

# Phương pháp nghiên cứu trong khoa học liên ngành - Các phương pháp định lượng

---

**Nguyễn Bích Ngọc**

Khoa các khoa học liên ngành, ĐHQGHN

# Thông tin lớp học

- Lý thuyết: 16/06 và sáng 23/06
- Thực hành: chiều 23/06
  - R
  - Máy tính cài sẵn R và Rstudio
  - <https://rstudio-education.github.io/hopr/starting.html>
- Bài thi:
  - 2 phần – 23/07/2024
    - Phân tích bảng số liệu
    - Đọc bài báo và nhận xét
  - Gạch đầu dòng ý chính
  - KHÔNG SAO CHÉP LẤN NHAU

# Mục tiêu lớp học

- Giới thiệu khái niệm cơ bản của các phương pháp định lượng và thống kê
- Xác định vấn đề và định hướng phương pháp sử dụng (tên phương pháp)

# Tài liệu tham khảo

- Cẩm nang nghiên cứu khoa học: từ ý tưởng đến công bố – Nguyễn Văn Tuấn (2<sup>nd</sup> edition, 2020)
- Từng bước nhập môn nghiên cứu khoa học xã hội – Phạm Hiệp & cộng sự (2022)
- Fundamentals of data visualization – Claus O. Wilke (<https://clauswilke.com/dataviz/index.html>)
- Applied statistics with R – David Dalpiaz (<https://book.stat420.org/>)
- The Scientist's Guide to Writing: How to Write More Easily and Effectively throughout Your Scientific Career – Stephen B. Heard (2<sup>nd</sup> 2022)
- Understanding research methods – Coursera (<https://www.coursera.org/learn/research-methods/home/info>)

# Nội dung

- Giới thiệu chung
- Dữ liệu và nguồn dữ liệu
- Phân tích dữ liệu thăm dò
- Phân tích dữ liệu khẳng định
- Các phương pháp nâng cao

# Giới thiệu chung

---

# 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

1. Dogs are better than cats

# Ví dụ

1. Dogs are better than cats
2. Dog owners are physically fitter than cat owners

# Nghiên cứu khoa học?

# Nghiên cứu khoa học?

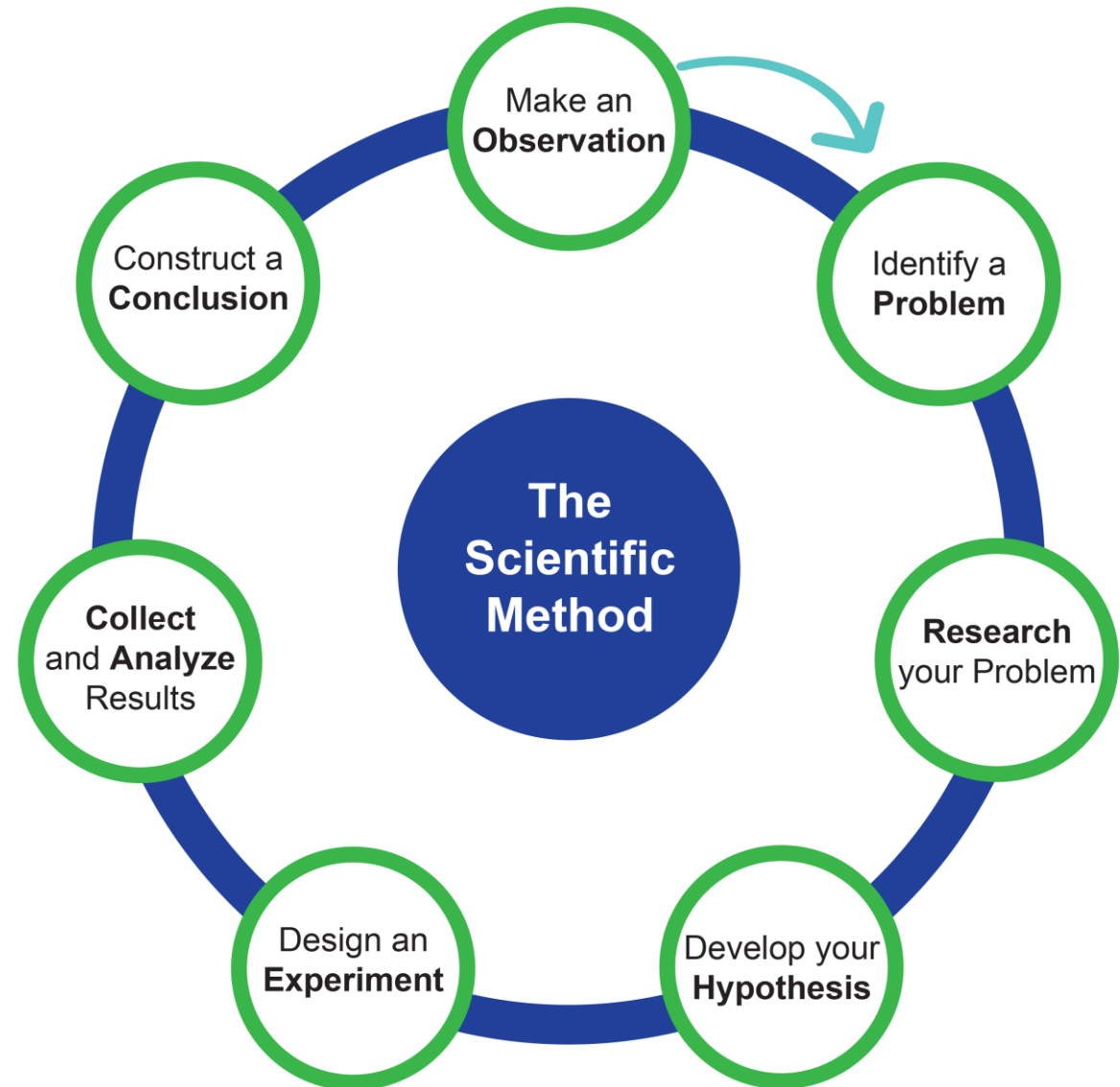
Research is **systematic inquiry** that helps to **make sense of the world** and that helps to make sensible the debates and interpretations that we have of issues of **contemporary significance**.

Professor Sandra Halperin

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

# Quá trình nghiên cứu

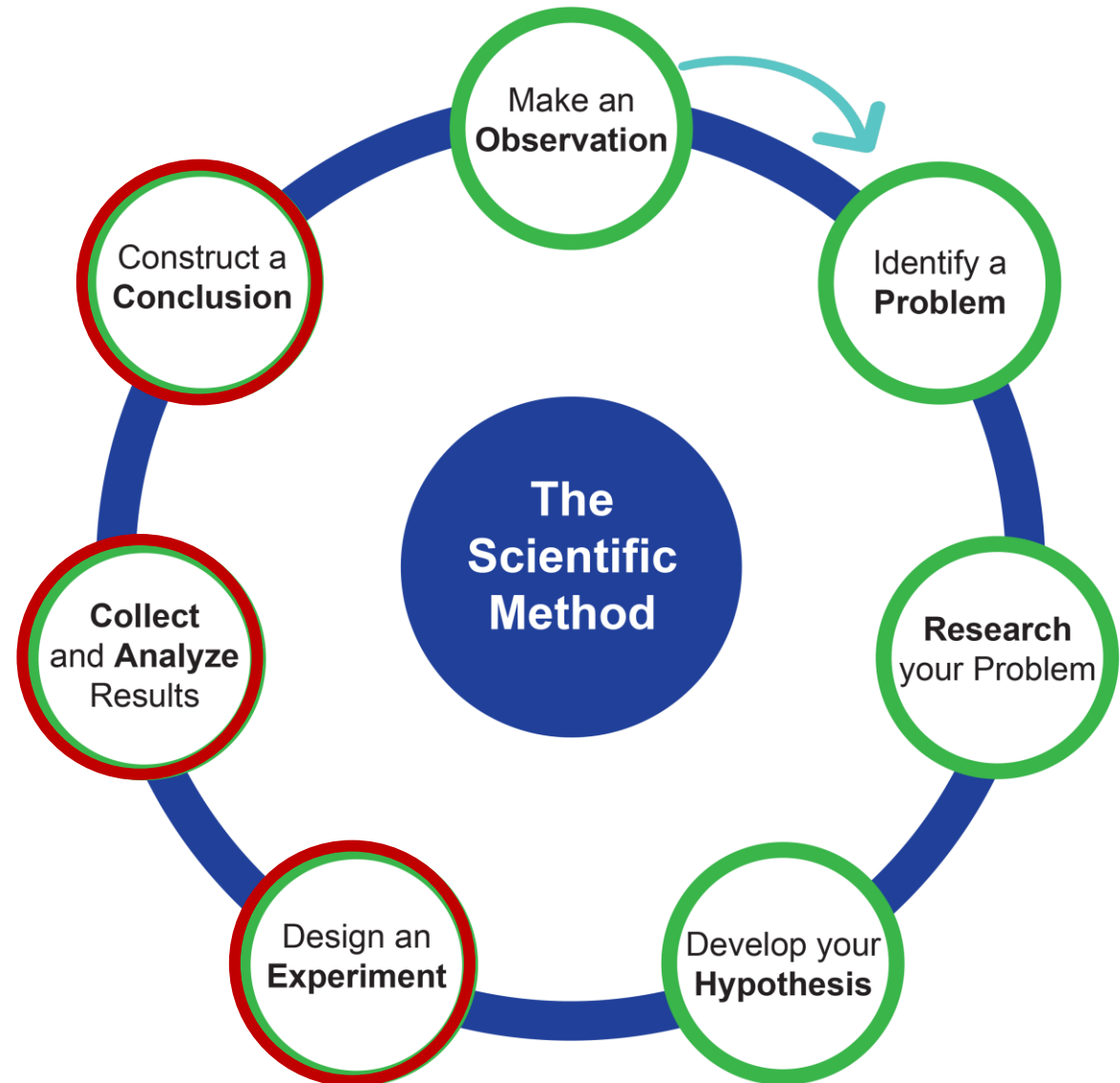
# Quá trình nghiên cứu



Source: <https://www.arfreethinkers.org/>



# Phương pháp nghiên cứu?



Source: <https://www.arfreethinkers.org/>

# Dữ liệu và nguồn dữ liệu

---

# Dữ liệu

- Dữ liệu là gì? (Dữ liệu vs. thông tin?)

# Dữ liệu

- Dữ liệu là gì? (Dữ liệu vs. thông tin?)
- Dữ liệu có cấu trúc và dữ liệu phi cấu trúc

# Dữ liệu

- Dữ liệu là gì? (Dữ liệu vs. thông tin?)
- Dữ liệu có cấu trúc và dữ liệu phi cấu trúc

|    | Gender | Age.Range         | Year   | Nationality | Q1 | Q2 | Q3 | Q4 | Q5 |
|----|--------|-------------------|--------|-------------|----|----|----|----|----|
| 1  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 5  | 7  | 7  | 7  |
| 2  | Female | 20 - 21 years old | Year 4 | Thai        | 6  | 5  | 7  | 5  | 6  |
| 3  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 7  | 7  | 7  | 7  |
| 4  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 2  | 7  | 6  | 7  |
| 5  | Female | 22 - 23 years old | Year 4 | Thai        | 6  | 6  | 7  | 7  | 7  |
| 6  | Male   | 20 - 21 years old | Year 3 | Thai        | 5  | 4  | 4  | 4  | 4  |
| 7  | Male   | 20 - 21 years old | Year 3 | Thai        | 6  | 4  | 5  | 7  | 6  |
| 8  | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 4  | 7  | 6  | 7  |
| 9  | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 5  | 7  | 7  | 7  |
| 10 | Male   | 20 - 21 years old | Year 3 | Thai        | 5  | 5  | 5  | 7  | 6  |
| 11 | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 5  | 7  | 7  | 7  |



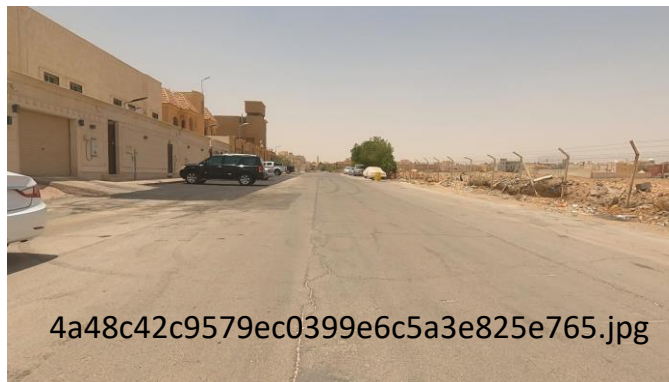
# Dữ liệu

Phi cấu trúc  
=> cấu trúc?

# Dữ liệu

Phi cấu trúc  
=> cấu trúc?

Các dữ liệu  
nào khác có  
thể bổ sung?



|   | A     | B                                    | C             |    |
|---|-------|--------------------------------------|---------------|----|
|   | class | image_path                           | name          | xt |
|   | 3     | 4a48c42c9579ec0399e6c5a3e825e765.jpg | GARBAGE       |    |
|   | 3     | 4a48c42c9579ec0399e6c5a3e825e765.jpg | GARBAGE       |    |
|   | 3     | 4a48c42c9579ec0399e6c5a3e825e765.jpg | GARBAGE       |    |
|   | 7     | ea906a663da6321bcef78be4b7d1afff.jpg | BAD_BILLBOARD |    |
|   | 8     | 1c7d48005a12d1b19261b8e71df7cafe.jpg | SAND_ON_ROAD  |    |
|   | 8     | 1c7d48005a12d1b19261b8e71df7cafe.jpg | SAND_ON_ROAD  |    |
|   | 8     | 8ca1b825716ea6755180fde347ac79c1.jpg | SAND_ON_ROAD  |    |
|   | 0     | 8ca1b825716ea6755180fde347ac79c1.jpg | GRAFFITI      |    |
| 0 | 0     | 8ca1b825716ea6755180fde347ac79c1.jpg | GRAFFITI      |    |
| 1 | 2     | e1f3026bc4b1689d81f03e92e9043c2b.jpg | POTHOLES      |    |
| 2 | 3     | c12b006174423ceb3e2e3563a8ca7751.jpg | GARBAGE       |    |
| 3 | 3     | 7fb40d10dde6d5643aa8e197b6b46c2e.jpg | GARBAGE       |    |

# Dữ liệu

- Dữ liệu là gì? (Dữ liệu vs. thông tin?)
- Dữ liệu có cấu trúc và dữ liệu phi cấu trúc
- Nguồn sơ cấp vs thứ cấp



## Sơ cấp

Thực nghiệm

Khảo sát/Bảng hỏi

Đo đạc ngoài thực địa



## Thứ cấp

Cơ sở dữ liệu mở

Báo cáo kỹ thuật của chính phủ

Báo cáo nội bộ



# Bài tập

- Dữ liệu trong nghiên cứu của bạn?
- Có cấu trúc, phi cấu trúc?
- Nguồn sơ cấp, thứ cấp?

# Dữ liệu

- Khảo sát nhận thức của sinh viên đại học Thái Lan đối với sự bền vững của môi trường
- 312 sinh viên
- Likert 1-7: Strongly Disagree – Strongly Agree

|    | Gender | Age.Range         | Year   | Nationality | Q1 | Q2 | Q3 | Q4 | Q5 |
|----|--------|-------------------|--------|-------------|----|----|----|----|----|
| 1  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 5  | 7  | 7  | 7  |
| 2  | Female | 20 - 21 years old | Year 4 | Thai        | 6  | 5  | 7  | 5  | 6  |
| 3  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 7  | 7  | 7  | 7  |
| 4  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 2  | 7  | 6  | 7  |
| 5  | Female | 22 - 23 years old | Year 4 | Thai        | 6  | 6  | 7  | 7  | 7  |
| 6  | Male   | 20 - 21 years old | Year 3 | Thai        | 5  | 4  | 4  | 4  | 4  |
| 7  | Male   | 20 - 21 years old | Year 3 | Thai        | 6  | 4  | 5  | 7  | 6  |
| 8  | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 4  | 7  | 6  | 7  |
| 9  | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 5  | 7  | 7  | 7  |
| 10 | Male   | 20 - 21 years old | Year 3 | Thai        | 5  | 5  | 5  | 7  | 6  |
| 11 | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 5  | 7  | 7  | 7  |

# Dữ liệu

Biến

- Khảo sát nhận thức của sinh viên đại học Thái Lan đối với sự bền vững của môi trường
- 312 sinh viên
- Likert 1-7: Strongly Disagree – Strongly Agree

|    | Gender | Age.Range         | Year   | Nationality | Q1 | Q2 | Q3 | Q4 | Q5 |
|----|--------|-------------------|--------|-------------|----|----|----|----|----|
| 1  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 5  | 7  | 7  | 7  |
| 2  | Female | 20 - 21 years old | Year 4 | Thai        | 6  | 5  | 7  | 5  | 6  |
| 3  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 7  | 7  | 7  | 7  |
| 4  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 2  | 7  | 6  | 7  |
| 5  | Female | 22 - 23 years old | Year 4 | Thai        | 6  | 6  | 7  | 7  | 7  |
| 6  | Male   | 20 - 21 years old | Year 3 | Thai        | 5  | 4  | 4  | 4  | 4  |
| 7  | Male   | 20 - 21 years old | Year 3 | Thai        | 6  | 4  | 5  | 7  | 6  |
| 8  | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 4  | 7  | 6  | 7  |
| 9  | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 5  | 7  | 7  | 7  |
| 10 | Male   | 20 - 21 years old | Year 3 | Thai        | 5  | 5  | 5  | 7  | 6  |
| 11 | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 5  | 7  | 7  | 7  |

Đối tượng  
quan sát/Mẫu

# Biến

- Các loại biến

- Định danh (nominal)

- Thứ bậc (ordinal)

- Liên tục/Định lượng (continuous variables)



Biến rời rạc/định tính  
(discrete variables)

# Biến định danh – nominal variables

- Ví dụ:
  - Giới tính
  - Tôn giáo
  - Quốc tịch

# Biến thứ bậc – ordinal variables

- Ví dụ:
  - Thang Likert: hoàn toàn không đồng ý – hoàn toàn đồng ý
  - Trình độ học vấn: THCS, THPT, trung cấp, đại học, sau đại học
  - Điều kiện kinh tế xã hội: thấp, trung bình, cao
  - 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

# Biến liên tục/định lượng – continuous variables

- Ví dụ:
  - GDP
  - Nhiệt độ không khí
  - $\text{PM}_{2,5}$  ( $\mu\text{g}/\text{m}^3$ )

# Bài tập

|    | Gender | Age.Range         | Year   | Nationality | Q1 | Q2 | Q3 | Q4 | Q5 |
|----|--------|-------------------|--------|-------------|----|----|----|----|----|
| 1  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 5  | 7  | 7  | 7  |
| 2  | Female | 20 - 21 years old | Year 4 | Thai        | 6  | 5  | 7  | 5  | 6  |
| 3  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 7  | 7  | 7  | 7  |
| 4  | Female | 20 - 21 years old | Year 4 | Thai        | 7  | 2  | 7  | 6  | 7  |
| 5  | Female | 22 - 23 years old | Year 4 | Thai        | 6  | 6  | 7  | 7  | 7  |
| 6  | Male   | 20 - 21 years old | Year 3 | Thai        | 5  | 4  | 4  | 4  | 4  |
| 7  | Male   | 20 - 21 years old | Year 3 | Thai        | 6  | 4  | 5  | 7  | 6  |
| 8  | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 4  | 7  | 6  | 7  |
| 9  | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 5  | 7  | 7  | 7  |
| 10 | Male   | 20 - 21 years old | Year 3 | Thai        | 5  | 5  | 5  | 7  | 6  |
| 11 | Female | 20 - 21 years old | Year 3 | Thai        | 7  | 5  | 7  | 7  | 7  |



# Bài tập

- csmptv: lượng nước cấp tiêu thụ trong 1 năm
- rwtank: có bể nước mưa
- iceqac2: thu nhập của gia đình
- hhs\_tot: số thành viên trong gia đình
- cfdiwq: sự tin tưởng vào chất lượng nước cấp
- livara: diện tích nhà ở/căn hộ

| id   | csmptv | rwtank | iceqac2    | hhs_tot | cfdiwq            | livara |
|------|--------|--------|------------|---------|-------------------|--------|
| 137  | 105    | no     | modest     | 3       | suspicious        | 129    |
| 431  | 56.99  | no     | average    | 2       | rather confident  | 120    |
| 655  | 122    | yes    | modest     | 5       | rather confident  | 130    |
| 730  | 74.57  | no     | average    | 2       | confident         | 132    |
| 780  | 30     | no     | average    | 1       | rather confident  | 70     |
| 781  | 66.36  | yes    | higher     | 2       | confident         | 162    |
| 1048 | 30.93  | yes    | modest     | 3       | suspicious        | 110    |
| 1403 | 100    | no     | higher     | 2       | confident         | 150    |
| 1405 | 52.95  | yes    | modest     | 4       | confident         | 100    |
| 1432 | 139    | no     | modest     | 3       | rather confident  | 100    |
| 1476 | 25     | yes    | average    | 2       | confident         | 100    |
| 1757 | 71.25  | yes    | average    | 3       | rather suspicious | 90     |
| 2183 | 69     | yes    | modest     | 2       | confident         | 150    |
| 2334 | 86.06  | no     | modest     | 2       | rather confident  | 90     |
| 2345 | 29.2   | yes    | precarious | 3       | confident         | 160    |
| 2375 | 33.46  | no     | average    | 2       | rather confident  | 100    |
| 2687 | 45.63  | no     | precarious | 1       | rather suspicious | 70     |
| 2704 | 126    | yes    | higher     | 4       | confident         | 200    |
| 2714 | 16.23  | yes    | modest     | 1       | confident         | 80     |
| 2752 | 105.09 | no     | higher     | 2       | confident         | 90     |

# Mẫu, quần thể, cỡ mẫu

- Quần thể? Mẫu?

- Tính đại diện?

[https://youtu.be/rxv\\_sB-wOkY](https://youtu.be/rxv_sB-wOkY)

- Cỡ mẫu?

