

# Phương pháp nghiên cứu liên ngành

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: 19:00-22:00, 26-29/10/2024
- Thực hành: 29/10/2024
  - R
  - Laptop cài sẵn R & RStudio
- Kiểm tra đánh giá: tiểu luận gồm 2 phần
  - Phần 1: Xử lý, phân tích và vẽ đồ thị cho tập dữ liệu cho trước
  - Phần 2: Đọc, tóm tắt, và phân tích bài báo

# Mục tiêu lớp học

- Giới thiệu khái niệm cơ bản
- Thảo luận xác định vấn đề nghiên cứu
- Thảo luận định hướng phương pháp sử dụng (tên phương pháp)
- Thực hành phân tích cơ bản với 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>)
- **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)
- 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)

# 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

# Khoa học liên ngành

# Nghiên cứu khoa học?

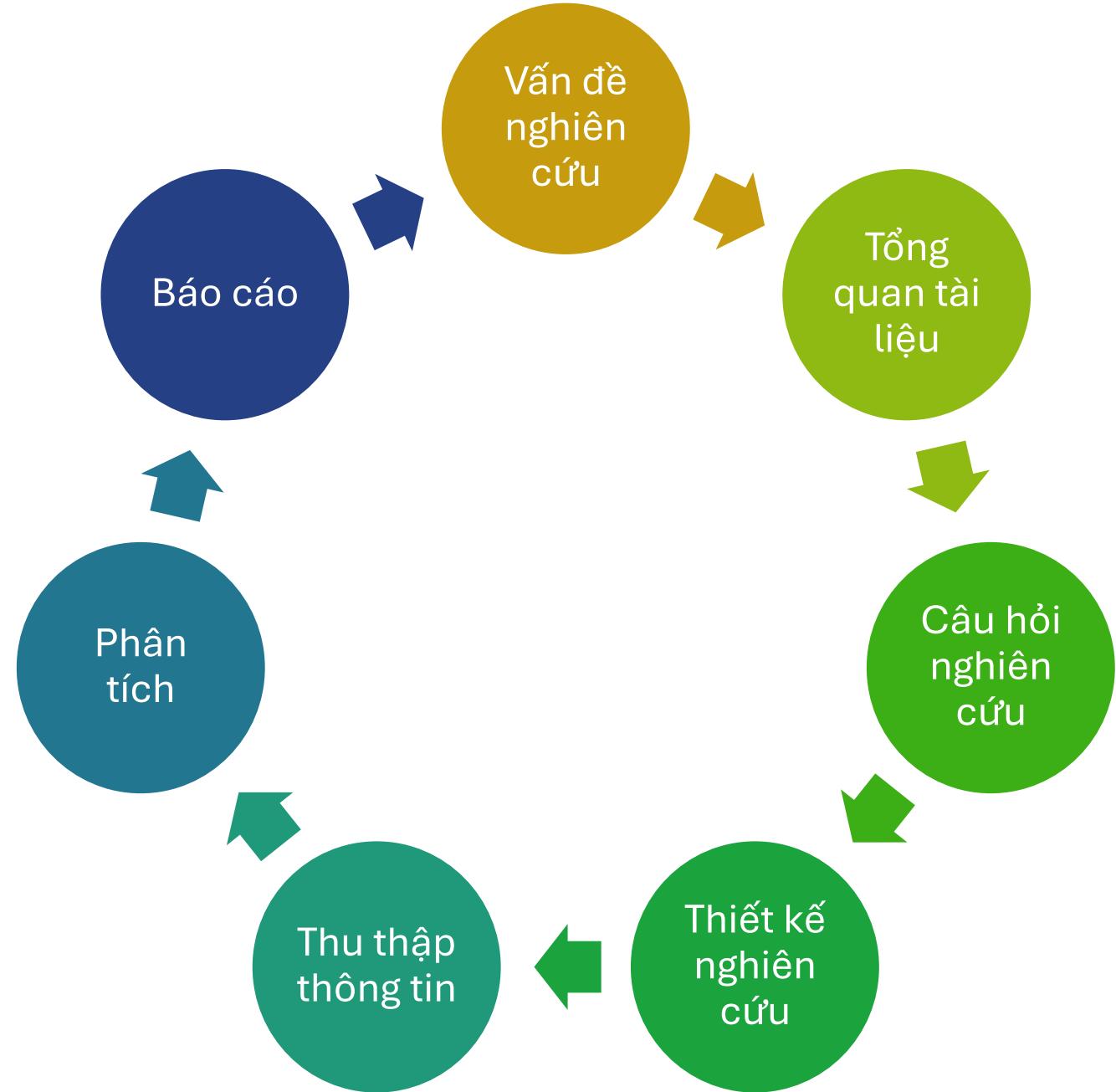
# Nghiên cứu khoa học?

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

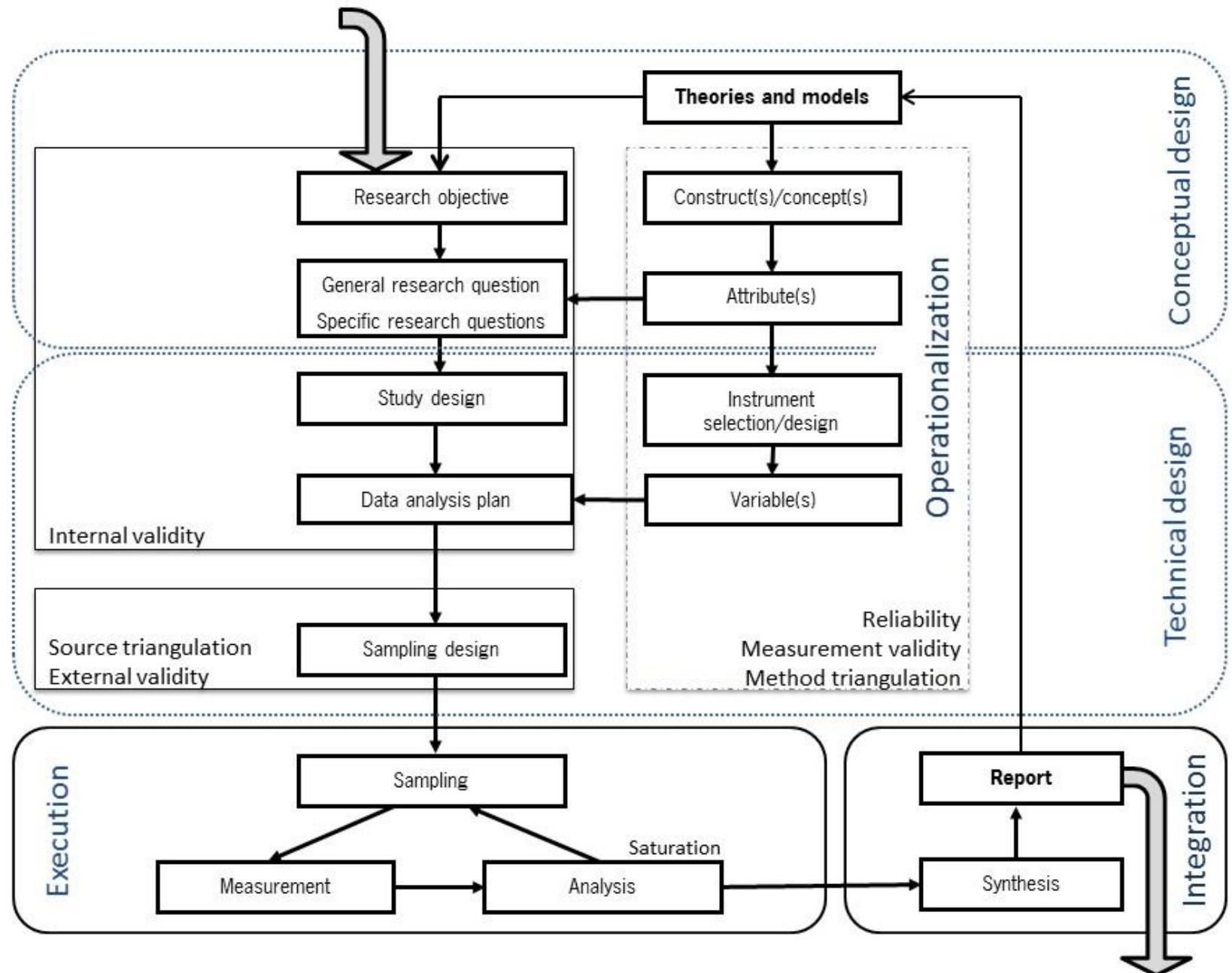
<https://www.coursera.org/learn/research-methods/home/info>

# Quá trình nghiên cứu



# Quá trình nghiên cứu

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>



# Chủ đề nghiên cứu

# Bài tập:

- Chủ đề nghiên cứu mà bạn quan tâm và định thực hiện trong nghiên cứu tiến sĩ của mình?
- Chủ đề này có liên quan đến cá nhân bạn hoặc công việc của bạn như thế nào?

# Chủ đề nghiên cứu đến từ đâu?

# Chủ đề nghiên cứu đến từ đâu?

- Thực tiễn cuộc sống
  - Trải nghiệm mối quan tâm cá nhân
  - Các vấn đề/câu hỏi thực tiễn đang được quan tâm
- Thực tiễn nghiên cứu
  - Câu hỏi mới phát sinh từ nghiên cứu được thực hiện trước
  - “Lỗ hổng” trong các nghiên cứu trước từ tổng quan tài liệu
- Đọc thêm chương Ý tưởng nghiên cứu đến từ đâu? – Cẩm nang nghiên cứu khoa học: Từ ý tưởng đến công bố - Nguyễn Văn Tuấn

# Thể nào là chủ đề nghiên cứu tốt?

# Thể nào là chủ đề nghiên cứu tốt?

## FINER: a research framework

- ✓ Feasible**  
Research questions should be answered under objective aspects like time, scope, resources, expertise, or funding.
- ✓ Interesting**  
Regardless of your own personal motivation about a subject, it is important to check if your question corresponds to more practical and broader interests.
- ✓ Novel**  
Answer to an existing gap in knowledge. Filling one of these gaps is important.
- ✓ Ethical**  
In empirical research, ethics is an absolute MUST.
- ✓ Relevant**  
Relevance can lead to real, visible changes in society.

Elsevier Author Services Blog

# Bài tập:

- Tính khả thi của chủ đề nghiên cứu bạn vừa đưa ra?
- Nó có thú vị với số đông mọi người?
- Tính mới của đề tài?
- Tính đạo đức?
- Tầm quan trọng?

# Tổng quan tài liệu

# Mục đích?

- Tìm hiểu về những nghiên cứu trước trong chủ đề nghiên cứu
- Xác định khoảng trống kiến thức (literature gaps)
- Thu hẹp chủ đề nghiên cứu thành các câu hỏi nghiên cứu và giả thuyết
- Xác định khung lý thuyết (theory framework) cho nghiên cứu hiện tại
- Định hướng phương pháp nghiên cứu
- Xây dựng hệ thống tài liệu tham khảo cho quá trình viết

# Các bước thực hiện

- Tìm, xác định các tài liệu liên quan
- Đọc nhanh, chọn lọc, phân loại
- Đọc sâu, **phân tích, tổng hợp**

Không dừng lại ở đọc và nhắc lại những nghiên cứu trước mà cần **xác định được điểm chung, xu hướng, và đưa ra các ý kiến tổng hợp**

# Các loại tài liệu

- Sách chuyên khảo
- Sách tổng hợp/chương sách
- Bài báo (có/không qua phản biện đồng cấp)
- Bài hội nghị
- Bản thảo tiền xuất bản (preprint)
- Luận án/luận văn
- Báo cáo tư vấn
- Văn bản luật

# Tìm kiếm tài liệu (1)

- Thông qua các cơ sở dữ liệu
  - Scopus (through library/academic library)
  - Google scholar

**Scopus**      TITLE-ABS-KEY (("Climat\* chang\*" OR "Climat\* risk\* OR " climat\* AND variabilit\* " OR " climat\* AND extrem\* " OR " climate AND variability\* OR "climat\* uncertain\*") OR "global warming\*" OR "temperature ris\*" OR "sea level ris\*" OR "el-nino" OR "la-nina") AND ("Adapt\* abilit\*" OR "adapt\* strateg\*" OR "adapt\* capacit\*" OR "adapt\* capabilit\*" OR "adapt\* strength\*" OR "adapt\* potential\*" OR "adopt\* abilit\*" OR "adopt\* capacity\*" OR "adopt\* capabilit\*" OR "Adopt\* potential\*" OR adopt\* AND strategy\*)) AND (farmer\*)

Source: Shaffril et al. (2018)

# Tìm kiếm tài liệu (2)

- Tài liệu tham khảo
- Các bài trích dẫn lại

 Scholar About 519,000 results (0.17 sec)

---

[HTML] Large-scale **temperature** inferences from **tree rings**: a **review**  
KR Briffa, TJ Osborn, FH Schweingruber - Global and planetary change, 2004 - Elsevier  
... them, several require further study. Tree-ring density and tree-ring width data will continue  
to enhance our detailed knowledge of past **temperature** and other **climate** changes, but a ...

☆ Save 99 Cite Cited by 442 Related articles All 6 versions

# Đọc nhanh, chọn lọc, phân loại

- EndNote
- Mendeley
- Zotero

# Đọc nhanh, chọn lọc, phân loại

The screenshot shows the Zotero application window titled "My Library - Zotero". The menu bar includes File, Edit, View, Tools, and Help. The toolbar features icons for adding items, editing, deleting, and searching. A search bar at the top right allows filtering by Title, Creator, and Year. The main area displays a table of search results with columns for Title, Creator, Item Type, Year, Publication, and Date Added. The results list various academic publications, such as books and journal articles, from 2011 to 2024. On the left sidebar, under "My Library", there are categories like "Reviewing papers", "To read", "My Publications", "Duplicate Items", "Unfiled Items", and "Trash". Below the sidebar, there is a list of search terms: <i>Aedes aegypti</i>, <i>Aedes albopictus</i>, 3D, 3D city model, 3D city models, 3D tiles, 20th century, Abnormal behavior.

Title	Creator	Item Type	Year	Publication	Date Added
Statistics for people who (think they) ...	Salkind a...	Book	2020		10/26/2024, 1...
bookdown: authoring books and tech...	Xie	Book	2016		10/17/2024, 4...
Smart city for the preservation of urb...	Thi Hai Yen	Thesis	2020		10/17/2024, 3...
An urban biodiversity assessment fra...	Li et al.	Journal Arti...	2019	Frontiers in E...	10/17/2024, 3...
Understanding uneven urban expansi...	Long et al.	Journal Arti...	2018	Landscape an...	10/17/2024, 3...
Research design: the methodology fo...	Tobi and ...	Journal Arti...	2018	Quality & Qu...	10/17/2024, 2...
A hypothesis is a liability	Yanai and...	Journal Arti...	2020	Genome Biol...	10/13/2024, 5...
Urbanization and land use change: a s...	Tuan	Journal Arti...	2022	Environment...	10/10/2024, 1...
Surface subsidence in urbanized coas...	Duffy et al.	Journal Arti...	2020	Remote Sensi...	10/10/2024, 1...
Utilizing publicly available satellite da...	Goldblatt...	Journal Arti...	2018	Development...	10/10/2024, 1...
Urbanization, economic development...	Fan et al.	Journal Arti...	2019	Landscape an...	10/10/2024, 1...
State of the Vietnamese coast—assess...	Lappe et al.	Journal Arti...	2022	Remote Sensi...	10/10/2024, 1...
Degradation of coastlines under the p...	Petrişor e...	Journal Arti...	2020	Land	10/10/2024, 1...
Estimating land-use change using ma...	Giang et al.	Journal Arti...	2022	Sustainability	10/10/2024, 1...
Examining the spatial variations of la...	Liang et al.	Journal Arti...	2022	Land	10/10/2024, 1...
Impacts of urbanization and land tran...	Tu et al.	Journal Arti...	2021	Regional Stu...	10/10/2024, 1...
The potential of open-access data for ...	Scheiber ...	Journal Arti...	2023	Natural Hazar...	10/10/2024, 1...
Effect of the urban land cover types o...	Kim and ...	Journal Arti...	2012	Korean Journ...	10/9/2024, 5...
Prediction of daily water consumptio...	Li et al.	Journal Arti...	2023	Water	10/9/2024, 2...
Data mining: concepts and techniques	Han et al.	Book	2023		10/7/2024, 2...
Ontologies in urban development pro...	Falquet e...	Book	2011		10/7/2024, 12...
Chocolate consumption, cognitive fu...	Messerli	Journal Arti...	2012	New England...	10/7/2024, 12...
Analysis of patterns of spatial occupa...	Marušić	Journal Arti...	2011	URBAN DESI...	10/4/2024, 3...

# Đọc nhanh, chọn lọc, phân loại

- Cấu trúc một bài báo - **IMRaD**
  - Abstract
  - Introduction
  - Methods
  - Results
  - Discussion
  - (Conclusion)

“It is impossible to be a good writer without being a good reader first”

# Đọc nhanh, chọn lọc, phân loại

The screenshot shows a Zotero library window with a tab for 'Policy note: a universal equity-efficiency model for pricing water - Beecher - 2020 - Zotero'. The main content area displays the first page of the document, which discusses water's intrinsic value and its place in Maslow's hierarchy of needs. A handwritten note 'vital / crucial things' is written next to the text. Below this, a section titled 'Policy Nook' contains a paragraph about utility user fees being regressive. This paragraph is highlighted with a green box and has several annotations: 'more than what can be considered their fair share (Tomer 2018)', 'Like sales taxes, utility user fees are regressive and a form of structural inequity. A regressive cost takes a greater share of the low-income household budget; high fixed utility charges are particularly regressive as they are disproportionately burdensome and uncontrollable by the consumer. These affordability effects are cumulative across water, wastewater, stormwater, and other utility bills, accentuated by inequalities of income and living conditions.' The right-hand panel of the Zotero interface shows the document title, author, and a summary of the research findings.

Policy note: a universal equity-efficiency model for pricing water - Beecher - 2020 - Zotero

File Edit View Go Tools Help

My Libr... Applied dat... X Methods in ... X A hands-on ... X Fundament... X The practice... X Social resear... X Policy note: ... X

2 of 29

Water is intrinsic not just to human life but to the quality of life. It occupies the physiological base (along with air, food, and shelter) on Maslow's (1943) hierarchy of needs. It is difficult to conceive of a public service more important than water; the social imperative to provide safe drinking water and sanitation services to individuals, households, and communities should be noncontroversial.

vital / crucial things

2071001-1

Policy Nook

When it comes to public infrastructure and utilities, low-income households pay more than what can be considered their fair share (Tomer 2018). Like sales taxes, utility user fees are regressive and a form of structural inequity. A regressive cost takes a greater share of the low-income household budget; high fixed utility charges are particularly regressive as they are disproportionately burdensome and uncontrollable by the consumer. These affordability effects are cumulative across water, wastewater, stormwater, and other utility bills, accentuated by inequalities of income and living conditions.

Affordable access to essential water services has recently gained in position on the research, policy, and civil rights agendas, brought to light by high rates of service disconnection (water shutoffs) in cities like Detroit and Flint, Michigan,

Policy note: a universal equity-efficiency model for pricing w...

Beecher\_2020\_Policy note

SKIMMED => need to reread

Summary:

- > review of current problems in water tariff
- >> a message repeated throughout: current tariff prefers economic equity over social equity, care more about willingness-to-pay than ability-to-pay
- > suggest new model for pricing with fixed fee based on property values with 5 components
- >> For me:
- > aargh, too many new economic terms I

Related: [click here]

Tags: [click here]

# Đọc sâu, phân tích, và tổng hợp

- Chỉ với những bài quan trọng
- Tổng quan cập nhật tình hình nghiên cứu của vấn đề đang quan tâm
- Đặt nền tảng của khung lý thuyết định sử dụng
- Câu hỏi hoặc thiết kế nghiên cứu rất gần với nghiên cứu đang định thực hiện

# Đọc sâu, phân tích, và tổng hợp

- Câu hỏi nghiên cứu/thông điệp chính của bài báo là gì?
- Tại sao câu hỏi đó cần được nghiên cứu?
- Dữ liệu nào cần để trả lời câu hỏi đó?
- Phương pháp sử dụng để thu thập dữ liệu?
- Phương pháp nào cần thiết để phân tích dữ liệu và trả lời câu hỏi chính?
- Kết quả chính của bài báo là gì?
- Kết quả này đóng góp vào việc trả lời câu hỏi chính như thế nào?
- **Mối liên quan giữa bài này và các bài báo khác?**
- **Từ kết quả của bài báo có thể gợi mở ra những gì cho lĩnh vực nghiên cứu?**

# Bài tập

- Đọc tóm tắt của bài sau và trả lời các câu hỏi trong slide trước

Strielkowski, W., Štreimikienė, D., & Bilan, Y. (2017). Network charging and residential tariffs: A case of household photovoltaics in the United Kingdom. *Renewable and Sustainable Energy Reviews*, 77, 461–473. <https://doi.org/10.1016/j.rser.2017.04.029>

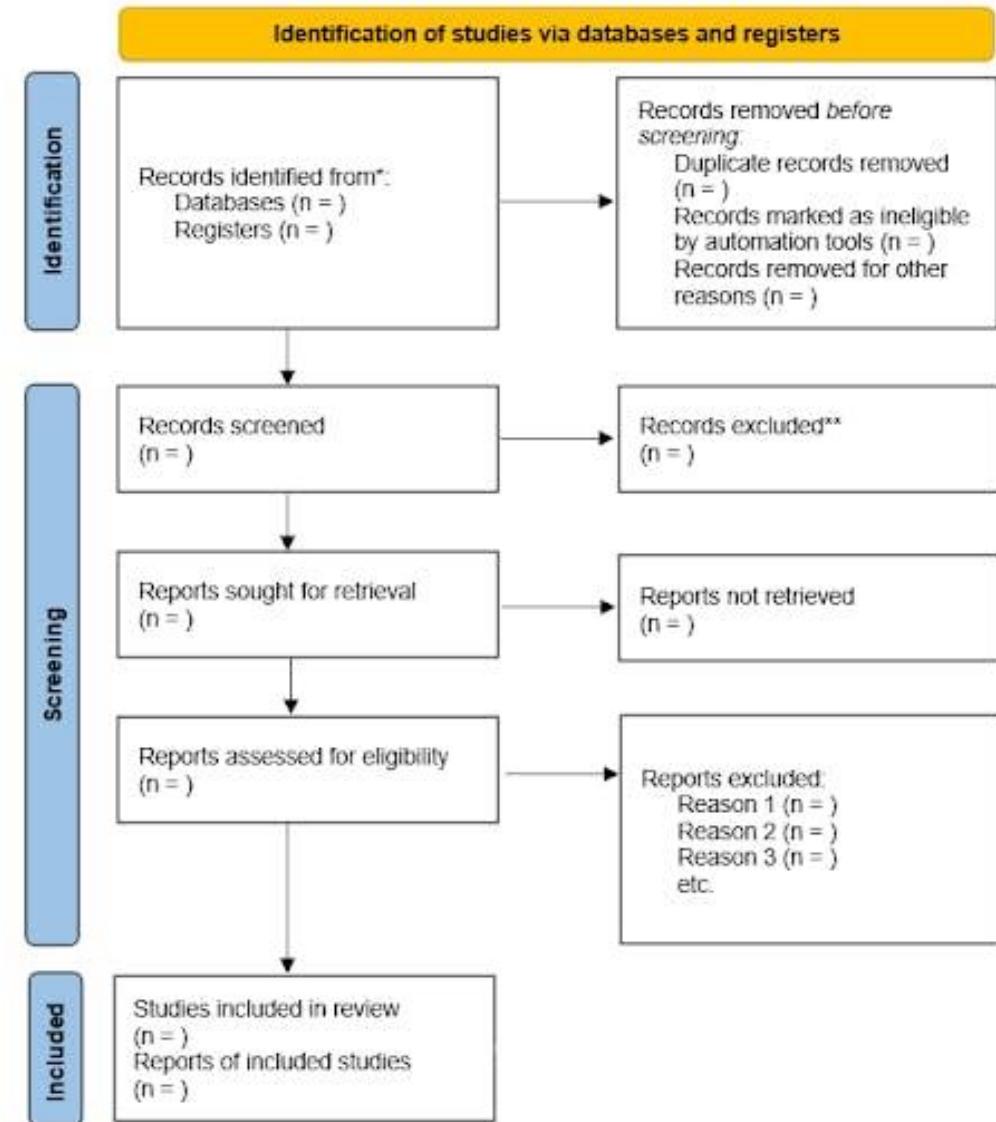
# Những điều cần tránh

- Đọc tất cả mọi thứ
- Chỉ đọc không viết
- Không lưu lại thông tin của tài liệu

# Tổng quan hệ thống (systematic reviews)

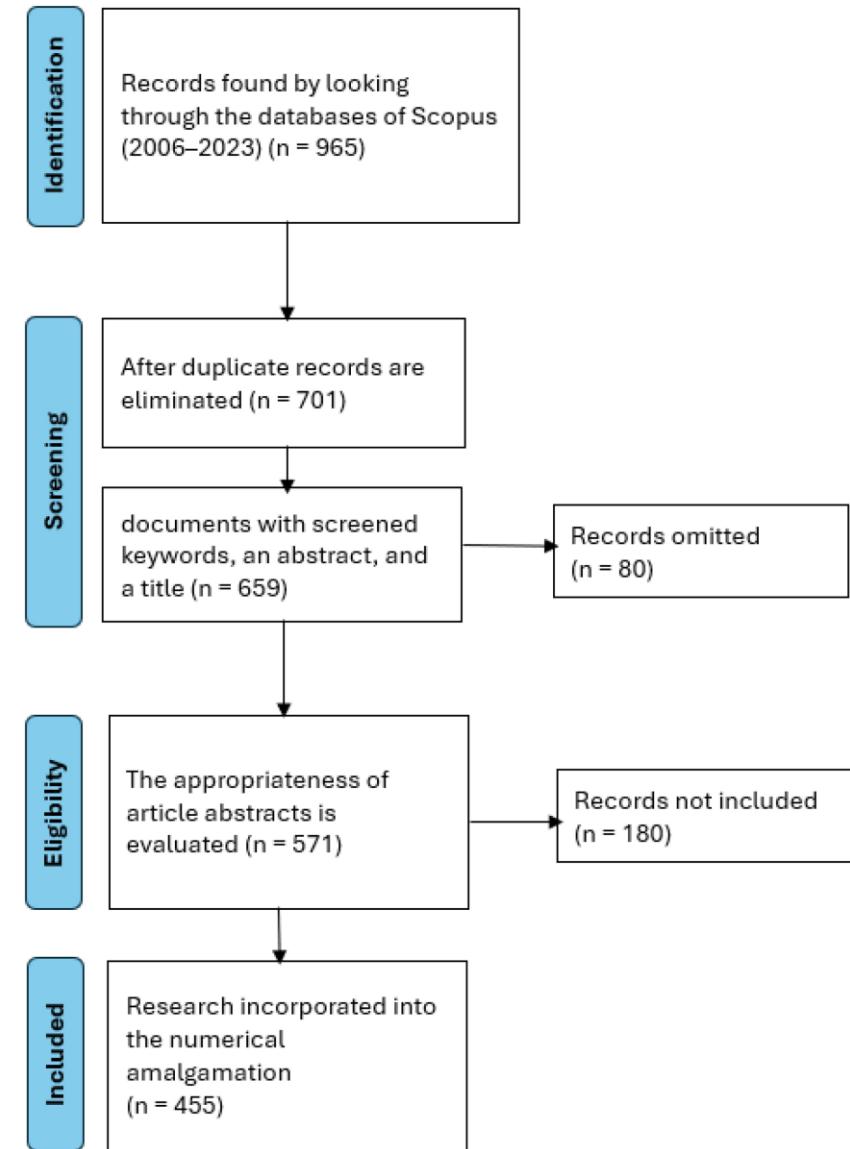
PRISMA

Preferred Reporting Items for Systematic reviews and Meta-Analyses.



