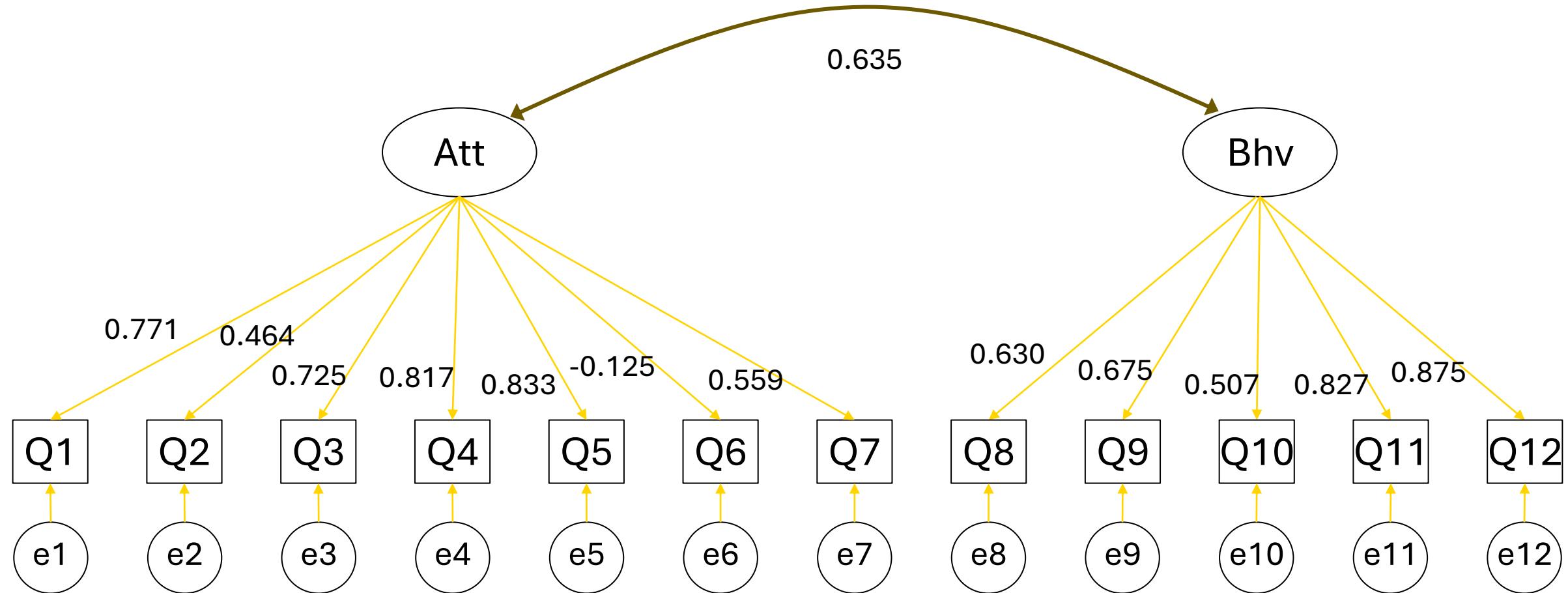
# Phân tích nhân tố khẳng định (CFA)

Attitude	
Column I	Question 1 In my opinion, it is important to protect the environment.
Column J	Question 2 I actively practice environmental sustainability at home (e.g., energy conservation, recycling).
Column K	Question 3 Everyone is responsible for caring for the environment
Column L	Question 4 I am concerned about the long-term future of the environment.
Column M	Question 5 In my opinion, it is important to conserve natural resources.
Column N	Question 6 I think that environmental sustainability is a waste of time and effort.
Column O	Question 7 I am a passionate advocate of environmental sustainability.
Perceived behavioral control	
Column P	Question 8 It is easy for me to perform environmentally sustainable activities (e.g., energy conservation, recycling).
Column Q	Question 9 I have control over my actions to support the environment.
Column R	Question 10 It is my decision whether or not to perform environmentally sustainable activities.
Column S	Question 11 I have the ability to carry out environmentally sustainable activities.
Column T	Question 12 I have control over performing environmentally sustainable activities.