[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%A1n%20c%E1%BB%A1%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%A1n%20c%E1%BB%A1%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)
  - 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 ngạch (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 ngạch (Quota sample)
  - Mẫu có mục đích (Judgement (or purposive) sample)
  - Mẫu bóng tuyết (Snowball sample)

Sử dụng  
trọng số để  
khắc phục  
tính không  
đại diện

# Bảng hỏi

**Table 1**  
Description of the characteristics in the dataset.

| Column   | Data label     | Explanation  |
|----------|----------------|--|
| Column A | Student Status | Degree student; Exchange student   |
| Column B | Institution    | Prince of Songkla University   |
| Column C | Faculty        | Faculty of Hospitality and Tourism; College of Computing; Faculty of International Studies     |
| Column D | Gender         | Male; Female; I do not wish to say; Other  |
| Column E | Age Range      | 18–19 years old; 20–21 years old; 22–23 years old; 24 years or above                           |
| Column F | Year           | Year 1; Year 2; Year 3; Year 4   |
| Column G | Nationality    | Thai; Foreign  |
| Column H | Probe          | Have you heard about environmental sustainability before?<br>[Answer options: (Yes) and (No)]. |

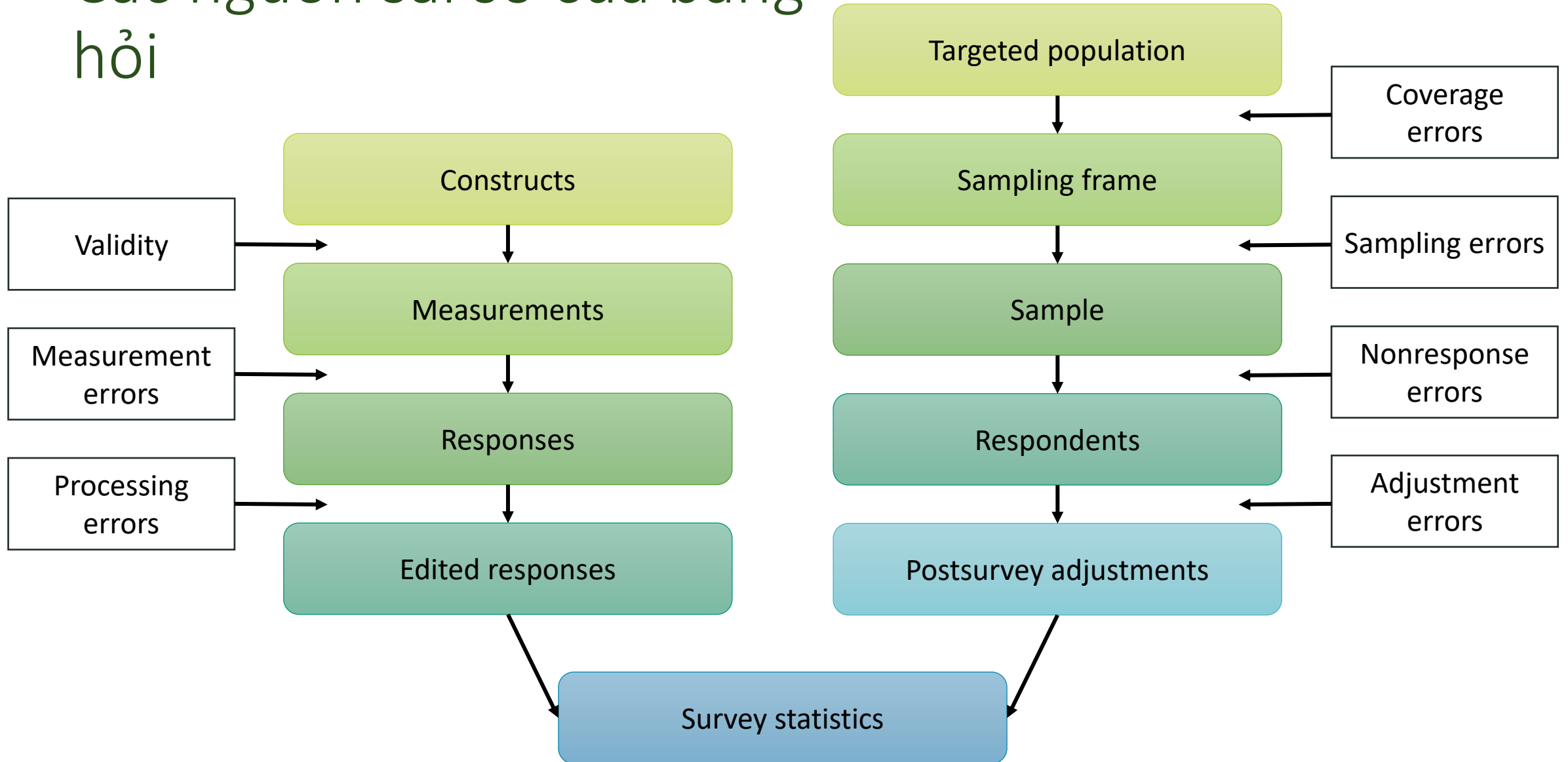
**Table 2**  
Questionnaire organized by their respective factor.

| Column                              | Data label  | Explanation   |
|-------------------------------------|-------------|---|
| <i>Attitude</i>                     |             |   |
| 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.   |
| <i>Perceived behavioral control</i> |             |   |
| 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.                                      |

# Thiết kế bảng hỏi

- Phạm trù (Construct) cần quan tâm
  - Là gì?
  - **Làm sao để đo?**
- Thiết kế bảng hỏi cần chú ý
  - Cách dùng từ
  - Tránh việc sử dụng chỉ một câu hỏi để đo lường cho 1 phạm trù
  - Tâm lý người hỏi và người trả lời
  - **Luôn luôn thử nghiệm trước** bộ câu hỏi

# Các nguồn sai số của bảng hỏi



# Phân tích dữ liệu thăm dò

---



# 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, và lý thuyết mới

a

| ID | steps | bmi  |
|----|-------|------|
| 3  | 15000 | 17.0 |
| 4  | 14861 | 17.2 |
| 5  | 14861 | 17.2 |

| ID | steps | bmi  |
|----|-------|------|
| 1  | 15000 | 16.9 |
| 2  | 15000 | 16.9 |
| 6  | 14861 | 16.8 |
| 7  | 14861 | 16.8 |
| 8  | 14699 | 17.3 |
| 10 | 14560 | 20.5 |
| 11 | 14560 | 20.6 |
| 13 | 14560 | 20.5 |
| 17 | 14560 | 20.4 |
| 18 | 14560 | 20.4 |
| 19 | 14560 | 19.8 |
| 20 | 14560 | 19.7 |
| 22 | 14560 | 19.7 |
| 24 | 14560 | 19.6 |
| 25 | 14560 | 19.6 |
| 27 | 14560 | 19.6 |
| 29 | 14560 | 17.4 |
| 30 | 14560 | 17.4 |
| 32 | 14398 | 20.9 |
| 37 | 14398 | 17.5 |
| 40 | 14398 | 17.1 |
| 42 | 14259 | 21.1 |
| 43 | 14259 | 21.1 |
| 44 | 14259 | 21.1 |
| 45 | 14259 | 21.1 |

- 2 groups of students
- bmi vs steps
- hypothesis?

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

a

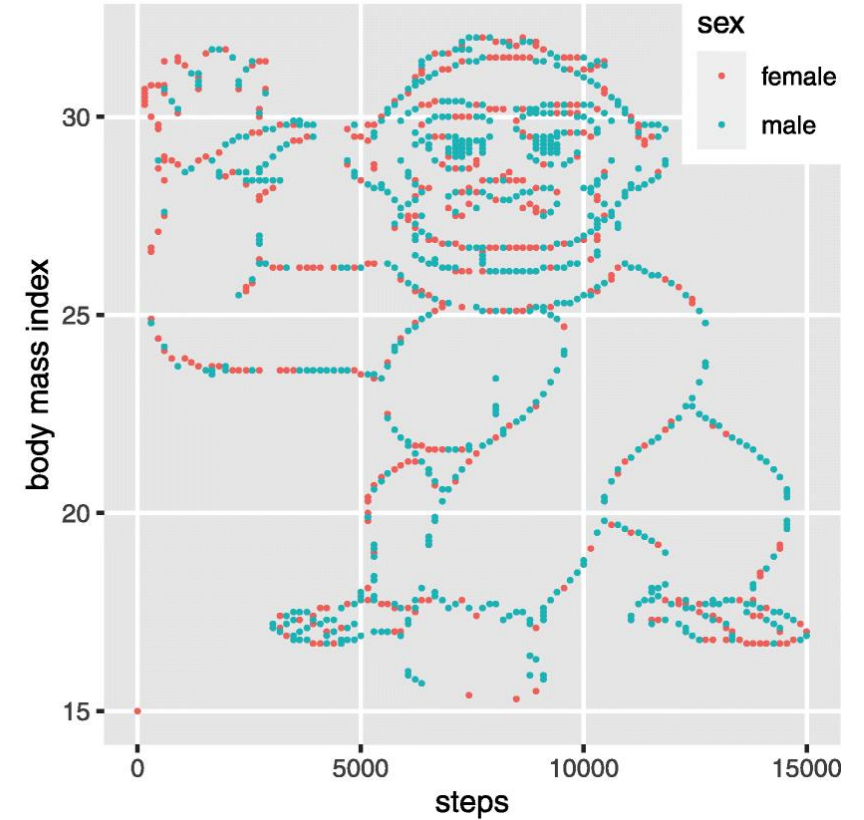
| ID | steps | bmi  |
|----|-------|------|
| 3  | 15000 | 17.0 |
| 4  | 14861 | 17.2 |

| ID | steps | bmi  |
|----|-------|------|
| 1  | 15000 | 16.9 |
| 2  | 15000 | 16.9 |
| 6  | 14861 | 16.8 |
| 7  | 14861 | 16.8 |
| 8  | 14699 | 17.3 |
| 10 | 14560 | 20.5 |
| 11 | 14560 | 20.6 |
| 13 | 14560 | 20.5 |
| 17 | 14560 | 20.4 |
| 18 | 14560 | 20.4 |
| 19 | 14560 | 19.8 |
| 20 | 14560 | 19.7 |
| 22 | 14560 | 19.7 |
| 24 | 14560 | 19.6 |
| 25 | 14560 | 19.6 |
| 27 | 14560 | 19.6 |
| 29 | 14560 | 17.4 |
| 30 | 14560 | 17.4 |
| 32 | 14398 | 20.9 |
| 37 | 14398 | 17.5 |
| 40 | 14398 | 17.1 |
| 42 | 14259 | 21.1 |
| 43 | 14259 | 21.1 |
| 44 | 14259 | 19.8 |

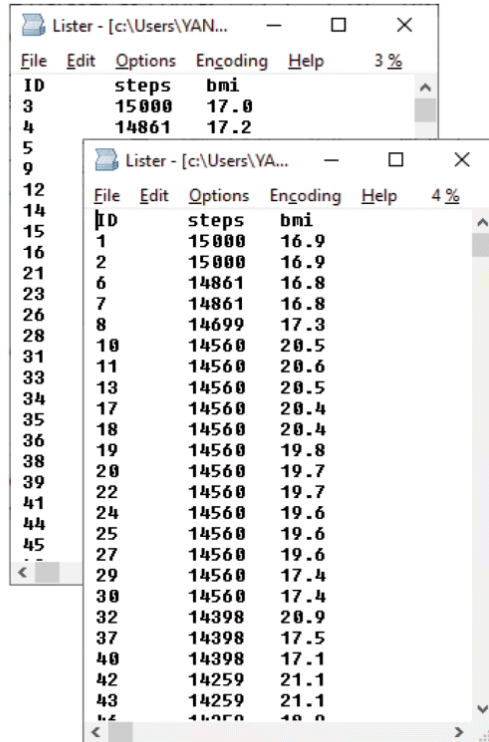
- 2 groups of students
- bmi vs steps
- hypothesis?

b



Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

a

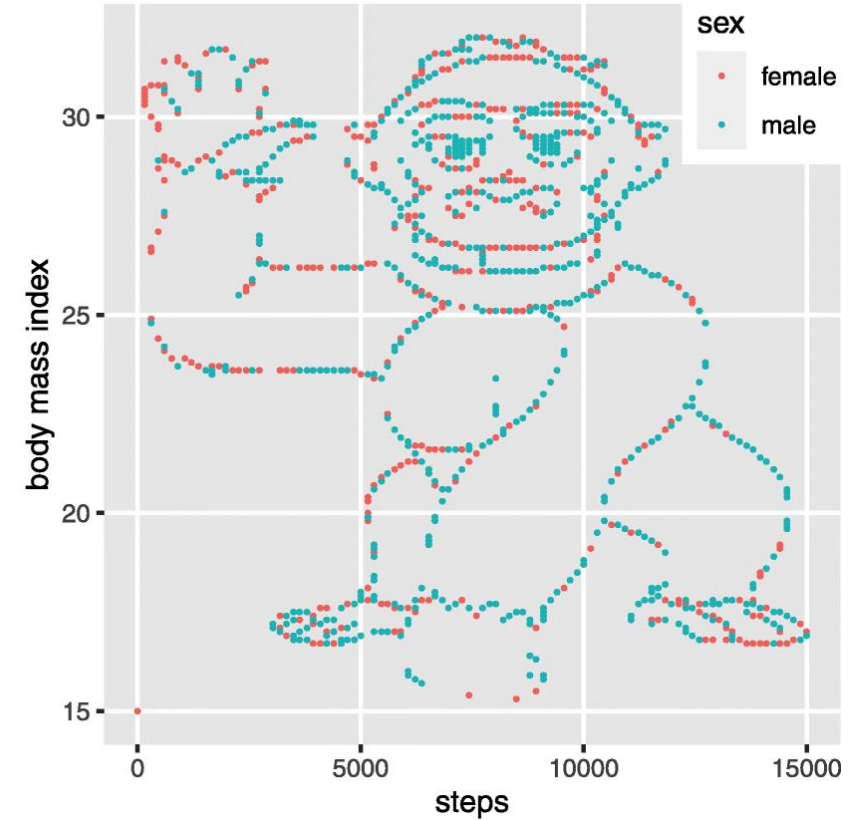


| ID | steps | bmi   |      |
|----|-------|-------|------|
| 3  | 15000 | 17.0  |      |
| 4  | 14861 | 17.2  |      |
| 5  |       |       |      |
| 6  |       |       |      |
| 7  |       |       |      |
| 8  | 14699 | 17.3  |      |
| 9  |       |       |      |
| 10 | 14560 | 20.5  |      |
| 11 | 14560 | 20.6  |      |
| 12 |       |       |      |
| 13 | 14560 | 20.5  |      |
| 14 |       |       |      |
| 15 |       |       |      |
| 16 | 1     | 15000 | 16.9 |
| 17 | 2     | 15000 | 16.9 |
| 18 | 6     | 14861 | 16.8 |
| 19 | 7     | 14861 | 16.8 |
| 20 | 8     | 14699 | 17.3 |
| 21 | 10    | 14560 | 20.5 |
| 22 | 11    | 14560 | 20.6 |
| 23 | 13    | 14560 | 20.5 |
| 24 | 17    | 14560 | 20.4 |
| 25 | 18    | 14560 | 20.4 |
| 26 | 19    | 14560 | 19.8 |
| 27 | 20    | 14560 | 19.7 |
| 28 | 22    | 14560 | 19.7 |
| 29 | 24    | 14560 | 19.6 |
| 30 | 25    | 14560 | 19.6 |
| 31 | 27    | 14560 | 19.6 |
| 32 | 29    | 14560 | 17.4 |
| 33 | 30    | 14560 | 17.4 |
| 34 | 32    | 14398 | 20.9 |
| 35 | 37    | 14398 | 17.5 |
| 36 | 40    | 14398 | 17.1 |
| 37 | 42    | 14259 | 21.1 |
| 38 | 43    | 14259 | 21.1 |
| 39 | 44    | 14259 | 21.1 |
| 40 | 45    | 14259 | 21.1 |

- 2 groups of students
- bmi vs steps
- hypothesis?

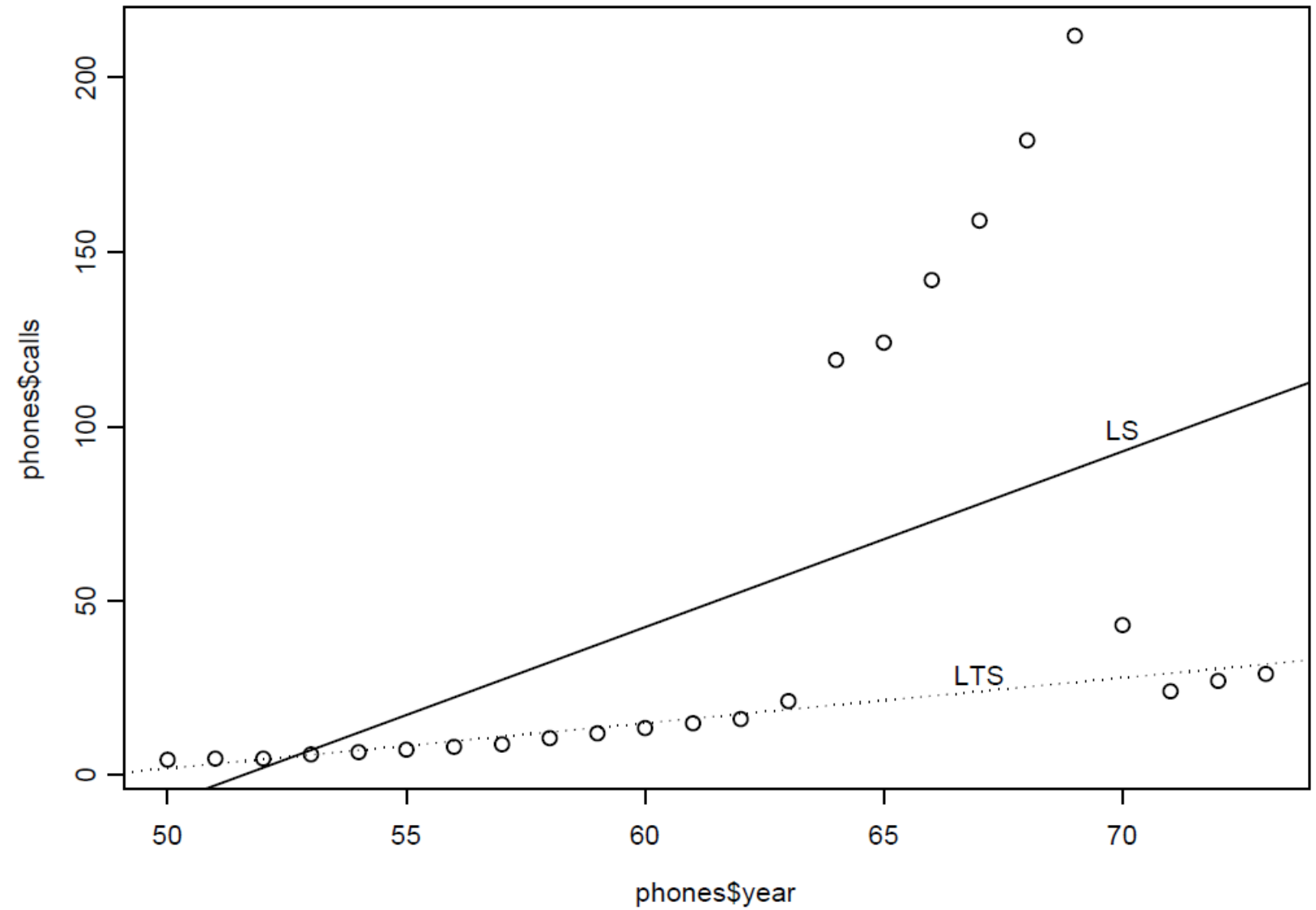
|                    | Gorilla <u>not</u> discovered | Gorilla discovered |
|--------------------|-------------------------------|--------------------|
| Hypothesis-focused | 14                            | 5                  |
| Hypothesis-free    | 5                             | 9                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

b



Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>

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



# Đồ thị

- Rõ ràng
- Chính xác
- Hiệu quả

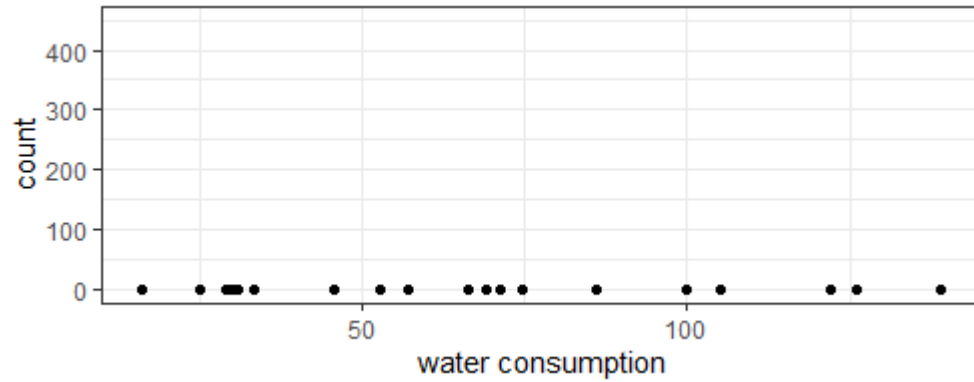
[https://www.ted.com/talks/hans\\_rosling\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen](https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen)

# Thống kê mô tả

- csmptv: lượng nước cấp tiêu thụ trong 1 năm
- rwtank: có bể nước mưa
- iceqac2: thu nhập của gia đình
- hhs\_tot: số thành viên trong gia đình
- cfdiwq: sự tin tưởng vào chất lượng nước cấp
- livara: diện tích nhà ở/căn hộ

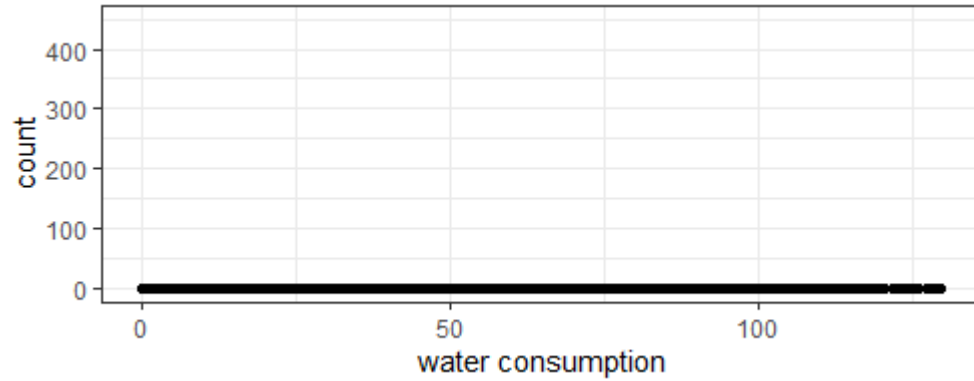
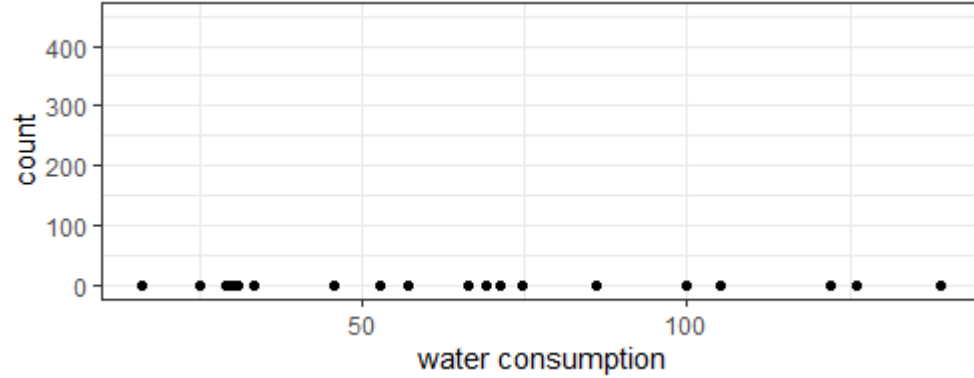
| id   | csmptv | rwtank | iceqac2    | hhs_tot | cfdiwq            | livara |
|------|--------|--------|------------|---------|-------------------|--------|
| 137  | 105    | no     | modest     | 3       | suspicious        | 129    |
| 431  | 56.99  | no     | average    | 2       | rather confident  | 120    |
| 655  | 122    | yes    | modest     | 5       | rather confident  | 130    |
| 730  | 74.57  | no     | average    | 2       | confident         | 132    |
| 780  | 30     | no     | average    | 1       | rather confident  | 70     |
| 781  | 66.36  | yes    | higher     | 2       | confident         | 162    |
| 1048 | 30.93  | yes    | modest     | 3       | suspicious        | 110    |
| 1403 | 100    | no     | higher     | 2       | confident         | 150    |
| 1405 | 52.95  | yes    | modest     | 4       | confident         | 100    |
| 1432 | 139    | no     | modest     | 3       | rather confident  | 100    |
| 1476 | 25     | yes    | average    | 2       | confident         | 100    |
| 1757 | 71.25  | yes    | average    | 3       | rather suspicious | 90     |
| 2183 | 69     | yes    | modest     | 2       | confident         | 150    |
| 2334 | 86.06  | no     | modest     | 2       | rather confident  | 90     |
| 2345 | 29.2   | yes    | precarious | 3       | confident         | 160    |
| 2375 | 33.46  | no     | average    | 2       | rather confident  | 100    |
| 2687 | 45.63  | no     | precarious | 1       | rather suspicious | 70     |
| 2704 | 126    | yes    | higher     | 4       | confident         | 200    |
| 2714 | 16.23  | yes    | modest     | 1       | confident         | 80     |
| 2752 | 105.09 | no     | higher     | 2       | confident         | 90     |