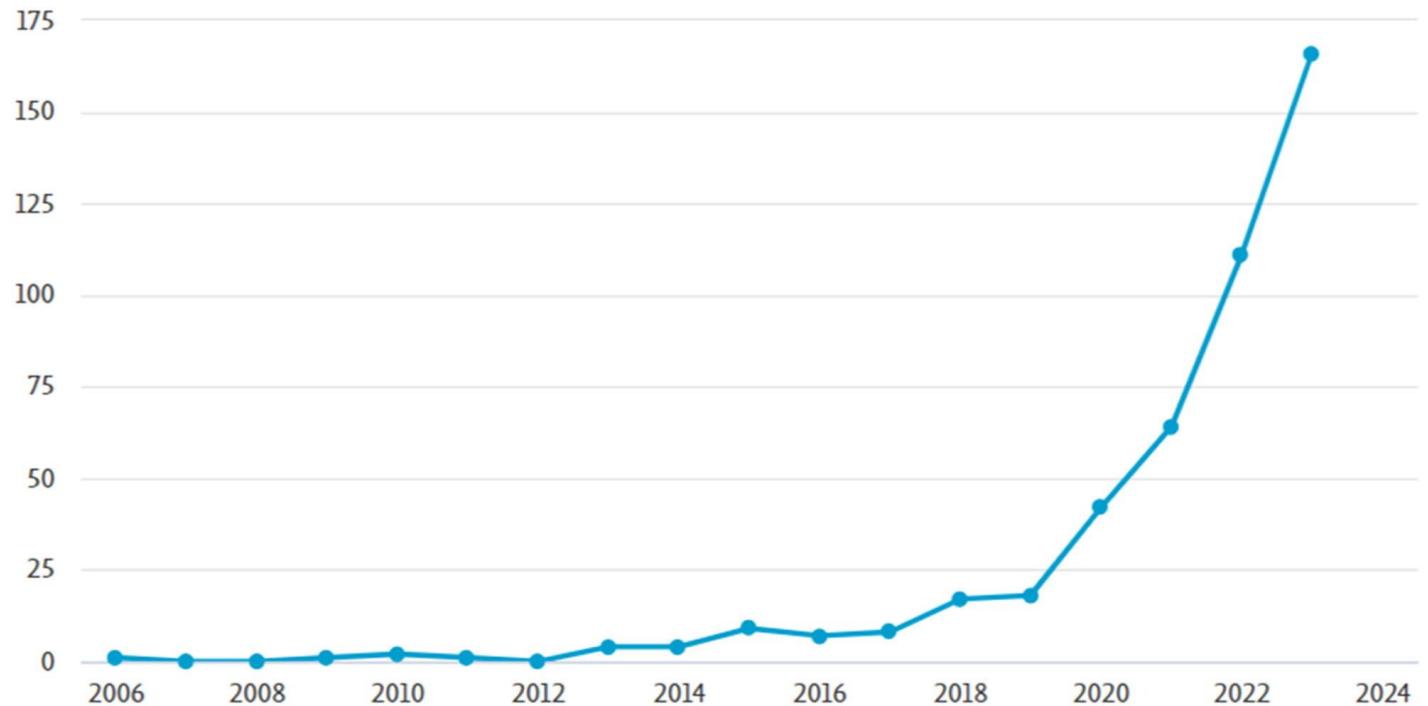
# Tổng quan hệ thống (systematic reviews)

Lindawati, A. S. L. & Meiryani. (2024). A bibliometric analysis on the research trends of global climate change and future directions. *Cogent Business & Management*, 11(1), 2325112.  
<https://doi.org/10.1080/23311975.2024.2325112>



# Tổng quan hệ thống (systematic reviews)

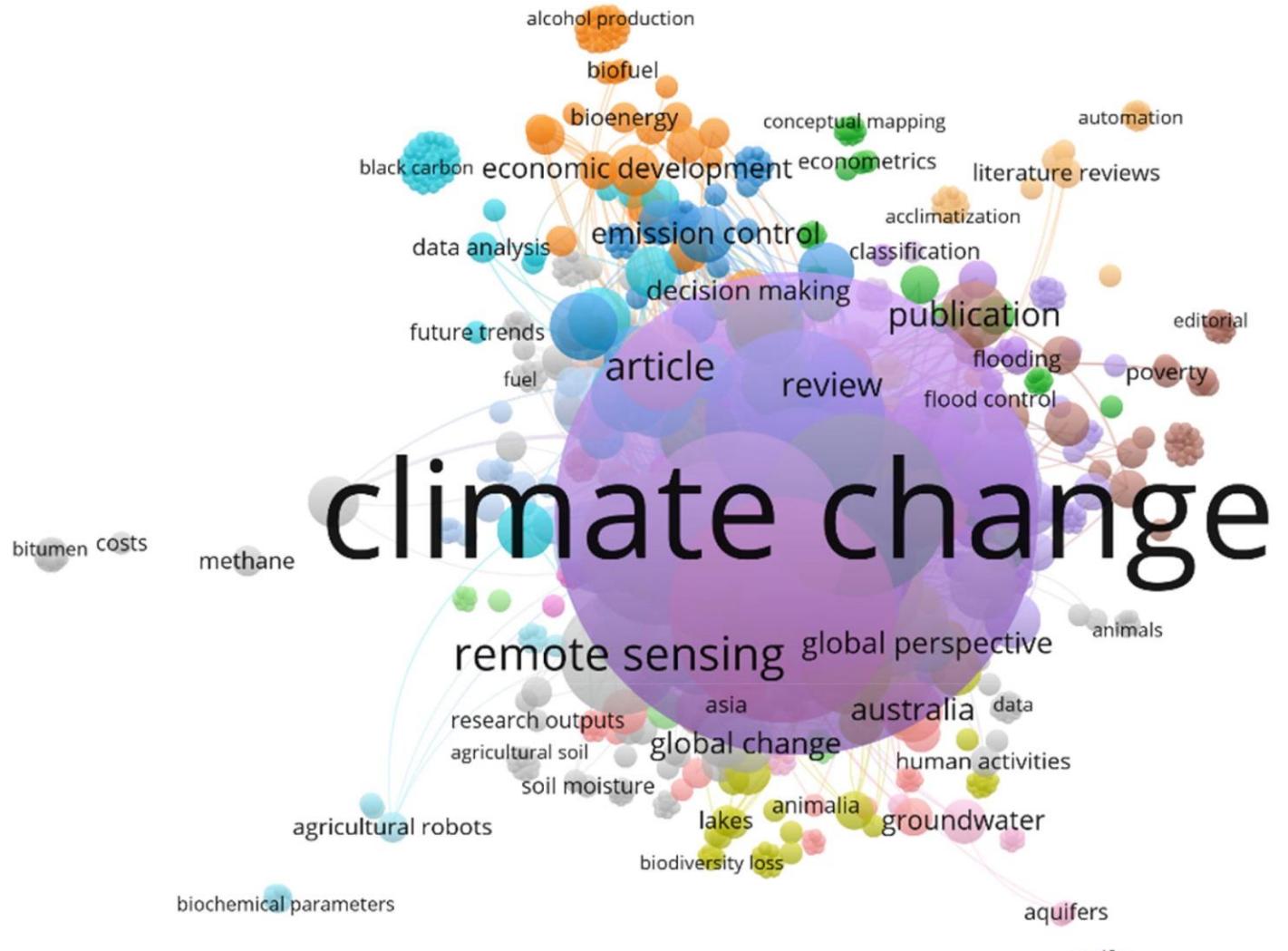
Lindawati, A. S. L. & Meiryani. (2024). A bibliometric analysis on the research trends of global climate change and future directions. *Cogent Business & Management*, 11(1), 2325112.  
<https://doi.org/10.1080/23311975.2024.2325112>



# Tổng quan hệ thống (systematic reviews)

# Phân tích Bibliometric

Lindawati, A. S. L. & Meiryani. (2024). A bibliometric analysis on the research trends of global climate change and future directions. *Cogent Business & Management*, 11(1), 2325112.  
<https://doi.org/10.1080/23311975.2024.2325112>

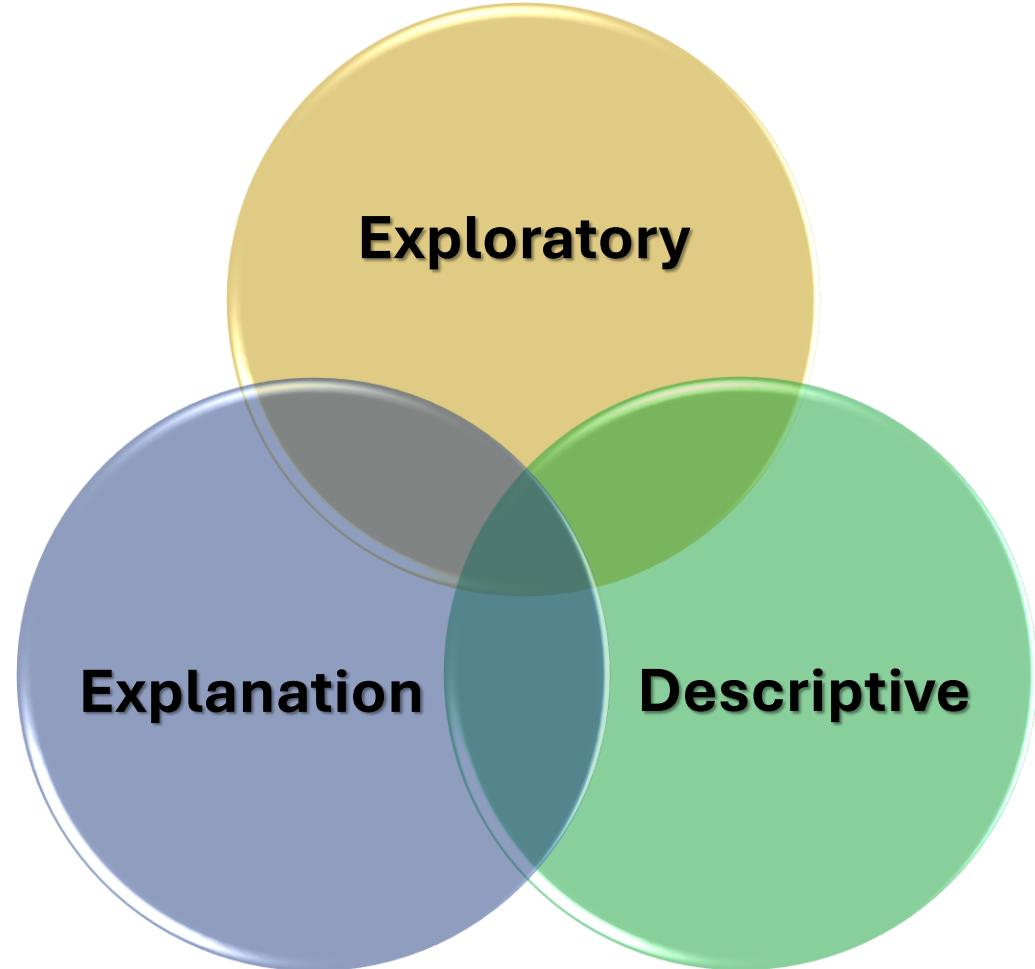


# Thiết kế nghiên cứu

# 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



# Các mục đích của nghiên cứu

- Khám phá mối quan hệ giữa biến đổi khí hậu và hệ sinh thái rừng nhiệt đới
- Mô tả sự thay đổi nhiệt độ, lượng mưa, và đa dạng sinh học của rừng nhiệt đới trong 50 năm vừa qua
- Nghiên cứu tác động của thay đổi lượng mưa do biến đổi khí hậu đến sự phát triển của một số loài thực vật trong rừng nhiệt đới

# 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

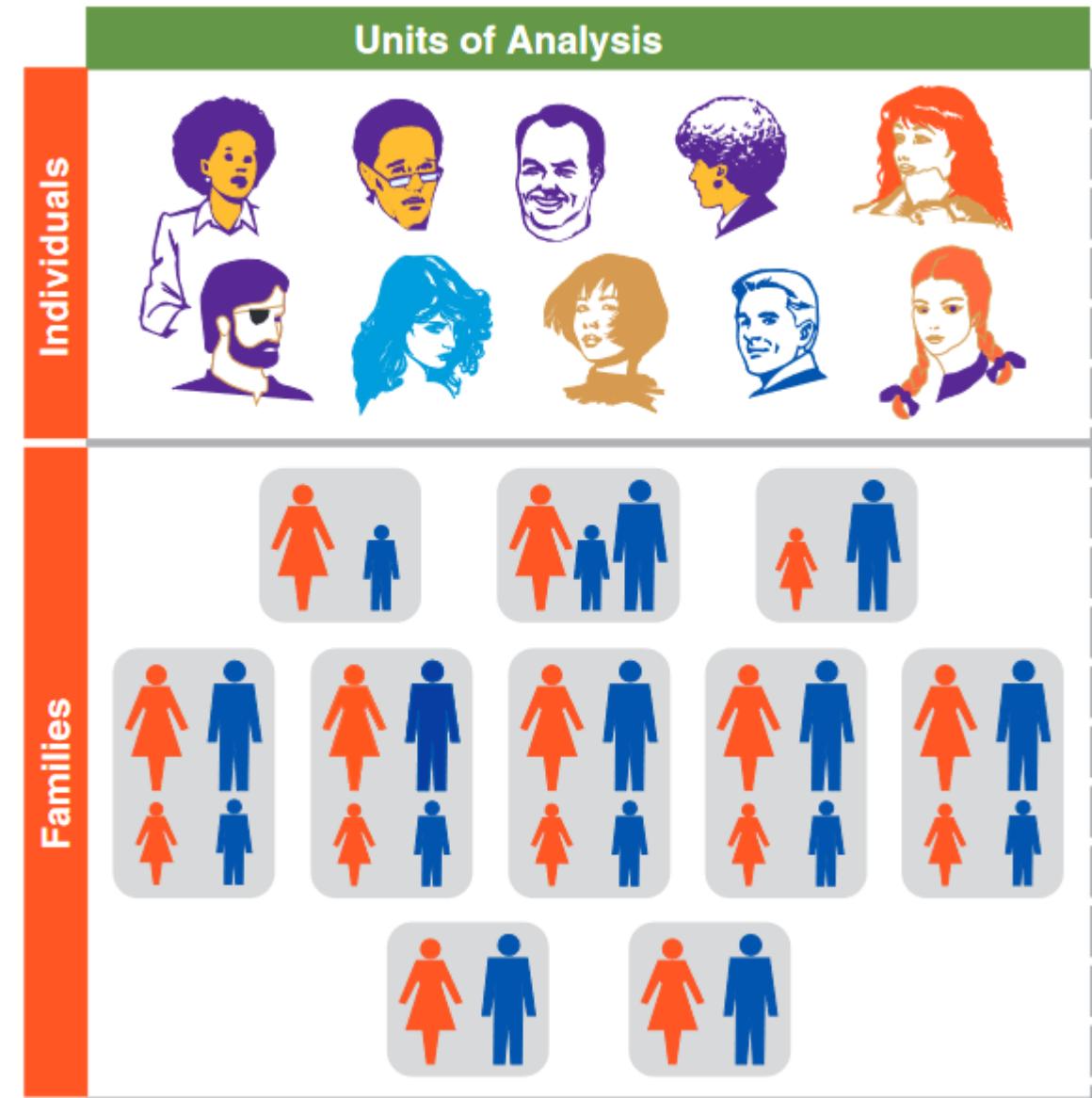
## Định tính

- 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

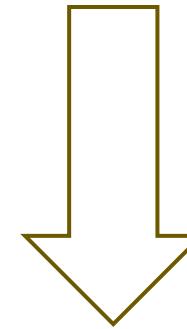
- Kết hợp cả 2 phương pháp
- Cân bằng giữa phân tích sâu về đối tượng quan tâm và tổng quan hóa bối cảnh

# 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%A0o%20t%C3%ADnh%20to%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%A0o%20t%C3%ADnh%20to%C3%A1n%20c%E1%BB%A1%20m%E1%BA%ABu_revised%20l%E1%BA%A7n%201_5.8.2020_0.pdf)

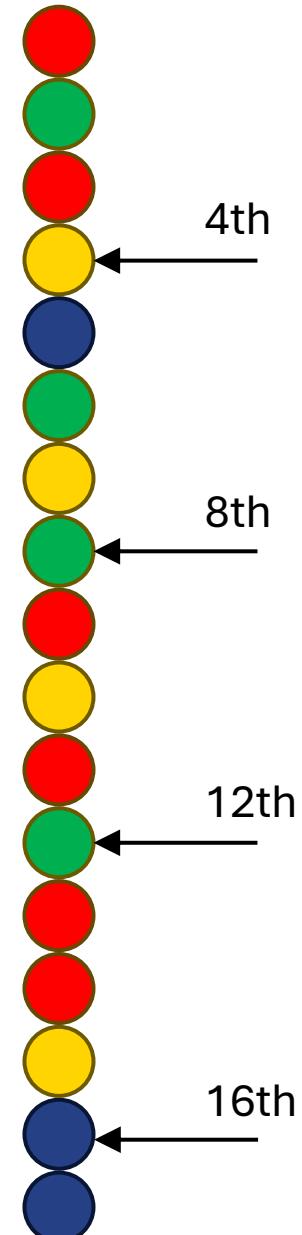
# Phương pháp lấy mẫu

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



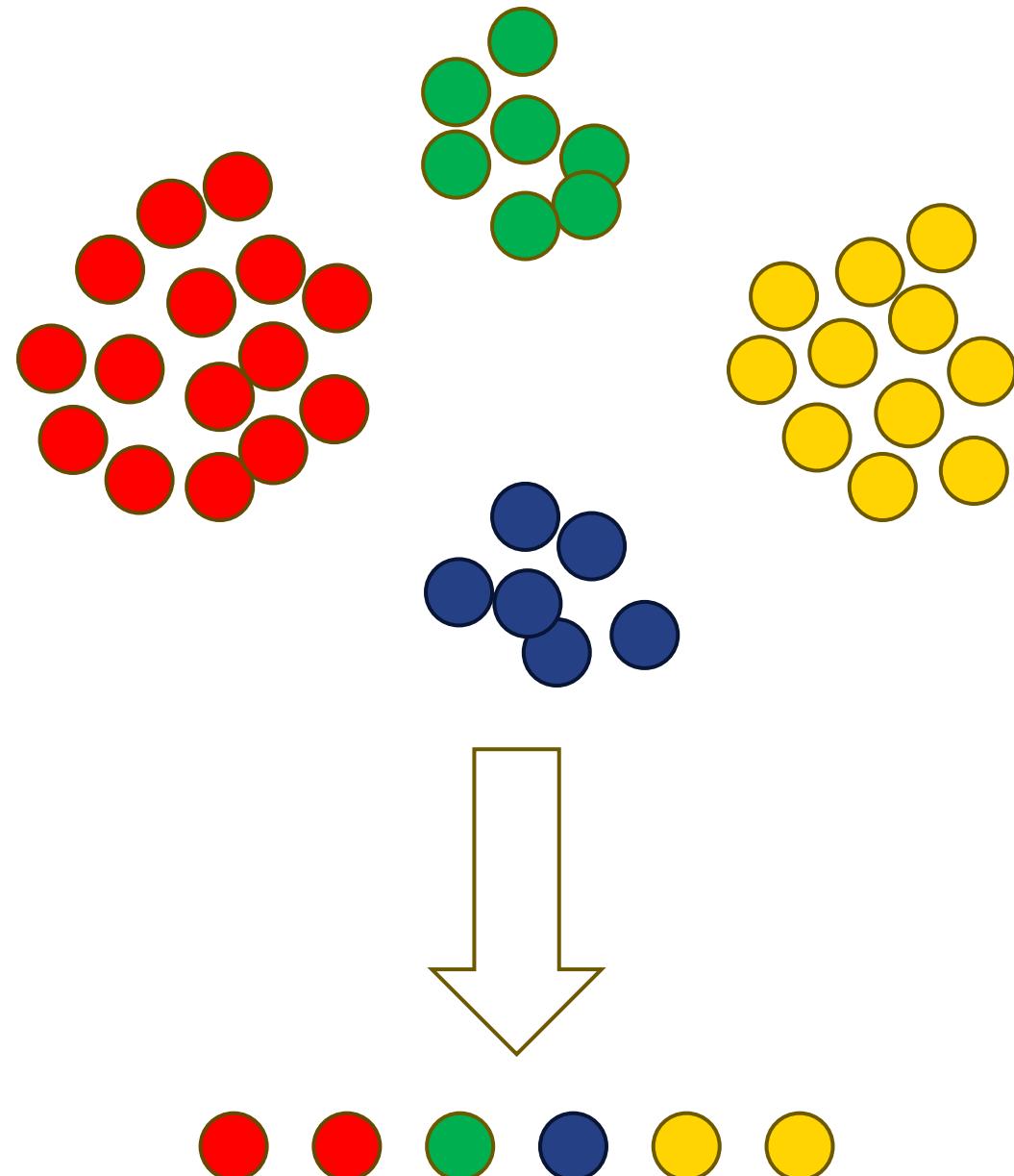
# Phương pháp lấy mẫu

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



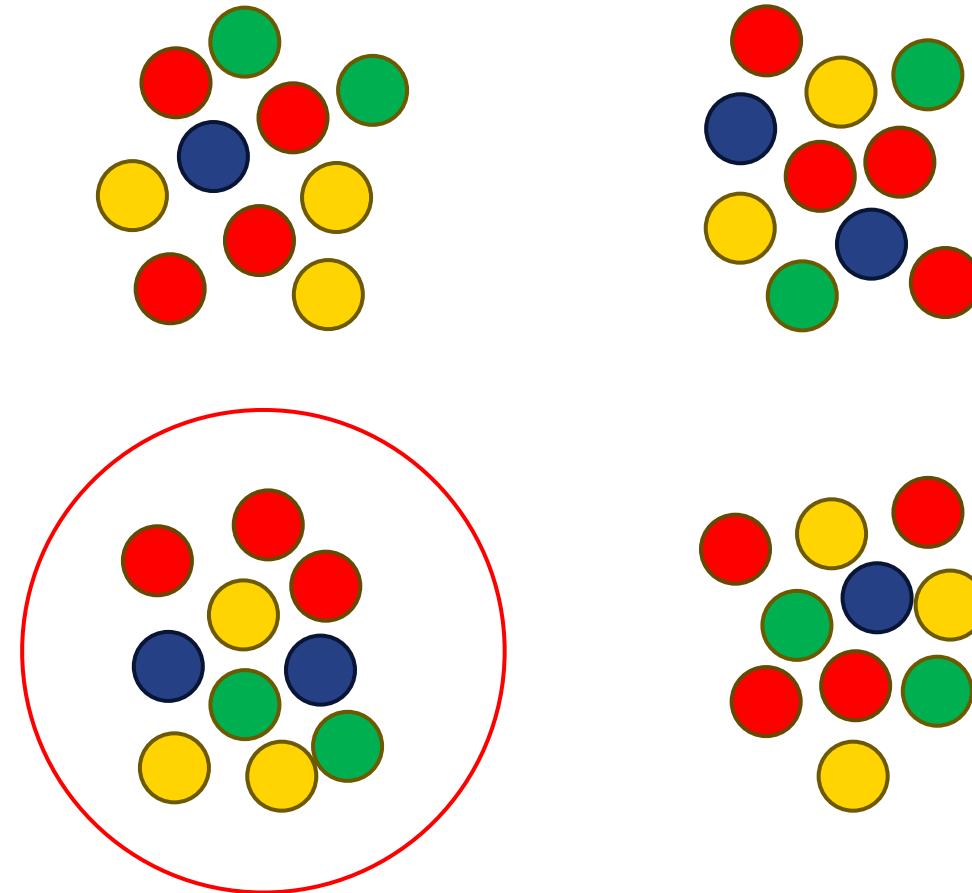
# Phương pháp lấy mẫu

- 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)



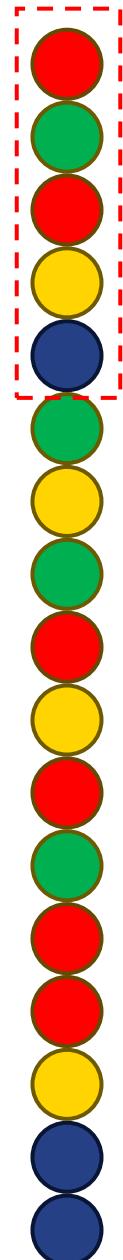
# Phương pháp lấy mẫu

- 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)



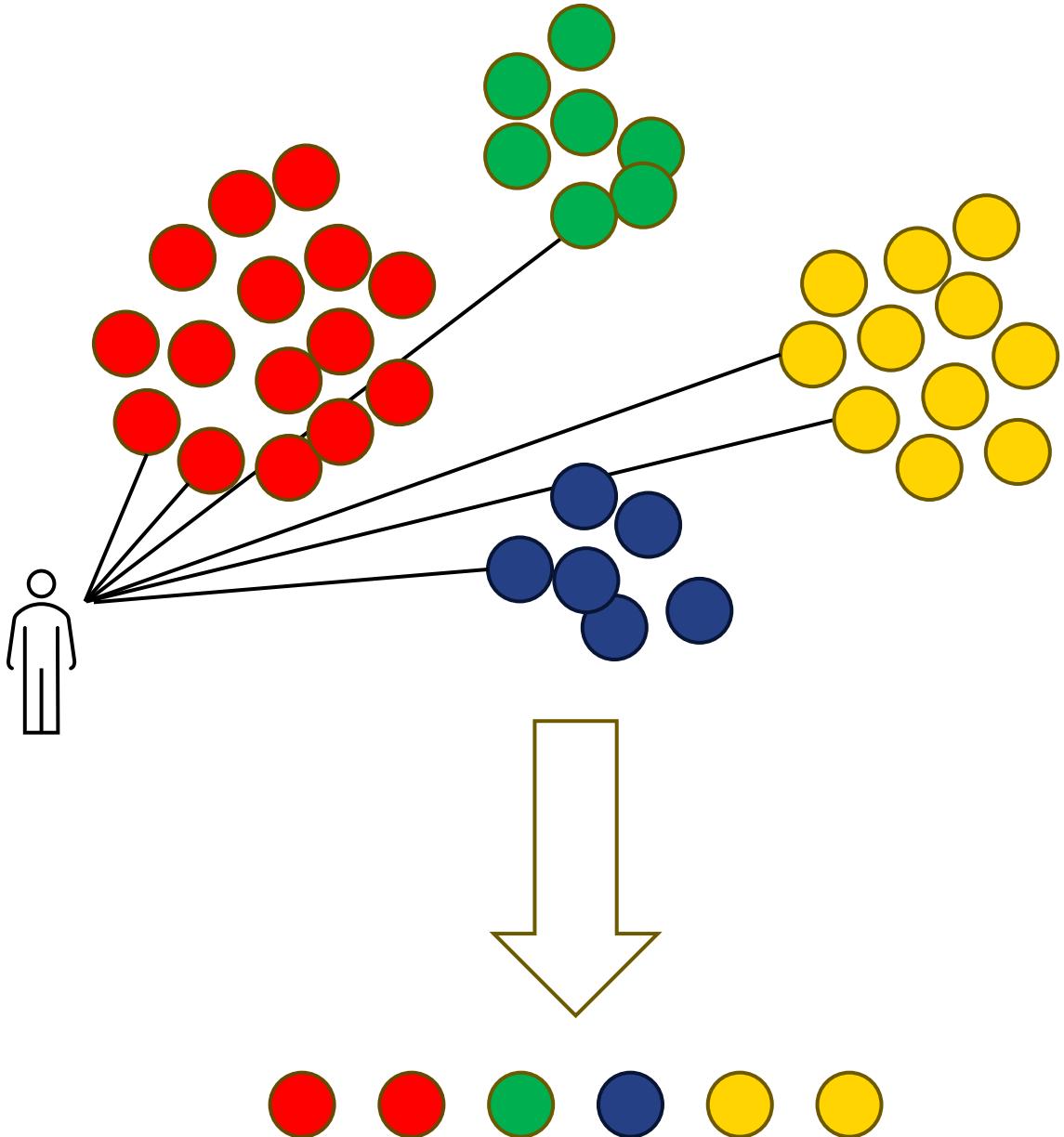
# Phương pháp lấy mẫu

- 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)