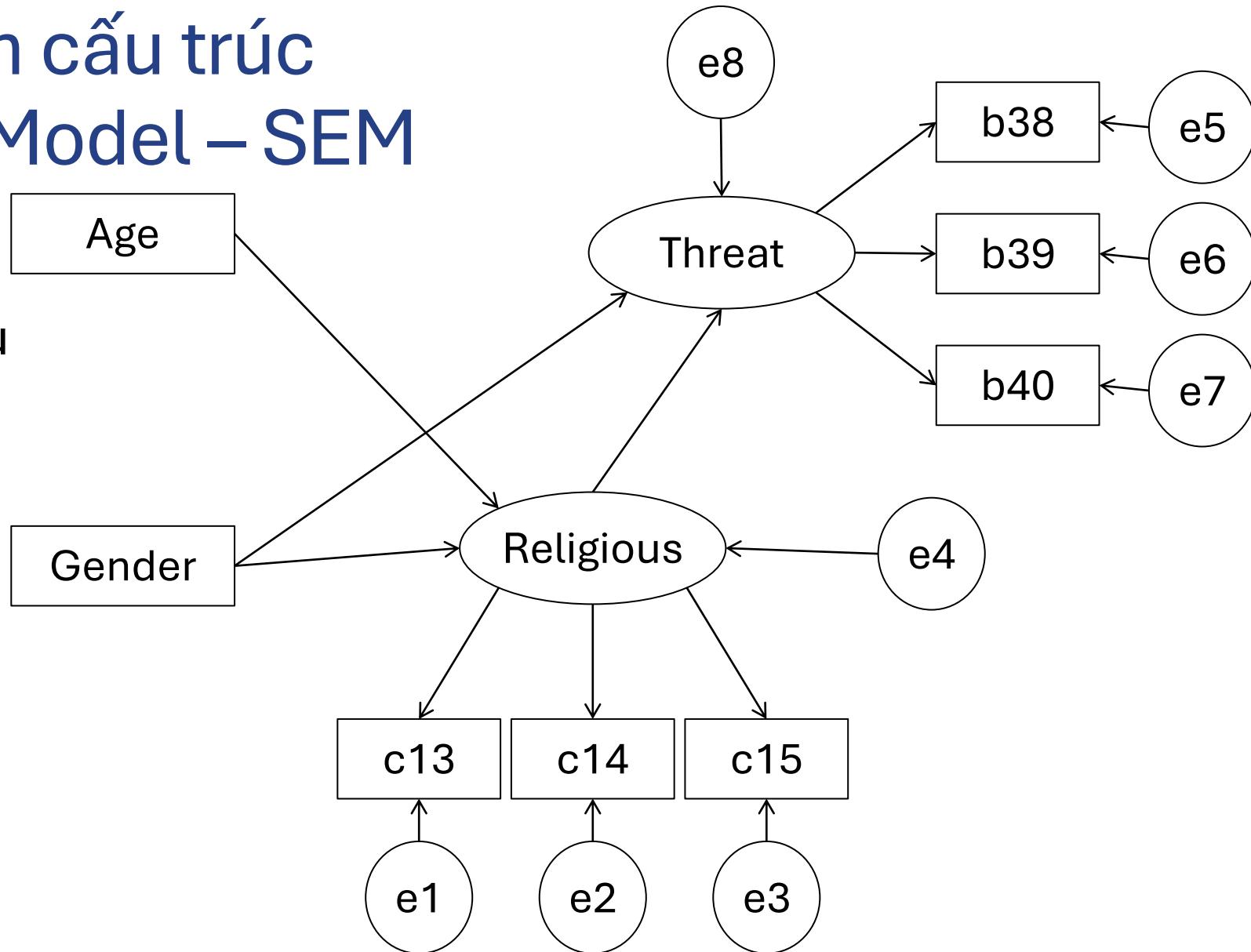
12

# Phân tích nhân tố khẳng định (CFA)



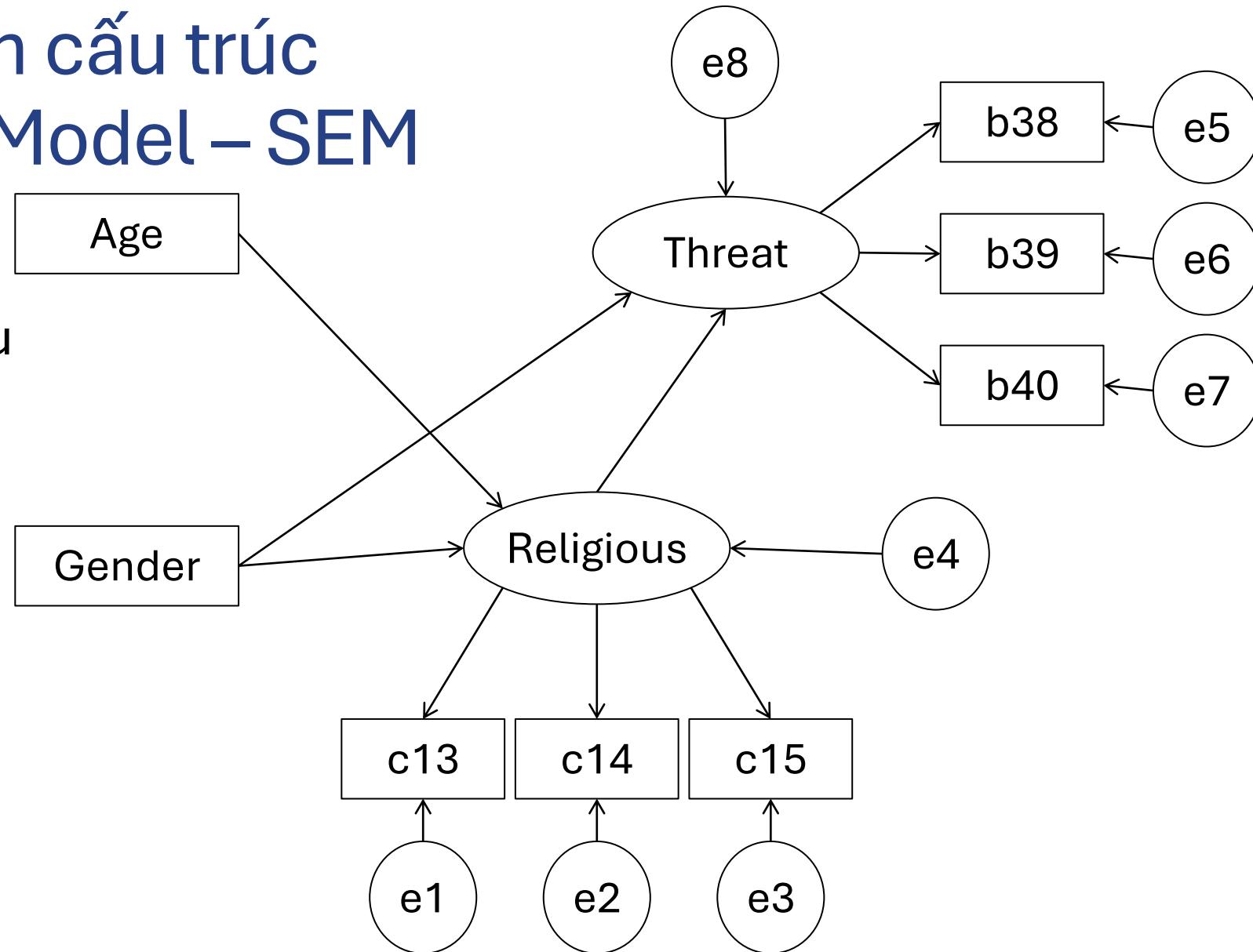
# Mô hình phương trình cấu trúc Structural Equation Model – SEM

- Tiềm ẩn (latent) và biểu hiện (manifest)