# Thống kê mô tả - Biến liên tục

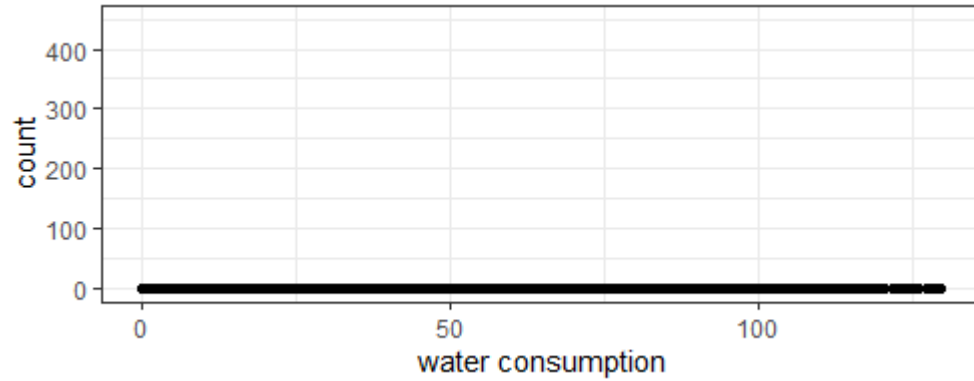
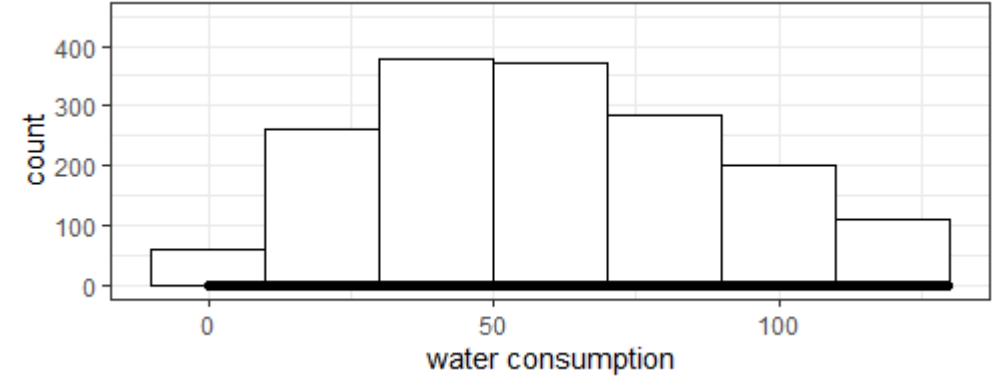
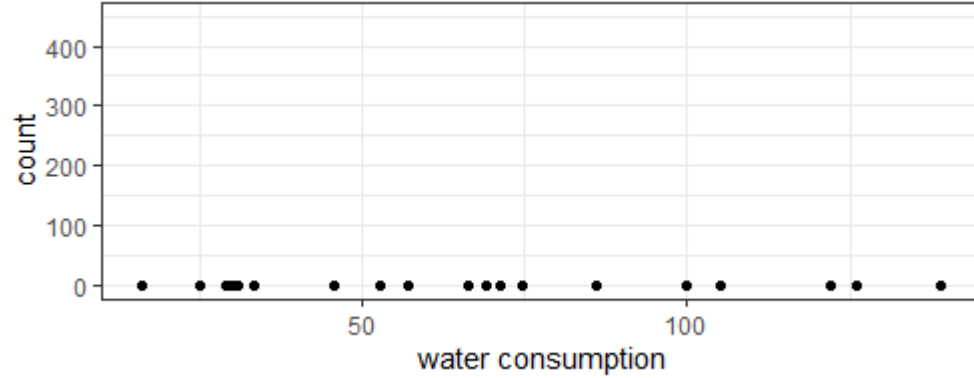




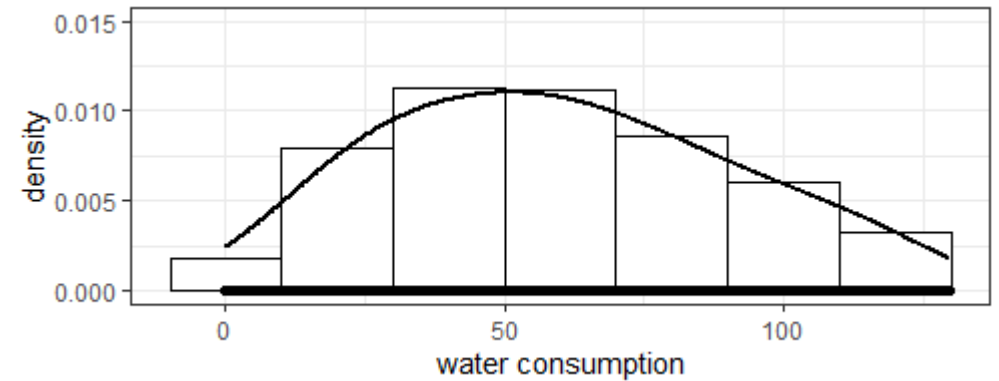
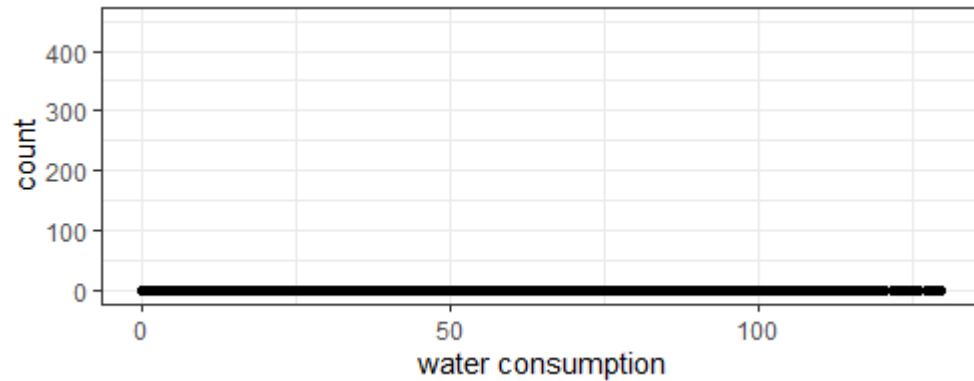
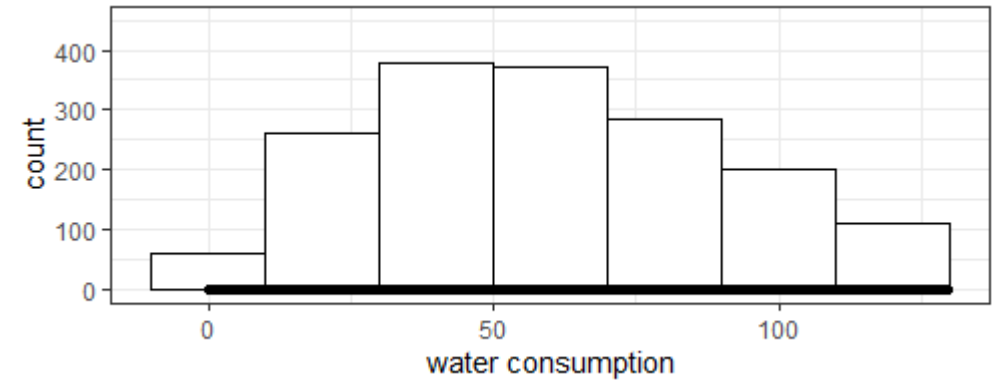
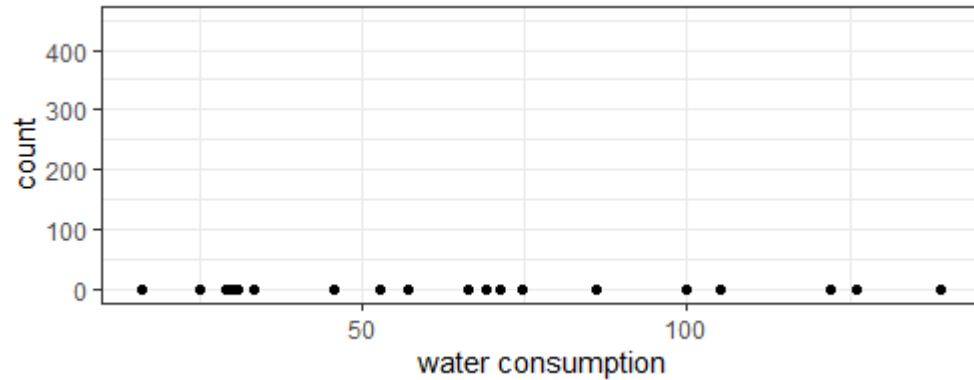
# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

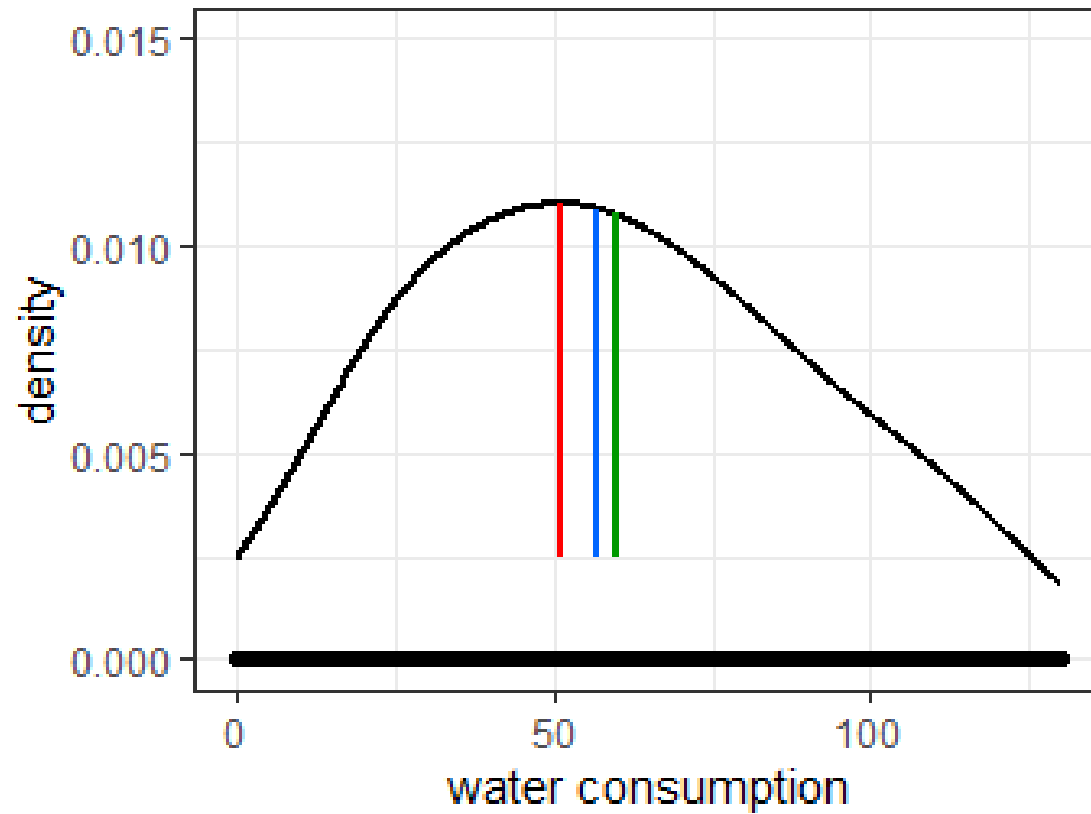


# Thống kê mô tả - Biến liên tục

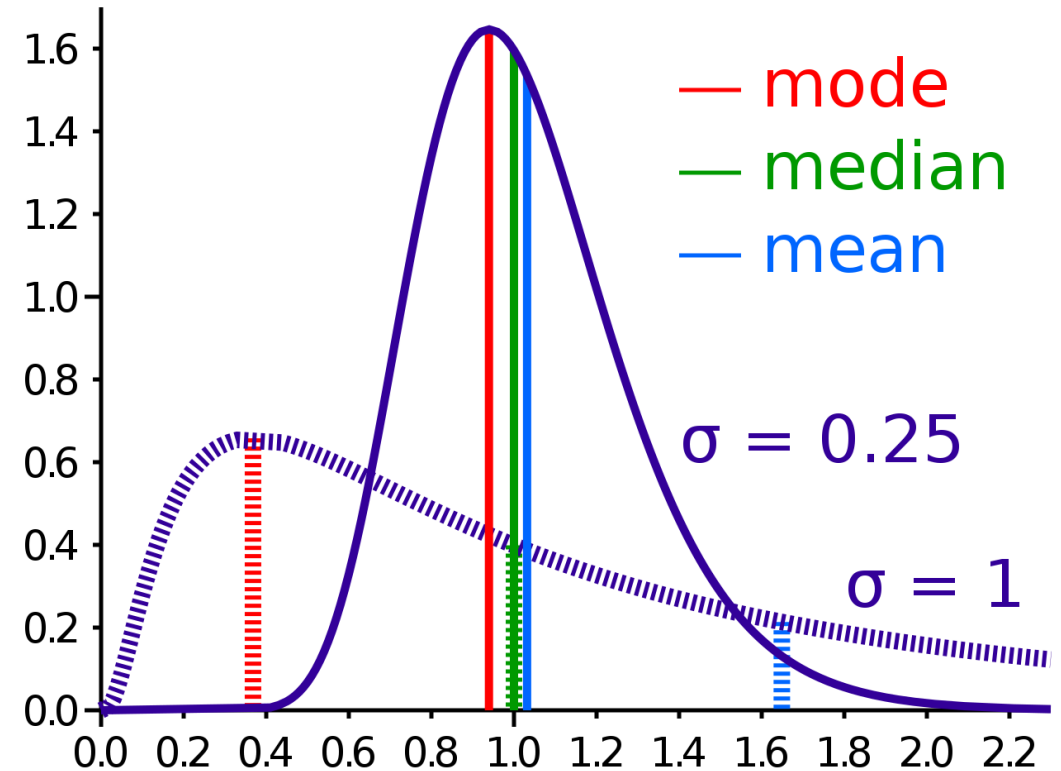
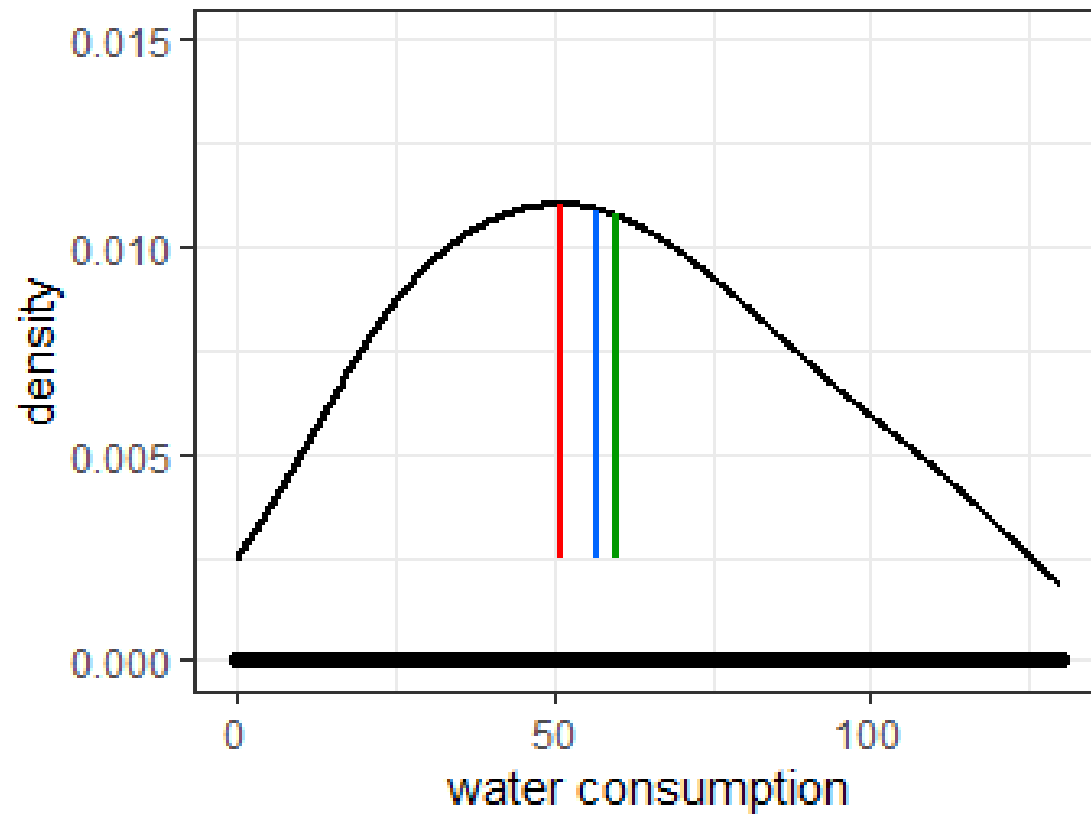


Biểu đồ tần suất/mật độ

# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

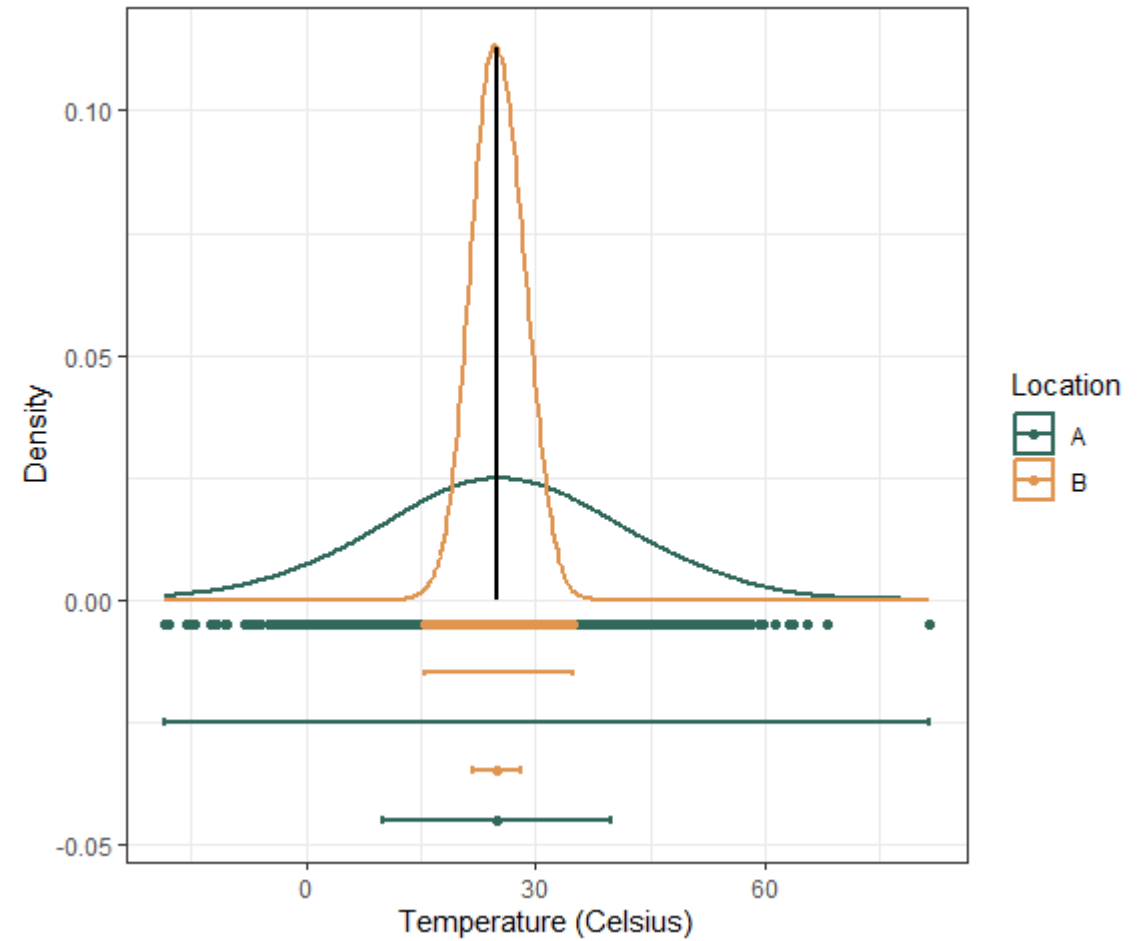


This Photo by Unknown Author is licensed under [CC BY-SA](#)

# Thống kê mô tả - Biến liên tục

Khoảng biến thiên

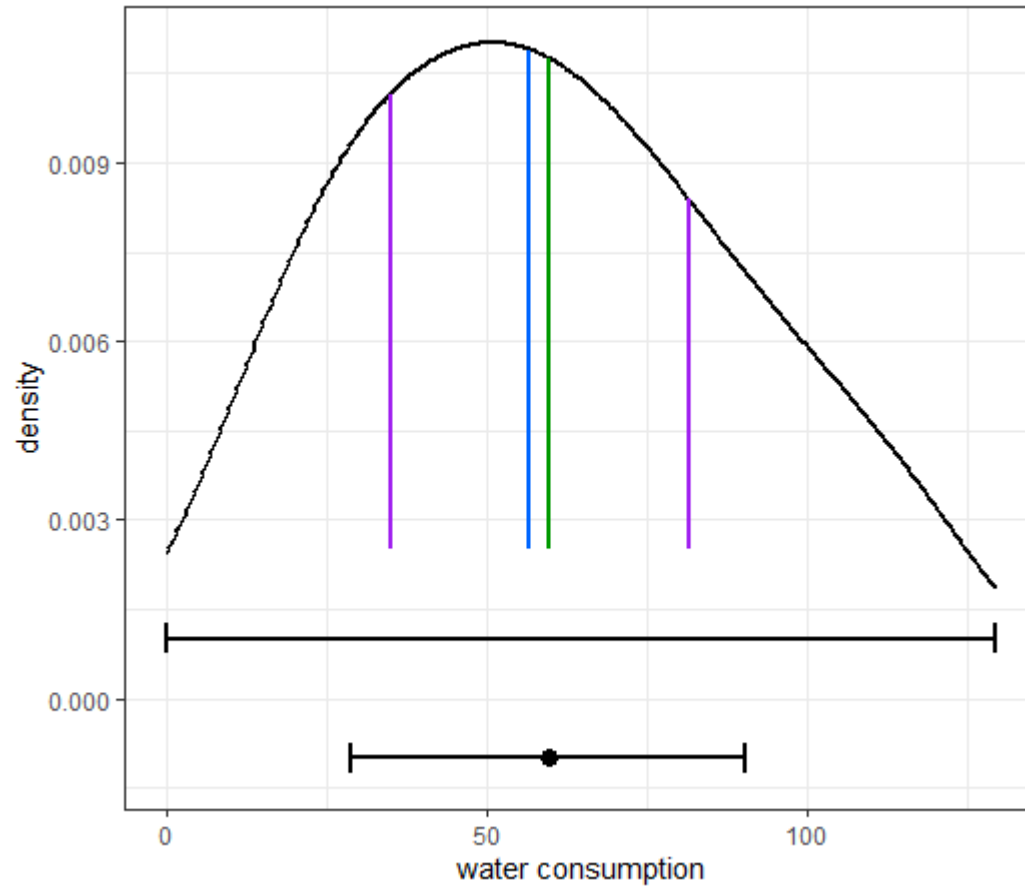
Độ lệch chuẩn  
(Phương sai)