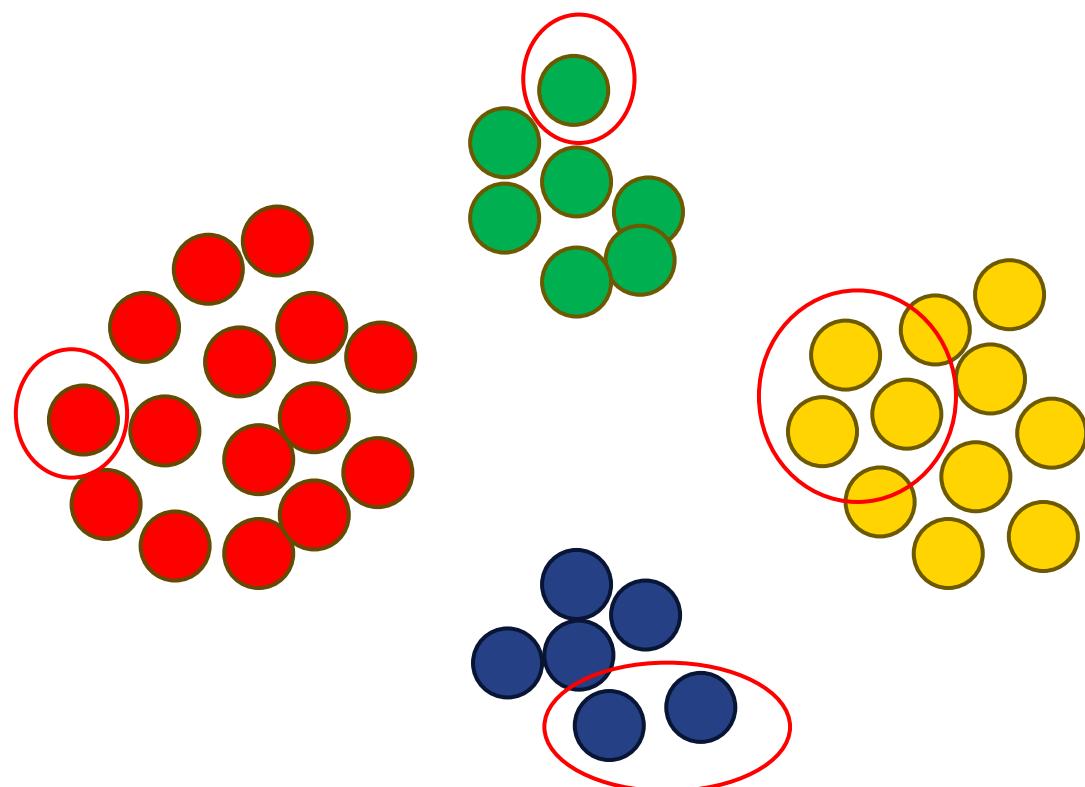
# Phương pháp lấy mẫu

- 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

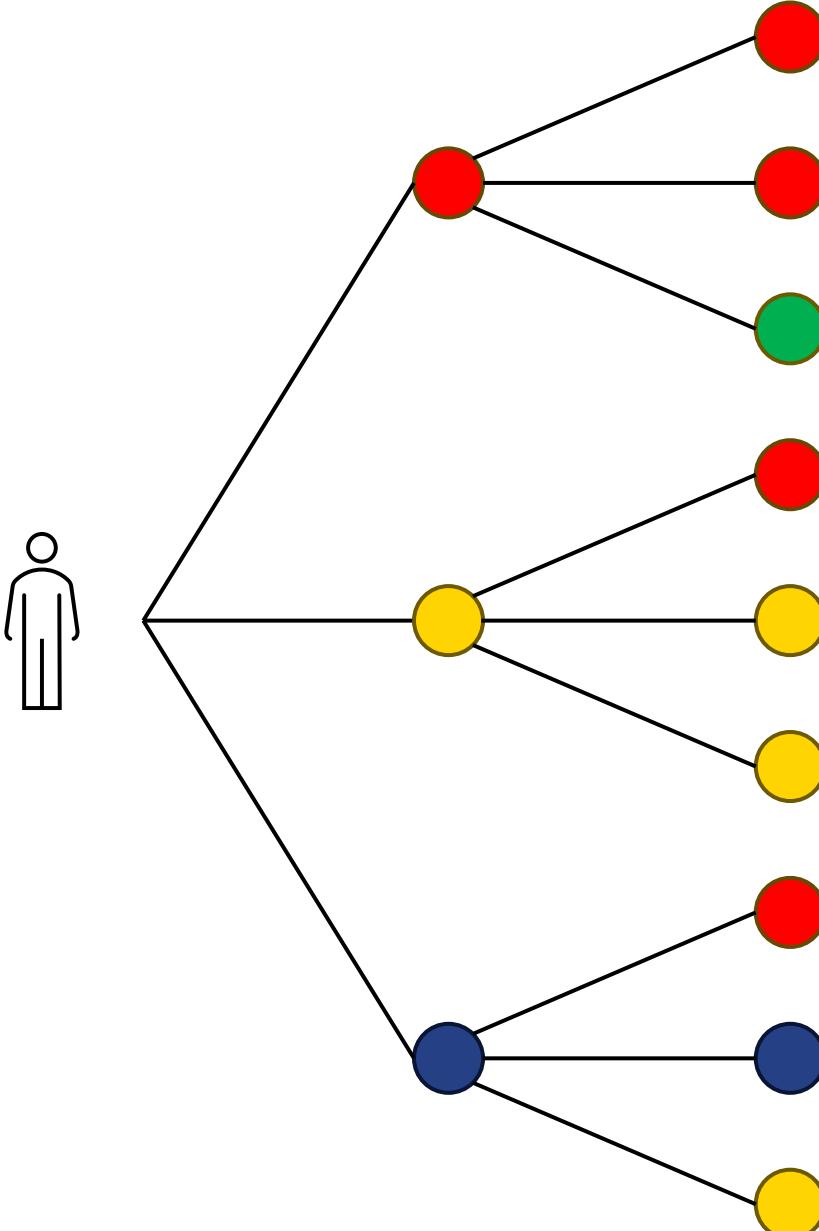
- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)
  - Mẫu ngẫu nhiên cụm (Cluster sample)
- Mẫu không ngẫu nhiên  
(Nonprobability sample)
  - Mẫu thuận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

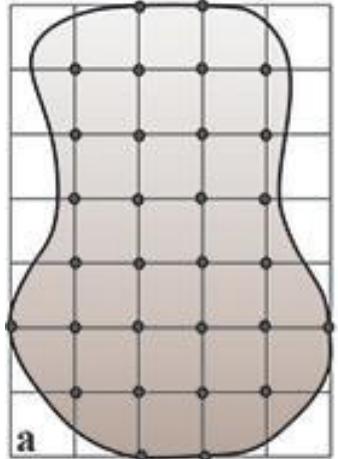
- Mẫu ngẫu nhiên  
(Probability/Random sample)
  - Mẫu ngẫu nhiên đơn giản  
(Simple random sample)
  - Mẫu ngẫu nhiên hệ thống  
(Systematic sample)
  - Mẫu ngẫu nhiên phân loại  
(Stratified sample)
  - Mẫu ngẫu nhiên cụm (Cluster sample)
- Mẫu không ngẫu nhiên  
(Nonprobability sample)
  - Mẫu thuận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



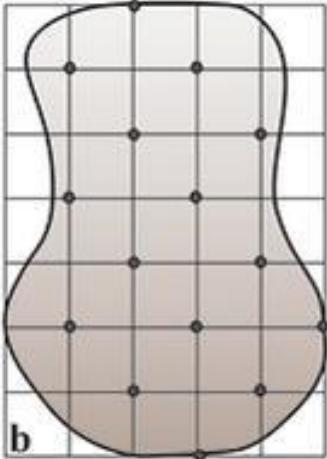
# Phương pháp lấy mẫu

- 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

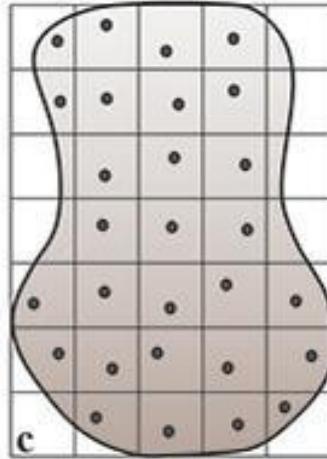
# Mẫu môi trường



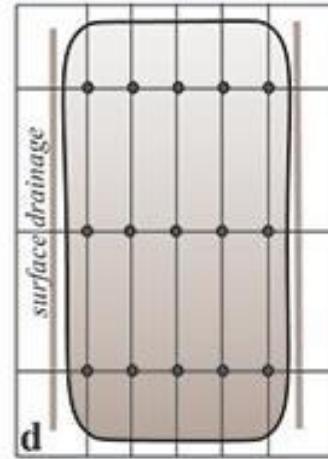
**Uniform Grid**  
• *more information  
greater cost*



**Offset Grid**  
• *economical*

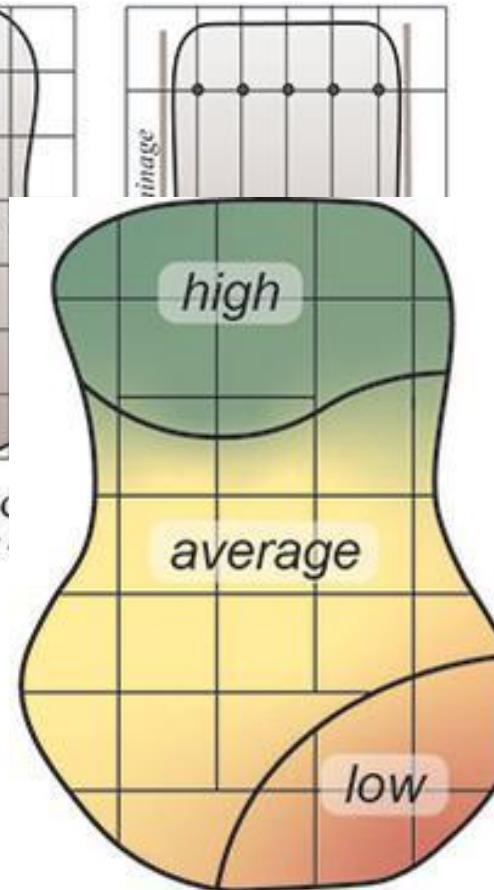
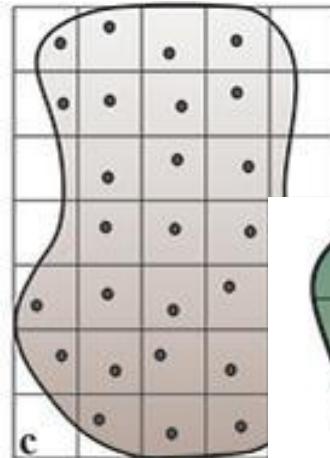
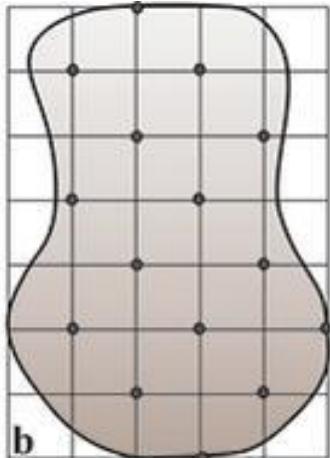
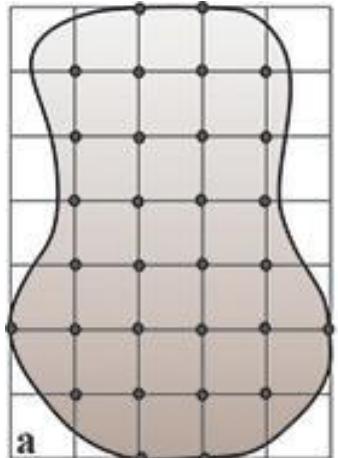


**Stratified Random**  
• *avoid systematic bias*

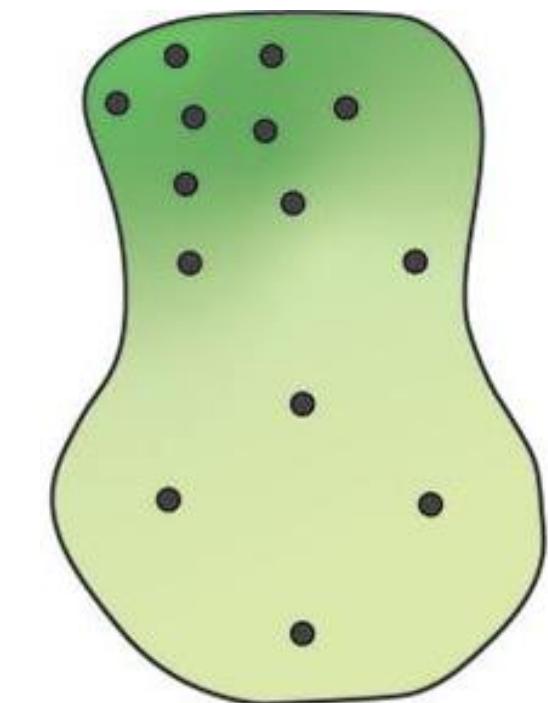


**Rectangular Grid**  
• *capture directional bias*

# Mẫu môi trường



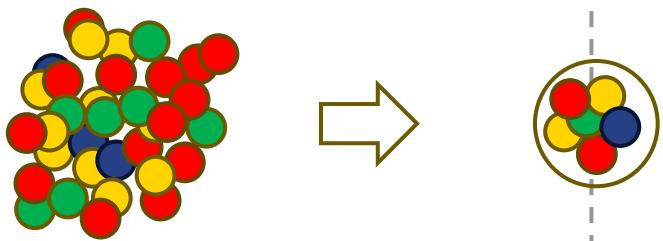
a. Hybrid Cell Sampling



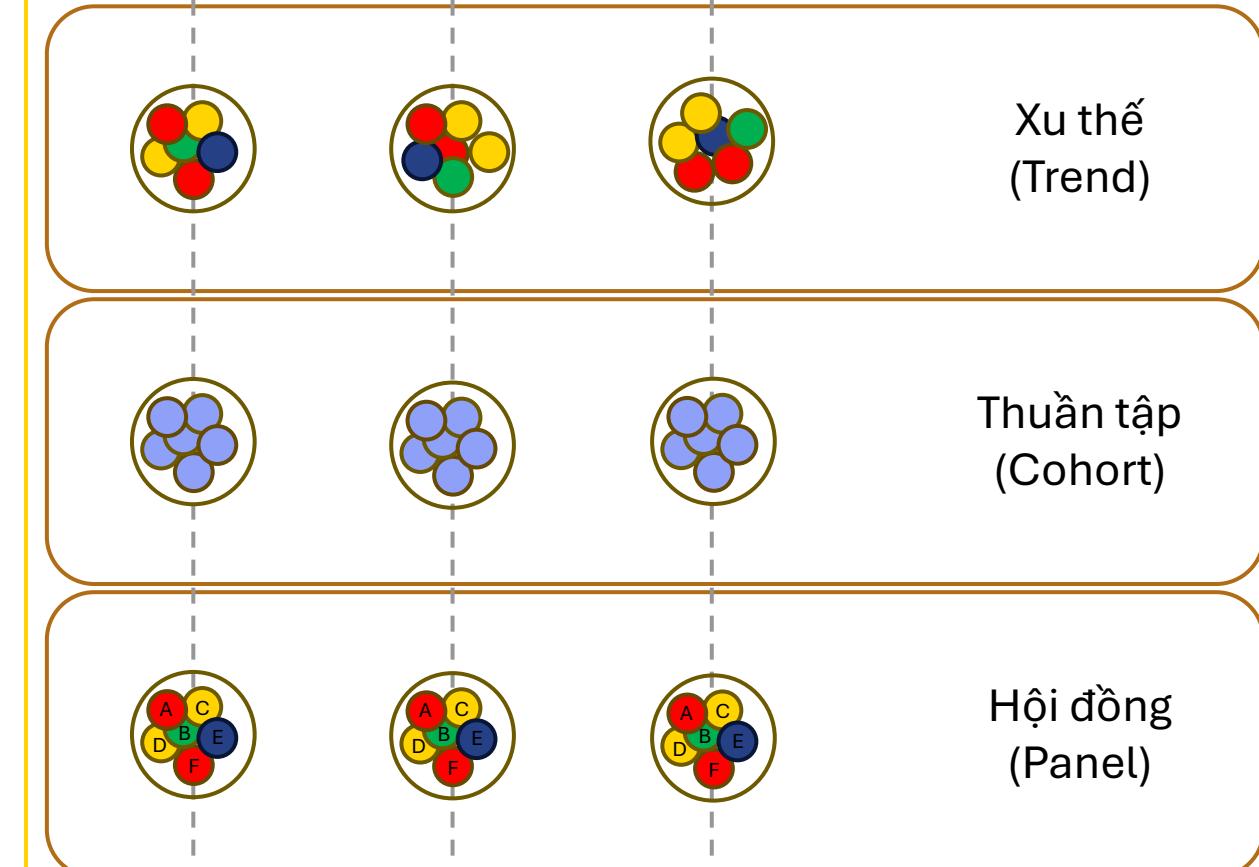
b. Hybrid Point Sampling

# 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

# Bài tập

- Đối tượng nghiên cứu?
- Đơn vị của đối tượng nghiên cứu?
- Bạn muốn quan sát hay đo lường gì ở đối tượng nghiên cứu?
- Quần thể trong nghiên cứu của bạn?
- Dự kiến lấy mẫu?
- Yếu tố thời gian?

# Thu thập dữ liệu

# 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

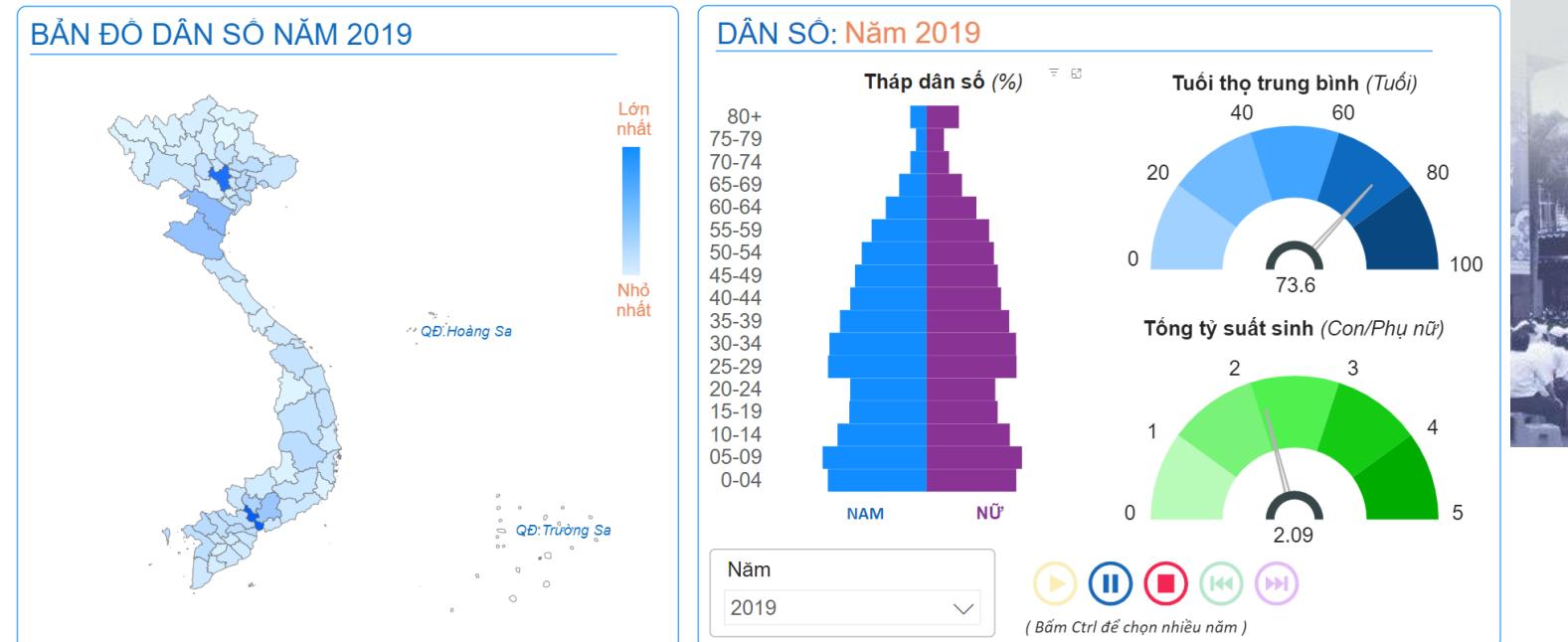
## Cơ hội

- 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?". Below it is the site's logo and name "data.gov.vn". A sub-header reads "Điểm đầu mối công bố dữ liệu mở, cung cấp thông tin về chia sẻ dữ liệu của cơ quan nhà nước". On the left, there are two main categories: "DỮ LIỆU MỞ" (Open Data) and "Địa phương" (Local Government). On the right, there are icons representing various sectors like agriculture, environment, and economy. The main content area features a large title "Kho dữ liệu chuyển đổi số và Cổng dữ liệu mở thành phố Hà Nội" (Digital Data Warehouse and Open Data Portal of Hanoi City). Below this, there's a section titled "Kho dữ liệu (Data warehouse)" with a sub-description: "Data warehouse (DW) hay kho dữ liệu là một hệ thống lưu trữ dữ liệu từ nhiều nguồn, nhiều môi trường khác nhau." To the right, there's an illustration of people interacting with a large computer system.

The screenshot shows the homepage of the Ministry of Planning and Investment's open data portal. At the top, there's a search bar and a sub-header about the ministry's role in managing data. Below is the ministry's logo and name "BỘ KẾ HOẠCH VÀ ĐẦU TƯ CÔNG DỮ LIỆU CỦA BỘ KẾ HOẠCH VÀ ĐẦU TƯ". The main content area is titled "DANH MỤC DỮ LIỆU MỞ CỦA BỘ KẾ HOẠCH VÀ ĐẦU TƯ". It displays a grid of nine data categories: "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 tư", "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

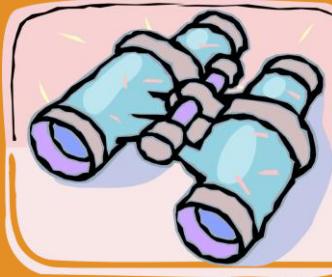
Phần trăm dân số sử dụng dịch vụ nước uống, công trình vệ sinh, và vệ sinh cơ bản, chia theo vùng/thành phố

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

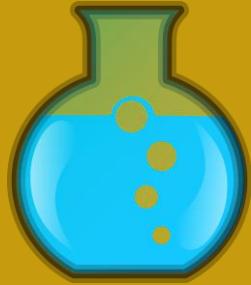
# Thu thập dữ liệu sơ cấp



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



# 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

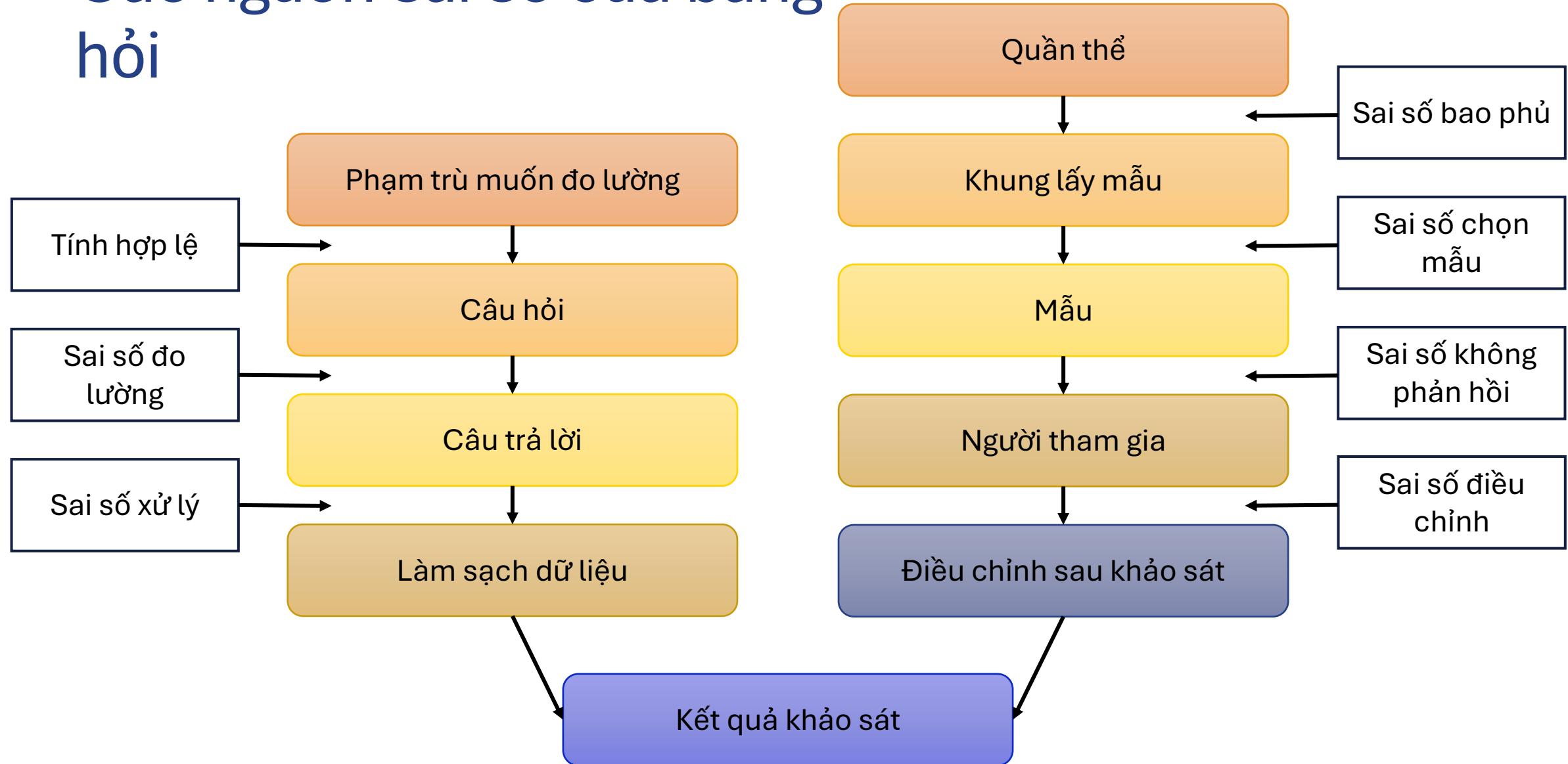
# Lưu ý khi thiết kế bảng hỏi

- 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

- Một số thông số sẽ đòi hỏi đo trực tiếp tại hiện trường (vd. T, pH, độ đục)
- Tuân thủ quy trình
- Chú ý về vấn đề lưu trữ và bảo quản



# Hiện trường

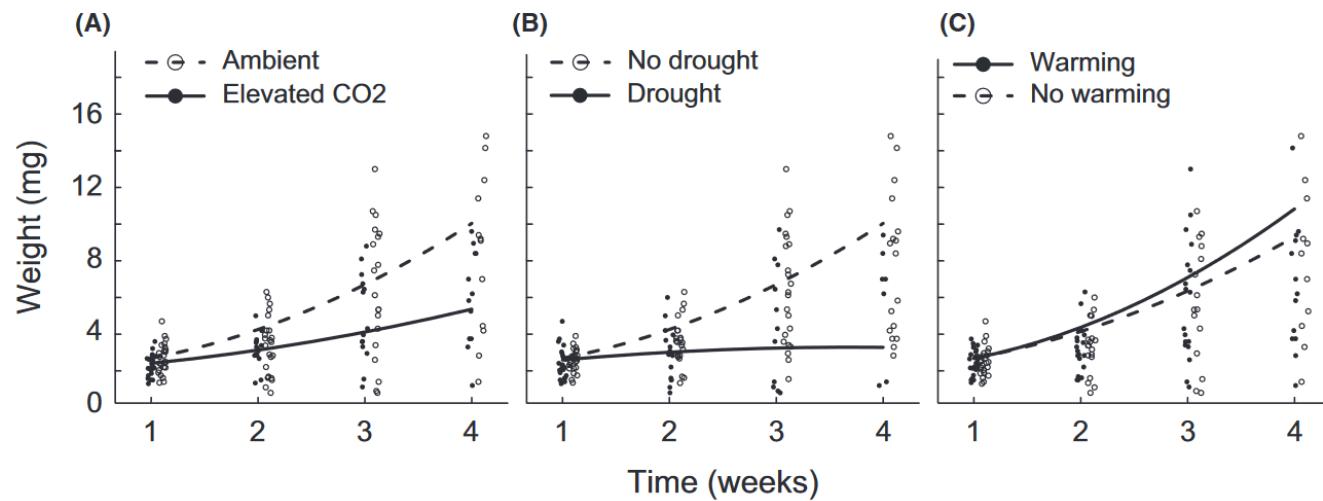
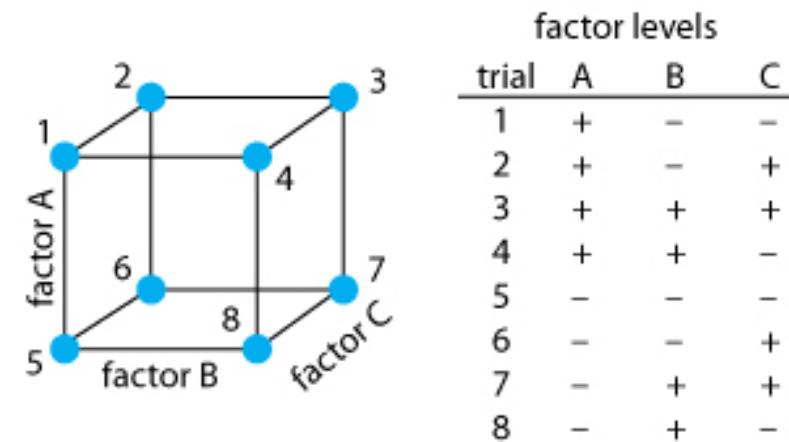
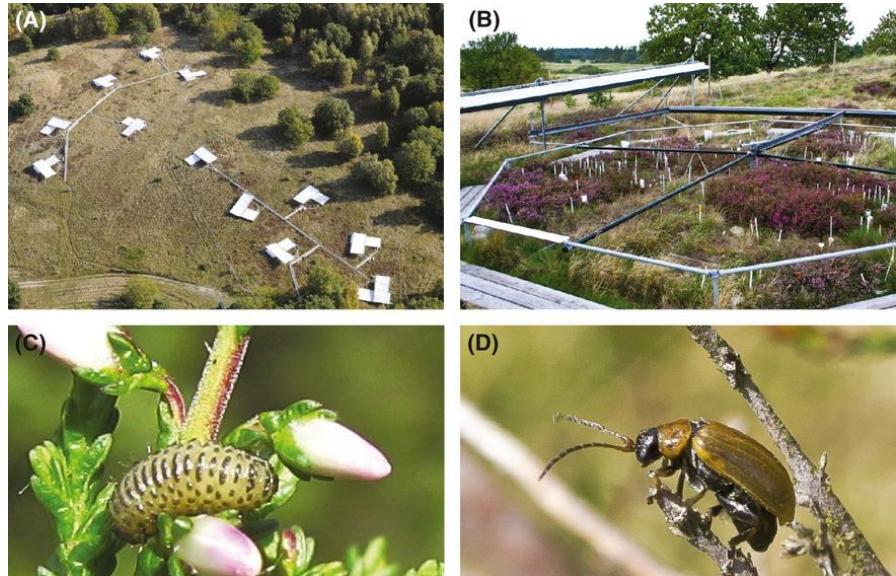