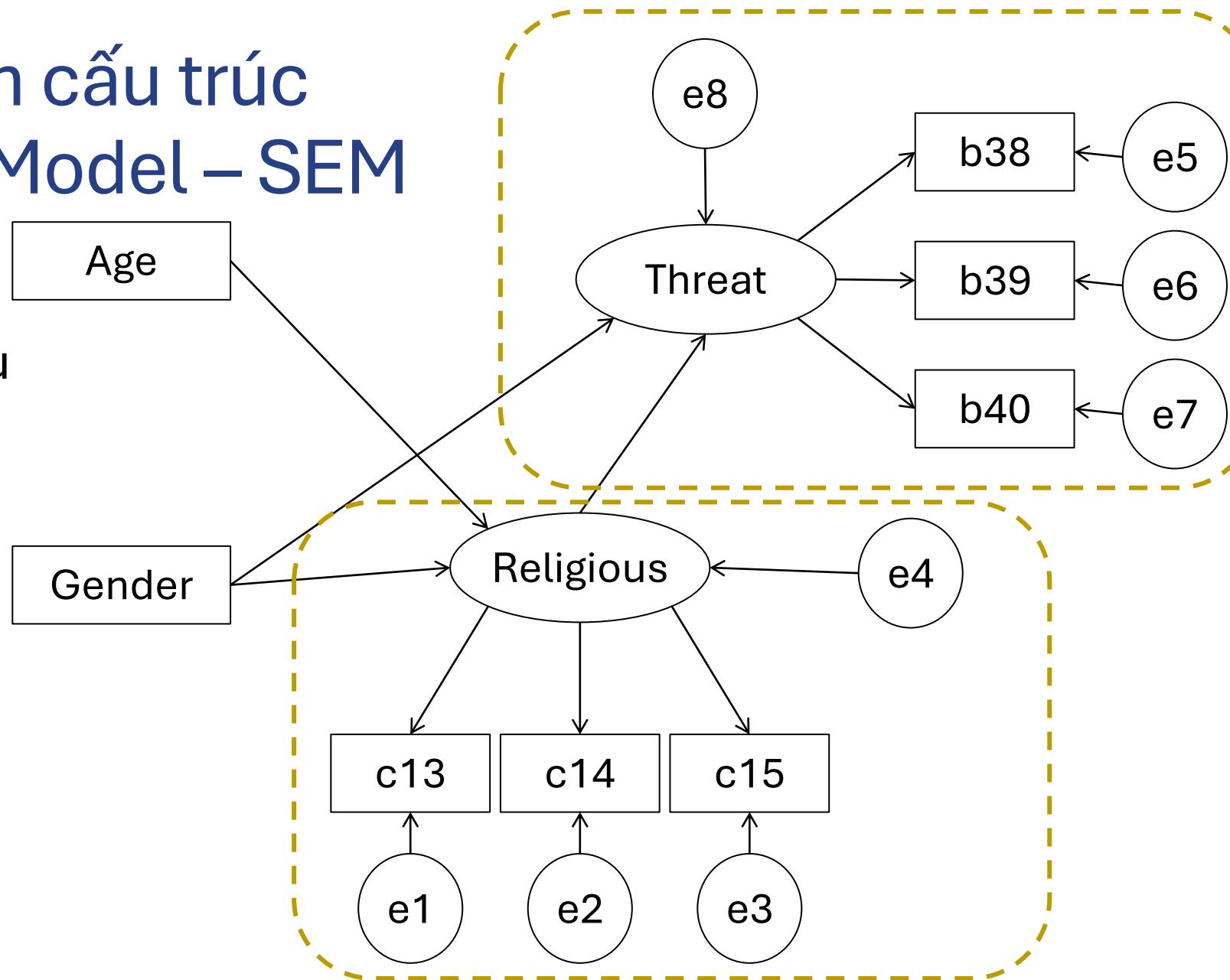
# Mô hình phương trình cấu trúc Structural Equation Model – SEM

- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)