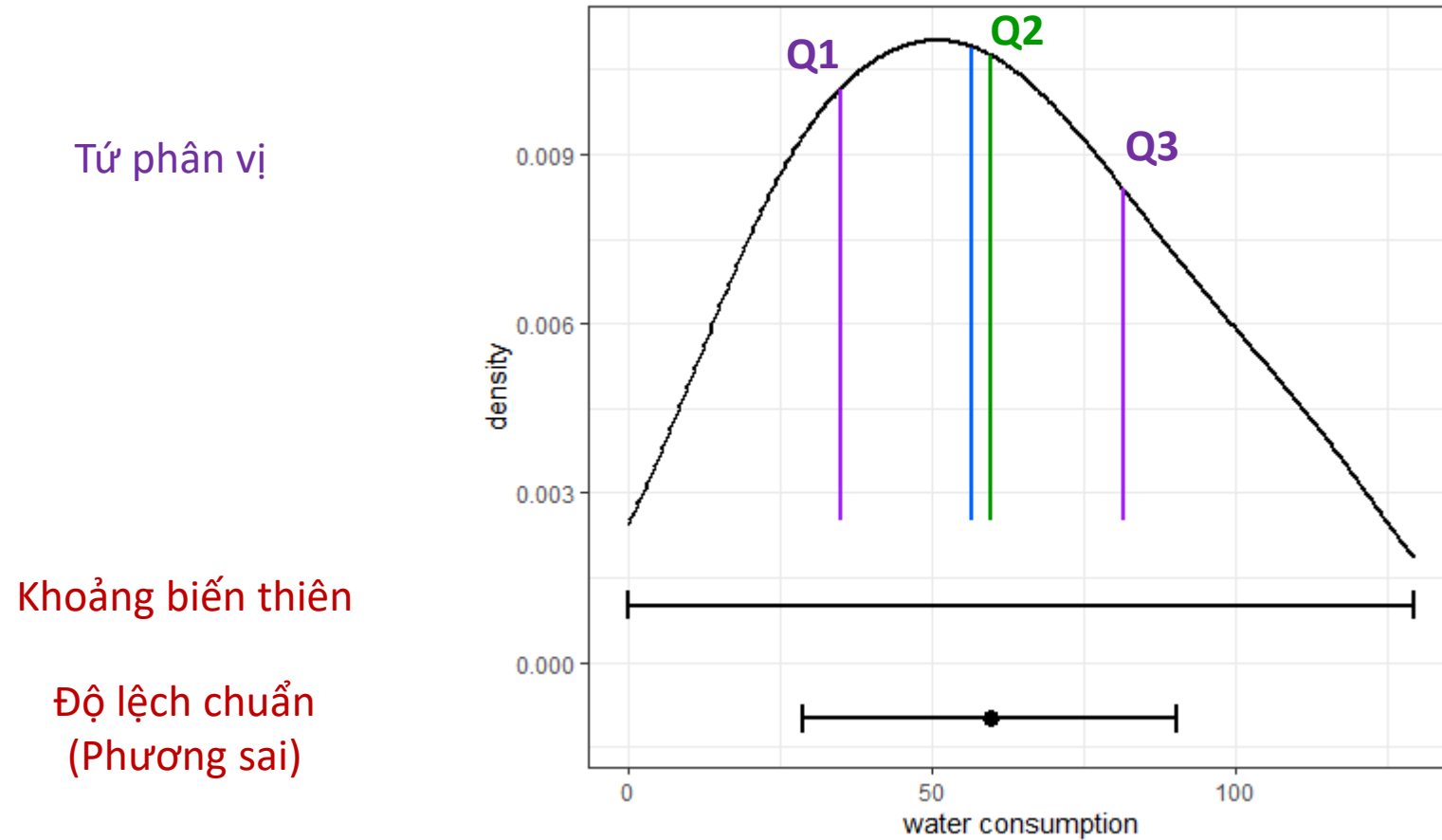
# Thống kê mô tả - Biến liên tục

Khoảng biến thiên

Độ lệch chuẩn  
(Phương sai)



# Thống kê mô tả - Biến liên tục



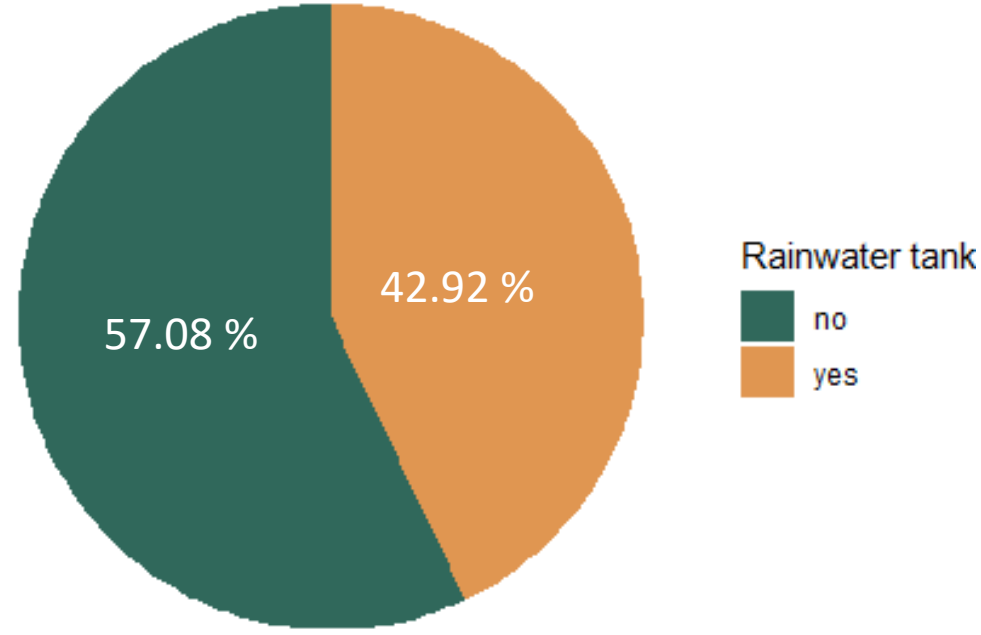


# Thống kê mô tả

- csmptv: lượng nước cấp tiêu thụ trong 1 năm
- rwtank: có bể nước mưa
- iceqac2: thu nhập của gia đình
- hhs\_tot: số thành viên trong gia đình
- cfdiwq: sự tin tưởng vào chất lượng nước cấp
- livara: diện tích nhà ở/căn hộ

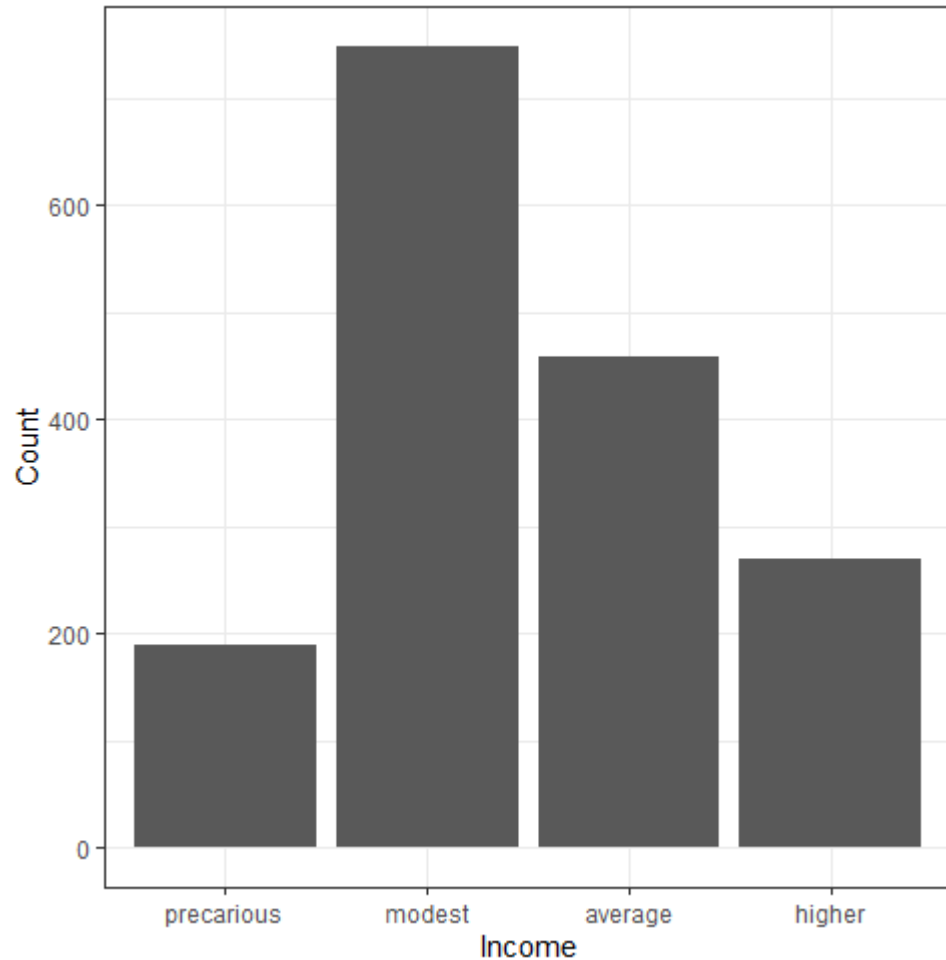
| id   | csmptv | rwtank | iceqac2    | hhs_tot | cfdiwq            | livara |
|------|--------|--------|------------|---------|-------------------|--------|
| 137  | 105    | no     | modest     | 3       | suspicious        | 129    |
| 431  | 56.99  | no     | average    | 2       | rather confident  | 120    |
| 655  | 122    | yes    | modest     | 5       | rather confident  | 130    |
| 730  | 74.57  | no     | average    | 2       | confident         | 132    |
| 780  | 30     | no     | average    | 1       | rather confident  | 70     |
| 781  | 66.36  | yes    | higher     | 2       | confident         | 162    |
| 1048 | 30.93  | yes    | modest     | 3       | suspicious        | 110    |
| 1403 | 100    | no     | higher     | 2       | confident         | 150    |
| 1405 | 52.95  | yes    | modest     | 4       | confident         | 100    |
| 1432 | 139    | no     | modest     | 3       | rather confident  | 100    |
| 1476 | 25     | yes    | average    | 2       | confident         | 100    |
| 1757 | 71.25  | yes    | average    | 3       | rather suspicious | 90     |
| 2183 | 69     | yes    | modest     | 2       | confident         | 150    |
| 2334 | 86.06  | no     | modest     | 2       | rather confident  | 90     |
| 2345 | 29.2   | yes    | precarious | 3       | confident         | 160    |
| 2375 | 33.46  | no     | average    | 2       | rather confident  | 100    |
| 2687 | 45.63  | no     | precarious | 1       | rather suspicious | 70     |
| 2704 | 126    | yes    | higher     | 4       | confident         | 200    |
| 2714 | 16.23  | yes    | modest     | 1       | confident         | 80     |
| 2752 | 105.09 | no     | higher     | 2       | confident         | 90     |

# Thống kê mô tả - Biền rời rạc



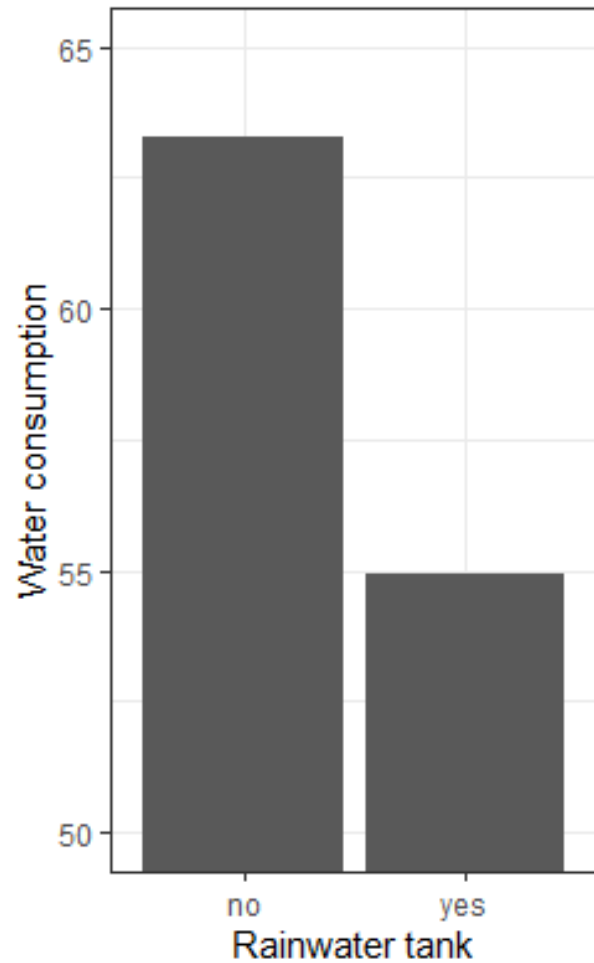
| Rainwater tank | Count | %      |
|----------------|-------|--------|
| No             | 951   | 57.08  |
| Yes            | 715   | 42.92  |
| Total          | 1666  | 100.00 |

# Thống kê mô tả - Biến rời rạc

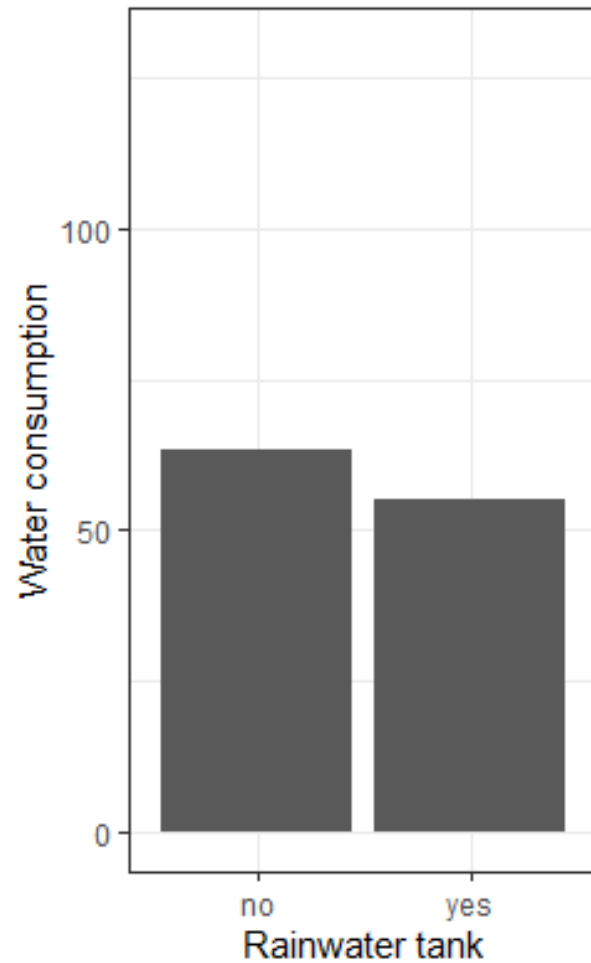
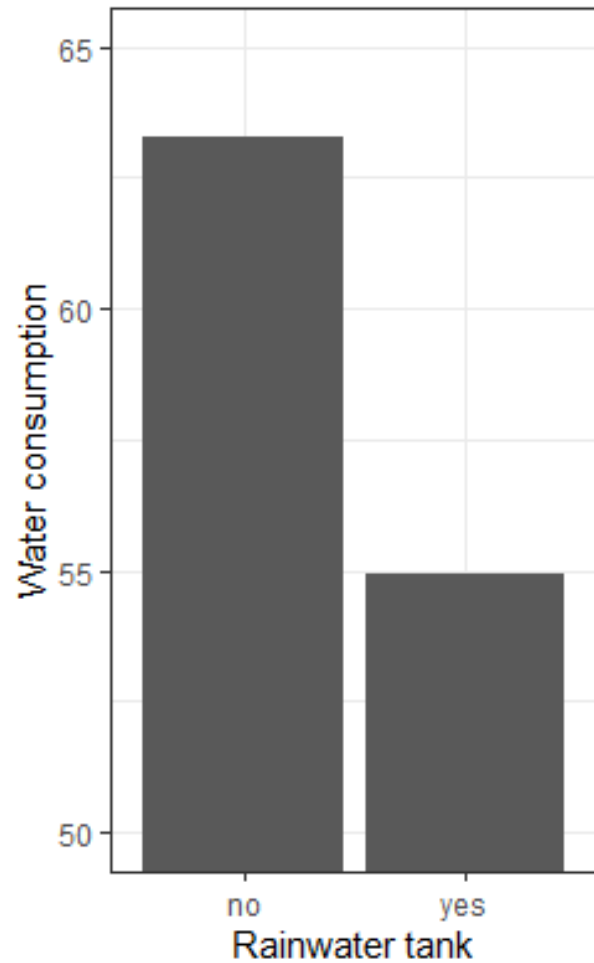


| Income     | Count | %      |
|------------|-------|--------|
| precarious | 189   | 11.34  |
| modest     | 749   | 44.96  |
| average    | 459   | 27.55  |
| higher     | 269   | 16.15  |
| Total      | 1666  | 100.00 |

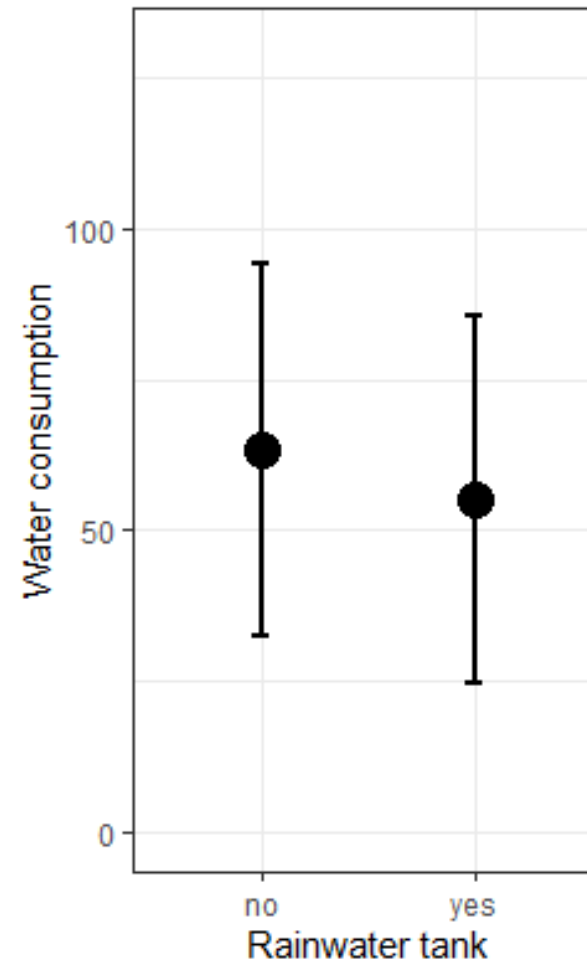
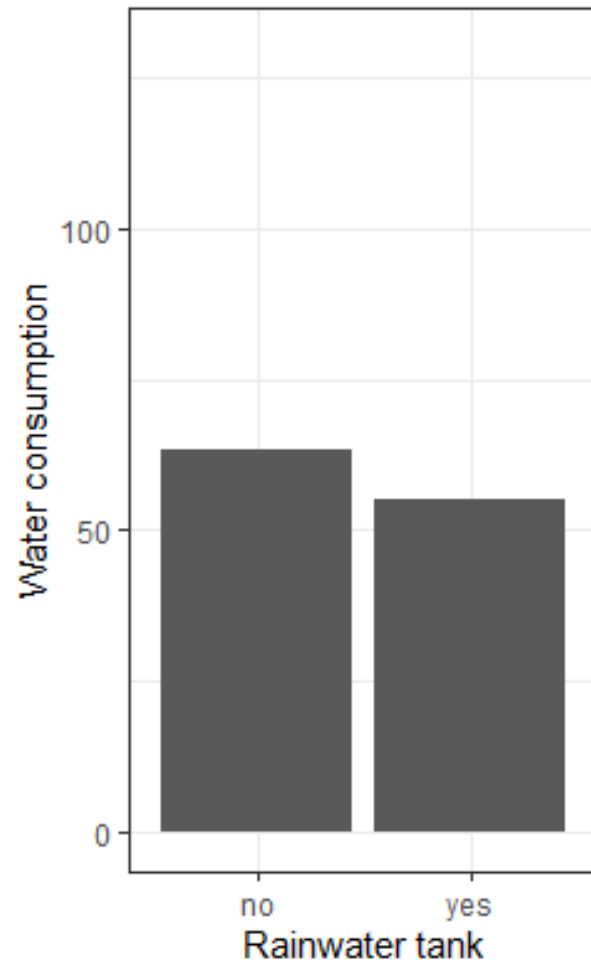
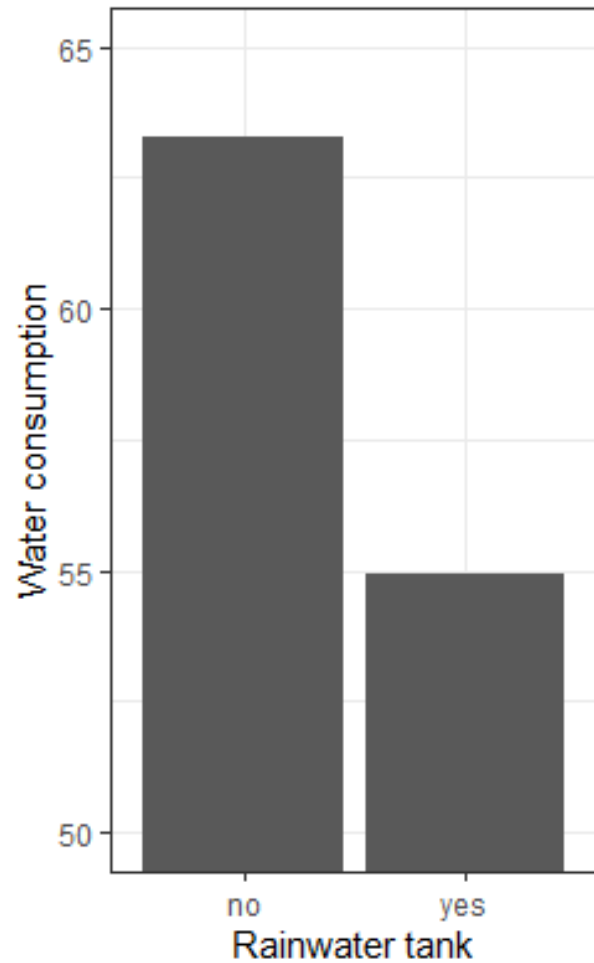
# Đồ thị - biến liên tục vs biến rời rạc