# Hiện trường

- 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

- Biến phụ thuộc
- Biến độc lập



Scherber, C., Gladbach, D. J., Stevnbak, K., Karsten, R. J., Schmidt, I. K., Michelsen, A., Albert, K. R., Larsen, K. S., Mikkelsen, T. N., Beier, C., & Christensen, S. (2013). Multi-factor climate change effects on insect herbivore performance. *Ecology and Evolution*, 3(6), 1449–1460. <https://doi.org/10.1002/ece3.564>

# Thực nghiệm



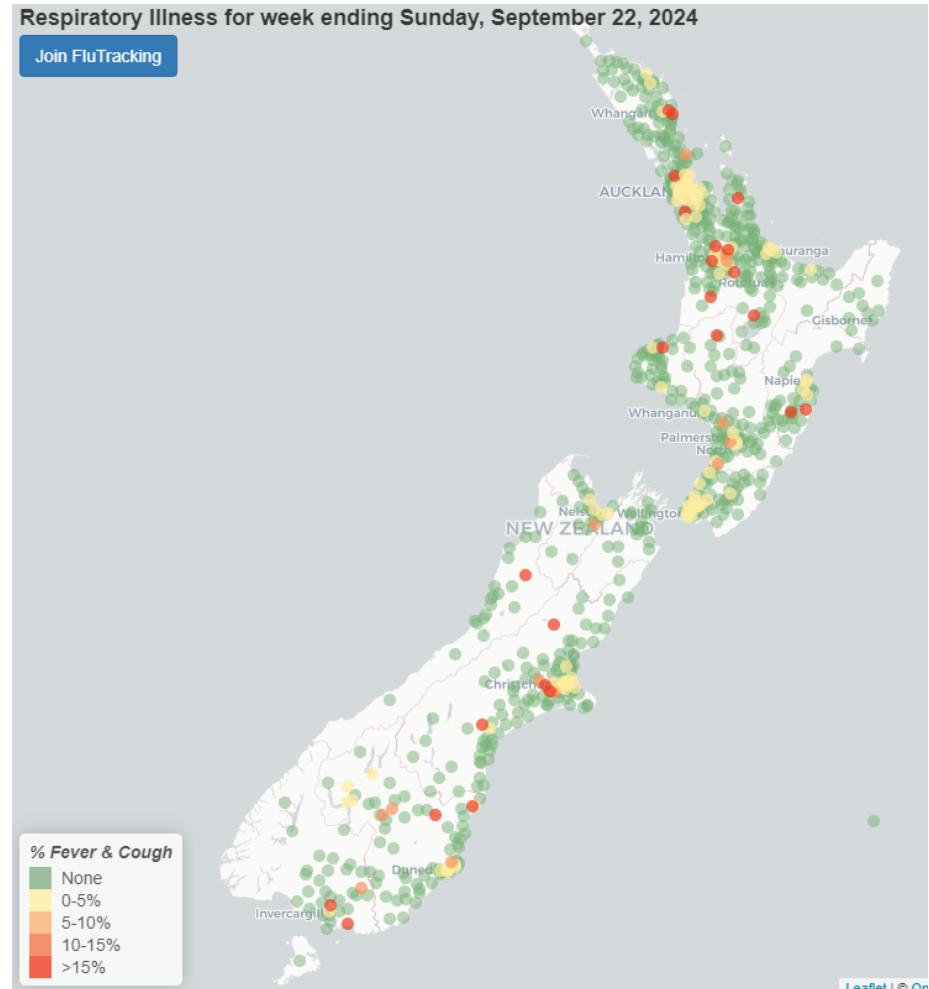
Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, 322(5908), 1681–1685.  
<https://doi.org/10.1126/science.1161405>

# 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

- <https://info.flutracking.net>
- 17 năm ở Úc, 5 năm ở New Zealand
- 110000 tình nguyện viên tham gia trả lời bảng hỏi mỗi tuần
- Trả lời hàng tuần các triệu chứng có thể của cúm/COVID
  - Sốt
  - Ho
  - ...



# Bài tập

- Dự kiến bạn sẽ cần dữ liệu nào cho nghiên cứu của bạn?
- Nguồn sẽ là sơ cấp hay thứ cấp?
- Các nguồn dữ liệu thứ cấp mà bạn có thể nghĩ đến?
- Dữ liệu sơ cấp có thể được thu thập theo cách nào?

# Phân tích

# 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.

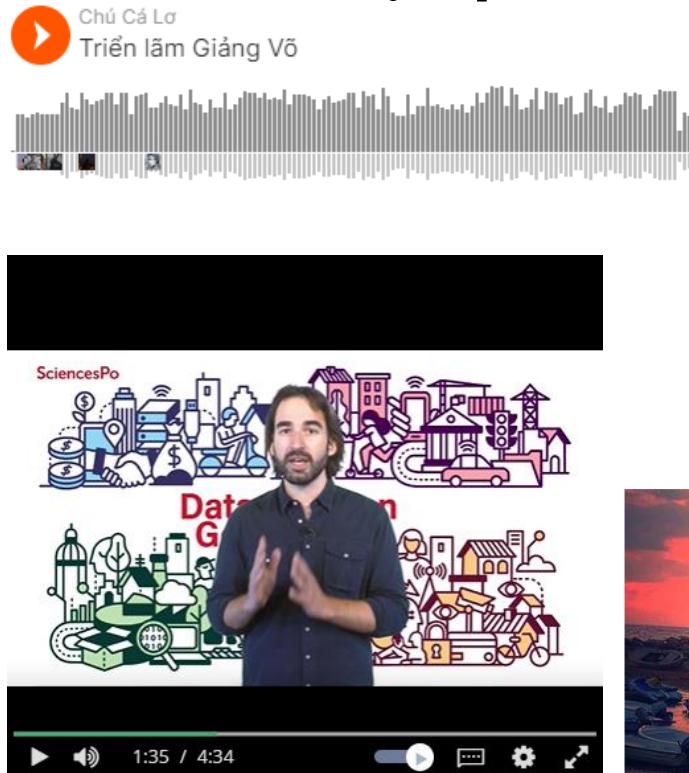
Nguồn: <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch1/definitions/5214853-eng.htm>

# Dữ liệu có cấu trúc và phi cấu trúc

## Dữ liệu có cấu trúc

STT	MSSV	Họ và tên	Ngày sinh
1	23090374	Nguyễn Thị Lan Anh	10/03/2004
2	23090375	Phạm Thị Lan Anh	24/05/2005
3	23090376	Hoàng Ngọc Anh	14/10/2005
4	23090377	Nguyễn Hoàng Phương Anh	27/06/2005
5	23090379	Hoàng Tâm Anh	11/12/2005
6	23090380	Nguyễn Tuấn Anh	14/10/2005
7	23090381	Nguyễn Quốc Bảo	27/07/2005

## Dữ liệu phi cấu trúc



3 years ago



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.

Sampson

I mean, an we be in choler we'll draw,

If angered (our swords)

Gregory

1

2

3

4

5

# 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
Vadimii 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
  - Nhiệt độ không khí
  - PM<sub>2,5</sub> ( $\mu\text{g}/\text{m}^3$ )

# Biến định tính/rời rạc – Biến định danh

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

# Biến định tính/rời rạc – Biến thứ bậc

- 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

# 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

	<b>id</b>	<b>csmptv</b>	<b>rwtank</b>	<b>iceqac2</b>	<b>hhs_tot</b>	<b>cfdiwq</b>	<b>livara</b>
	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

# 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 lặp dữ liệu

# Tiền xử lý dữ liệu

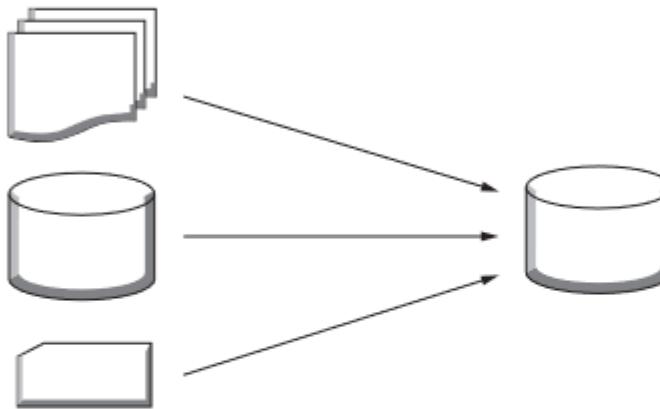
... là bước quan trọng để chuyển từ dữ liệu thô sang dạng dữ liệu  
**sẵn sàng sử dụng** cho bước phân tích

# Tiền xử lý dữ liệu

- Làm sạch dữ liệu



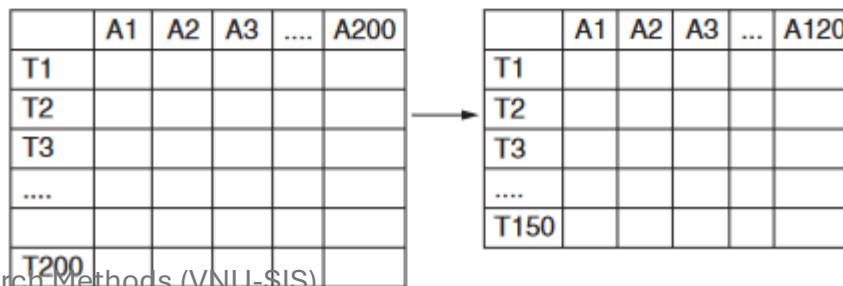
- Tích hợp dữ liệu



- 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

ID	steps	bmi
3	15000	17.0
4	14861	17.2
5		
9		
12		
14		
15	15000	16.9
16	15000	16.9
21	14861	16.8
23	14861	16.8
26	14699	17.3
28	14560	20.5
31	14560	20.6
33	14560	20.5
34	14560	20.4
35	14560	20.4
36	14560	19.8
38	14560	19.7
39	14560	19.7
41	14560	19.6
44	14560	19.6
45	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	20.0

# Biểu diễn dữ liệu

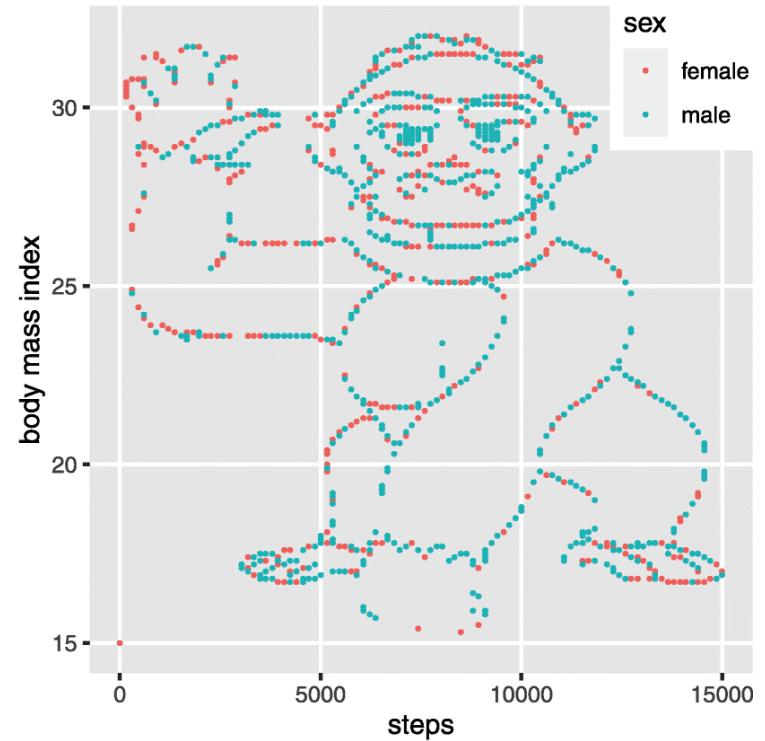
- BMI và số bước chân
  - Nam
  - Nữ
- 2 nhóm sinh viên
  - Giả thuyết: Tương quan giữa số bước chân và BMI
  - Có thể rút ra được gì từ tập dữ liệu này

a

The image shows two separate windows of the Lister application side-by-side. Both windows have a header bar with File, Edit, Options, Encoding, Help, and a progress bar indicating 3% and 4% completion respectively. The left window displays data for 'steps' and 'bmi' with rows 3 and 4 visible. The right window displays data for 'steps' and 'bmi' with rows 1 through 43 visible.

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
23	7	14861
26	8	14699
28	10	14560
31	11	14560
33	13	14560
34	17	14560
35	18	14560
36	19	14560
38	20	14560
39	22	14560
41	24	14560
44	25	14560
45	27	14560
29	29	14560
30	30	14560
32	32	14398
37	37	14398
40	40	14398
42	42	14259
43	43	14259
44	44	14259
45	45	14259

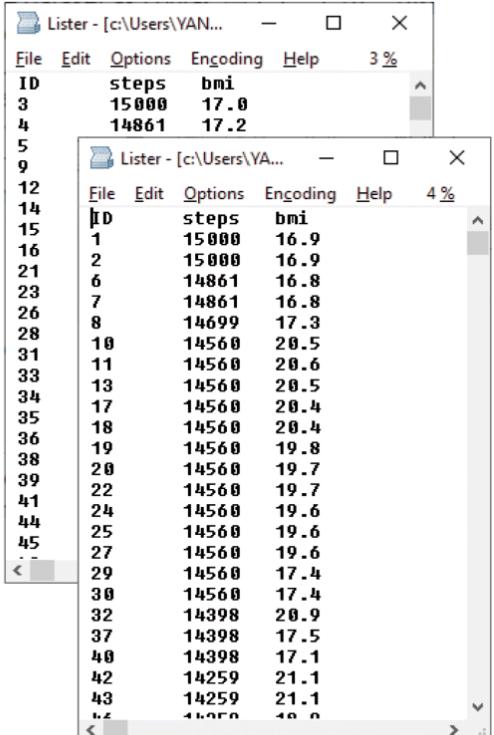
b



# Biểu diễn dữ liệu

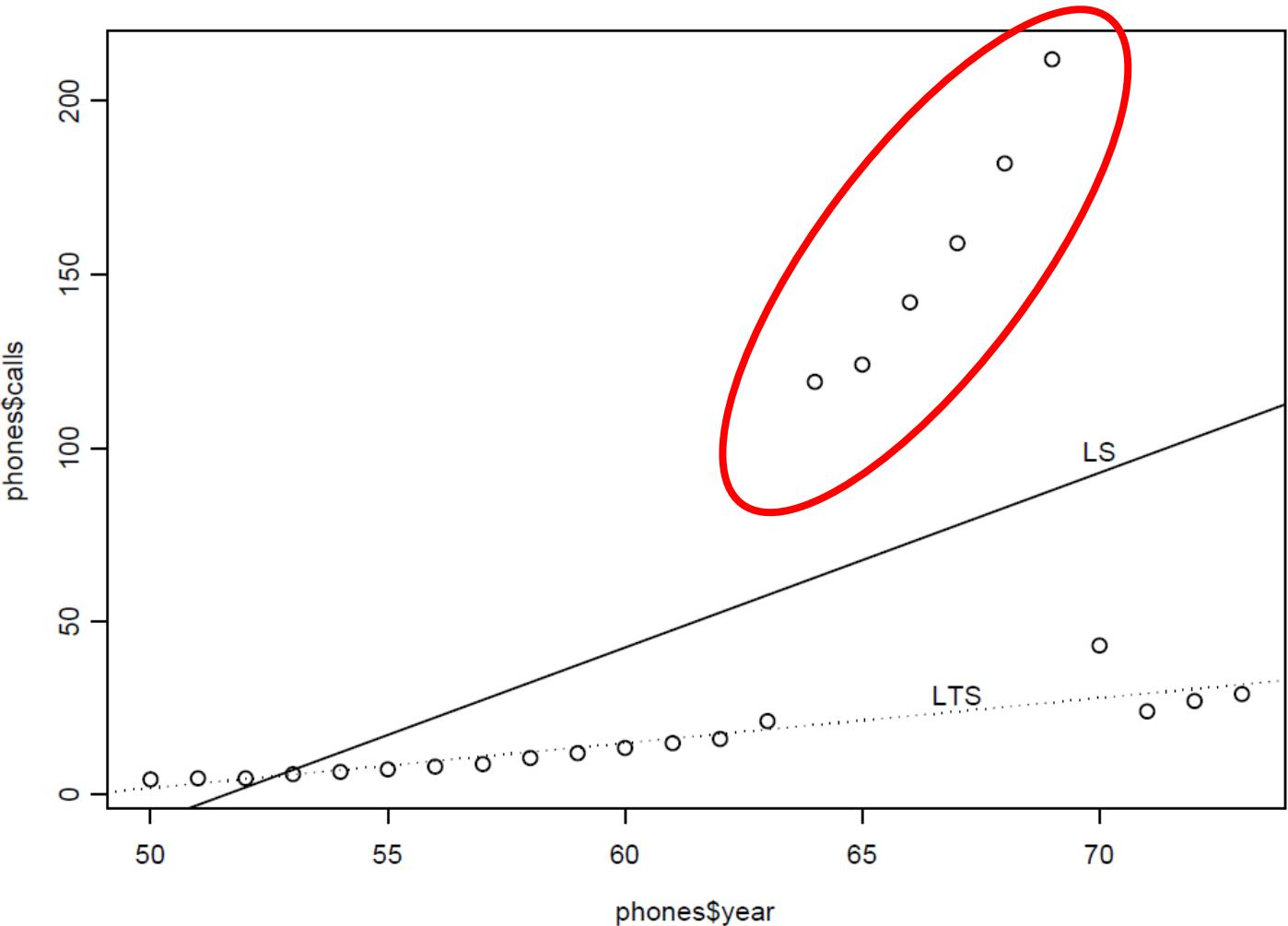
- BMI và số bước chân
  - Nam
  - Nữ
- 2 nhóm sinh viên
  - Giả thuyết: Tương quan giữa số bước chân và BMI
  - Có thể rút ra được gì từ tập dữ liệu này

a

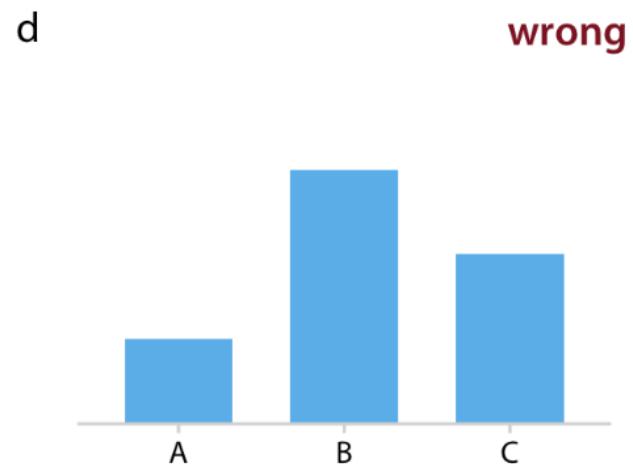
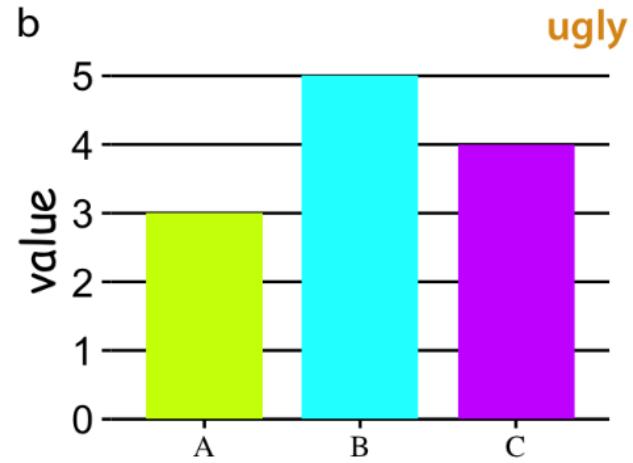
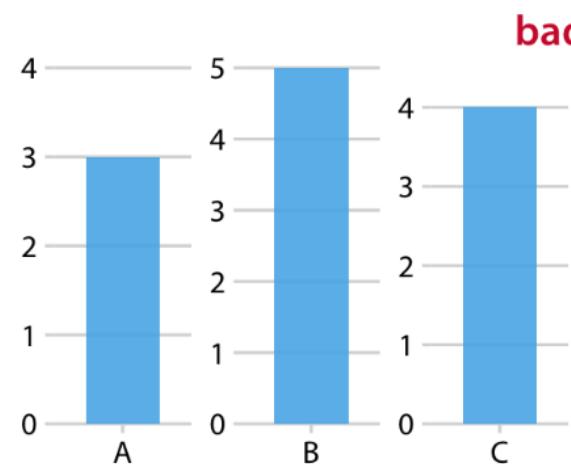
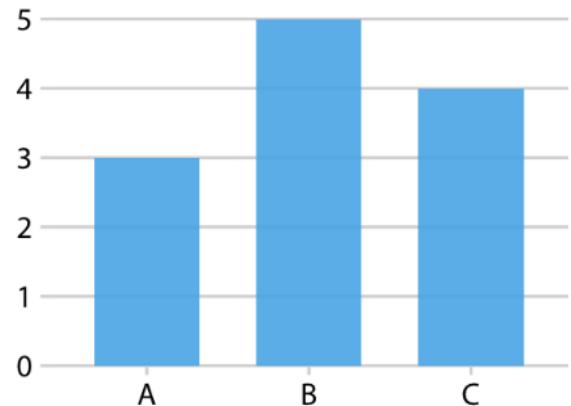


# Biểu diễn dữ liệu

- 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

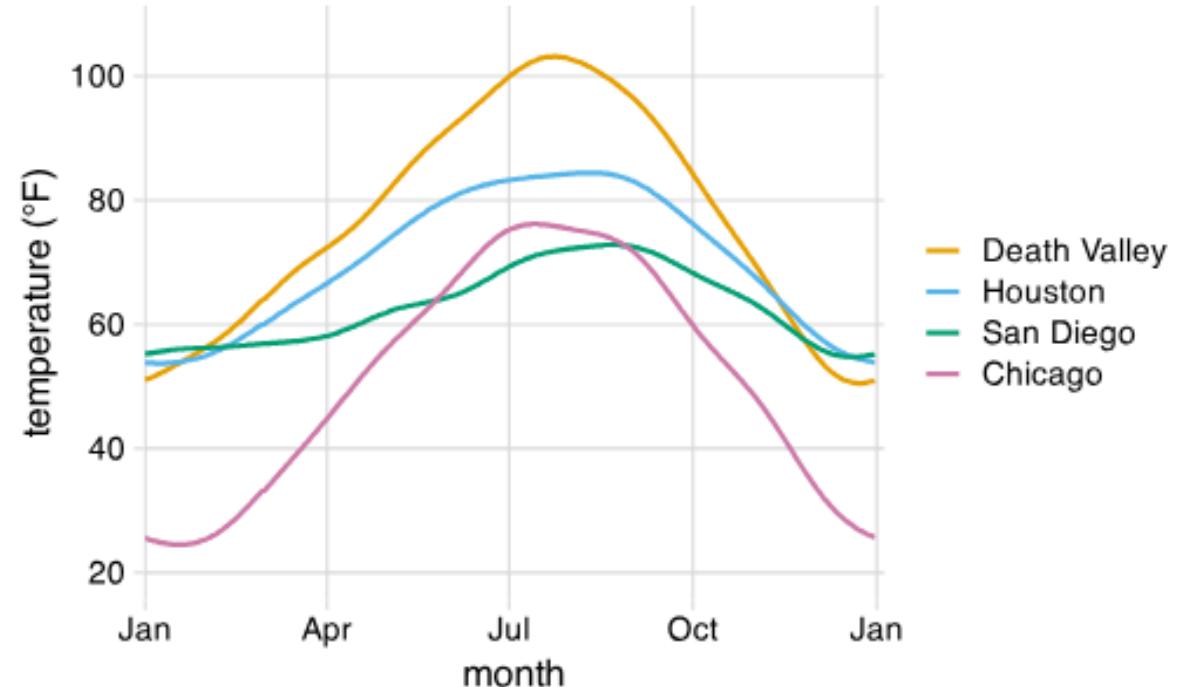


- Rõ ràng
- Chính xác, không gây hiểu nhầm
- Nêu bật thông điệp chính
- [https://www.ted.com/talks/hans\\_rosling\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen](https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen)

# Biểu diễn dữ liệu

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

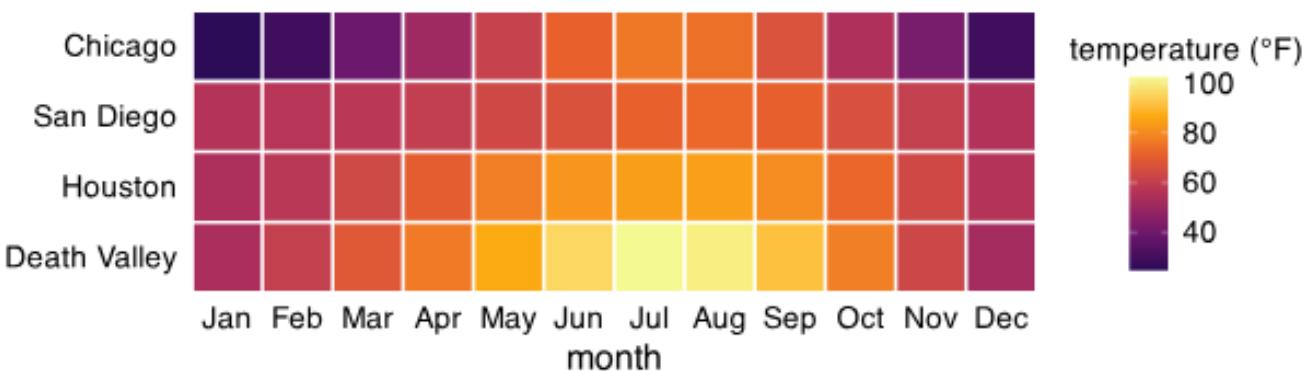
location	day_of_year	month	temperature
Death Valley	1	01	51.0
Death Valley	2	01	51.2
Death Valley	3	01	51.3
Death Valley	4	01	51.4
Death Valley	5	01	51.6
Death Valley	6	01	51.7
Death Valley	7	01	51.9
Death Valley	8	01	52.0
Death Valley	9	01	52.2
Death Valley	10	01	52.3
Death Valley	11	01	52.5



# Biểu diễn dữ liệu

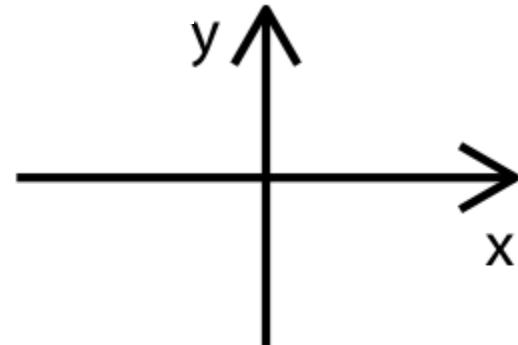
- Sử dụng **màu sắc** để thể hiện nhiệt độ