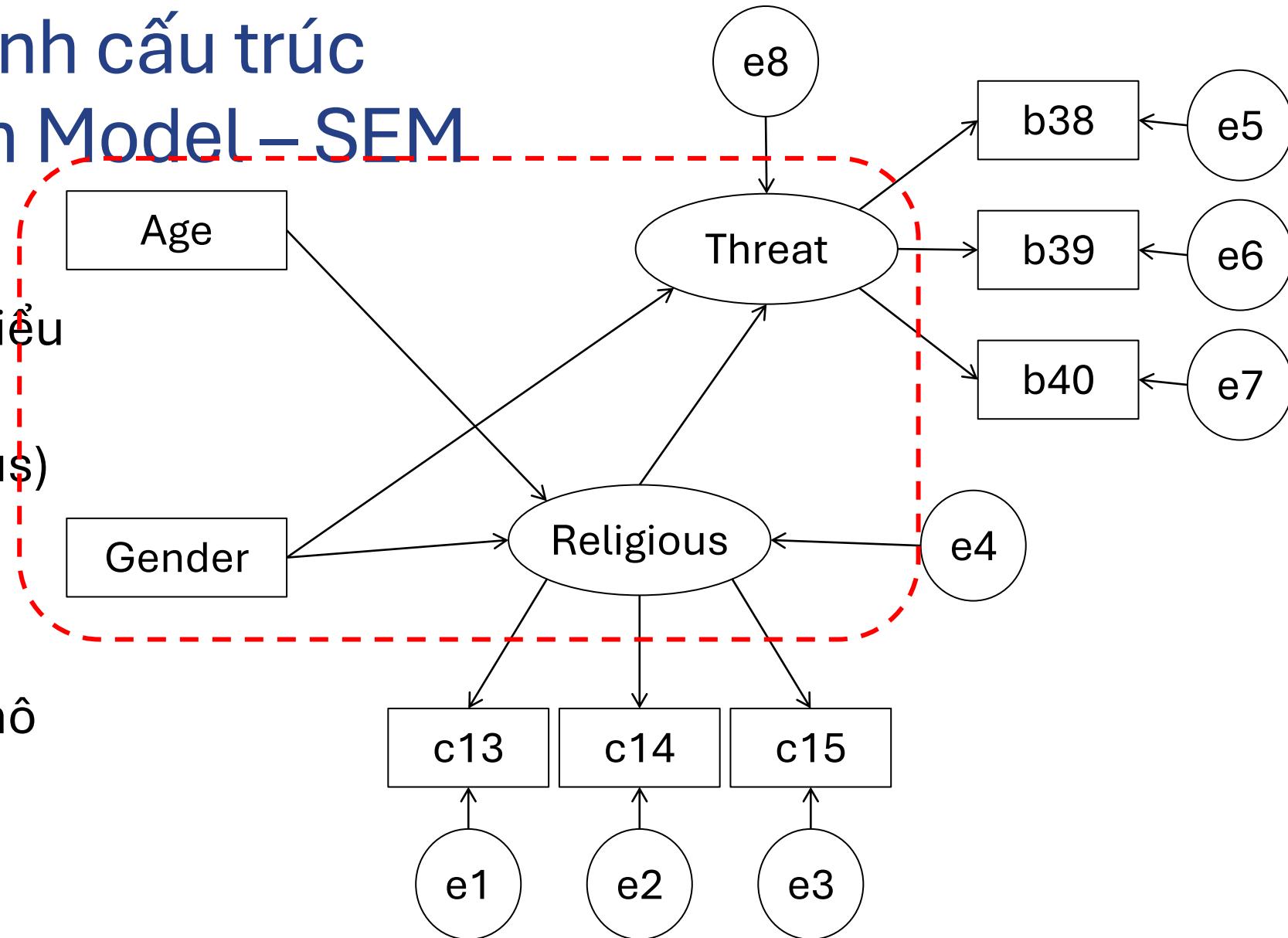
# Mô hình phương trình cấu trúc Structural Equation Model – SEM

- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement)