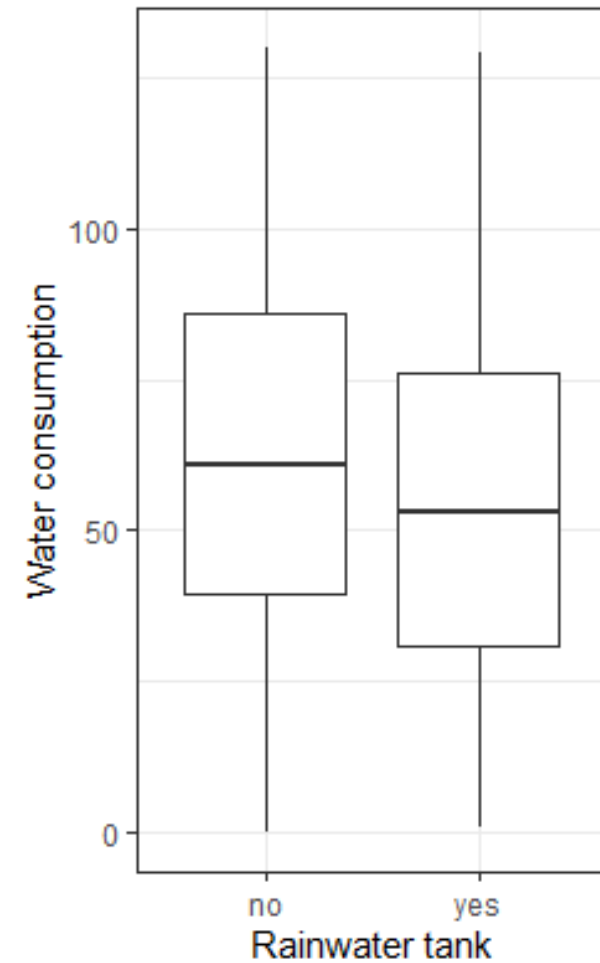
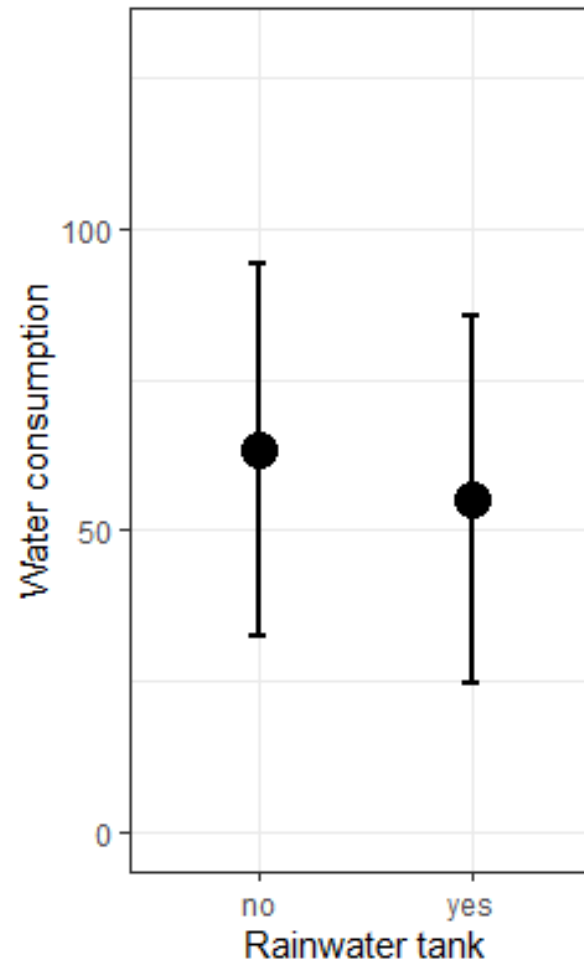
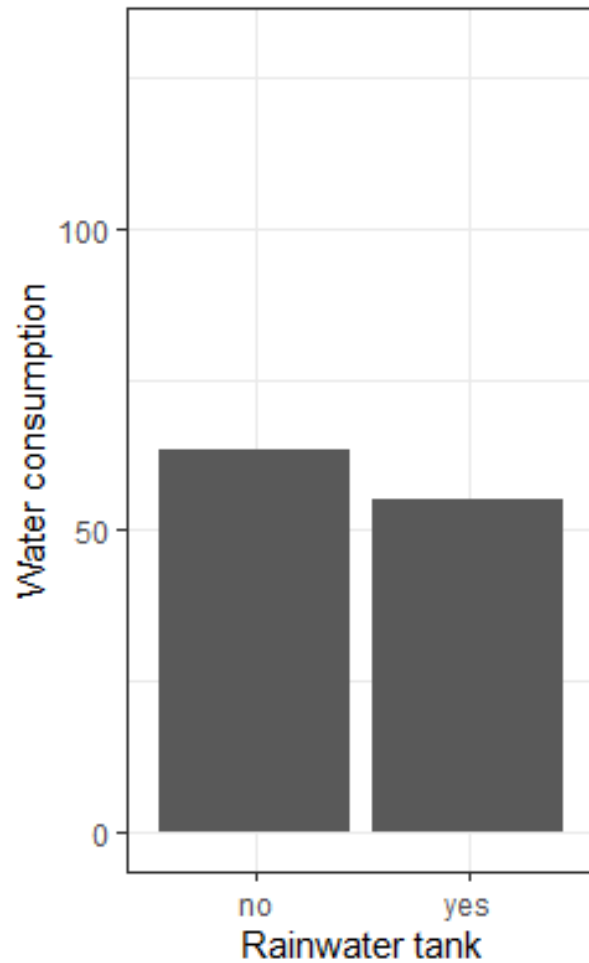
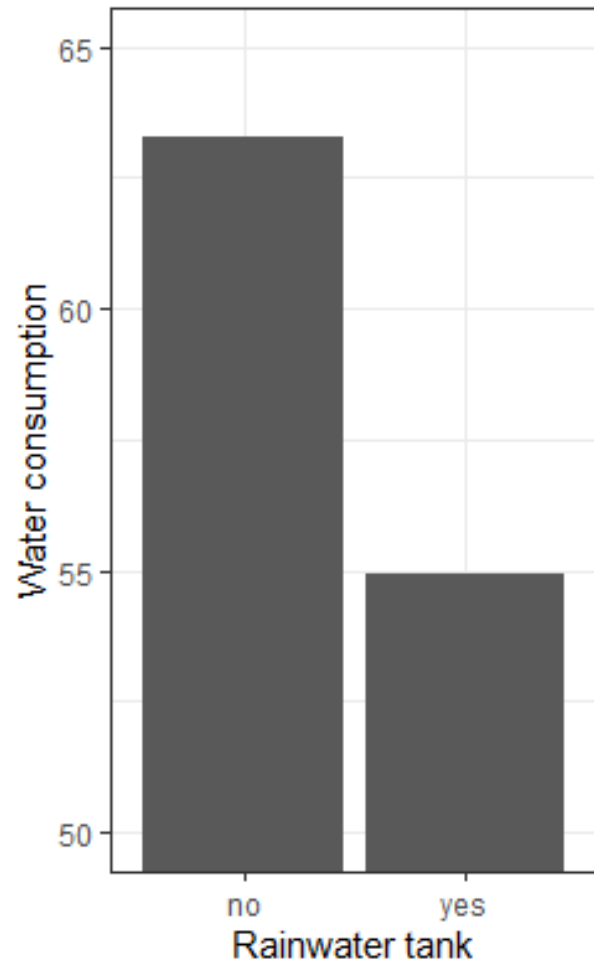
# Đồ thị - biến liên tục vs biến rời rạc



# Đồ thị - biến liên tục vs biến rời rạc

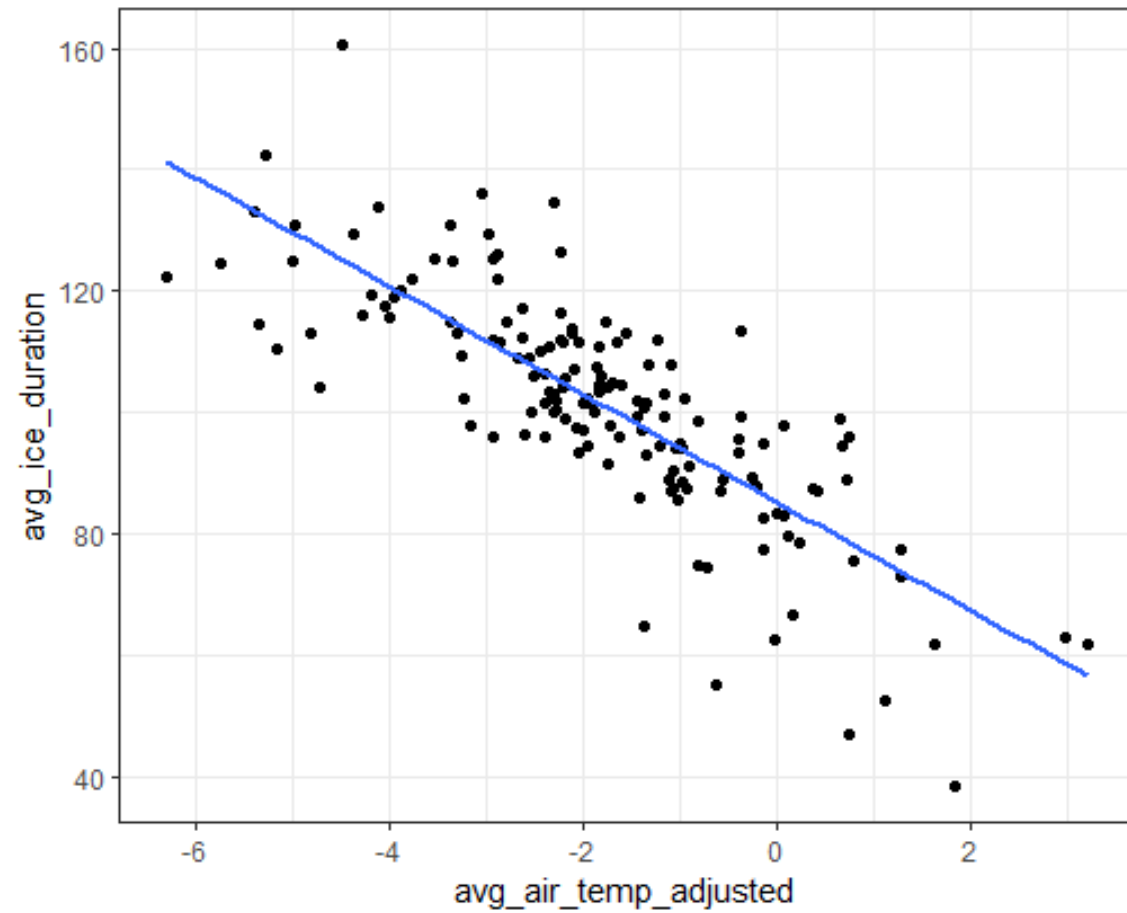


# Đồ thị - biến liên tục vs biến rời rạc



# Đồ thị - biến liên tục vs biến liên tục

- avg\_ice\_duration: số ngày mặt hồ đóng băng trong năm
- avg\_air\_temp\_adjusted: Nhiệt độ trung bình mùa đông

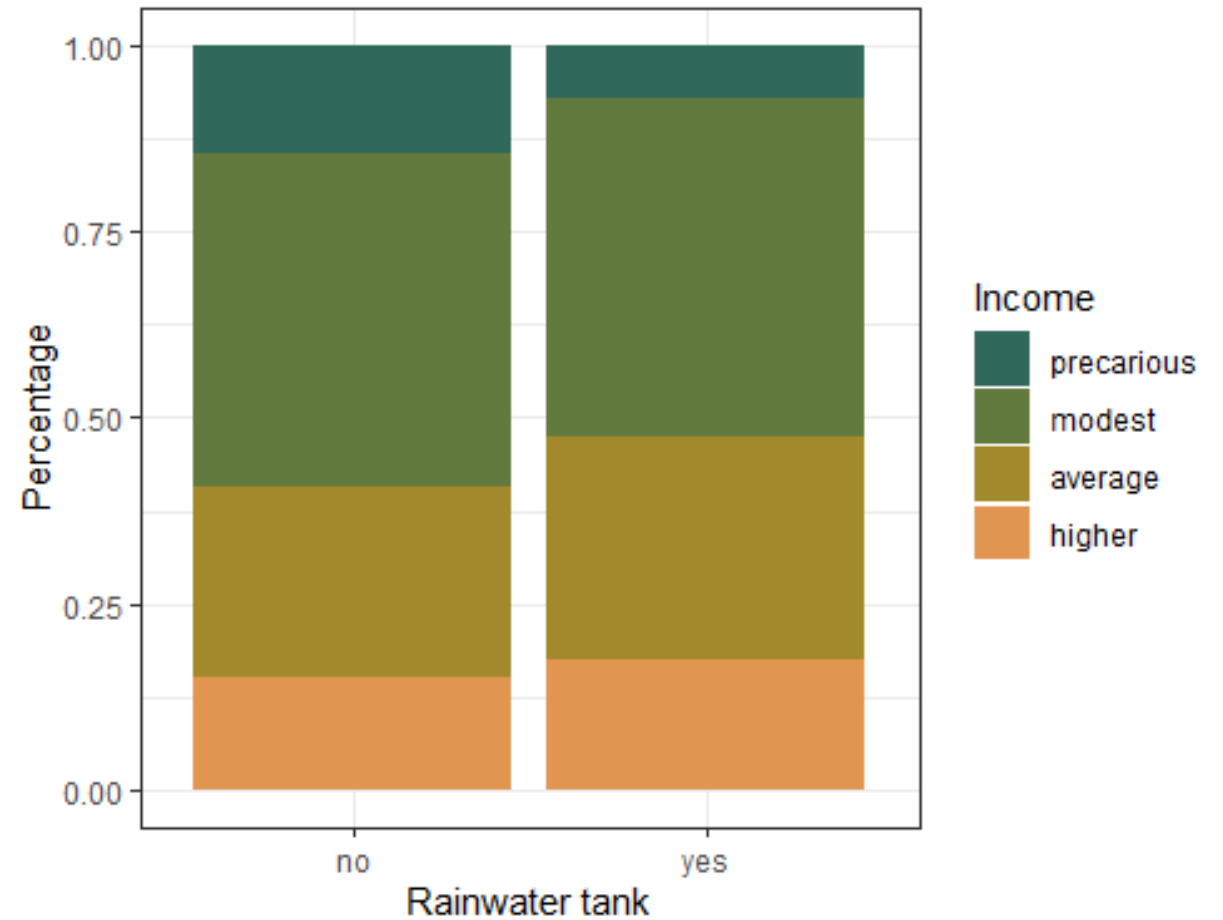




# Đồ thị - biến rời rạc vs biến rời rạc

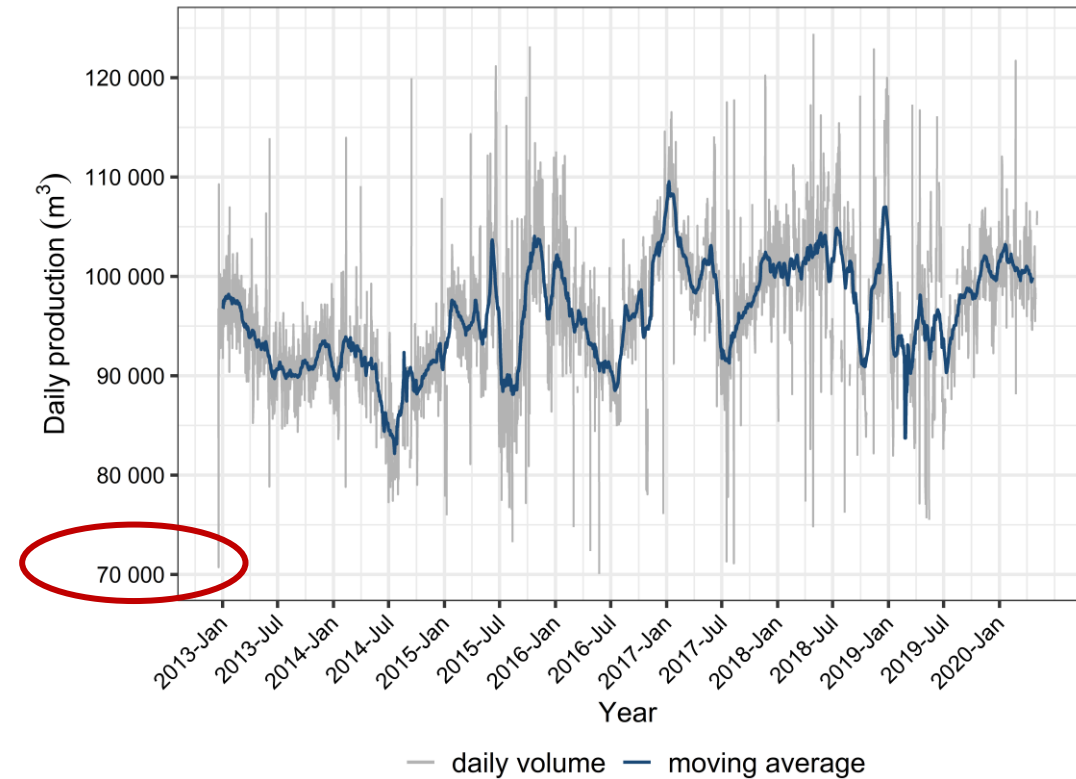
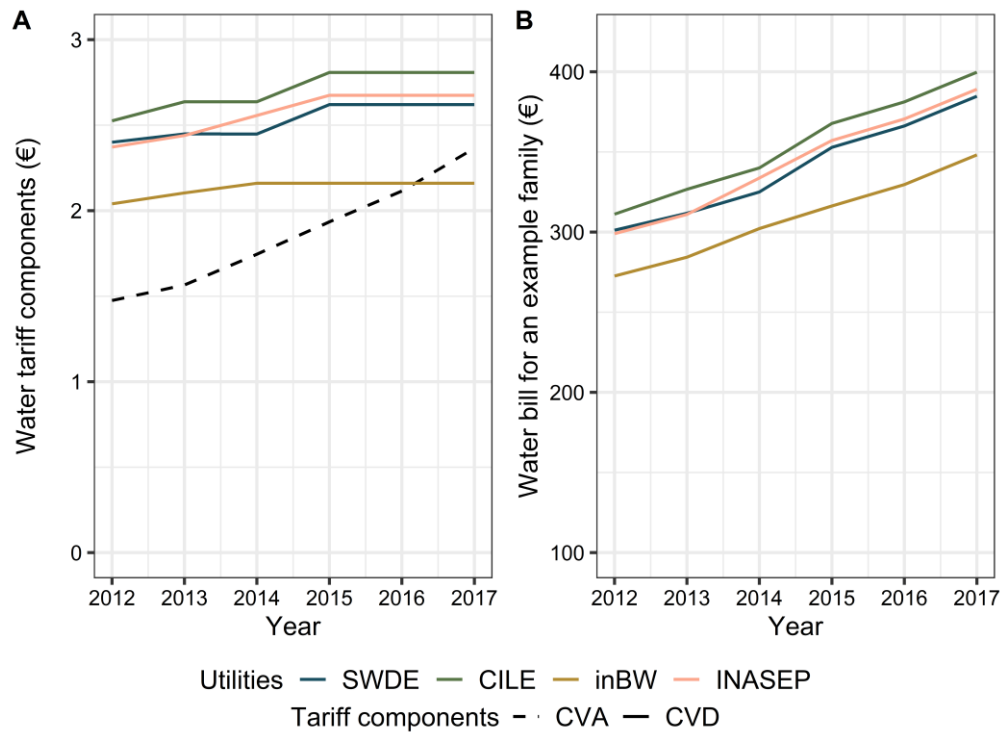
**Bảng phân phối đồng thời –  
Contingency table**

| Income     | Rainwater tank |     |
|------------|----------------|-----|
|            | no             | yes |
| precarious | 137            | 52  |
| modest     | 426            | 323 |
| average    | 245            | 214 |
| higher     | 143            | 126 |