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



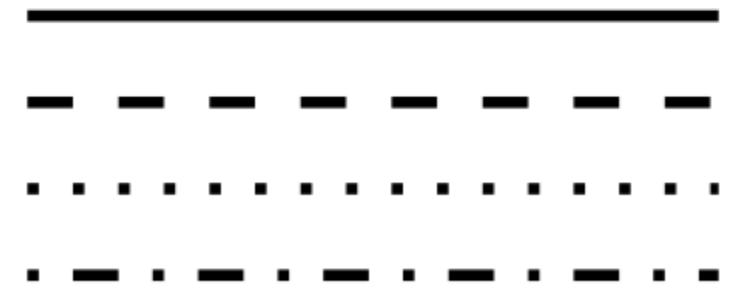
màu sắc



độ dày đường

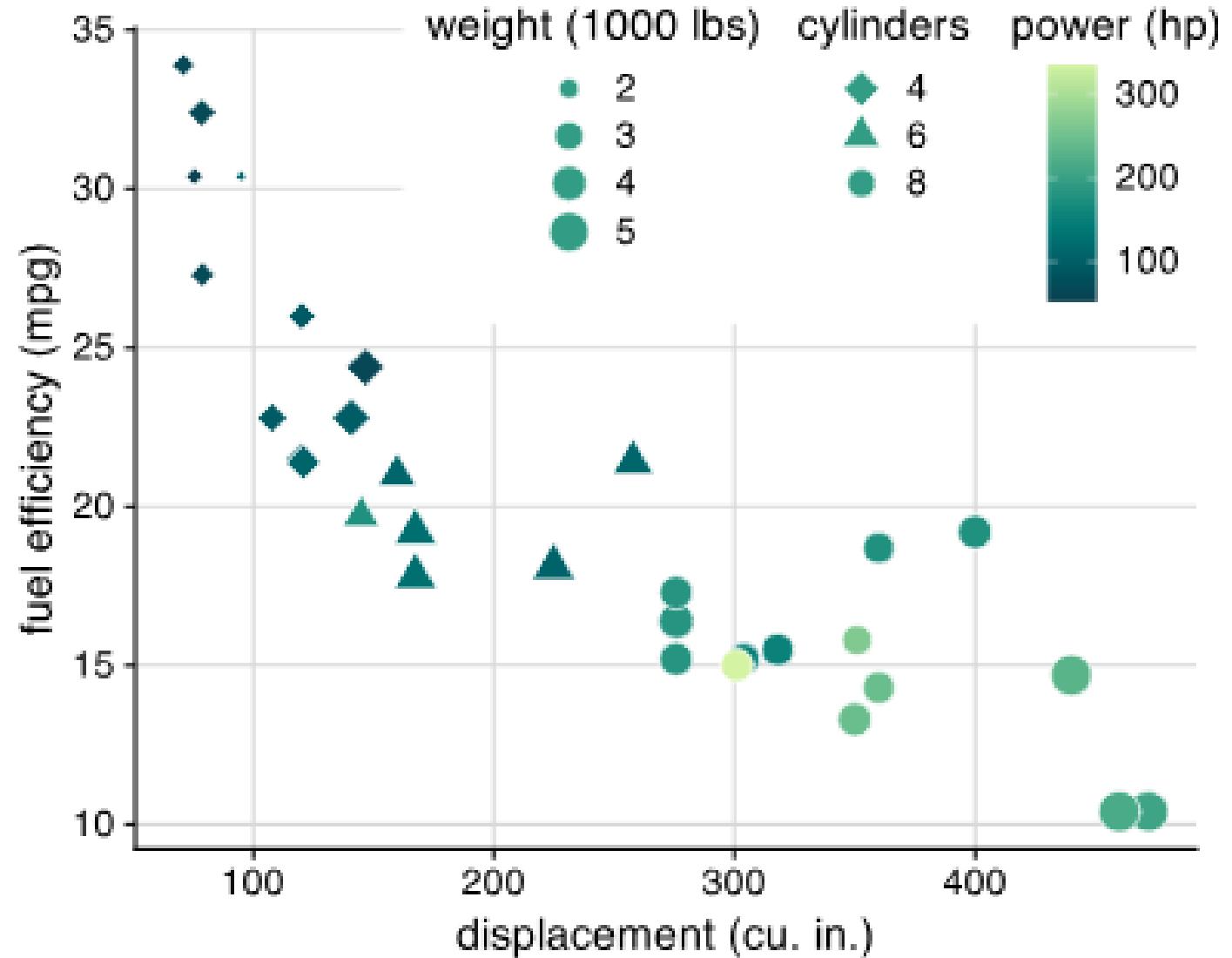


loại đường

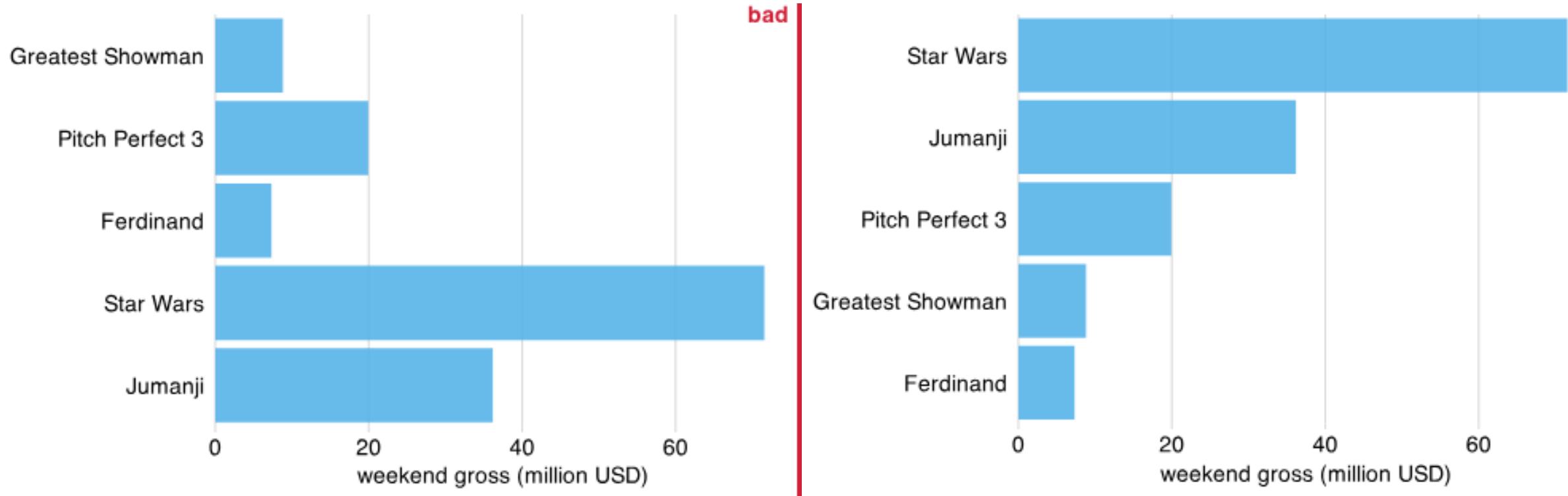


# Biểu diễn dữ liệu

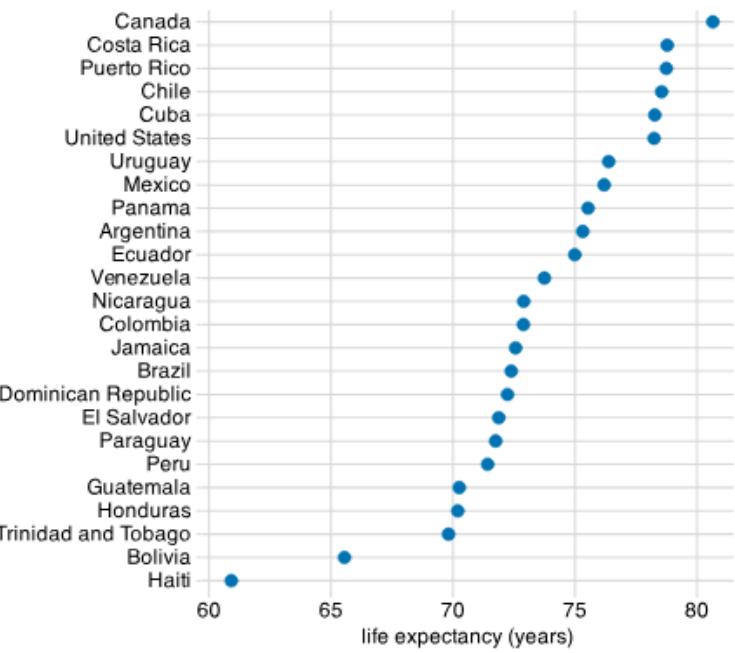
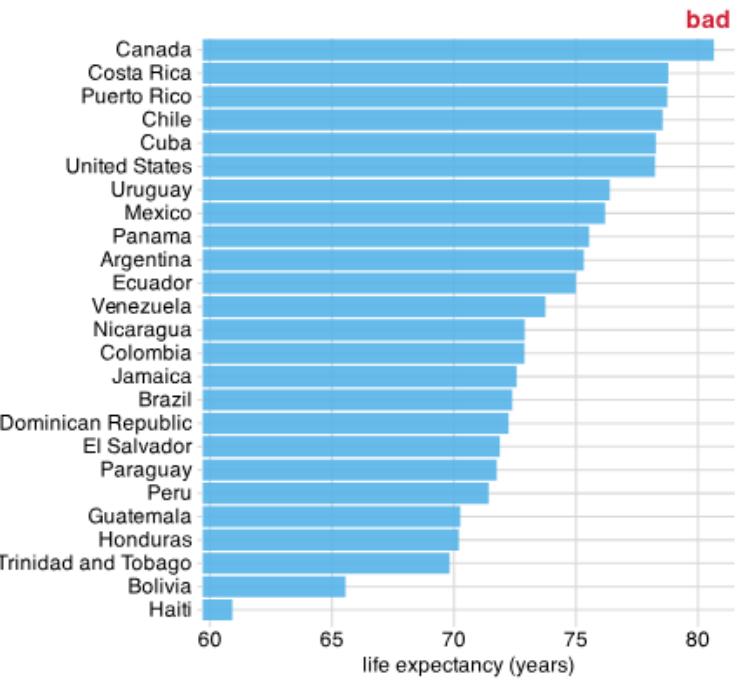
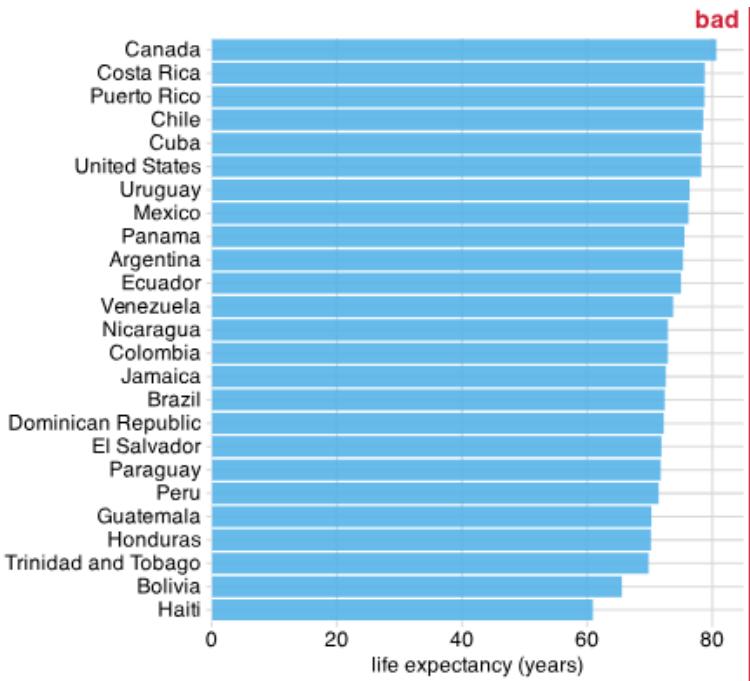
- Có thể sử dụng nhiều yếu tố hội họa trong cùng một đồ thị



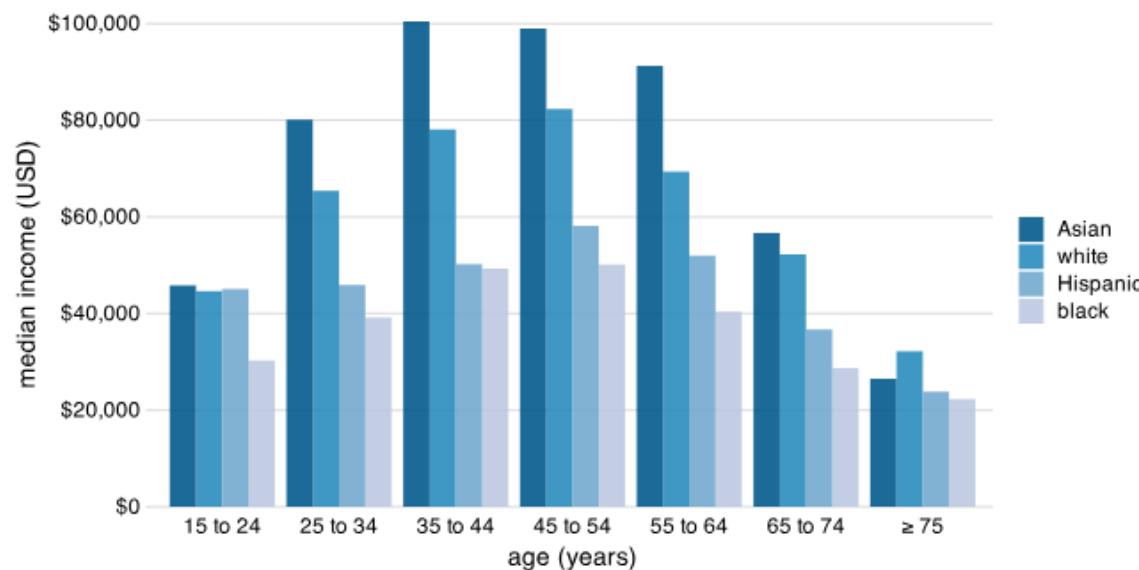
# Biểu diễn dữ liệu - lượng



# Biểu diễn dữ liệu - lượng



# Biểu diễn dữ liệu - lượng



median income (USD)

age (years)

26/10/2024

Research Methods (VNU-SIS)



median income (USD)

age (years)

120

# Biểu diễn dữ liệu – phân phối

- Xác định các khoảng
- Đếm số trường hợp

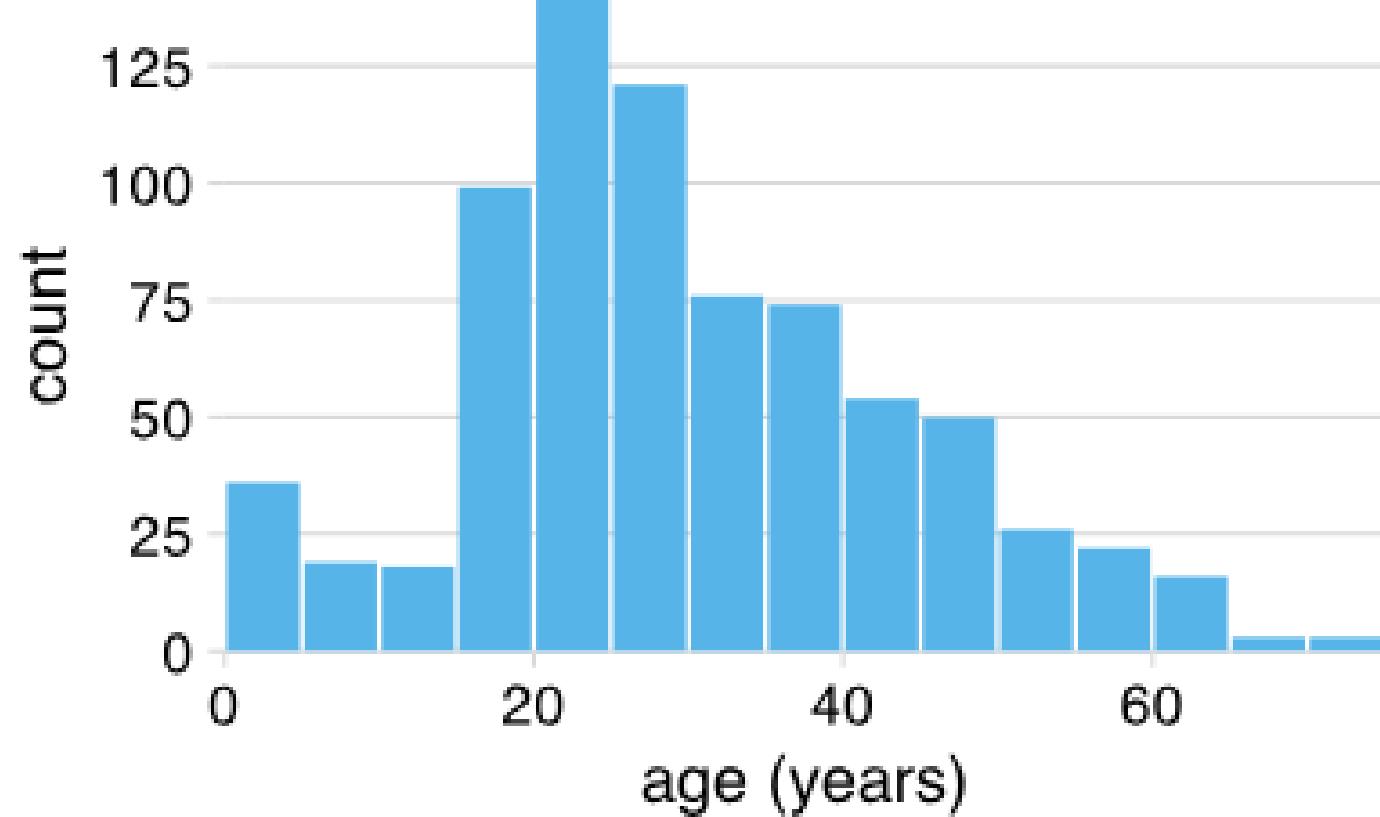
age range	count	age range	count
0-5	36	41-45	54
6-10	19	46-50	50
11-15	18	51-55	26
16-20	99	56-60	22
21-25	139	61-65	16
26-30	121	66-70	3
31-35	76	71-75	3
36-40	74	76-80	0

# Biểu diễn dữ liệu – phân phối

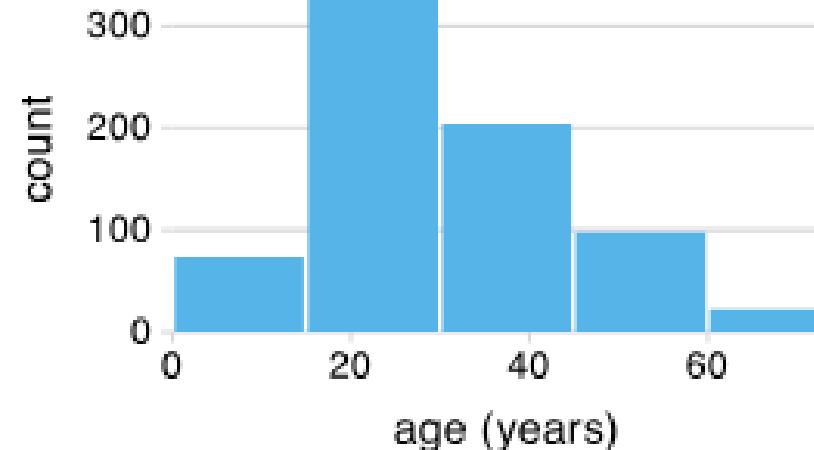
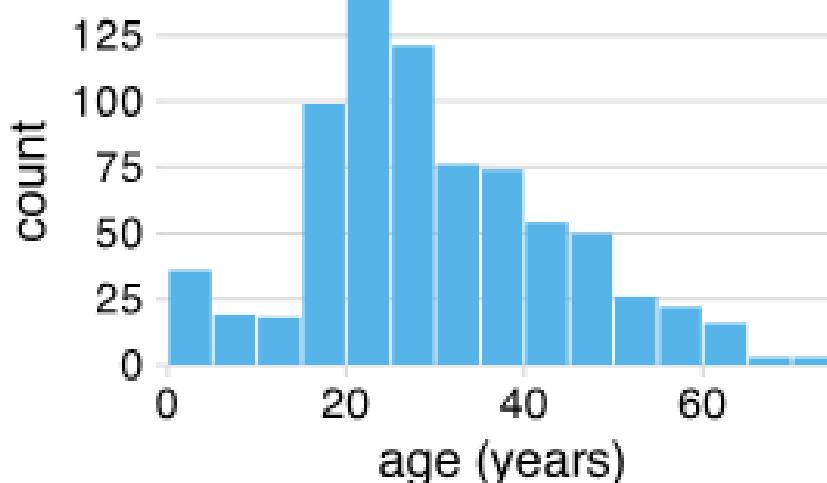
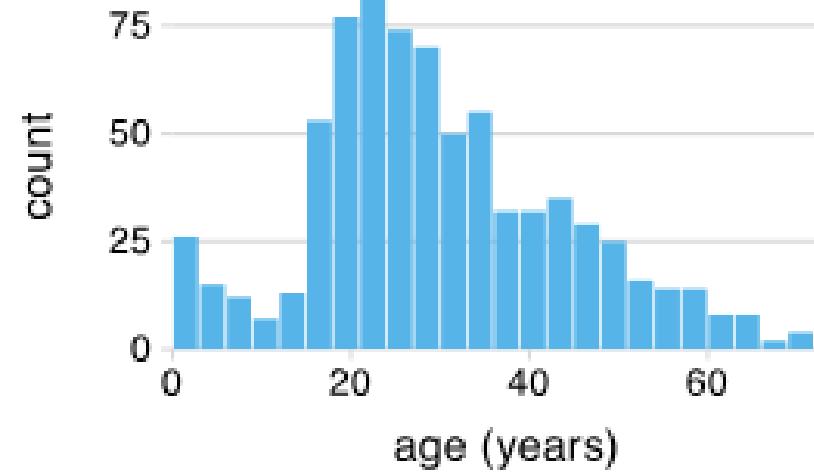
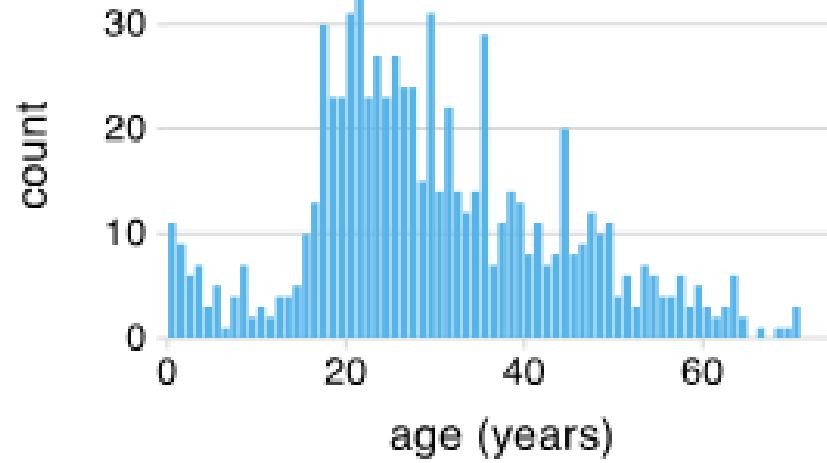
- Xác định các khoảng
- Đếm số trường hợp

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

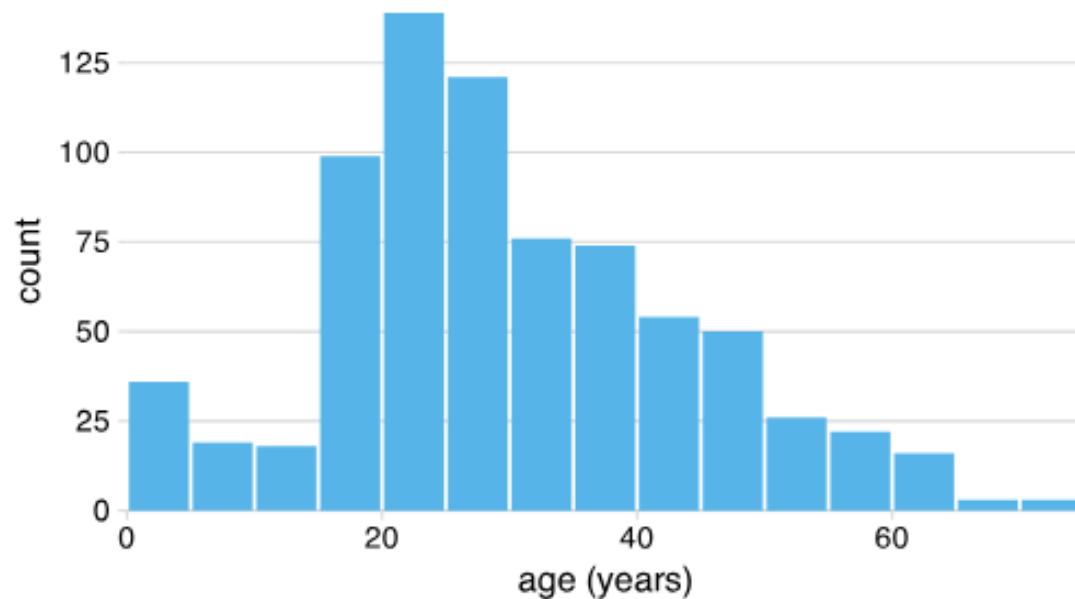


# Biểu diễn dữ liệu – phân phối

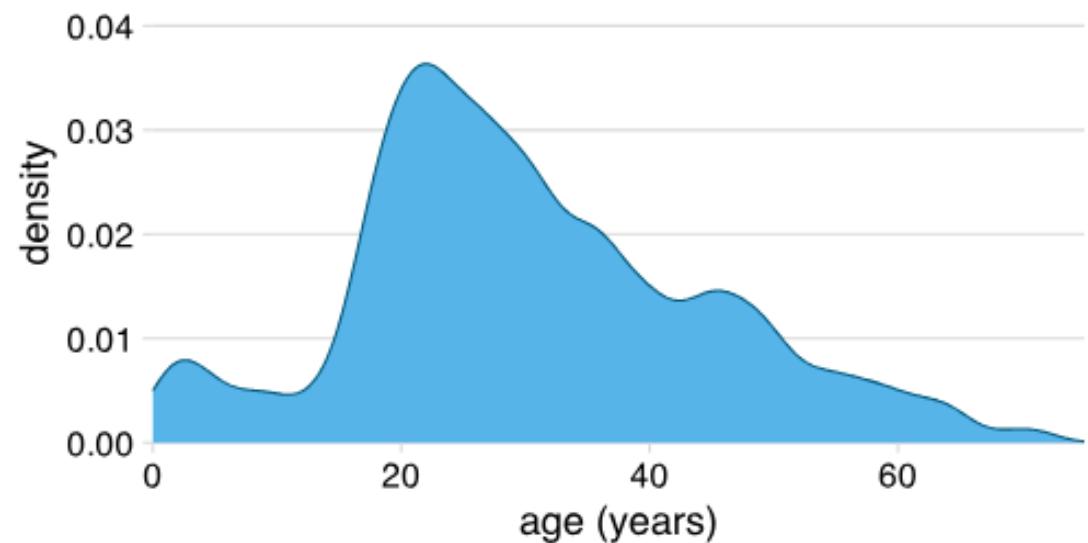


# Biểu diễn dữ liệu – phân phối

Histogram



Kernel density estimate



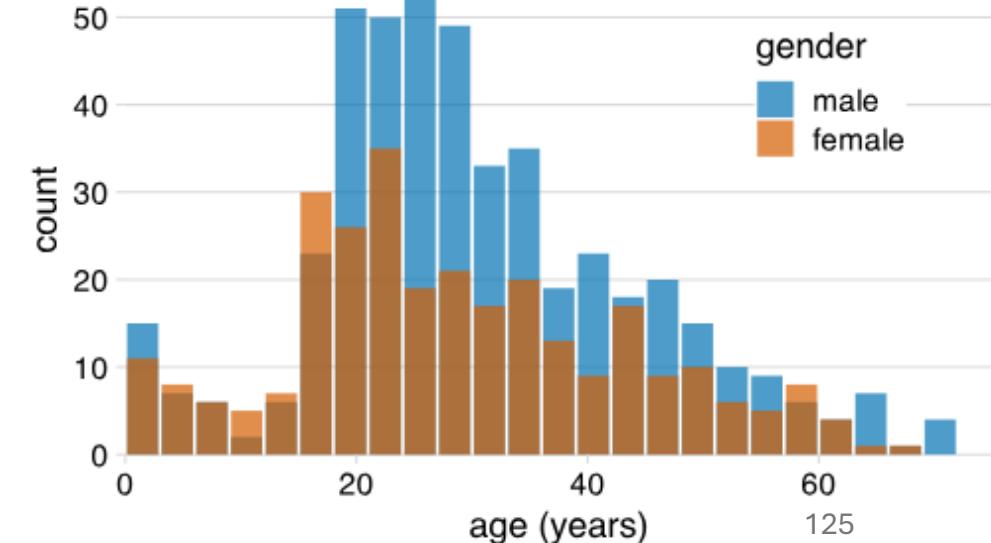
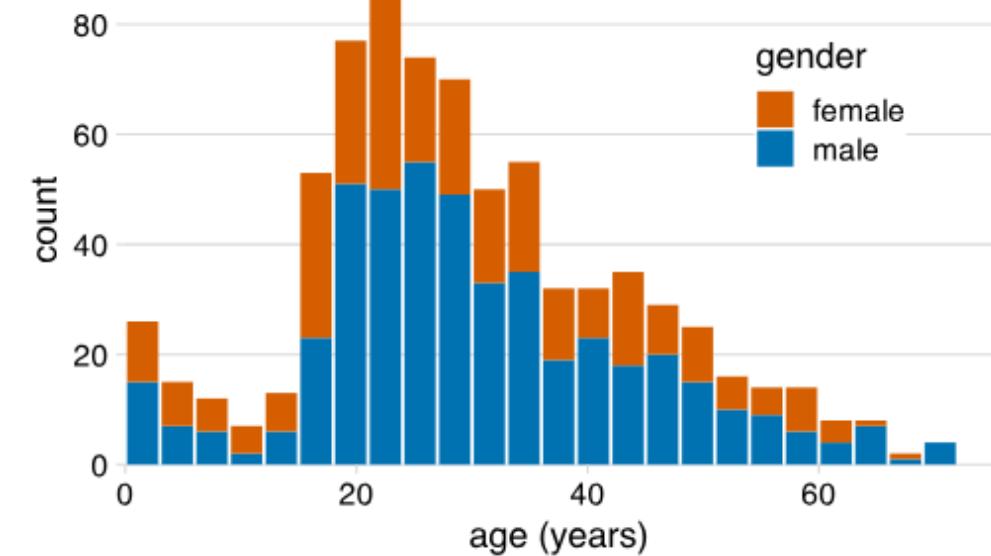
# Biểu diễn dữ liệu – phân phối