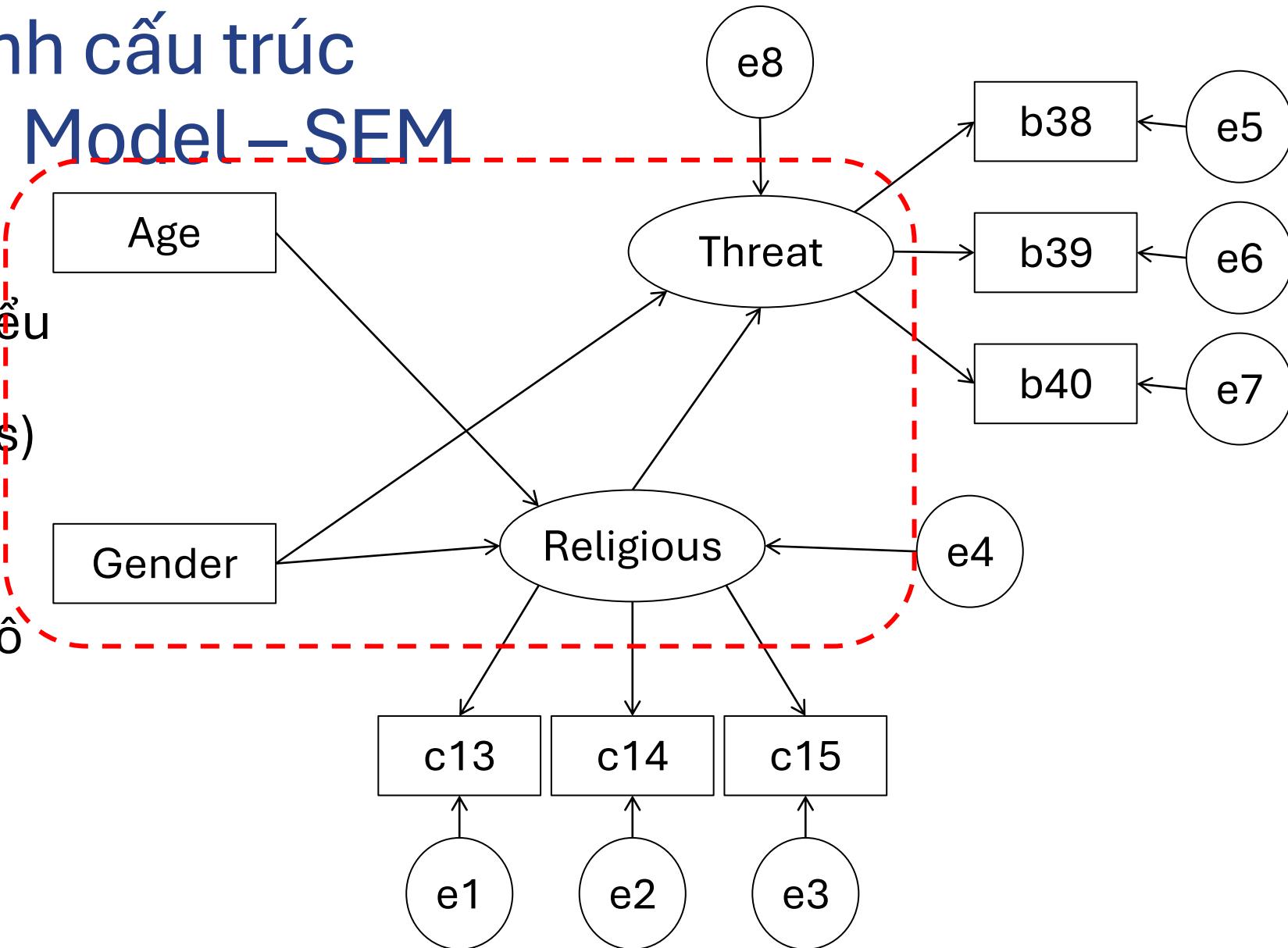
# Mô hình phương trình cấu trúc Structural Equation Model - SEM

- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement) và mô hình cấu trúc (structural)