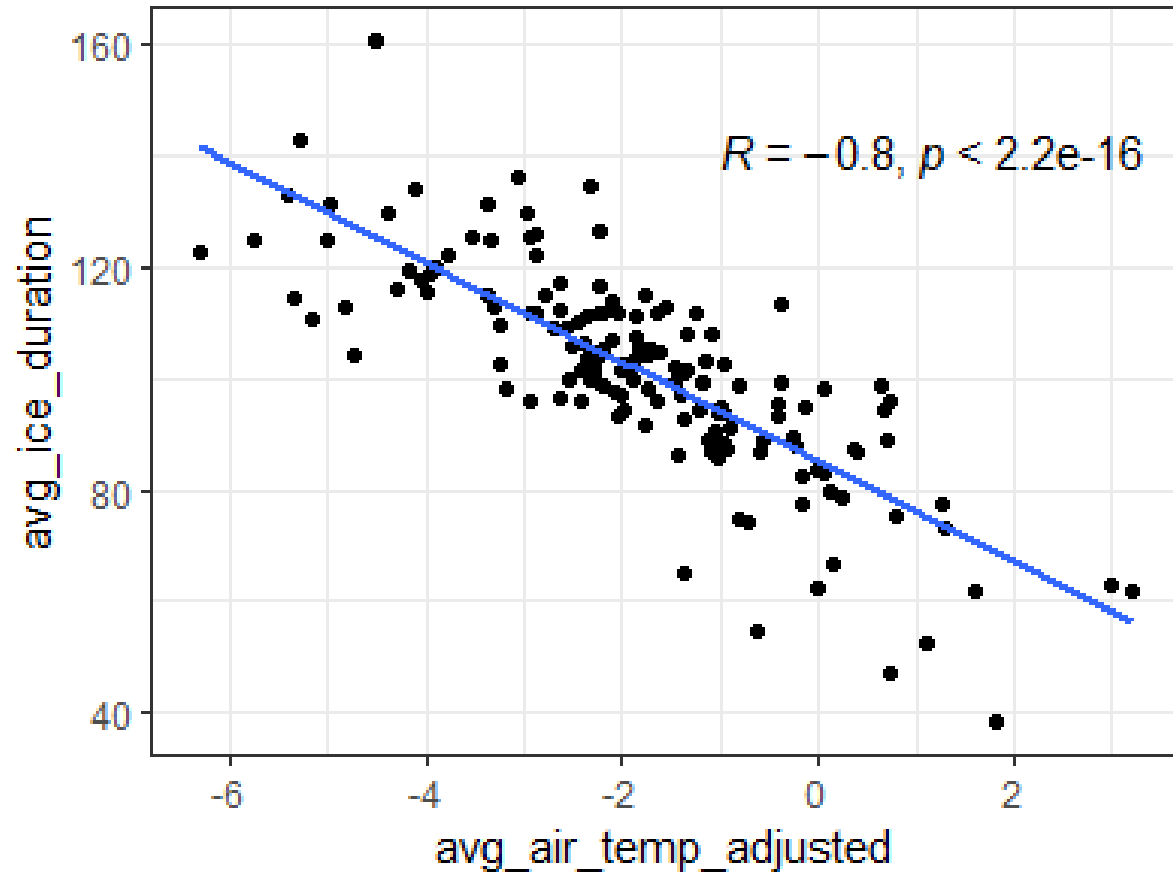
# Biến thiên theo thời gian

## Đồ thị đường – line charts

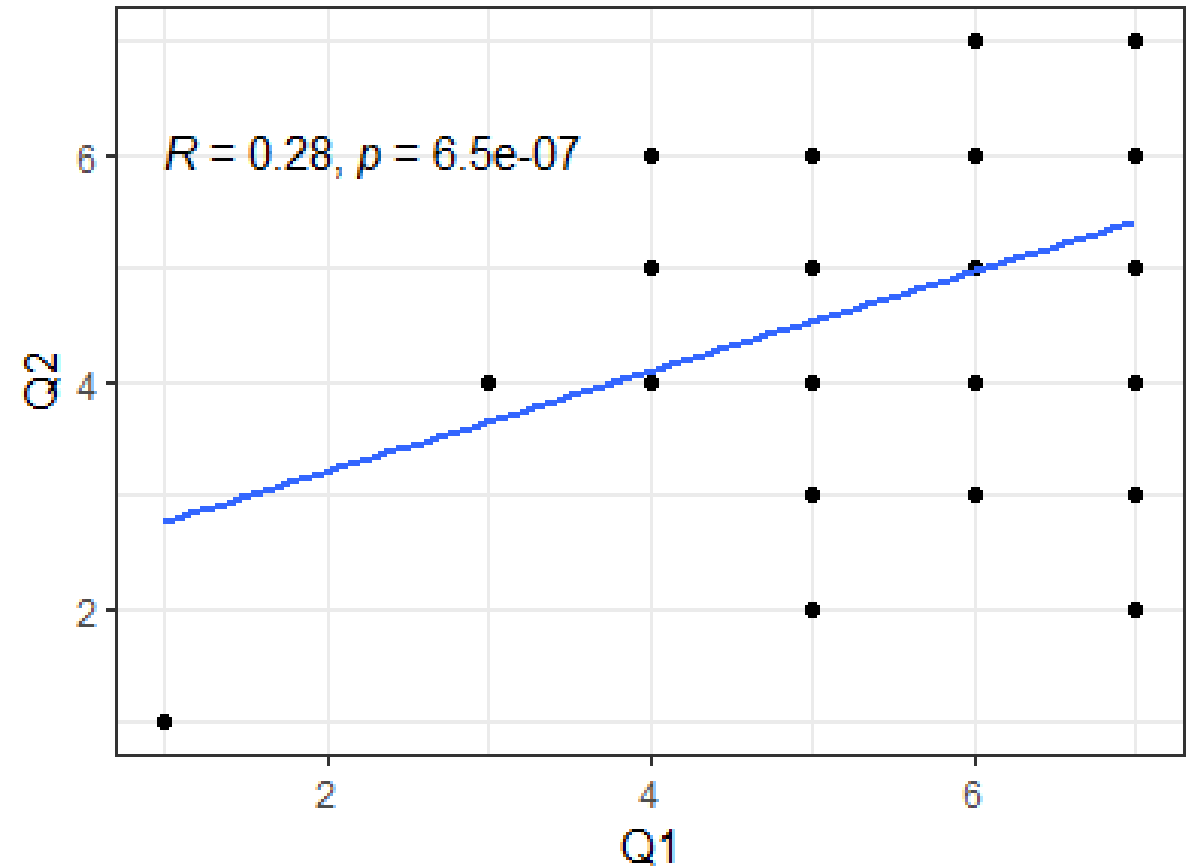


# Tương quan giữa hai biến

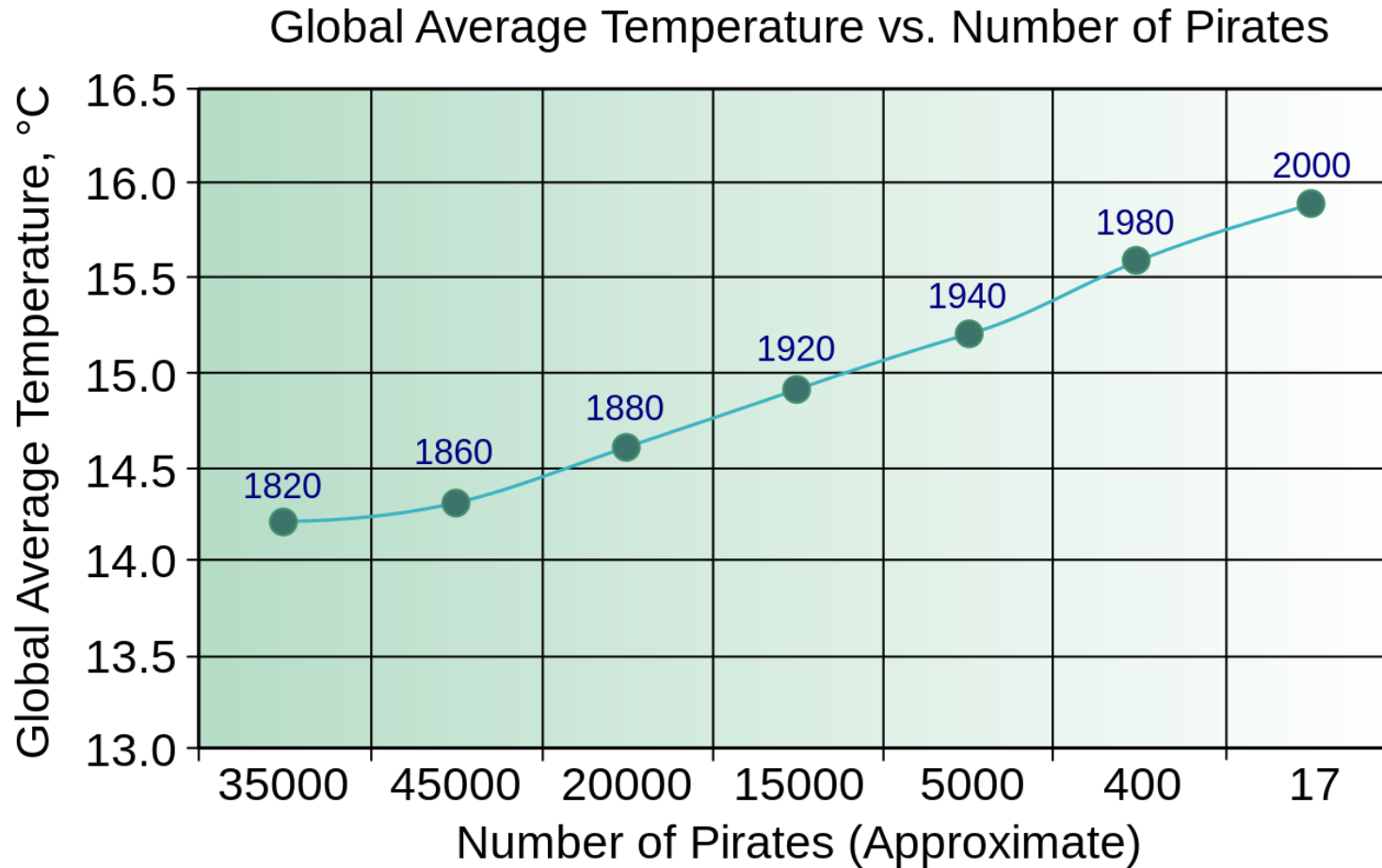
Pearson



Spearman



# Tương quan và quan hệ nhân quả

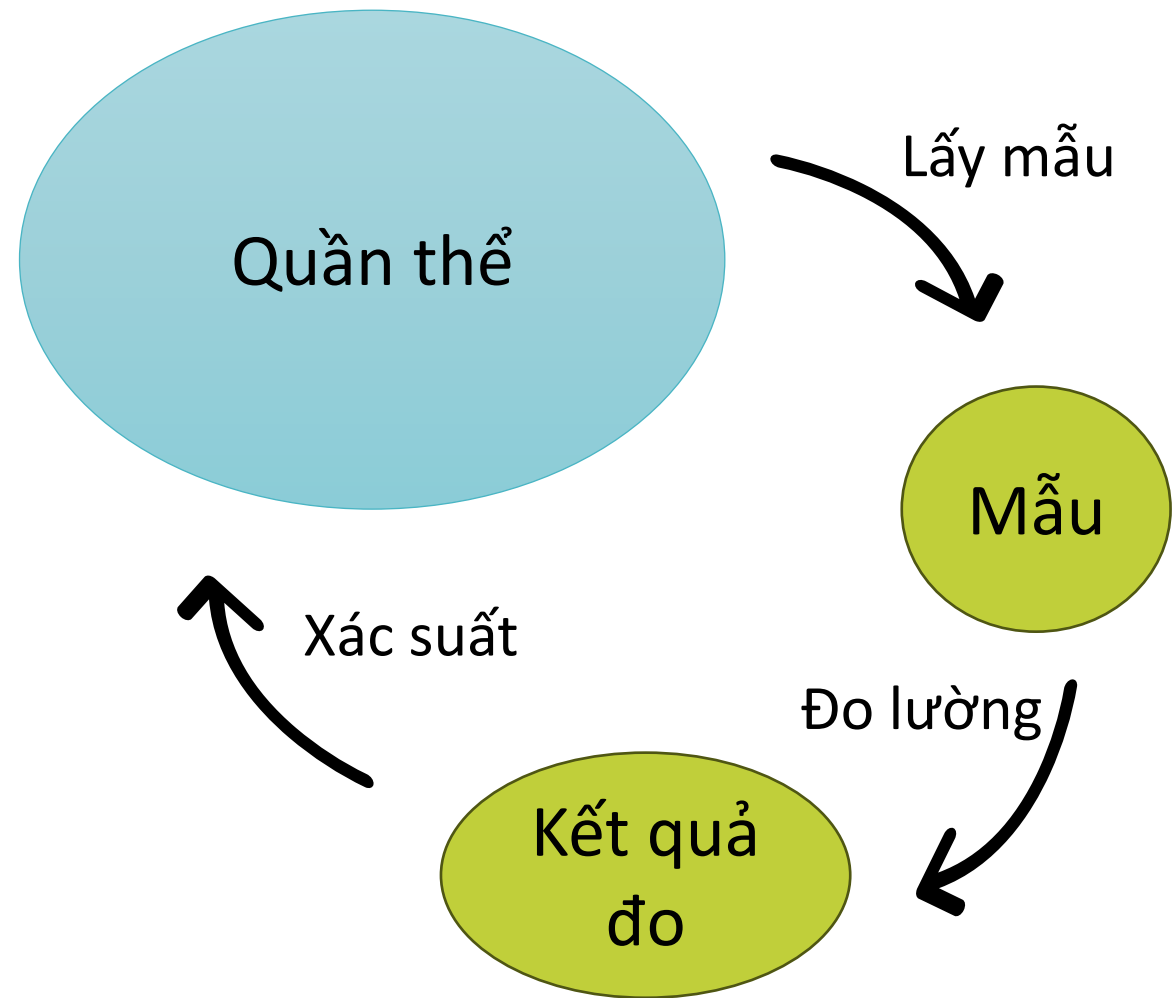


# Phân tích dữ liệu khẳng định

---

# 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



# Kiểm định thống kê (Hypothesis testing)

- Giả thuyết trống/không (Null hypothesis)
- Giả thuyết thay thế (Alternative hypothesis)

$$H_0: \mu = 0$$

$$H_a: \mu \neq 0$$

- Giả thuyết thú vị với nhà nghiên cứu luôn là giả thuyết thay thế
- Giả thuyết được kiểm định luôn là giả thuyết không/trống

# Bài tập

- Quá trình đô thị hóa làm tăng đáng kể nhiệt độ và thay đổi mô hình lượng mưa
- Rừng ngập mặn ven biển giữ cacbon nhiều hơn rừng trong đất liền và đóng vai trò quan trọng trong việc giảm nhẹ biến đổi khí hậu
- Các cộng đồng tích cực trong việc thực hiện chiến lược phục hồi và thích ứng sẽ trải qua ít tác động tiêu cực hơn từ các sự kiện liên quan đến biến đổi khí hậu



# Trị số p

- Xác suất để thu được kết quả tương tự hoặc cực đoan hơn khi giả thiết rằng giả thuyết trống là đúng
- Trị số  $p < 0.05$ : có ý nghĩa về mặt thống kê

|                    | $H_0$ đúng | $H_0$ sai   |
|--------------------|------------|-------------|
| Bác bỏ $H_0$       | Lỗi loại I | ✓           |
| Không bác bỏ $H_0$ | ✓          | Lỗi loại II |

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

# Ví dụ về kiểm định thống kê

- So sánh lượng nước trung bình tiêu thụ giữa nhóm hộ gia đình có bể chứa nước mưa và không

# Ví dụ về kiểm định thống kê

- So sánh lượng nước trung bình tiêu thụ giữa nhóm hộ gia đình có bể chứa nước mưa và không

Kiểm định t

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

| $\mu_1$ | $\mu_2$ | Trị số p             |
|---------|---------|----------------------|
| 63.26   | 54.95   | $4.3 \times 10^{-8}$ |

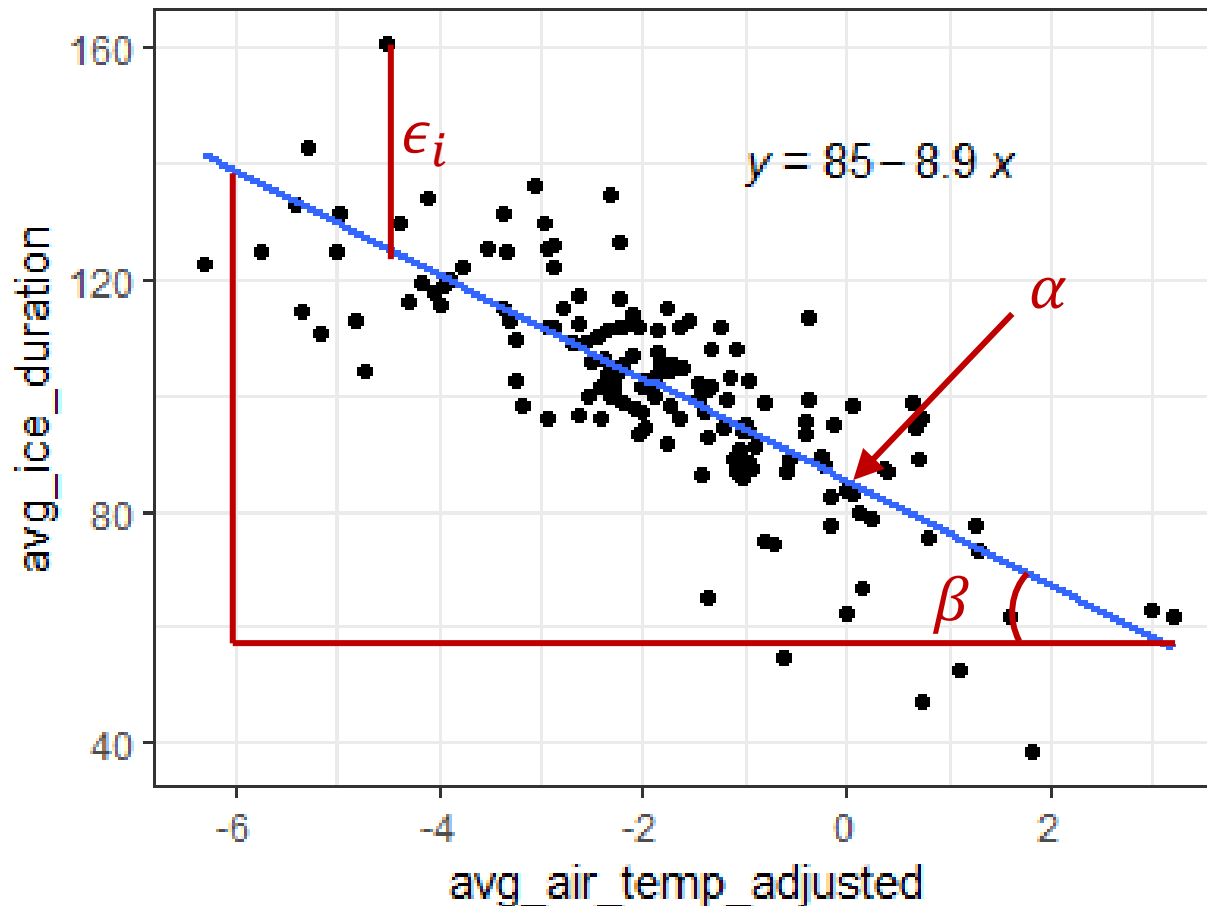
# Giới thiệu các kỹ thuật phân tích định lượng

---

# Một số kỹ thuật định lượng

- 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)
- ...

# Hồi quy tuyến tính đơn giản



- Nghiên cứu mối quan hệ giữa **biến độc lập X** (independent/ explanatory variable, predictor) và **biến phụ thuộc Y** (dependent/out come variable)
- Dự đoán giá trị của Y dựa trên giá trị của X
- Y là **biến liên tục**

$$Y_i = \alpha + \beta X_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

# Hồi quy tuyến tính đơn giản

Ví dụ: Mối quan hệ giữa **lượng nước tiêu thụ** và **diện tích nhà ở**

$$csmptv = \alpha + \beta * livara + \epsilon_i$$

$$\alpha = 45.31$$

$$\beta = 0.11$$

# Hồi quy tuyến tính đơn giản

Ví dụ: Mối quan hệ giữa **lượng nước tiêu thụ** và **diện tích nhà ở**

$$csmptv = \alpha + \beta * livara + \epsilon_i$$

$$\alpha = 45.31$$

$$\beta = 0.11$$

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

$$\text{Trị số } p = 2.24 * 10^{-15}$$

$$R^2 = 0.037$$



# Hồi quy tuyến tính đơn giản với biến rời rạc

Ví dụ: Mối quan hệ giữa **lượng nước tiêu thụ** và **sử dụng nước mưa**

$$csmptv = \alpha + \beta * rwtank + \epsilon_i$$

$$\alpha = 63.26$$

$$\beta = -8.31$$

$$\text{Trị số } p = 4.3 \times 10^{-8}$$

$$R^2 = 0.018$$

Kiểm định t

| $\mu_1$ | $\mu_2$ | Trị số p             |
|---------|---------|----------------------|
| 63.26   | 54.95   | $4.3 \times 10^{-8}$ |

# Hồi quy tuyến tính bội

Ví dụ: Giải thích sự biến thiên của **lượng nước tiêu thụ** bởi **diện tích nhà ở** và **sử dụng nước mưa**

$$csmptv = \alpha + \beta_1 * livara + \beta_2 * rwtank + \epsilon_i$$

# Hồi quy tuyến tính bội

Ví dụ: Giải thích sự biến thiên của **lượng nước tiêu thụ** bởi **diện tích nhà ở** và **sử dụng nước mưa**

$$csmptv = \alpha + \beta_1 * livara + \beta_2 * rwtank + \epsilon_i$$

$$\alpha = 63.26$$

$$\beta_1 = 0.12$$

$$\beta_2 = -10.54$$

$$\text{Trị số } p < 2 * 10^{-16}$$

$$\text{Trị số } p = 2.7 * 10^{-12}$$

$$R^2 = 0.065$$

# Hồi quy tuyến tính bội

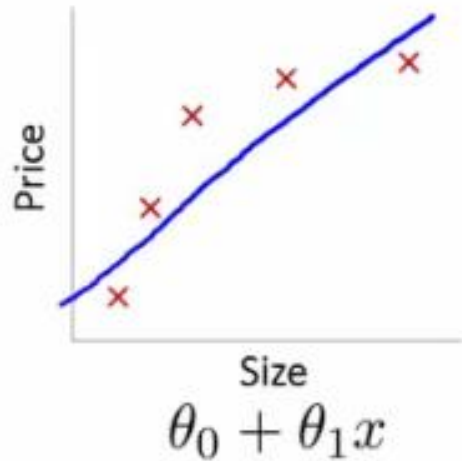
Ví dụ: Giải thích sự biến thiên của **lượng nước tiêu thụ** bởi **diện tích nhà ở** và **sử dụng nước mưa**

$csmptv$

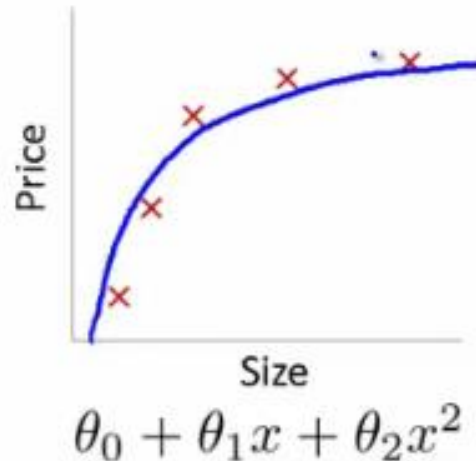
$$= \alpha + \beta_1 * livara + \beta_2 * livara^2 + \beta_3 * rwtank + \beta_4 * livara * rwtank + \epsilon_i$$



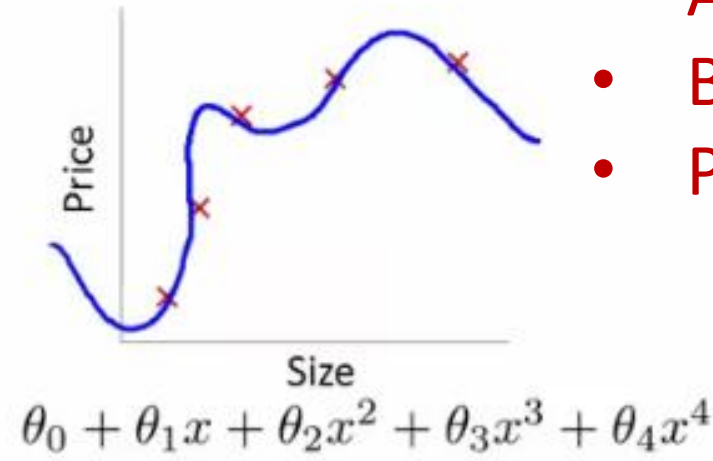
# Quá khớp hoặc thiếu khớp với số liệu



High bias  
(underfit)



"Just right"

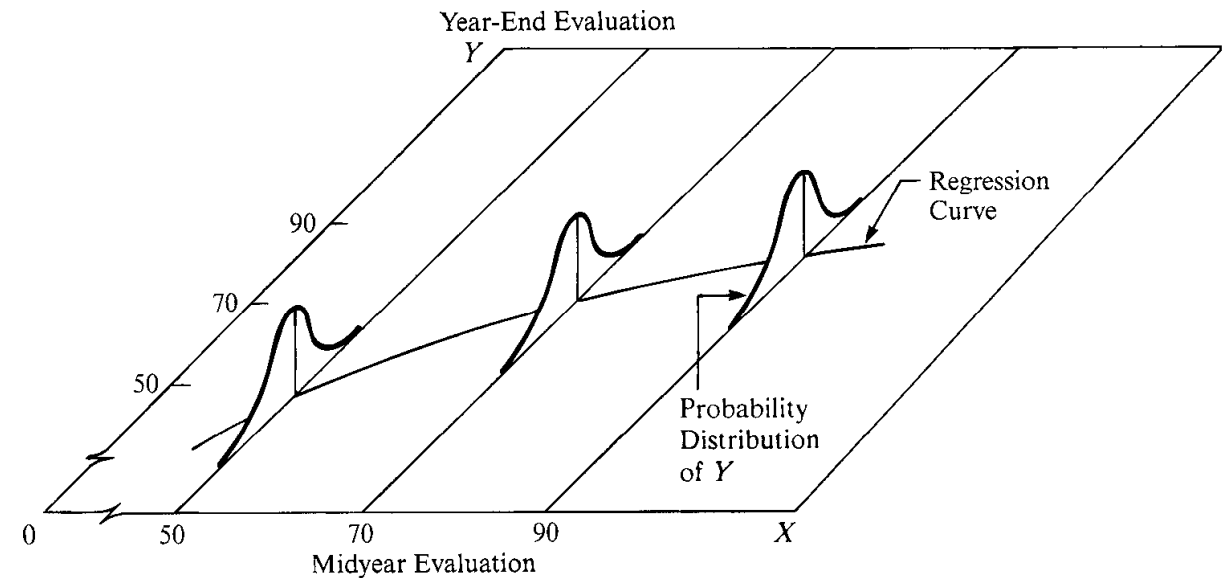


High variance  
(overfit)

- Adjusted  $R^2$
- AIC
- BIC
- Predictive power

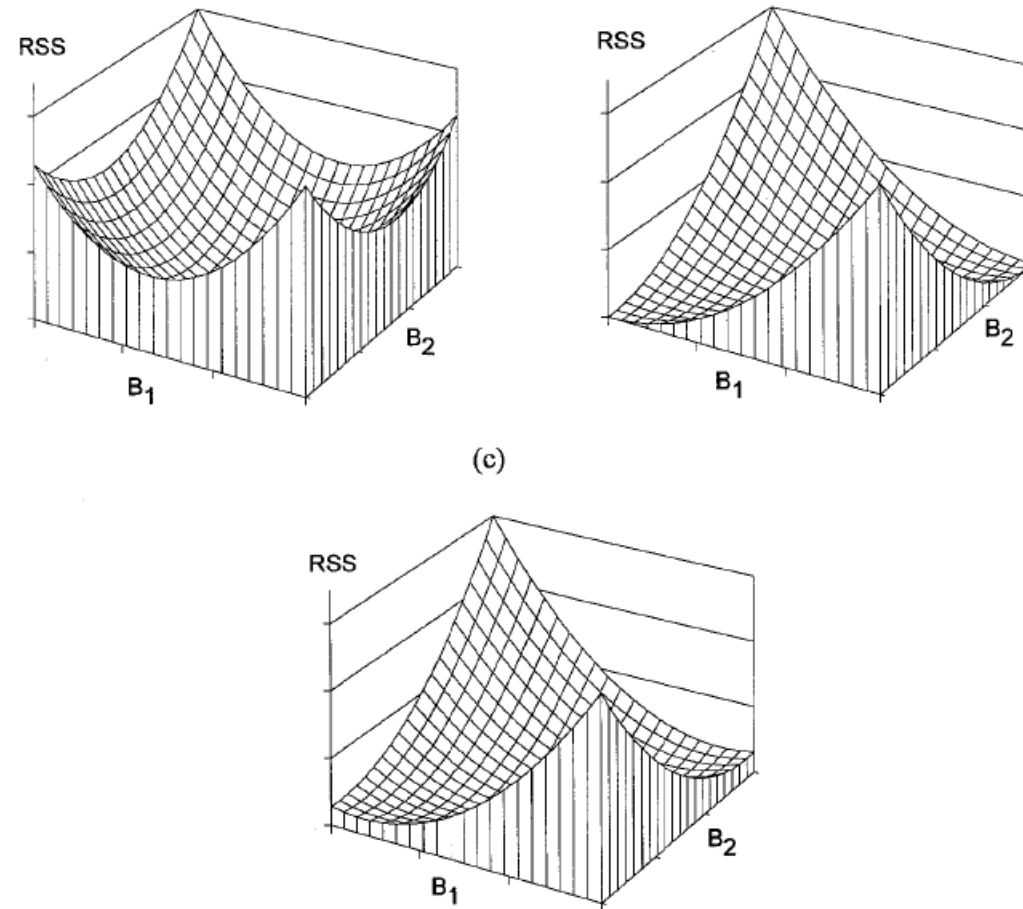
# Hồi quy tuyến tính – Các giả định

- Y là biến liên tục
- Mỗi 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)