**female**

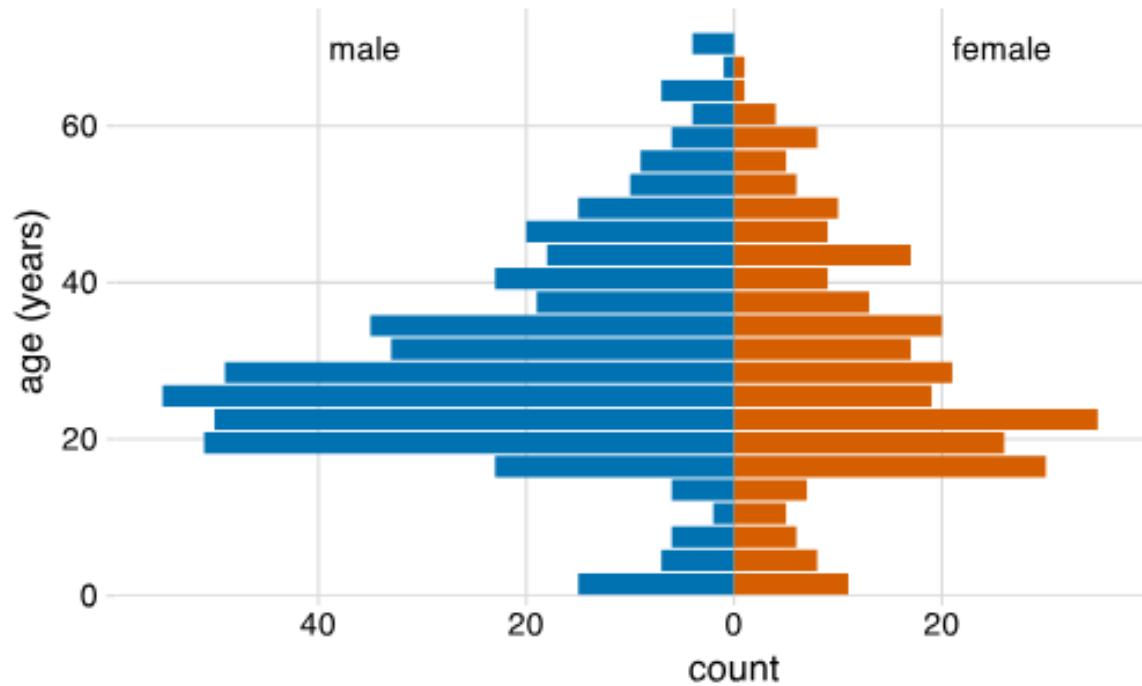
age_c	count
[0,3]	11
(3,6]	8
(6,9]	6
(9,12]	5
(12,15]	7
(15,18]	30
(18,21]	26
(21,24]	35
(24,27]	19
(27,30]	21

**male**

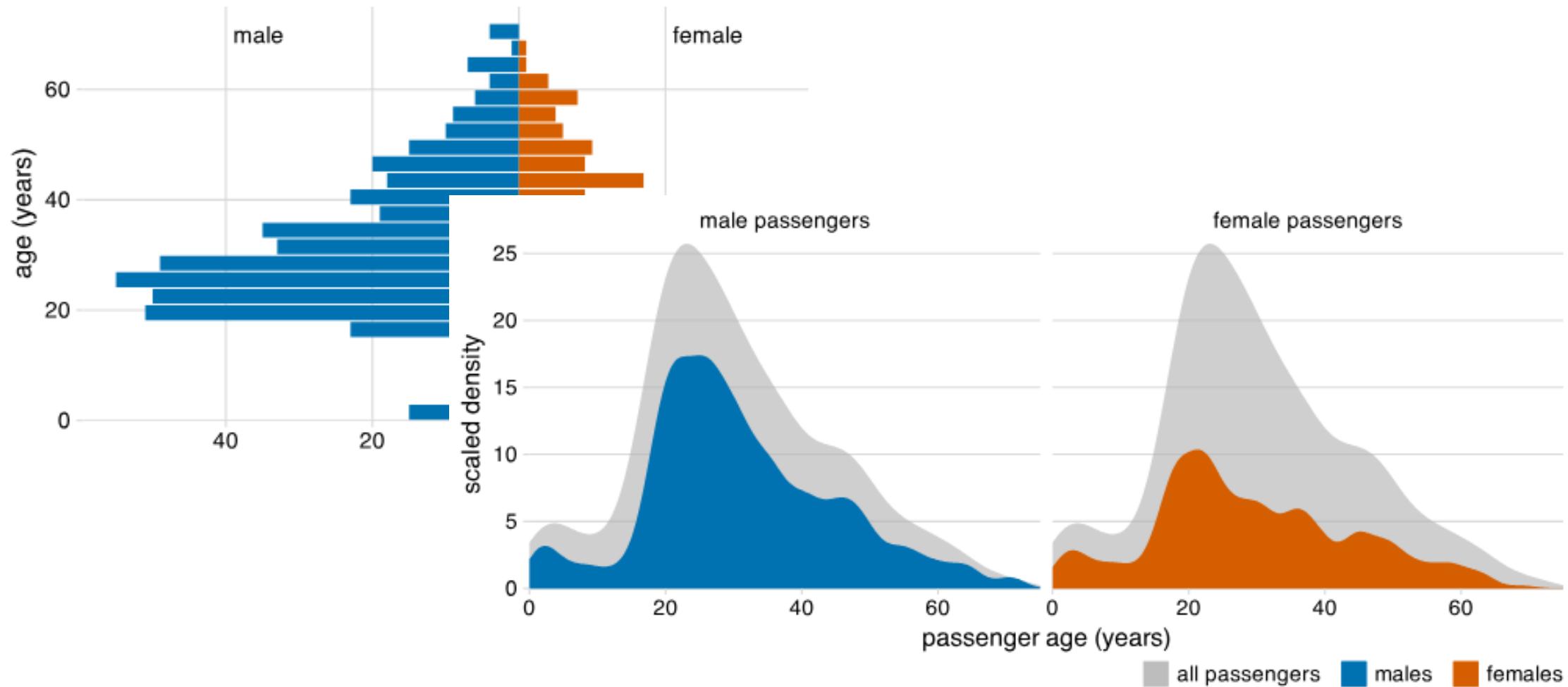
age_c	count
[0,3]	15
(3,6]	7
(6,9]	6
(9,12]	2
(12,15]	6
(15,18]	23
(18,21]	51
(21,24]	50
(24,27]	55
(27,30]	49



# Biểu diễn dữ liệu – phân phối

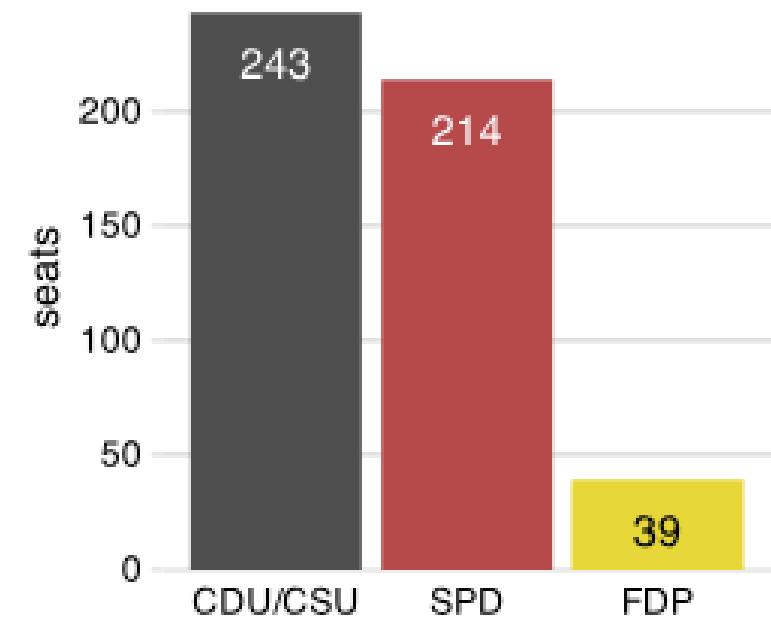
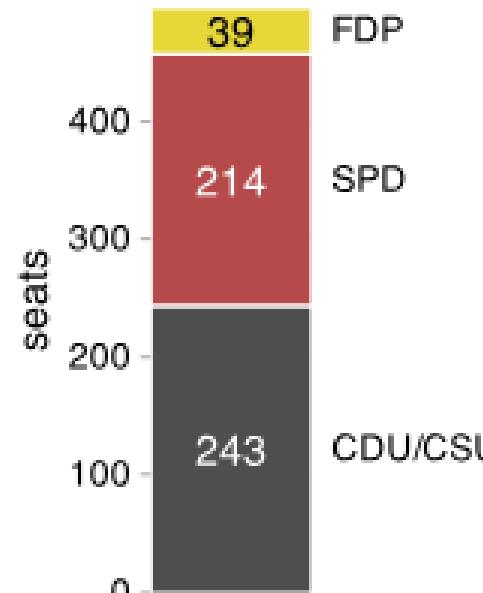
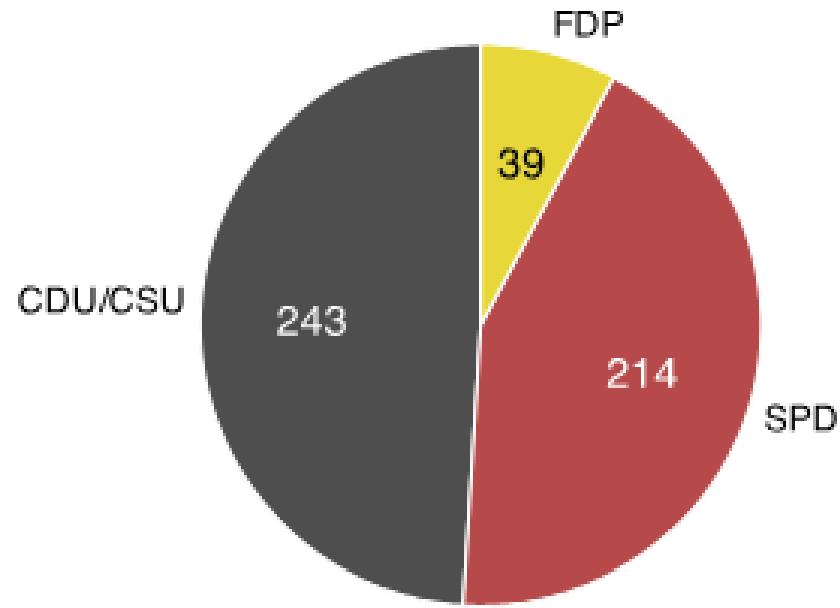


# Biểu diễn dữ liệu – phân phối

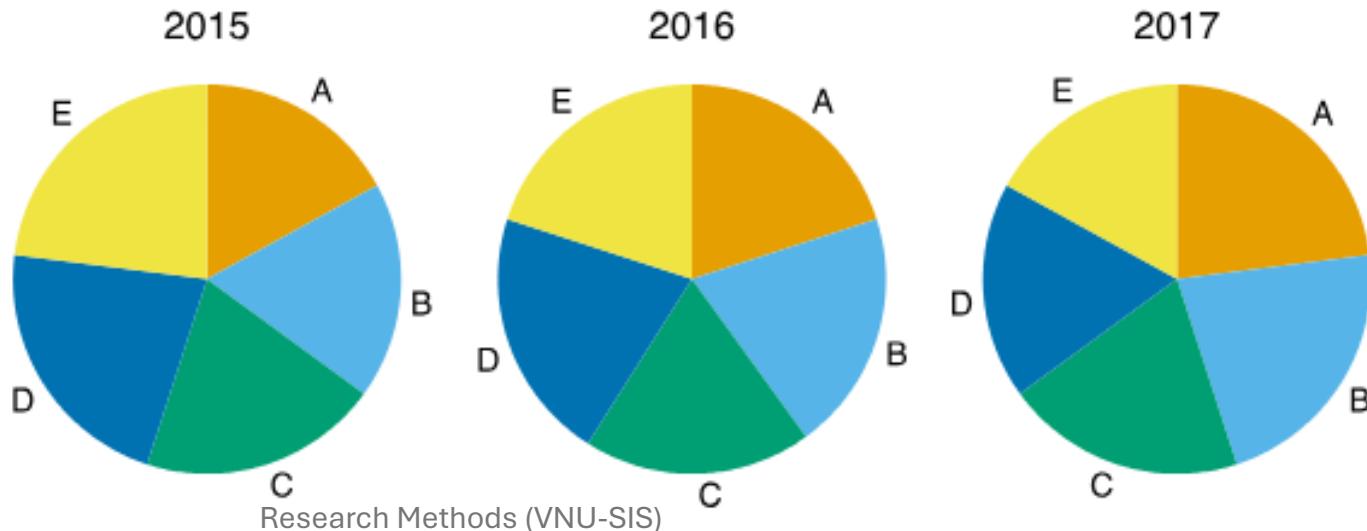
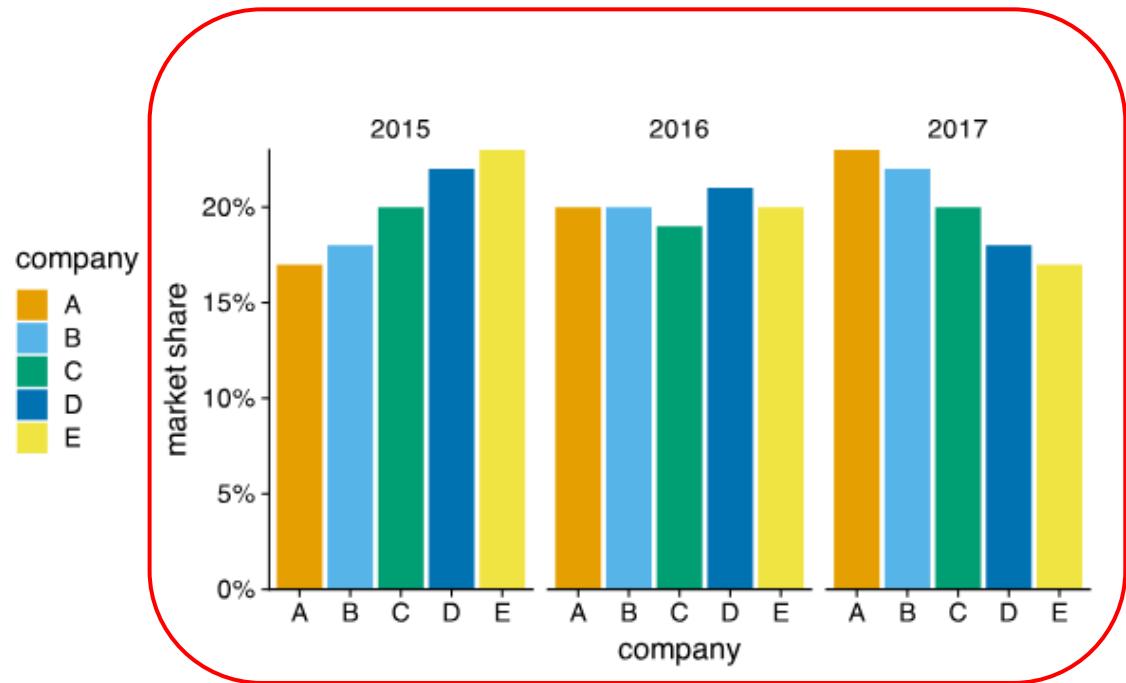
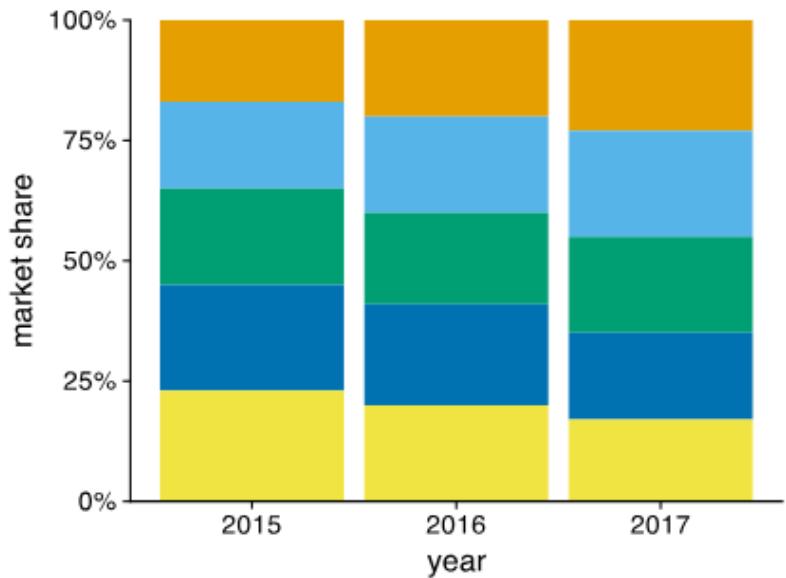


# Biểu diễn dữ liệu – tỷ lệ

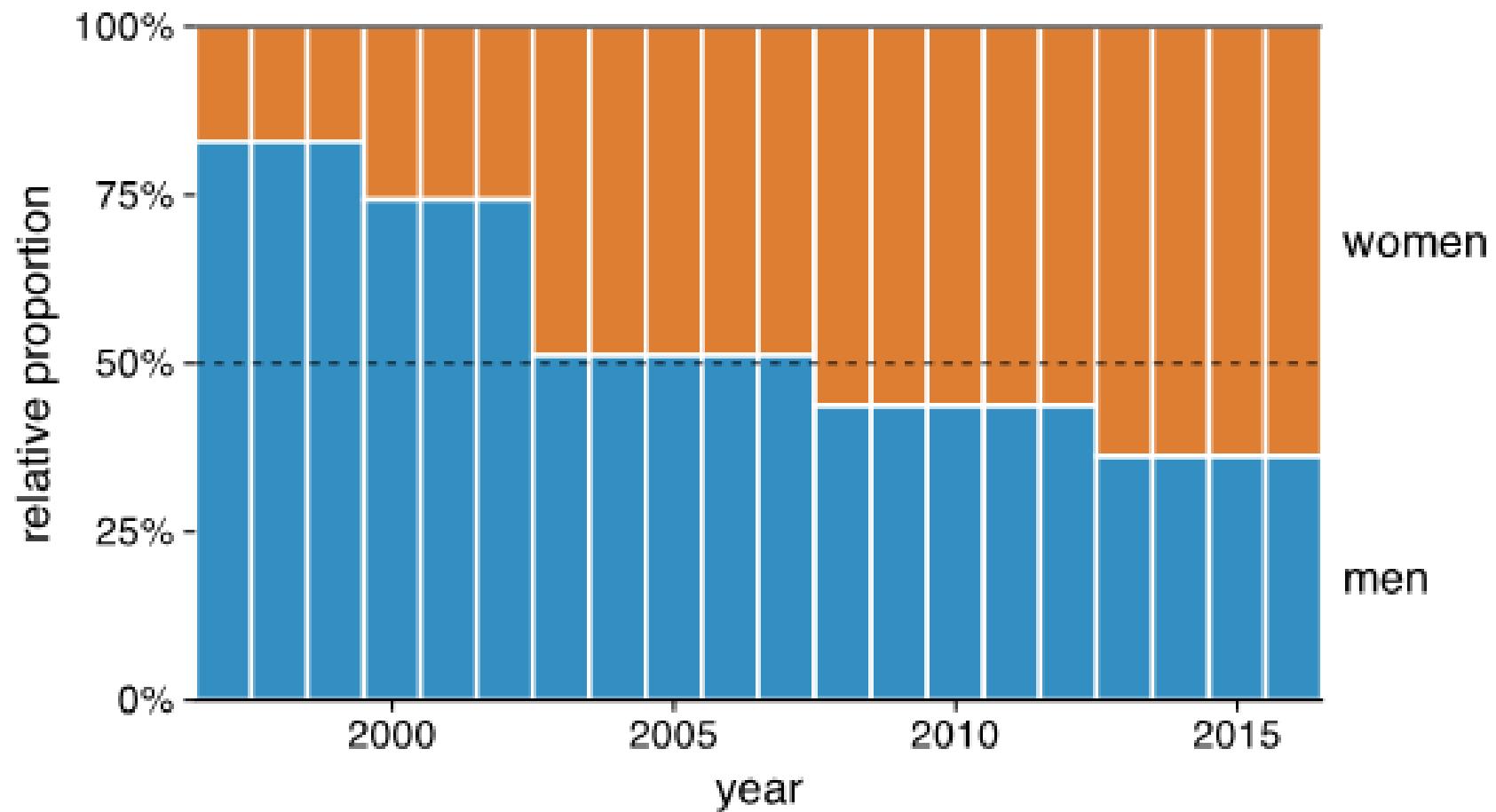
Thành phần đảng phái của Quốc hội Đức khóa 8,  
giai đoạn 1976–1980.



# Biểu diễn dữ liệu – tỷ lệ



# Biểu diễn dữ liệu – tỷ lệ



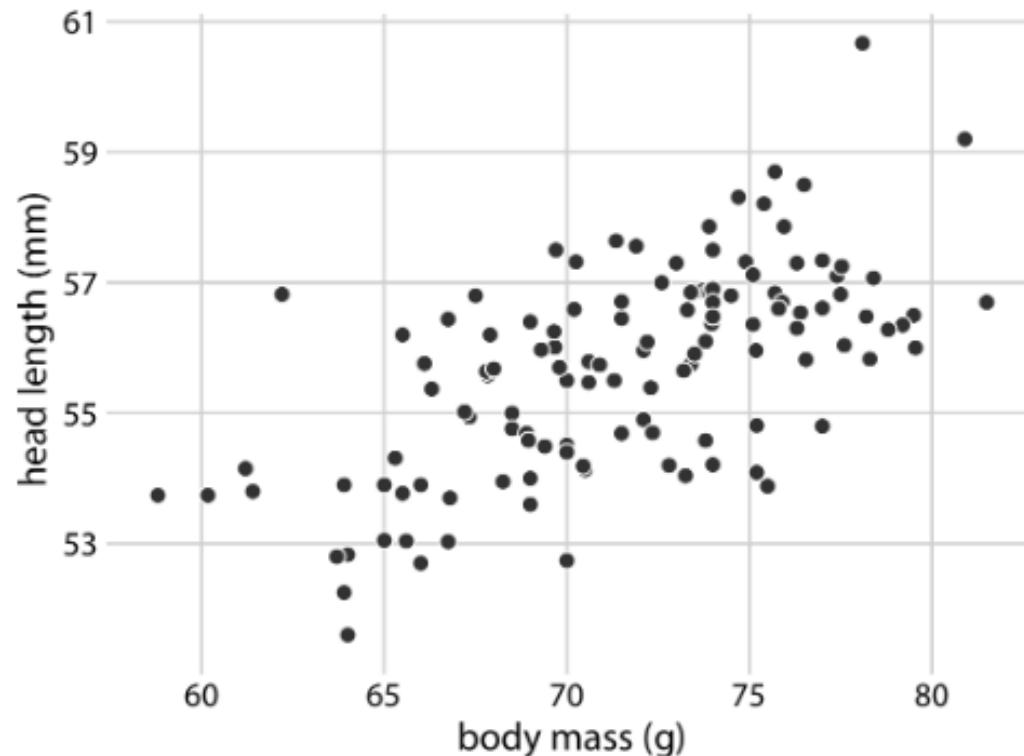
Thay đổi trong thành phần giới tính của quốc hội Rwanda  
từ năm 1997 đến năm 2016

# 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	✗	✓	✗

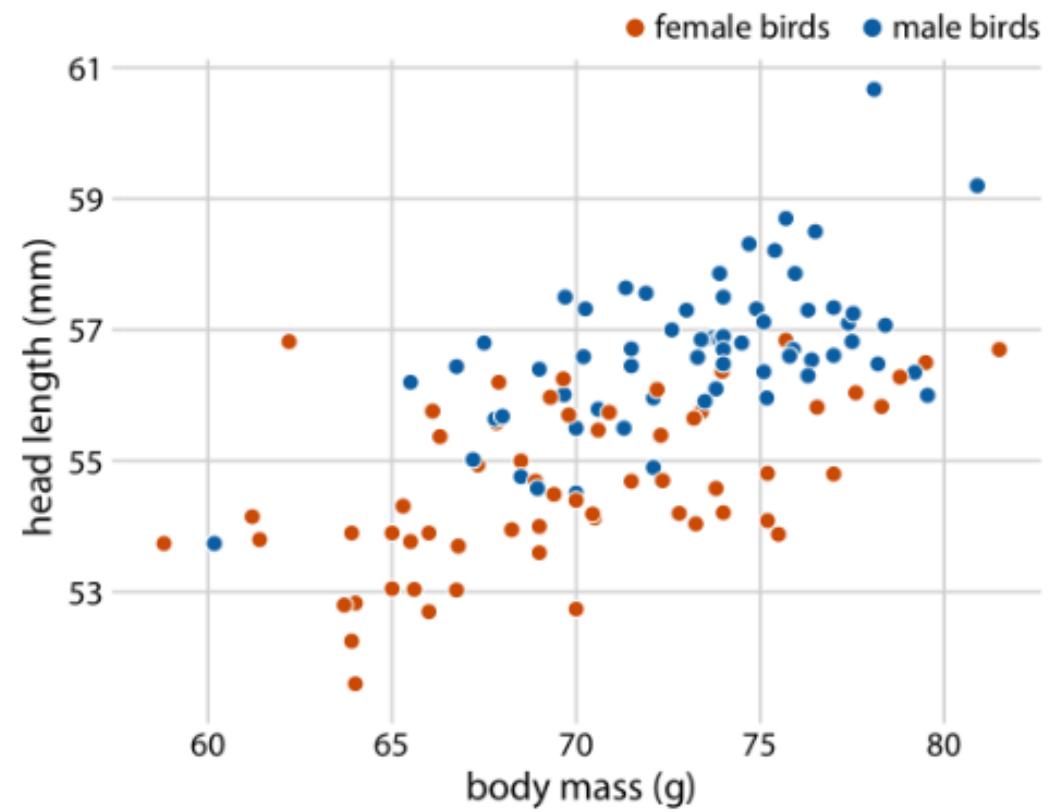
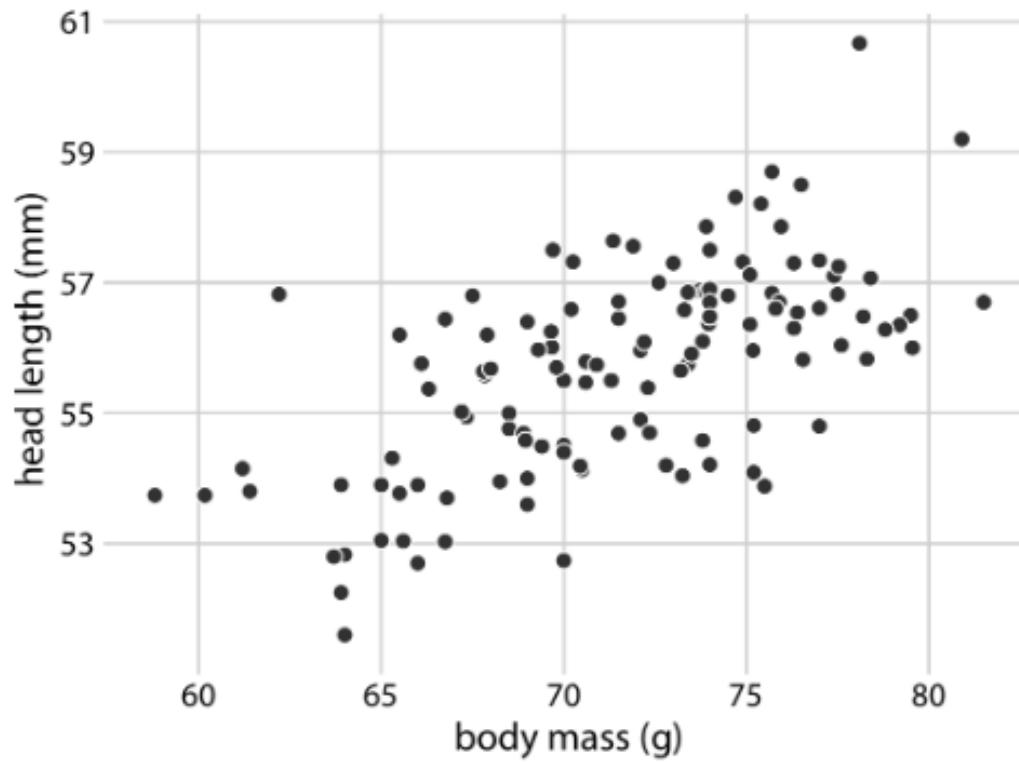
# Biểu diễn dữ liệu – quan hệ