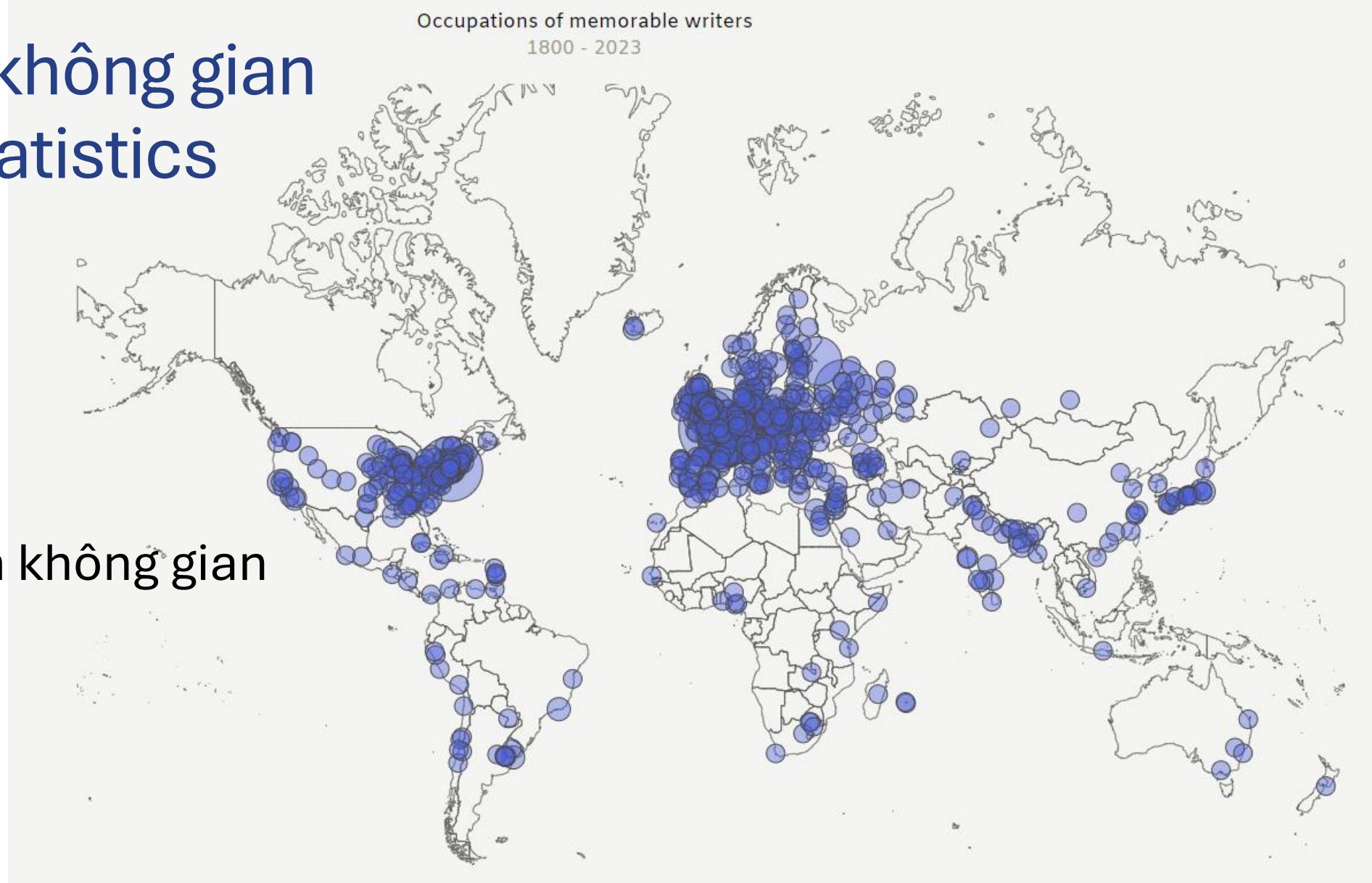
# Mô hình phương trình cấu trúc Structural Equation Model - SEM

- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement) và mô hình cấu trúc (structural)
- Tác động trực tiếp và gián tiếp (Direct vs indirect effects)



# Thống kê không gian Spatial Statistics

- Tương quan không gian
- GIS



# Lựa chọn phương pháp

Source: JASP Team (2024)  
JASP (Version 0.18.3)  
[Computer software].