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_1 X_{i1}^2 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

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)

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_1 X_{i1}^2 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Hồi quy tuyến tính bội

$Y$ : liên tục/định lượng

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

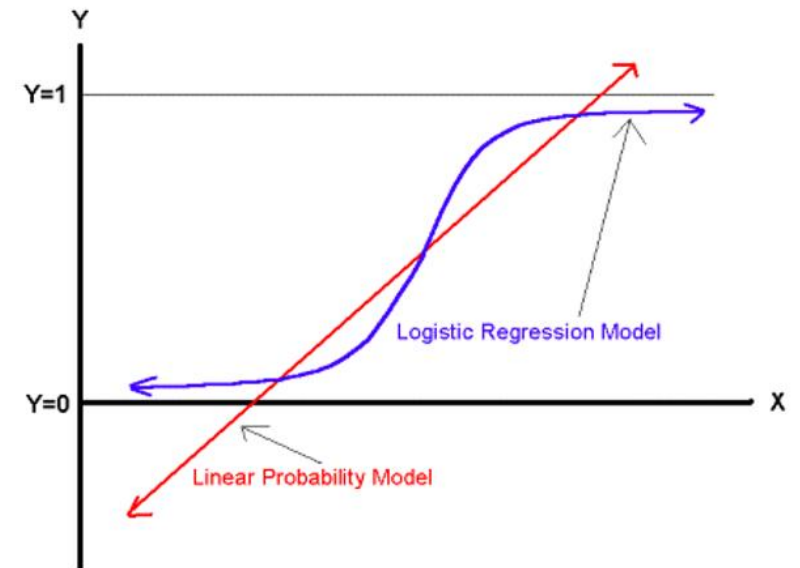
**Khi  $Y$  là biến rời rạc?**

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

Định danh: Multinomial logistic regression

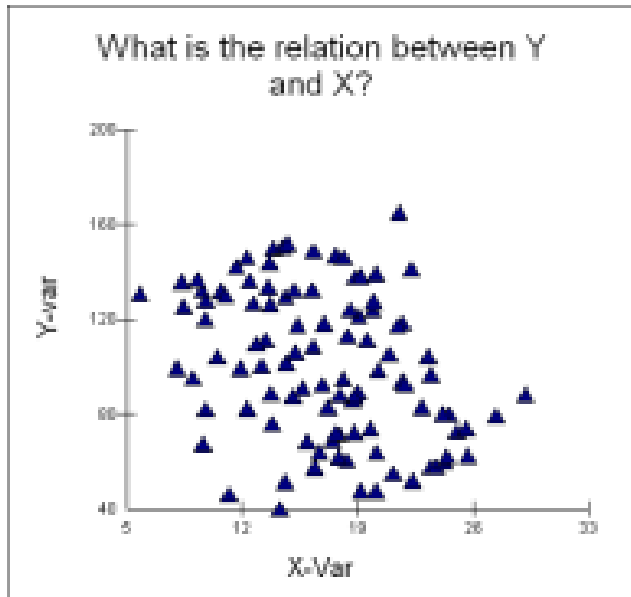
Thứ bậc: Cumulative logistic regression

Biến đếm: Poisson regression



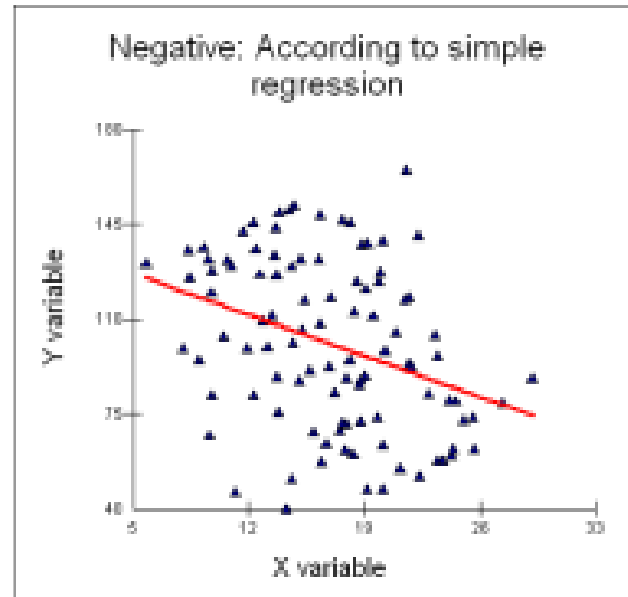
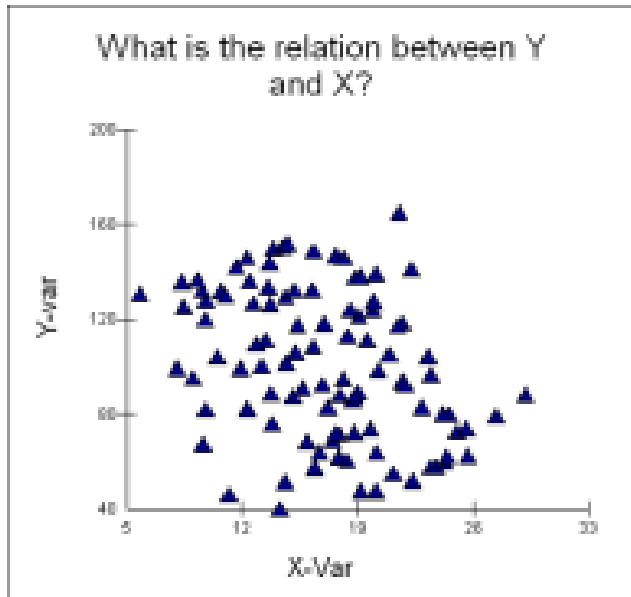
# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

## Multilevel model/Mixed effect model



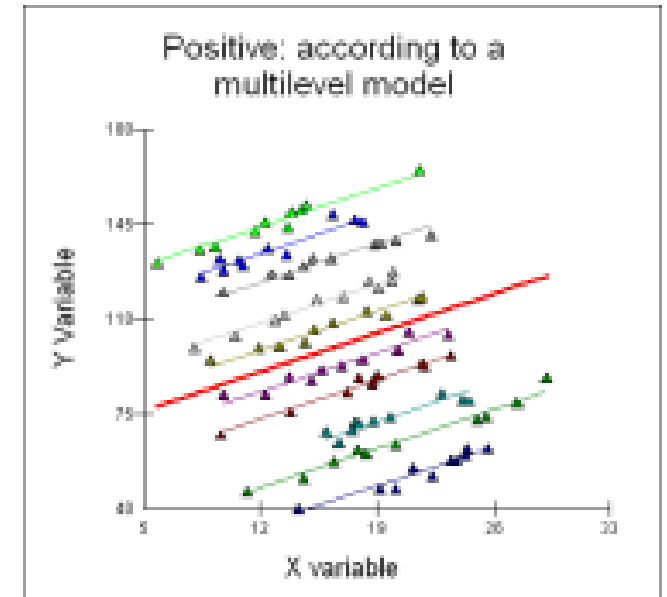
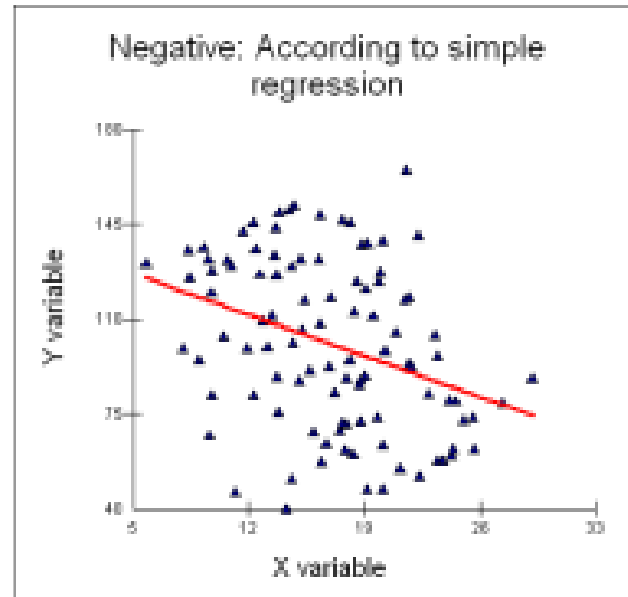
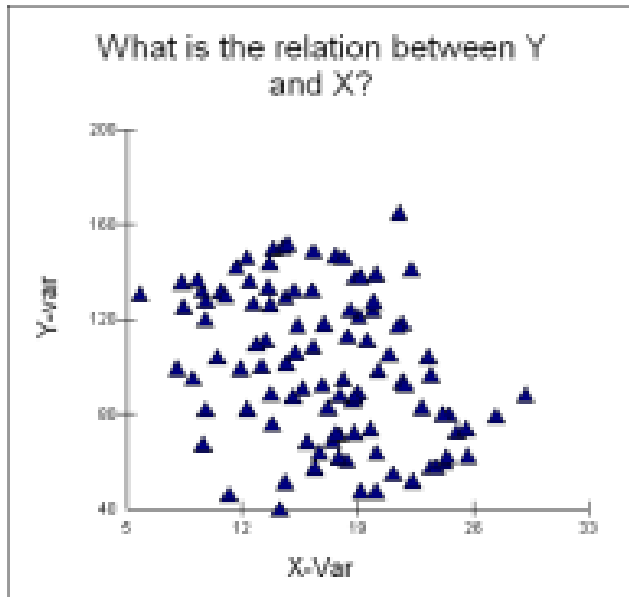
# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

## Multilevel model/Mixed effect model



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

## Multilevel model/Mixed effect model

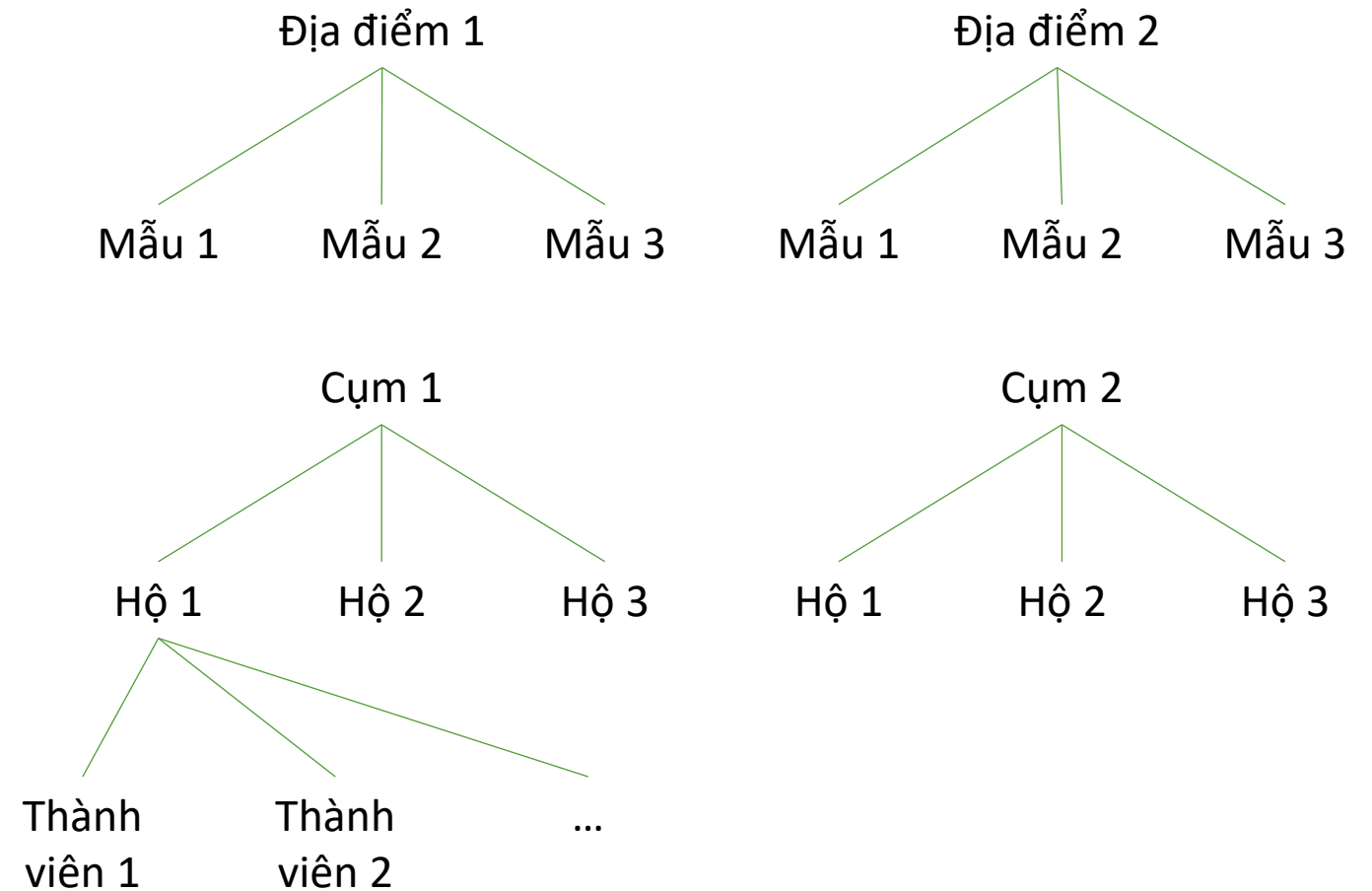


# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

## Multilevel model/Mixed effect model

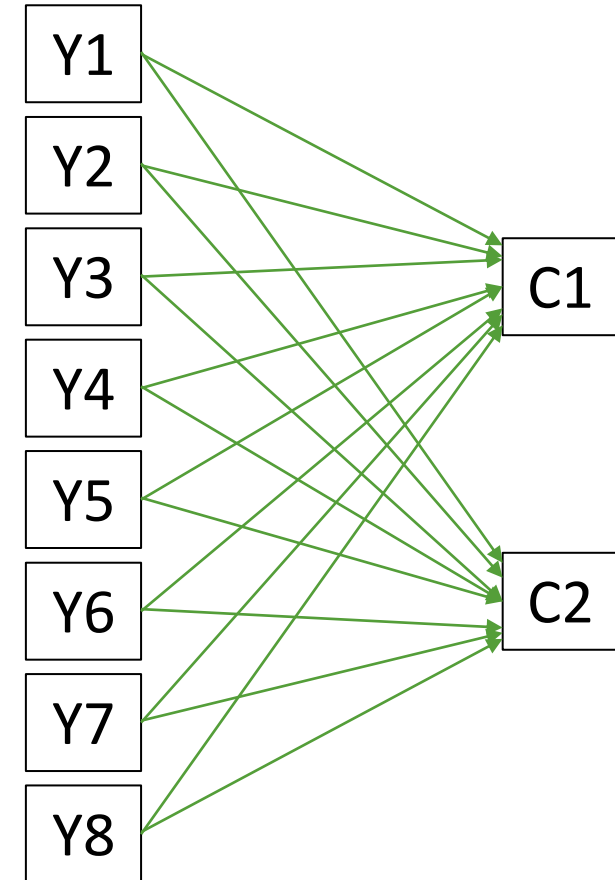


Image by [Chelsea Parlett-Pelleriti](#)



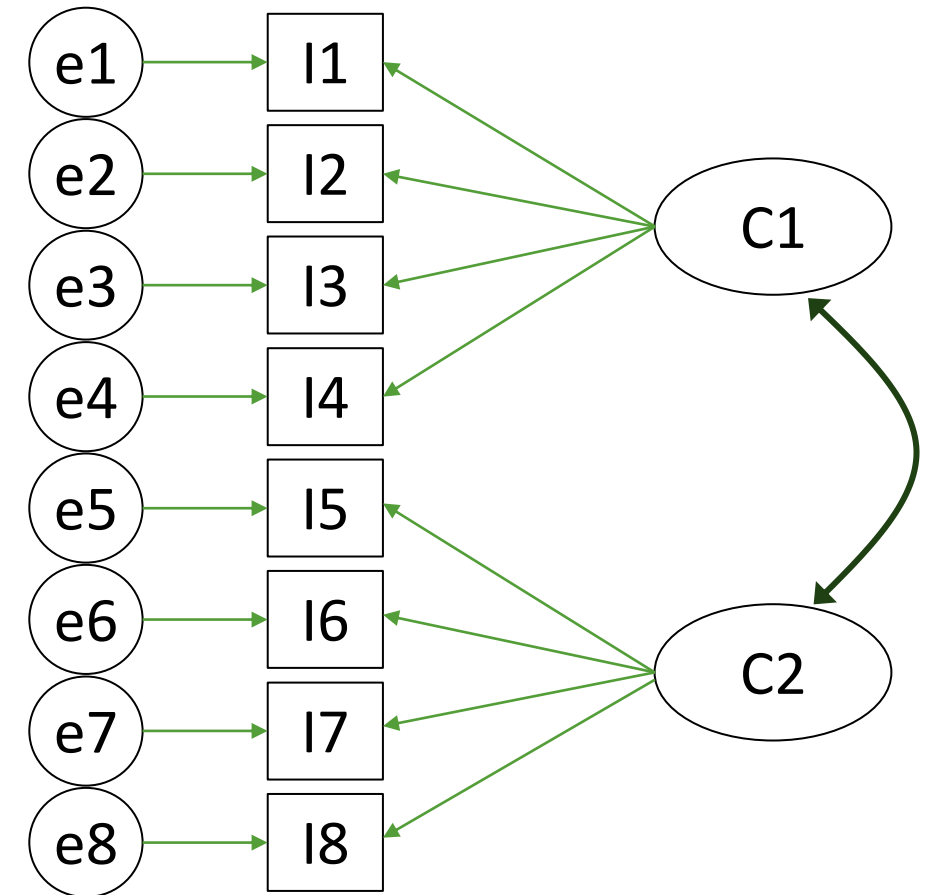
# Phân tích thành phần chính (PCA - Principal component analysis)

- 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)



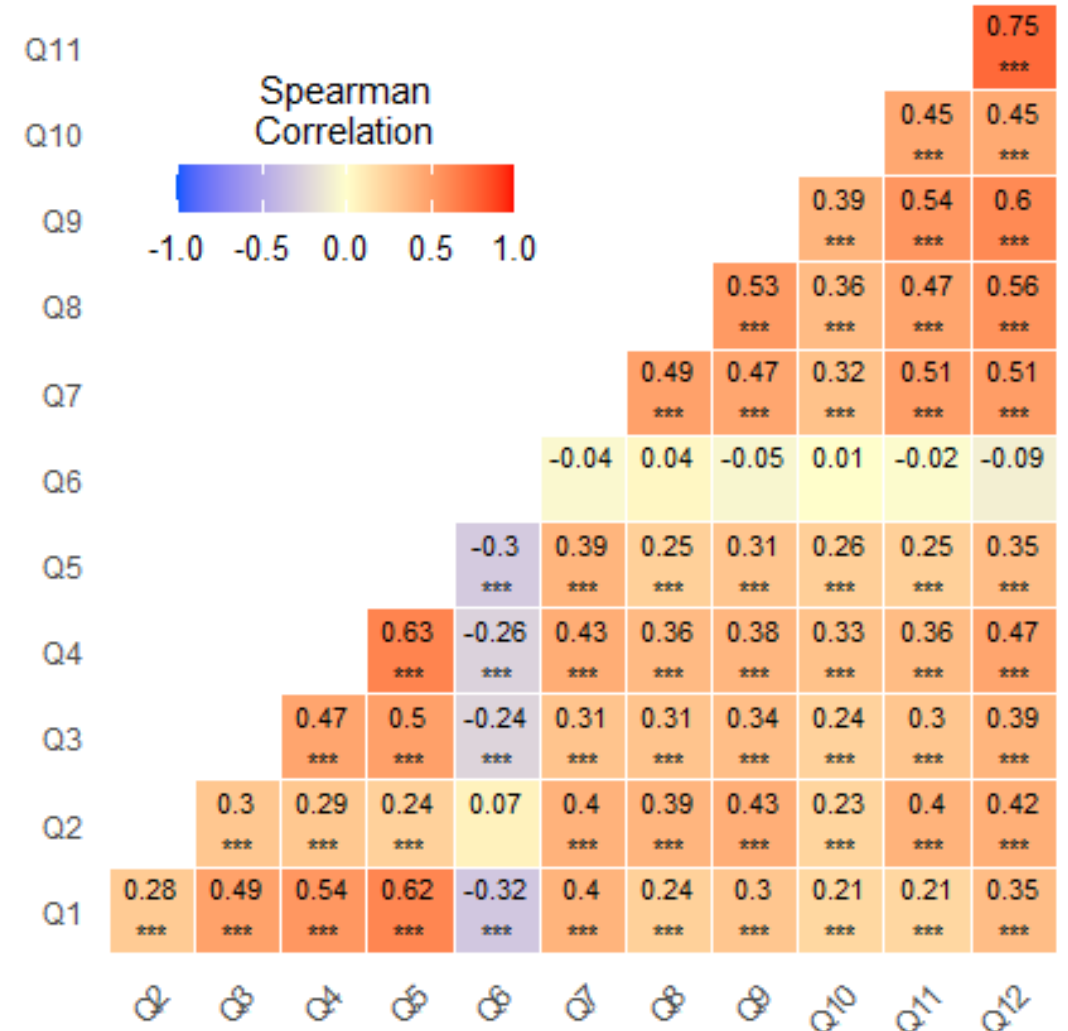
# Phân tích nhân tố khẳng định (CFA)

## 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.                                      |

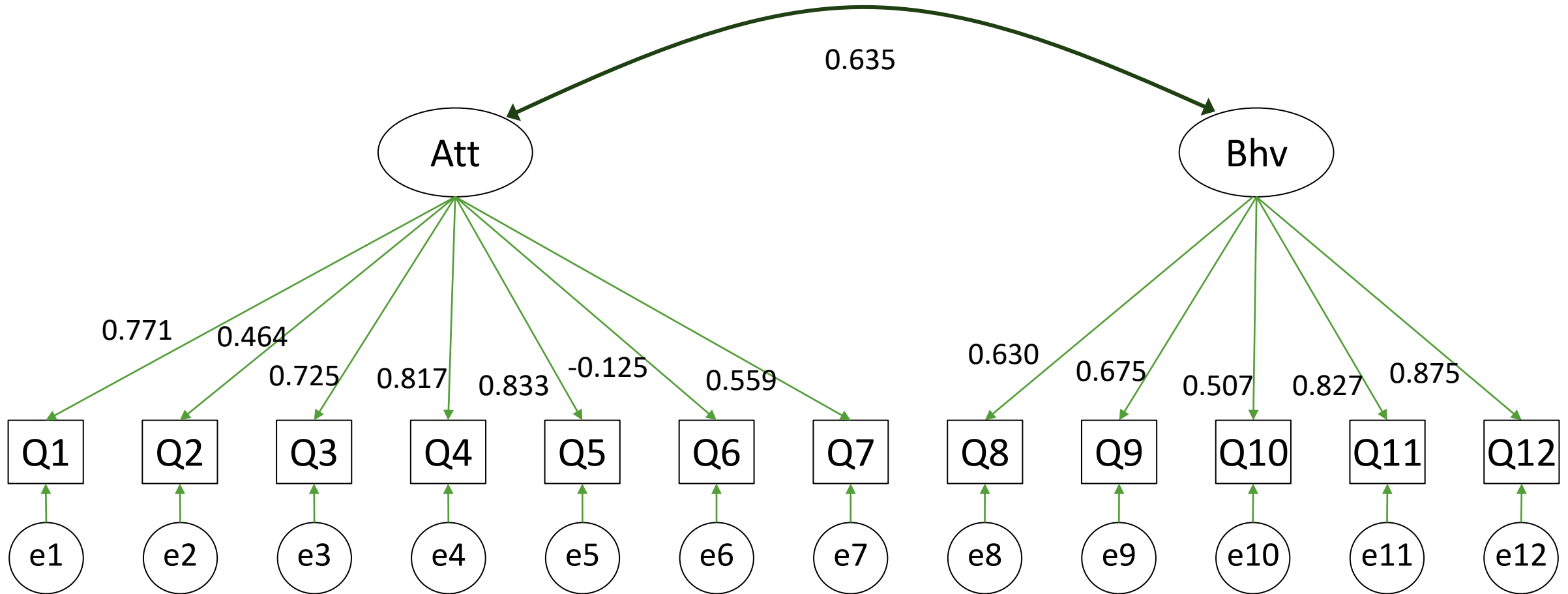
## Perceived behavioral control

|          |             |   |
|----------|-------------|---|
| 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.                                      |