123 chim giẻ cùi lam (blue jay)



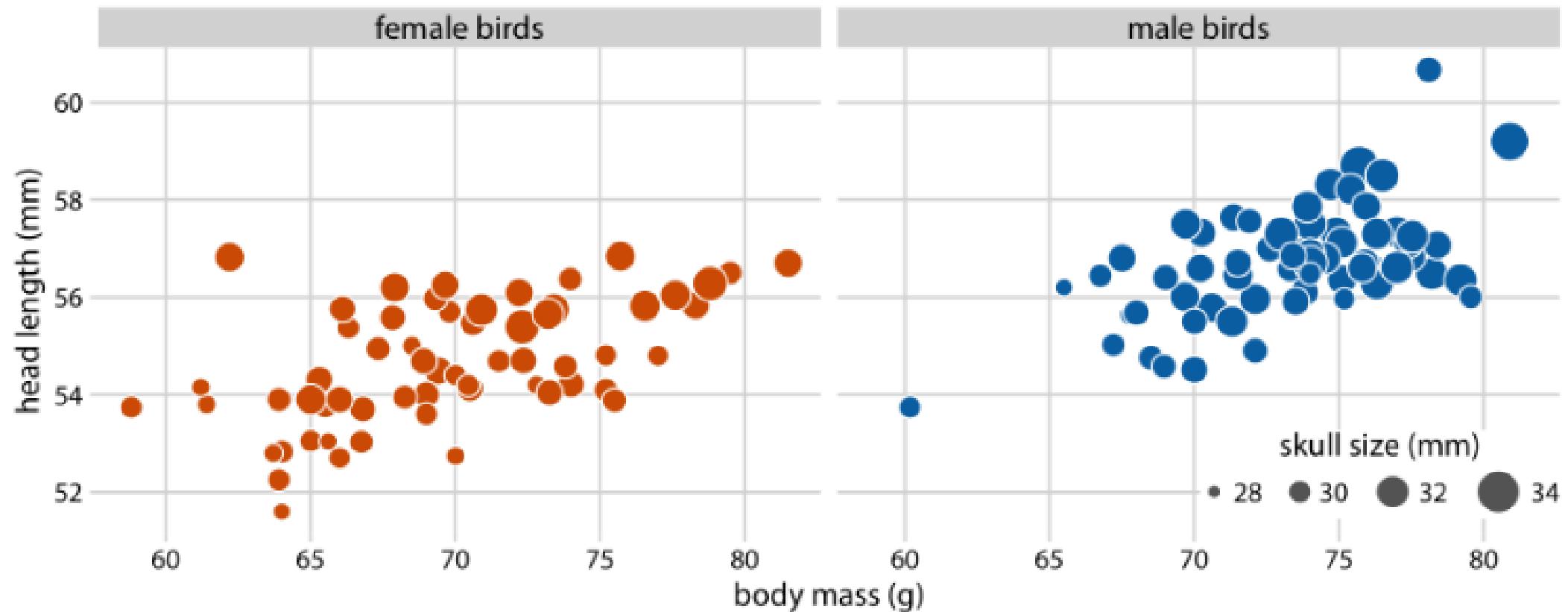
# Biểu diễn dữ liệu – quan hệ

123 chim giẻ cùi lam (blue jay)

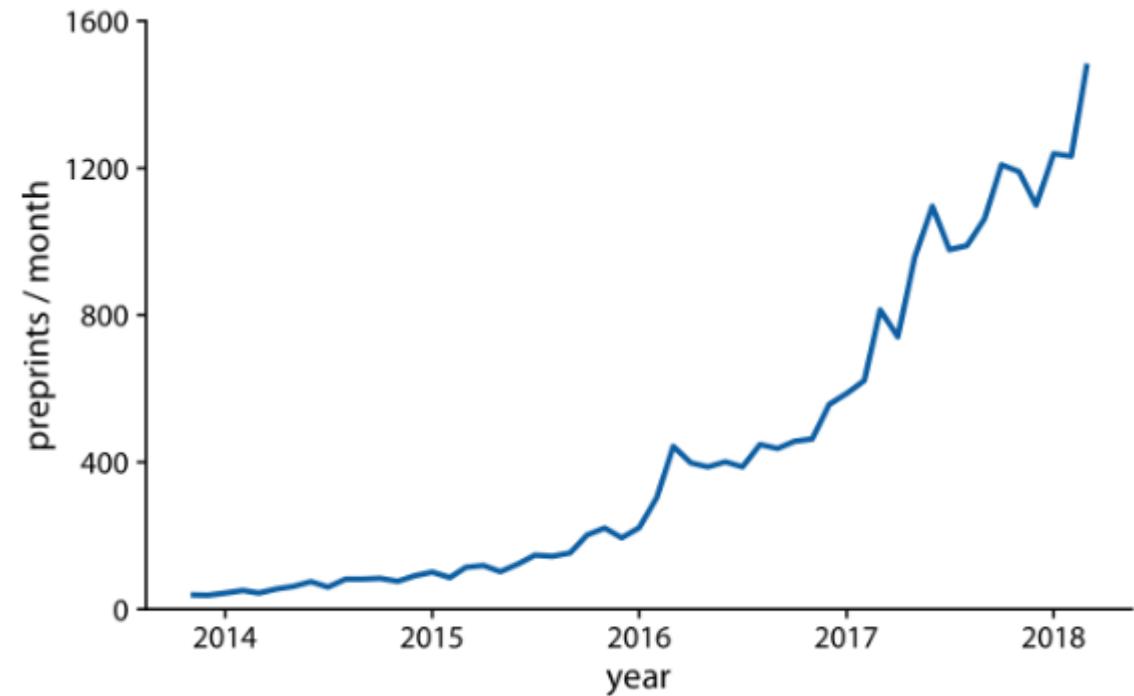
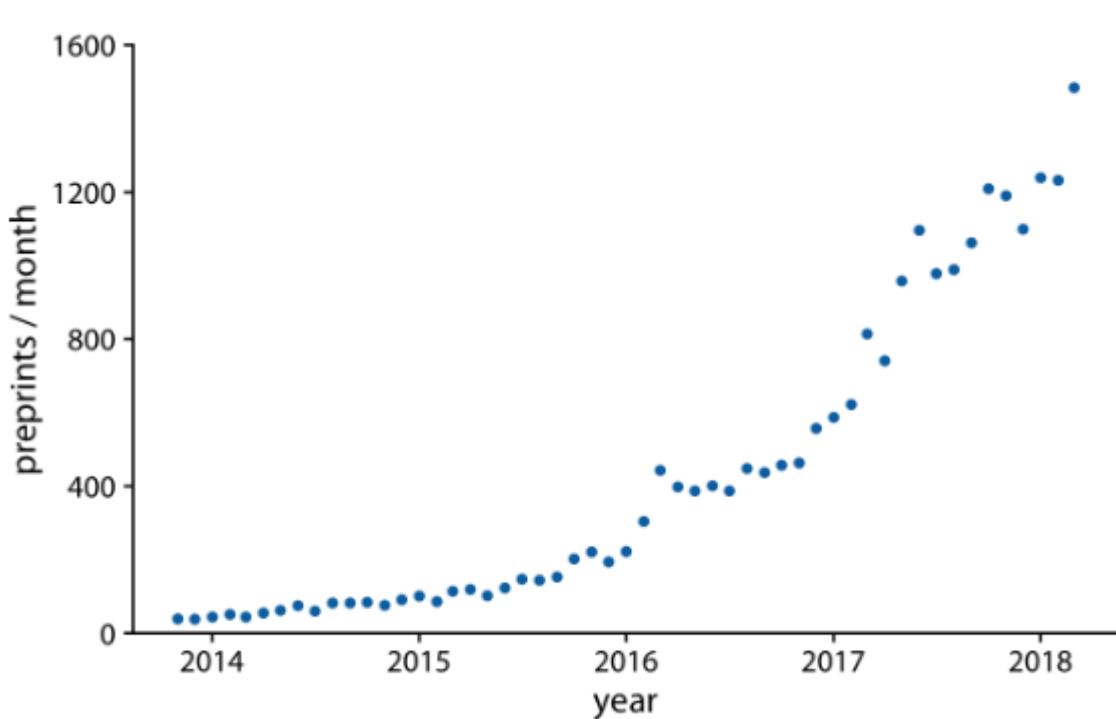


# Biểu diễn dữ liệu – quan hệ

123 chim giẻ cùi lam (blue jay)

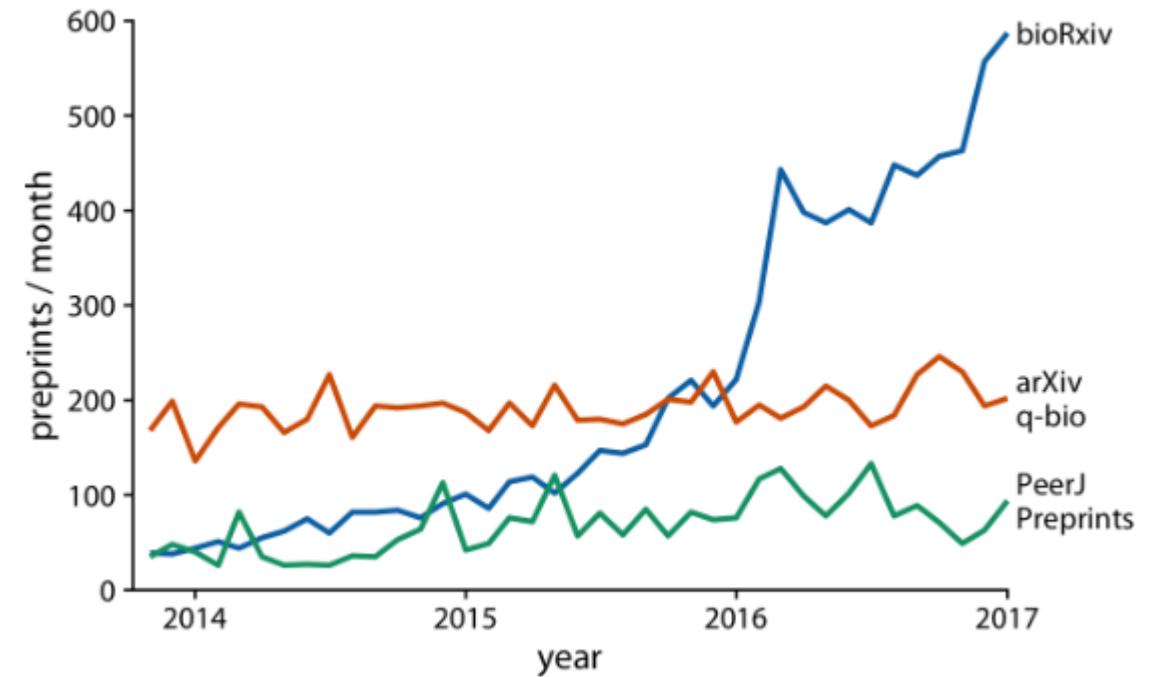
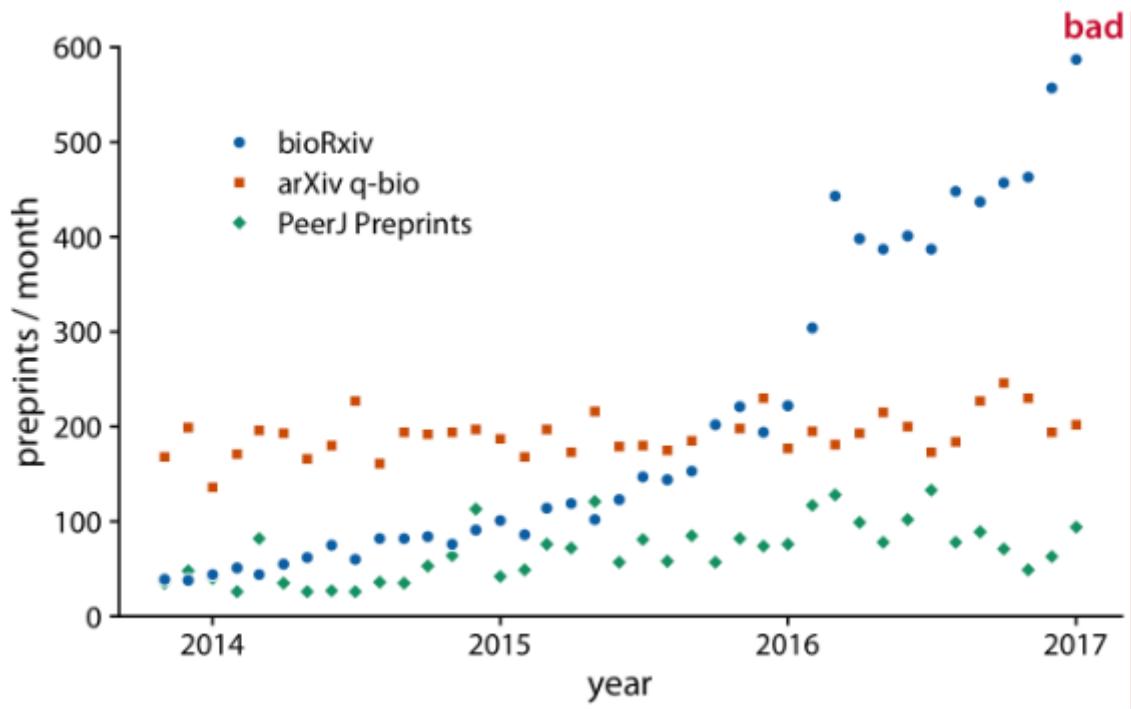


# Biểu diễn dữ liệu – xu hướng



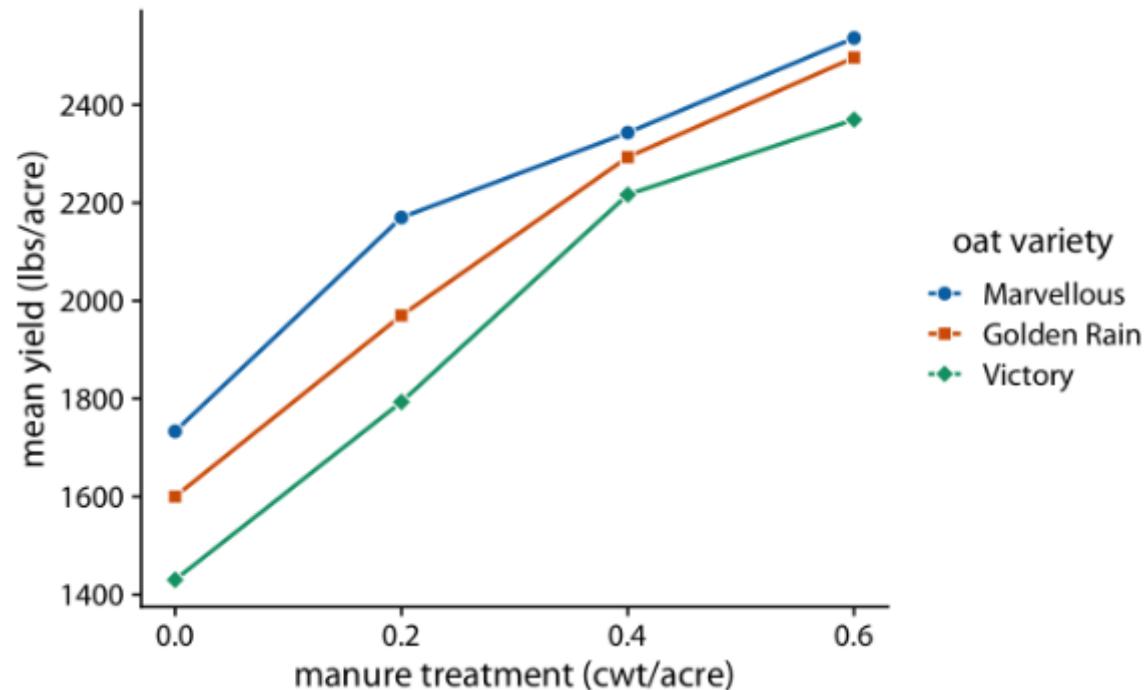
Số lượng bài báo nộp tính theo tháng lên bioRxiv

# Biểu diễn dữ liệu – xu hướng



Số lượng bài báo nộp tính theo tháng lên bioRxiv

# Biểu diễn dữ liệu – xu hướng



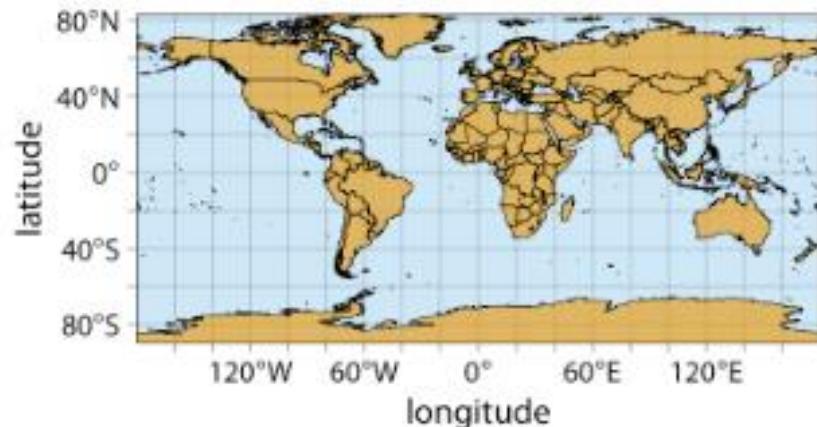
Sản lượng lúa mì trên hàm lượng phân bón sử dụng

# Biểu diễn dữ liệu – không gian

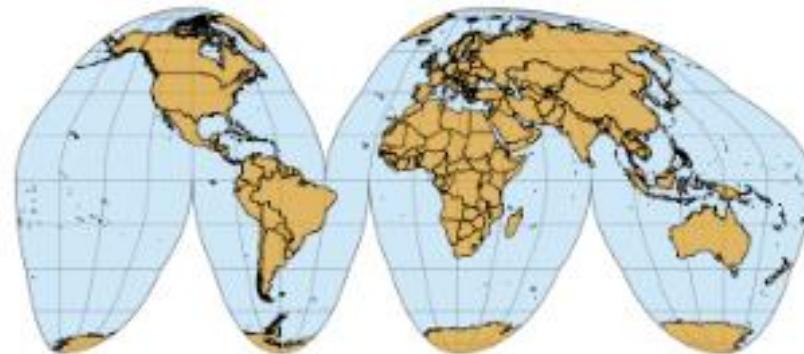


# Biểu diễn dữ liệu – không gian

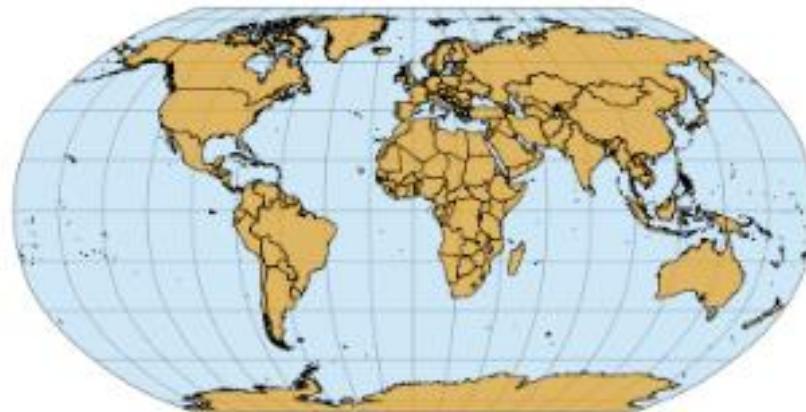
Cartesian longitude and latitude



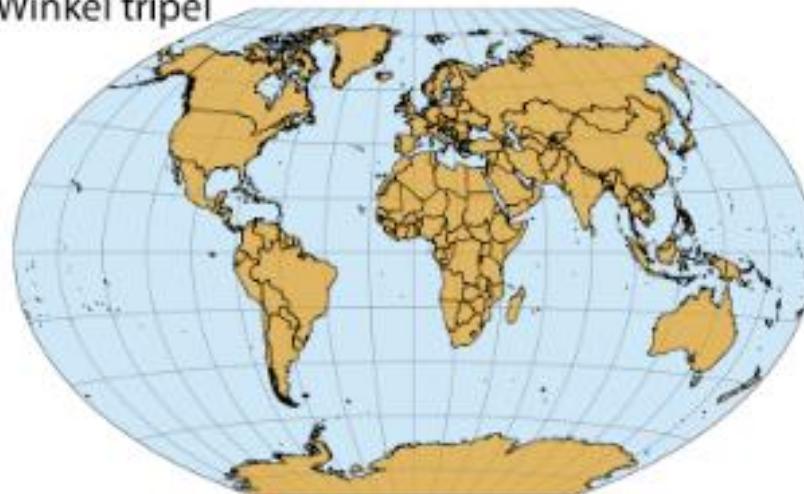
Interrupted Goode homolosine



Robinson



Winkel tripel



# Biểu diễn dữ liệu – không gian

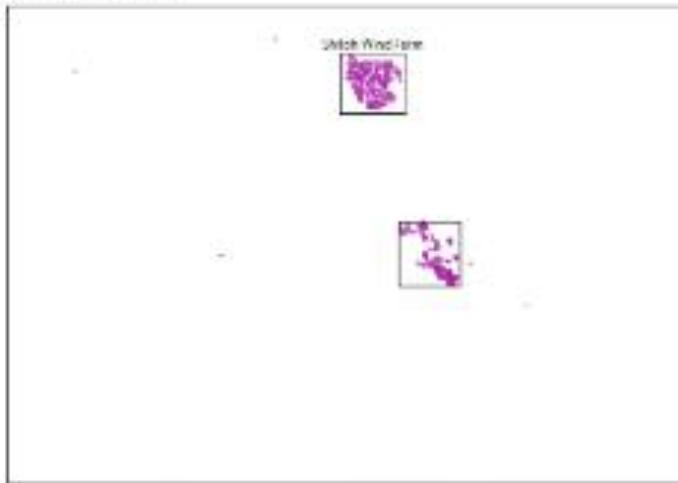
terrain



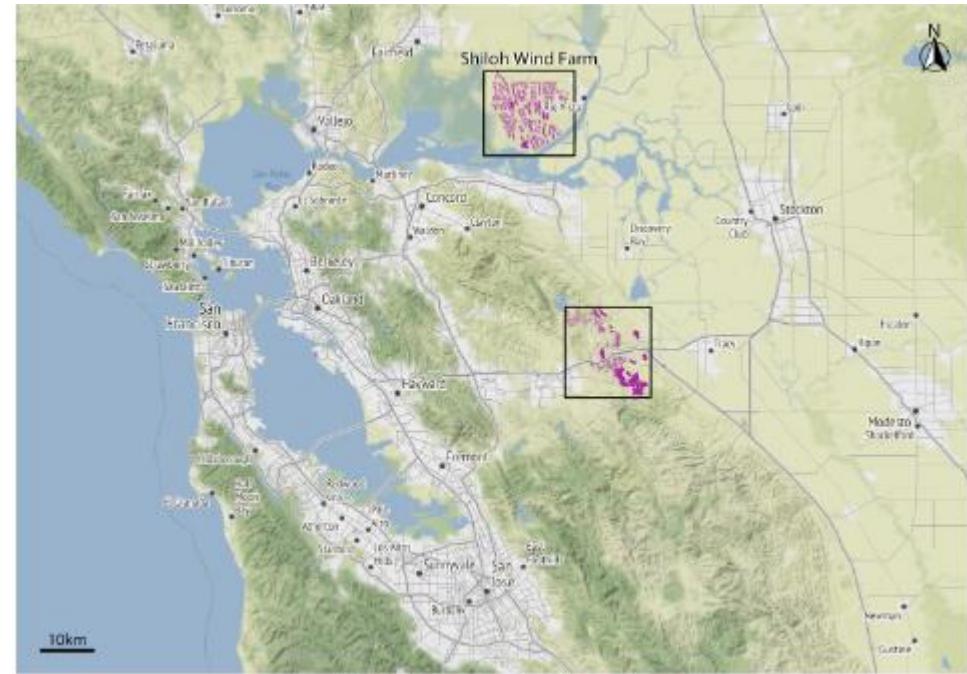
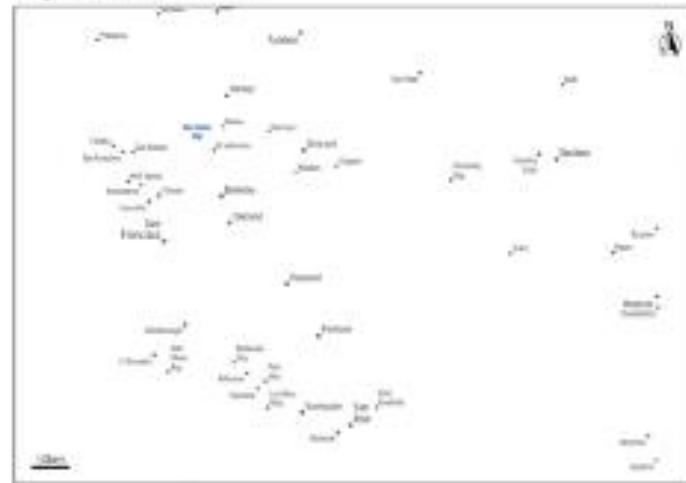
roads