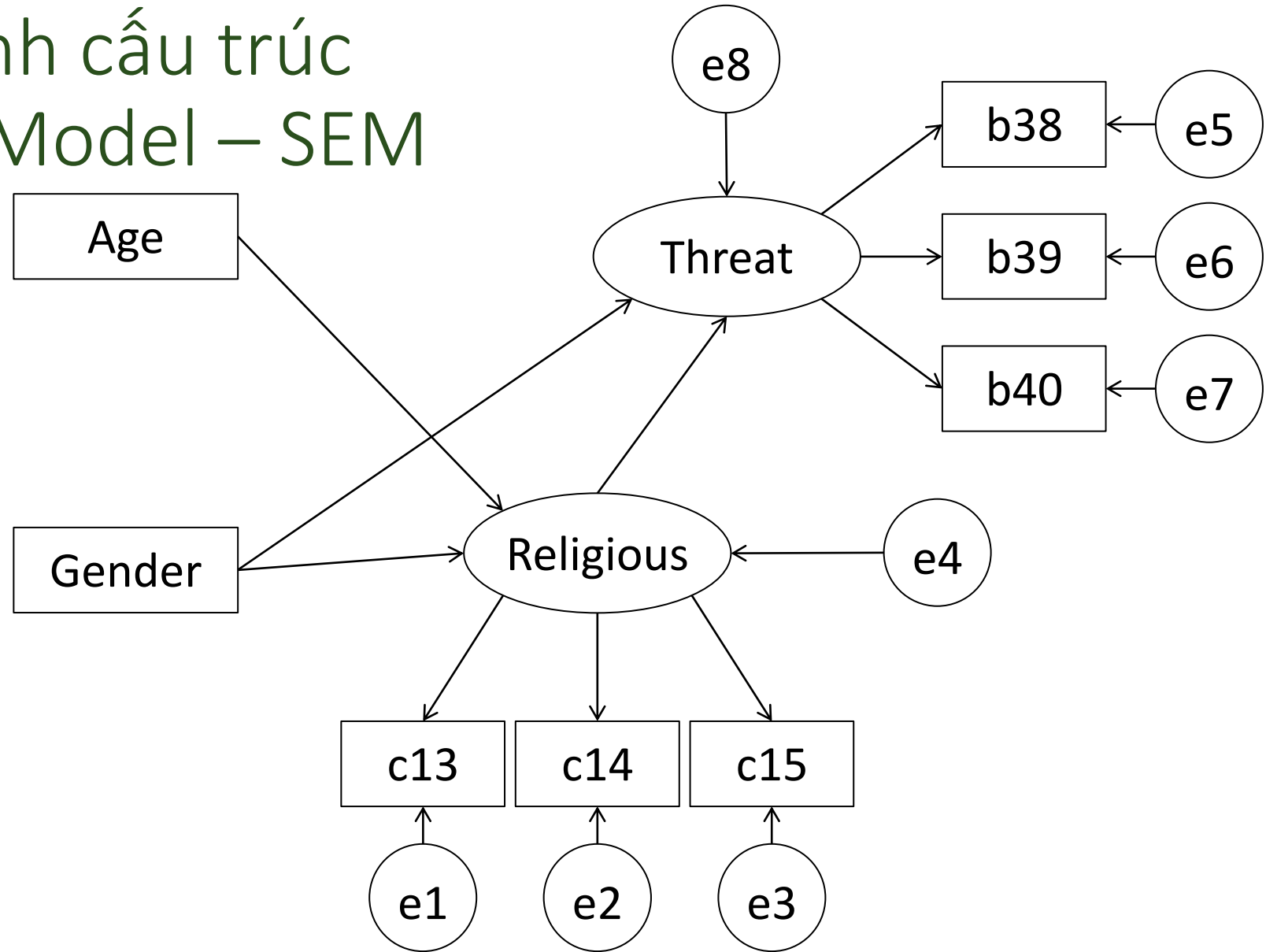
# Phân tích nhân tố khẳng định (CFA)



# Mô hình phương trình cấu trúc

## Structural Equation Model – SEM

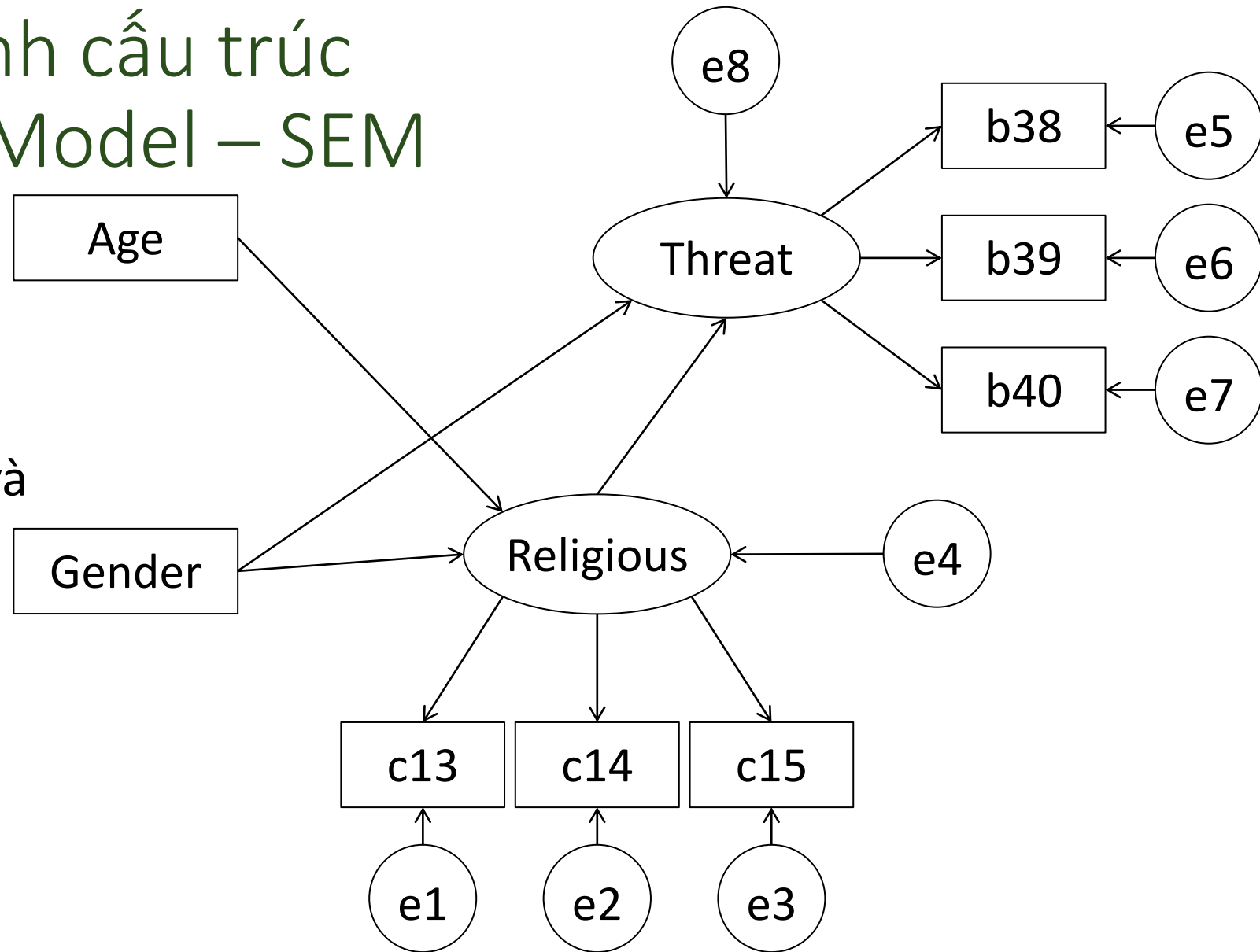
- Tiềm ẩn (latent) và biểu hiện (manifest)



# Mô hình phương trình cấu trúc

## Structural Equation Model – SEM

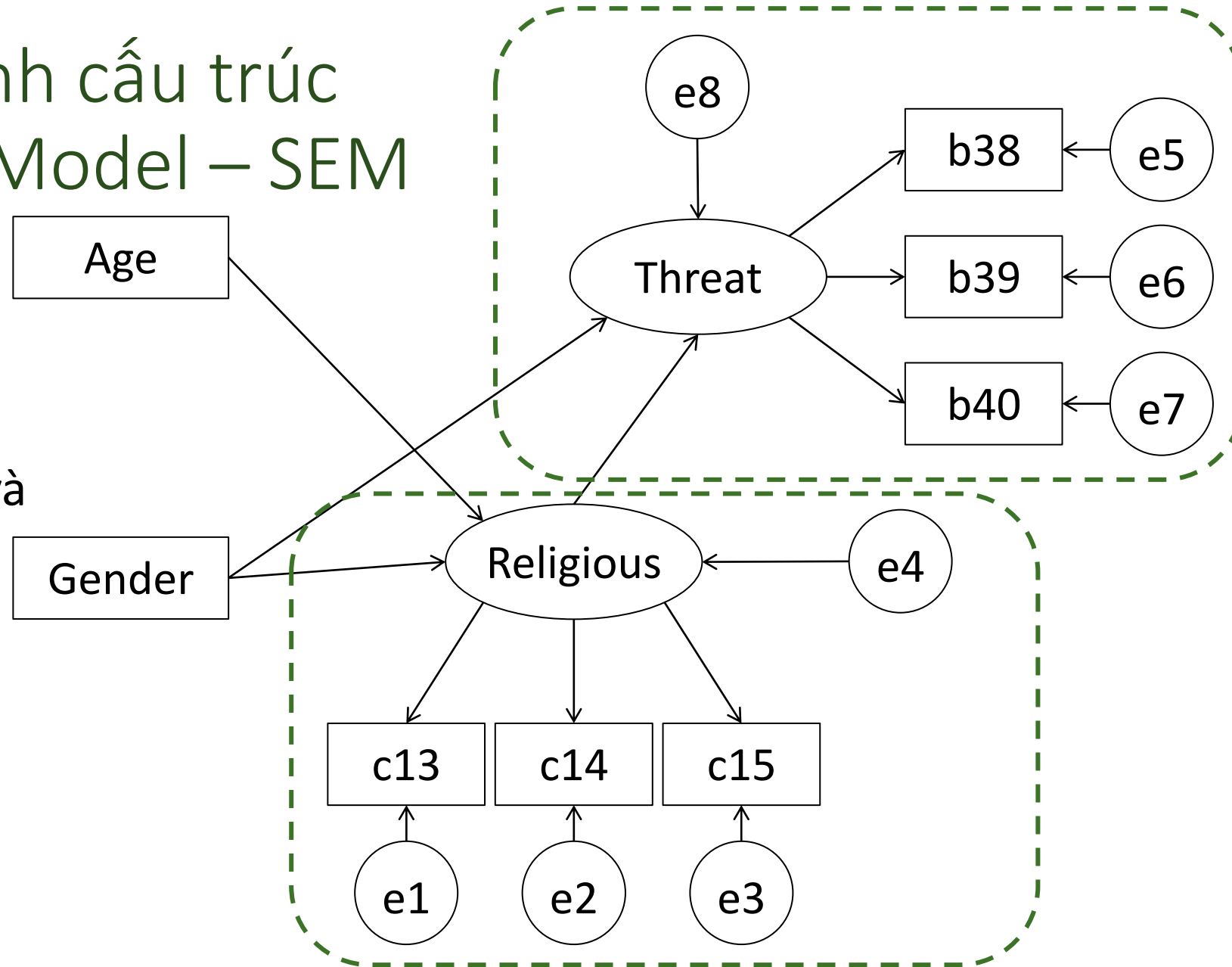
- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)



# Mô hình phương trình cấu trúc

## Structural Equation Model – SEM

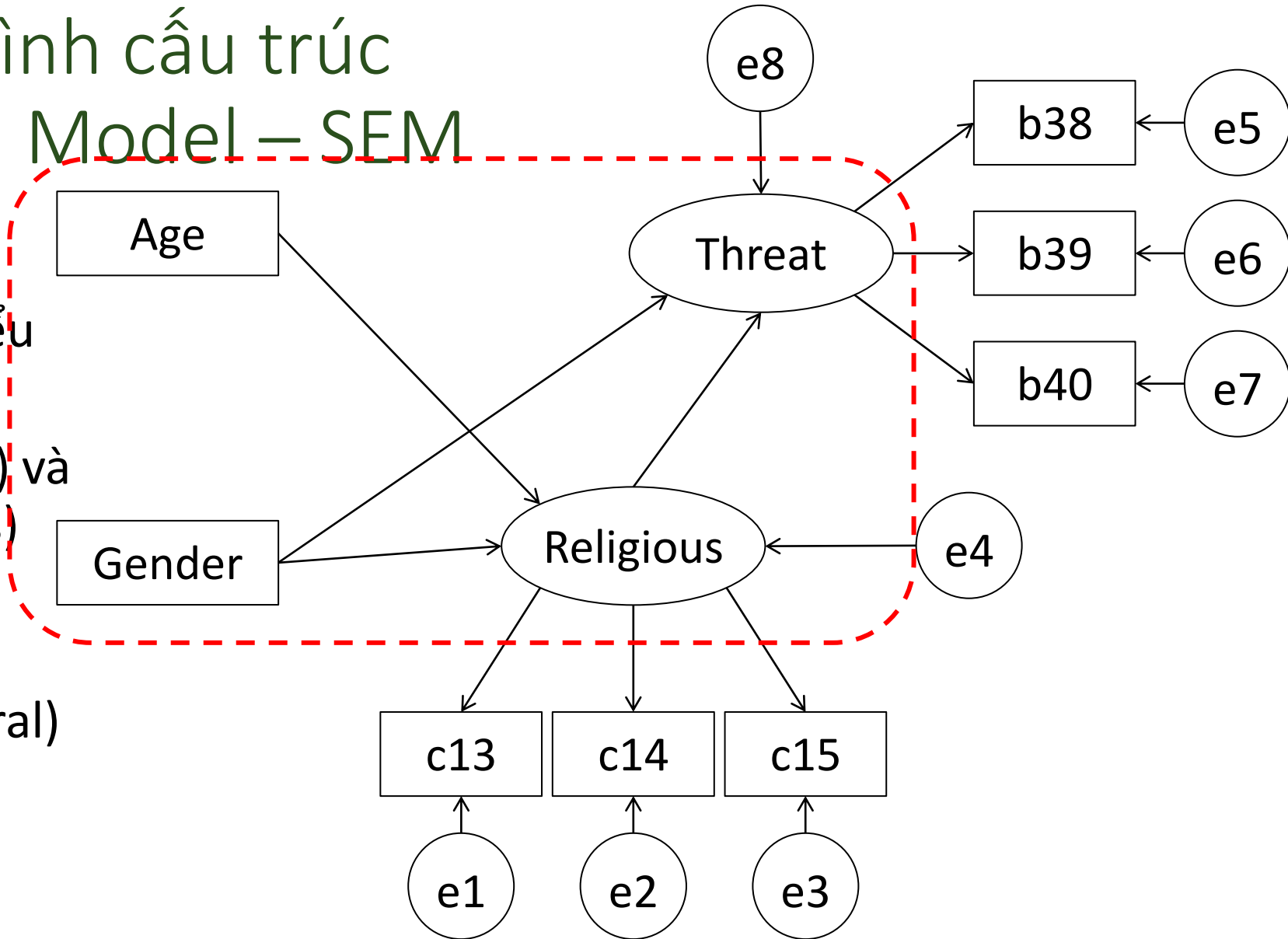
- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement)



# Mô hình phương trình cấu trúc

## Structural Equation Model – SEM

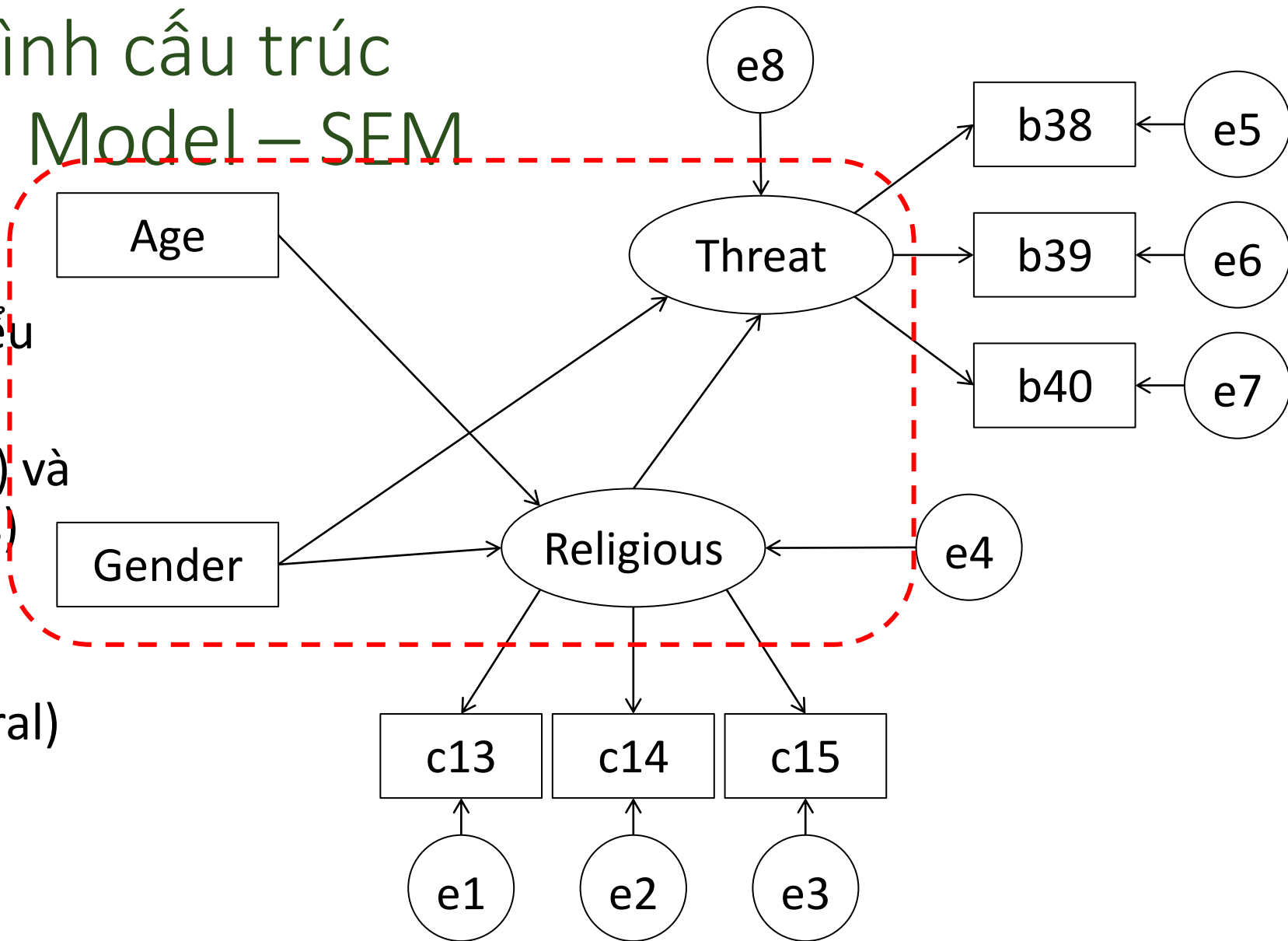
- Tiềm ẩn (latent) và biểu hiện (manifest)
- Nội sinh (endogenous) và ngoại sinh (exogenous)
- Mô hình đo lường (measurement) và mô hình cấu trúc (structural)



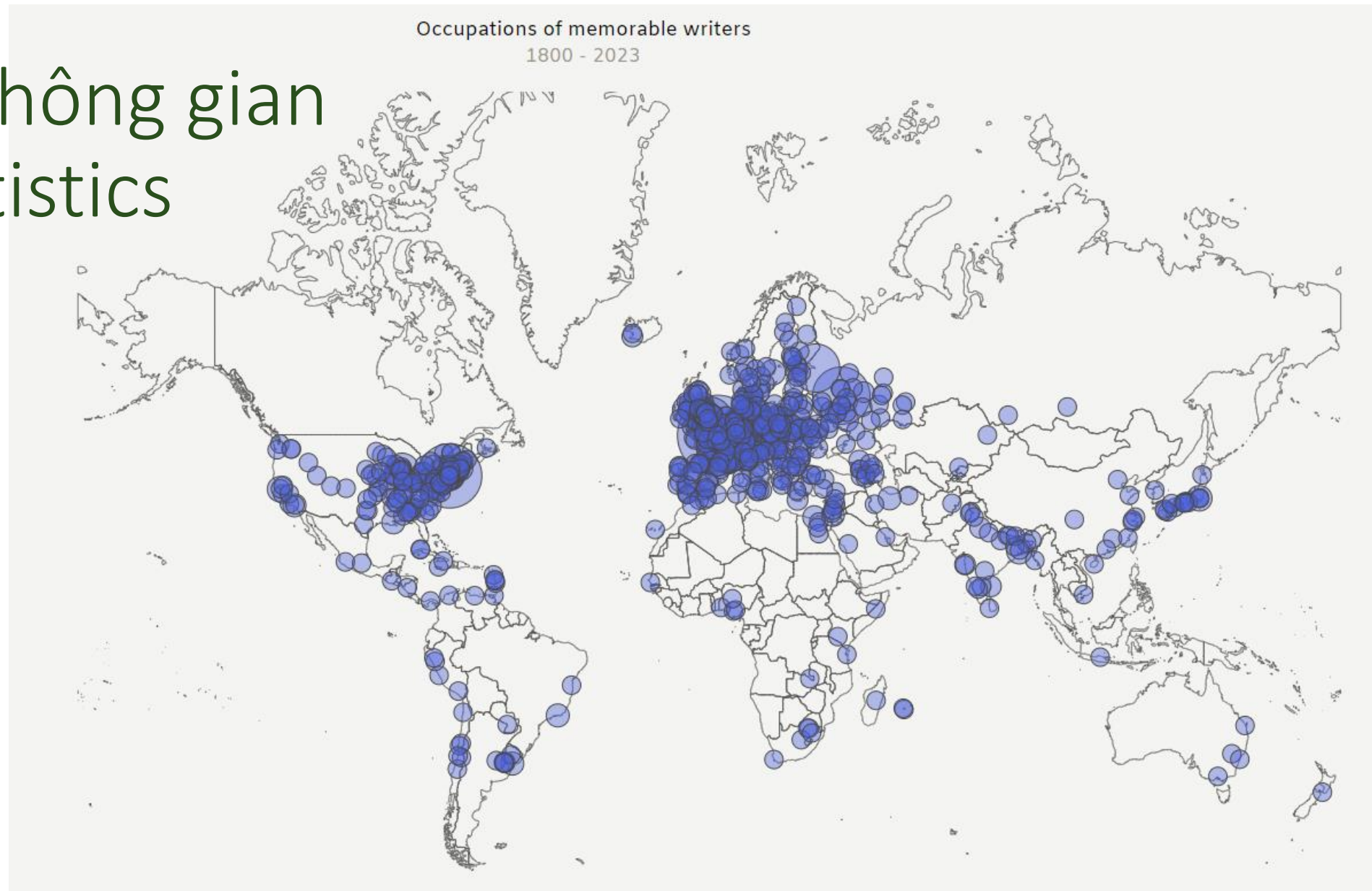
# Mô hình phương trình cấu trúc

## 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



# Lựa chọn phương pháp

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