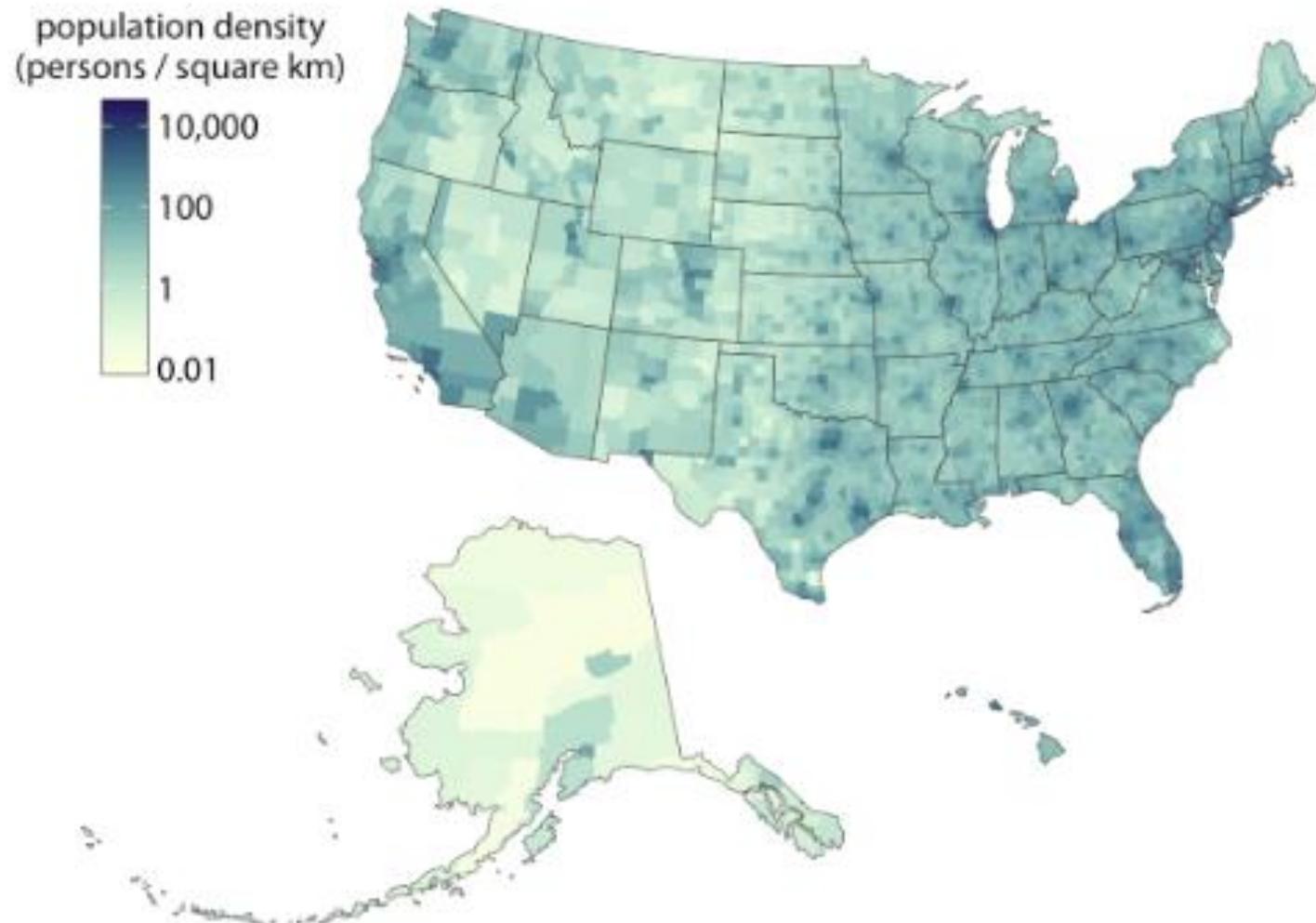
wind turbines



city labels, scale bar

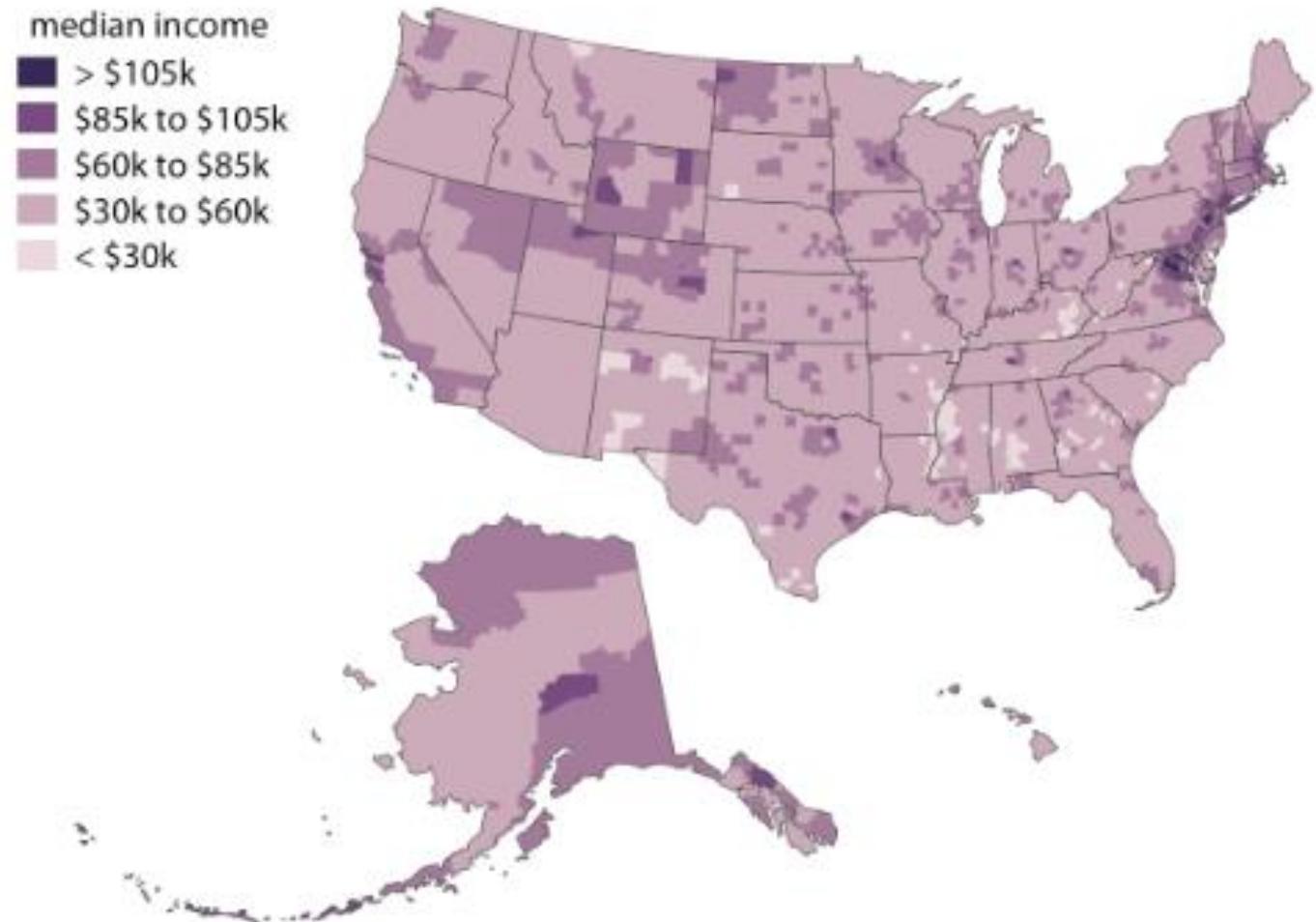


# Biểu diễn dữ liệu – không gian



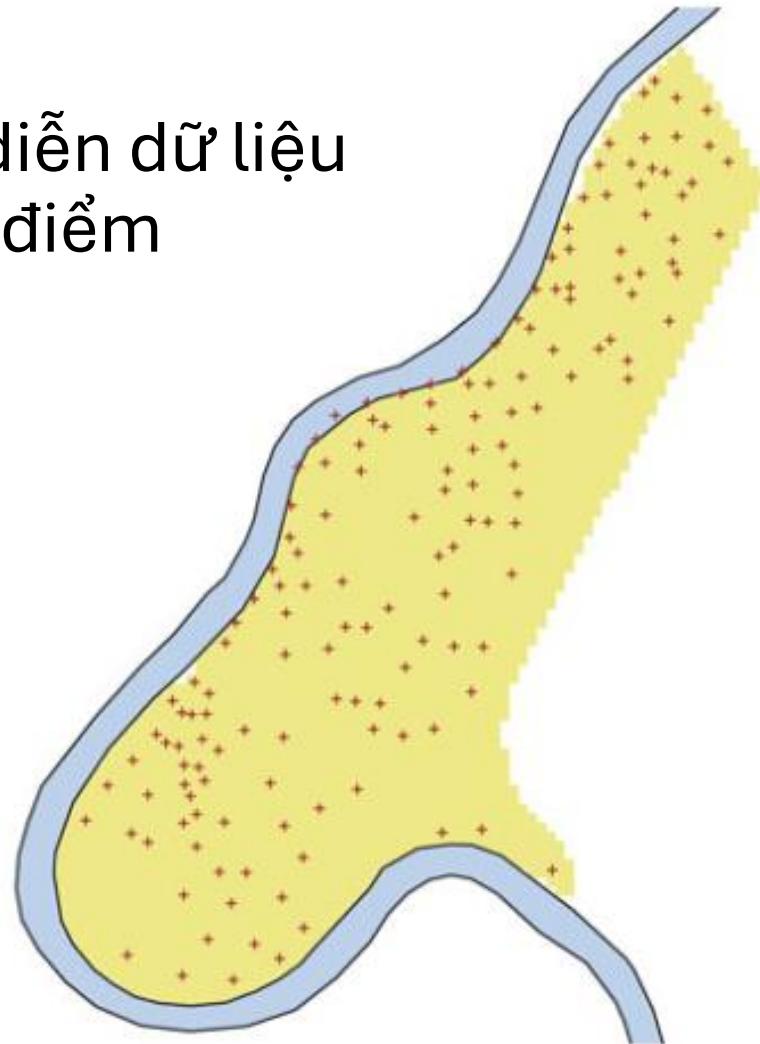
# Biểu diễn dữ liệu – không gian

Bản đồ phân màu



# Biểu diễn dữ liệu – không gian

Biểu diễn dữ liệu  
điểm



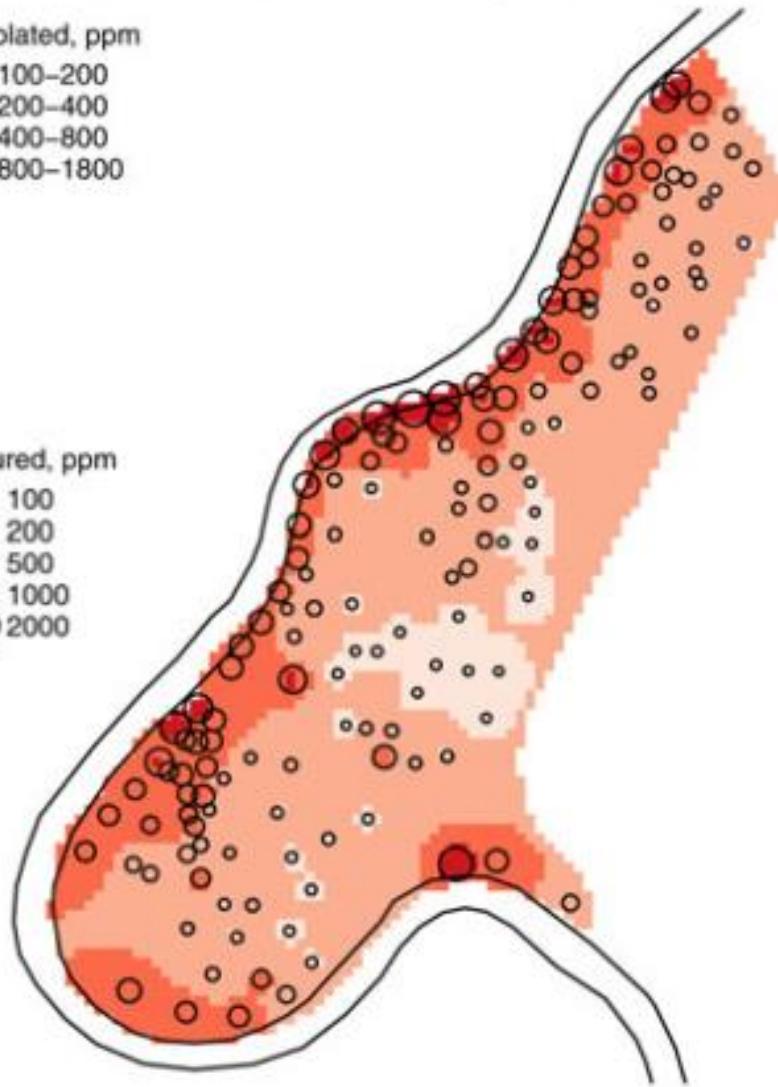
measured and interpolated zinc

interpolated, ppm

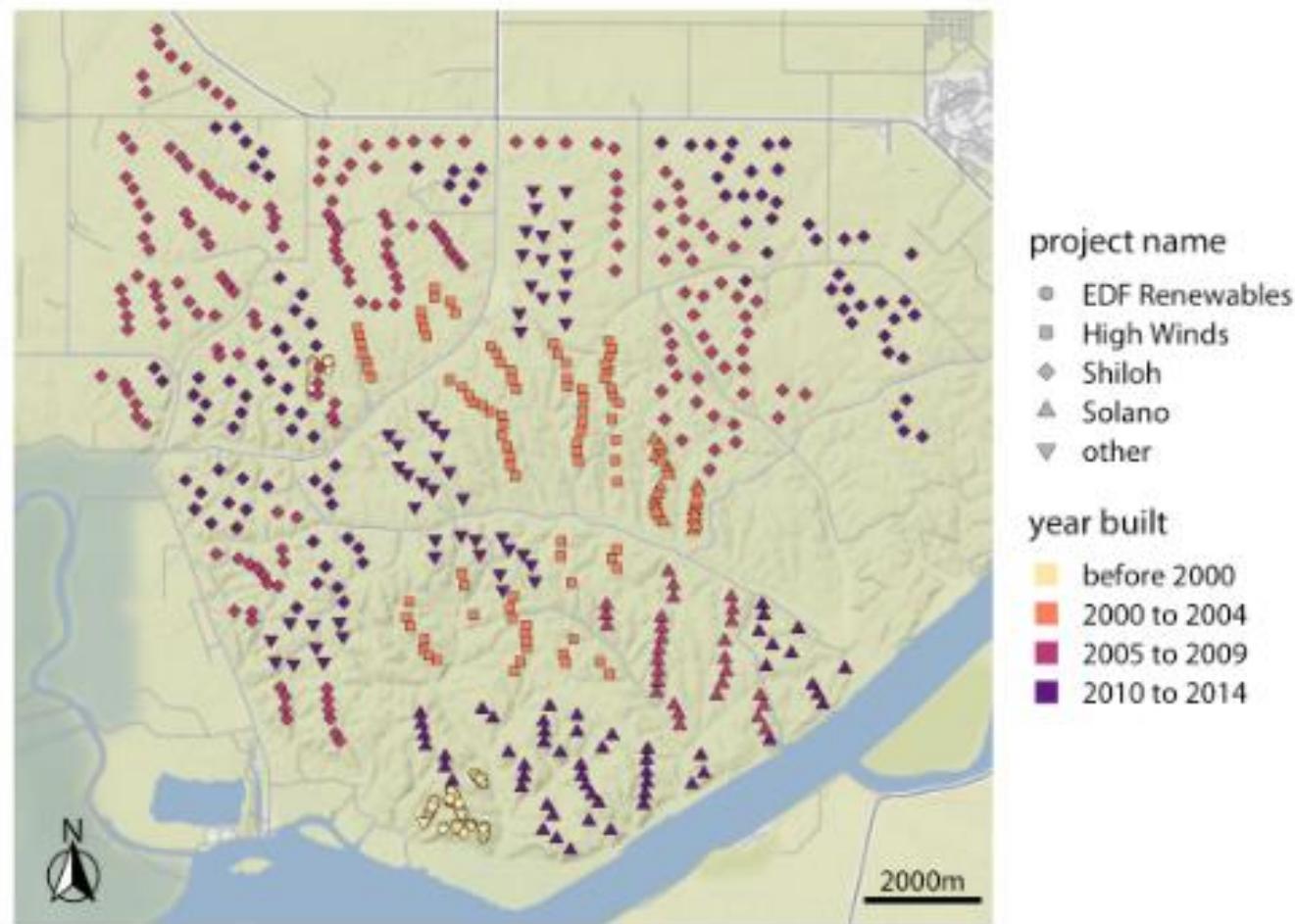
- 100–200
- 200–400
- 400–800
- 800–1800

measured, ppm

- 100
- 200
- 500
- 1000
- 2000



# Biểu diễn dữ liệu – không gian

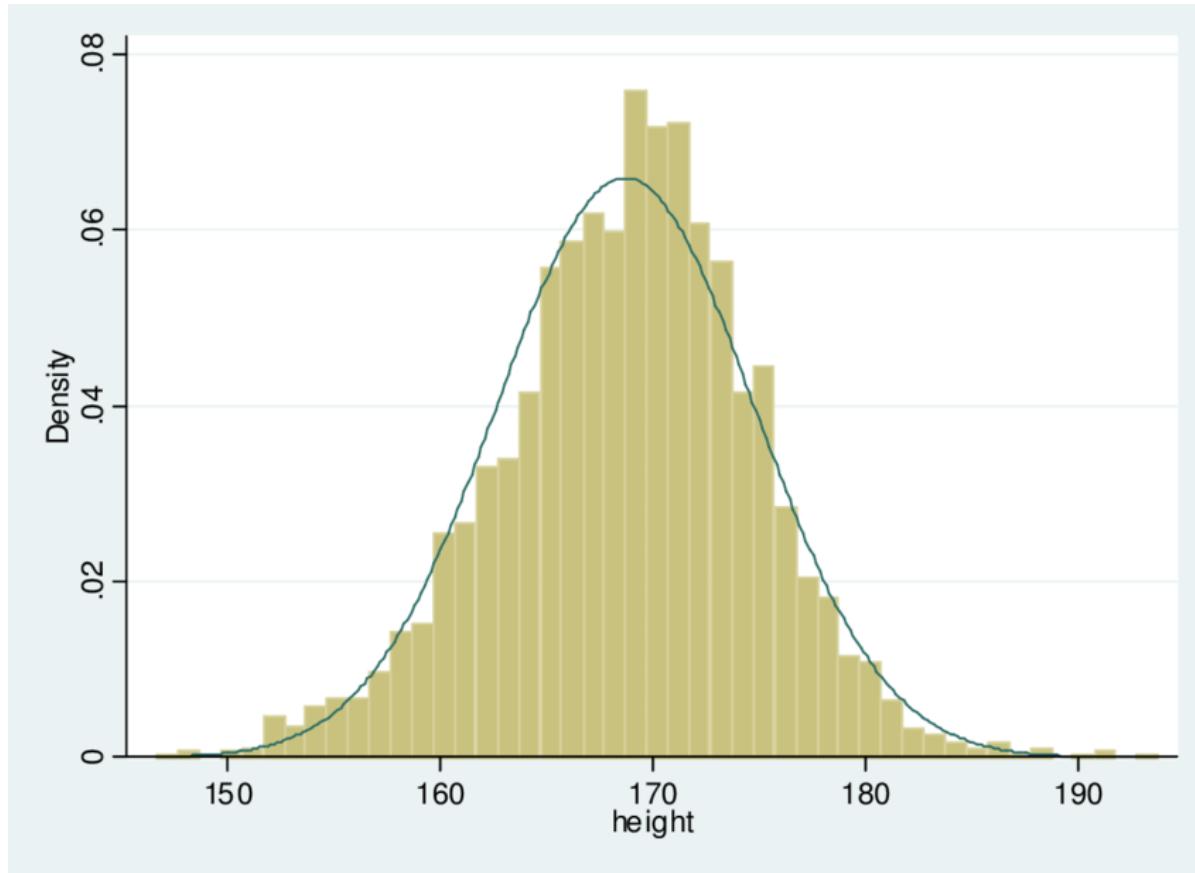


Vị trí tua bin gió ở Trang trại gió Shiloh

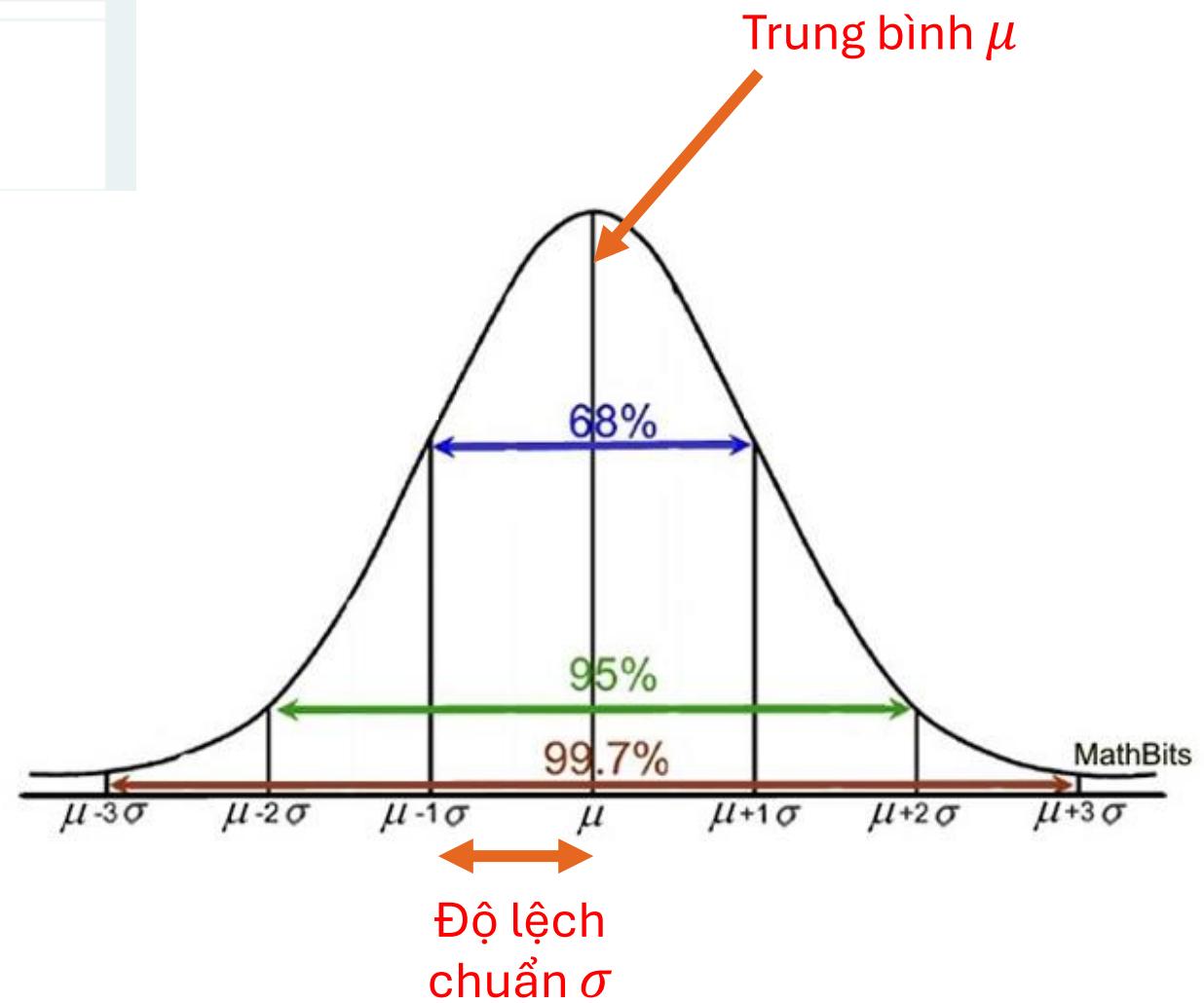
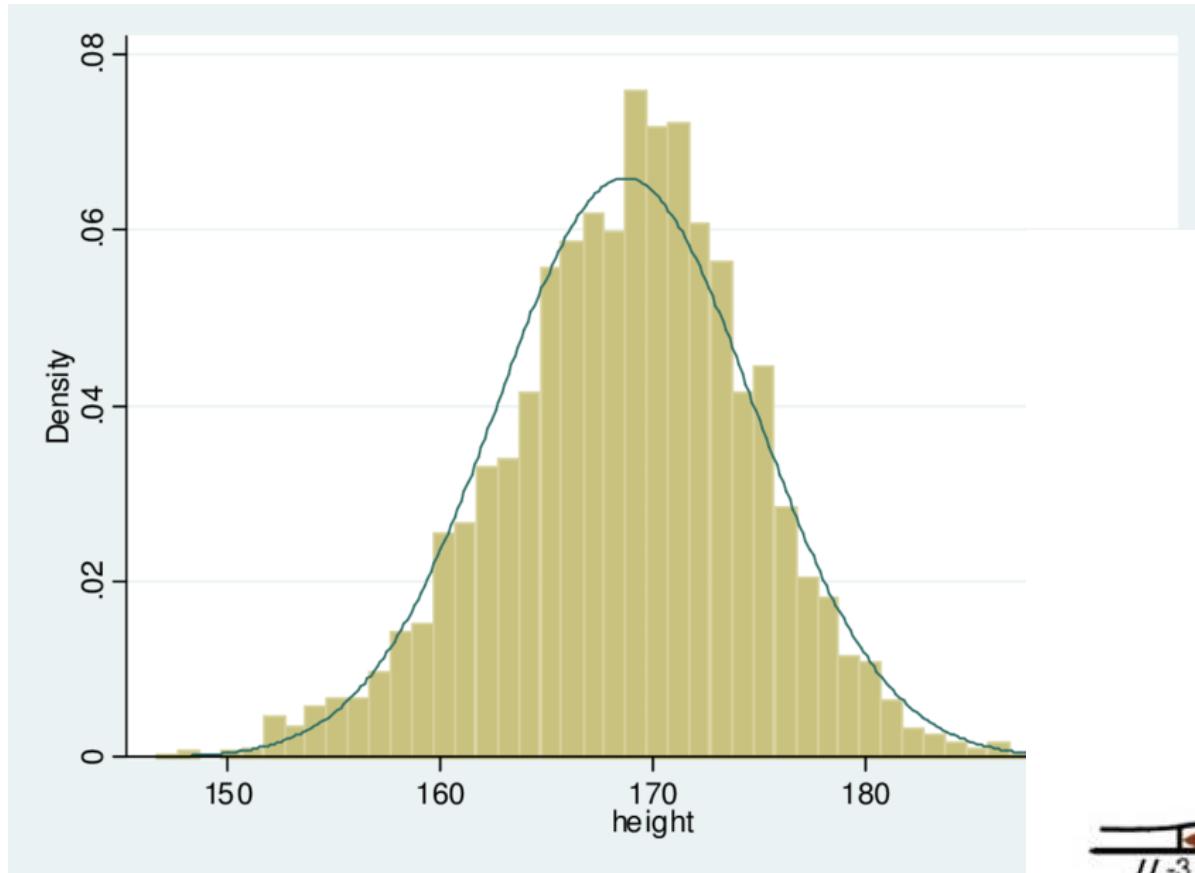
# Các bước phân tích

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

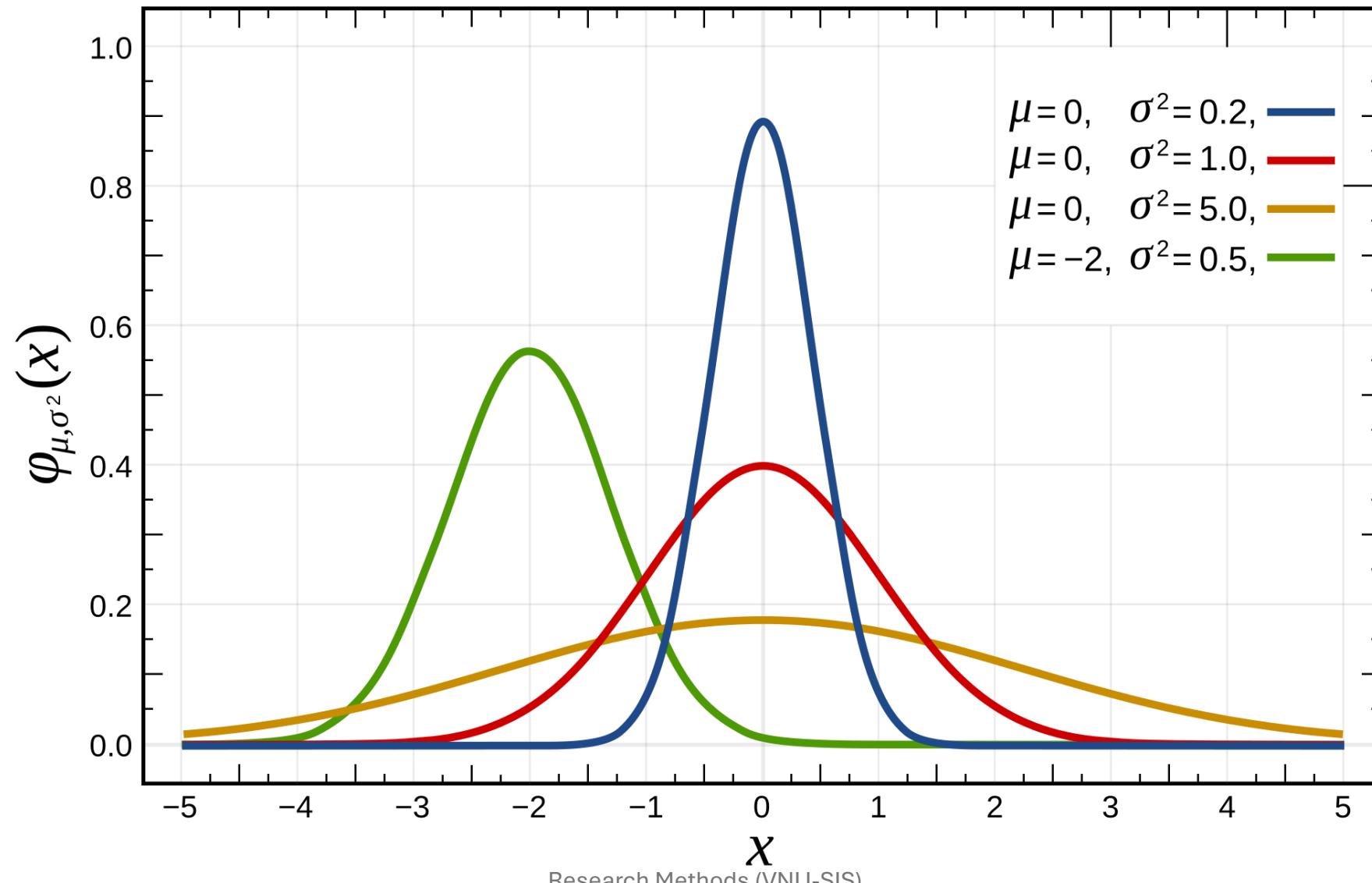
# Thống kê mô tả - Phân phối



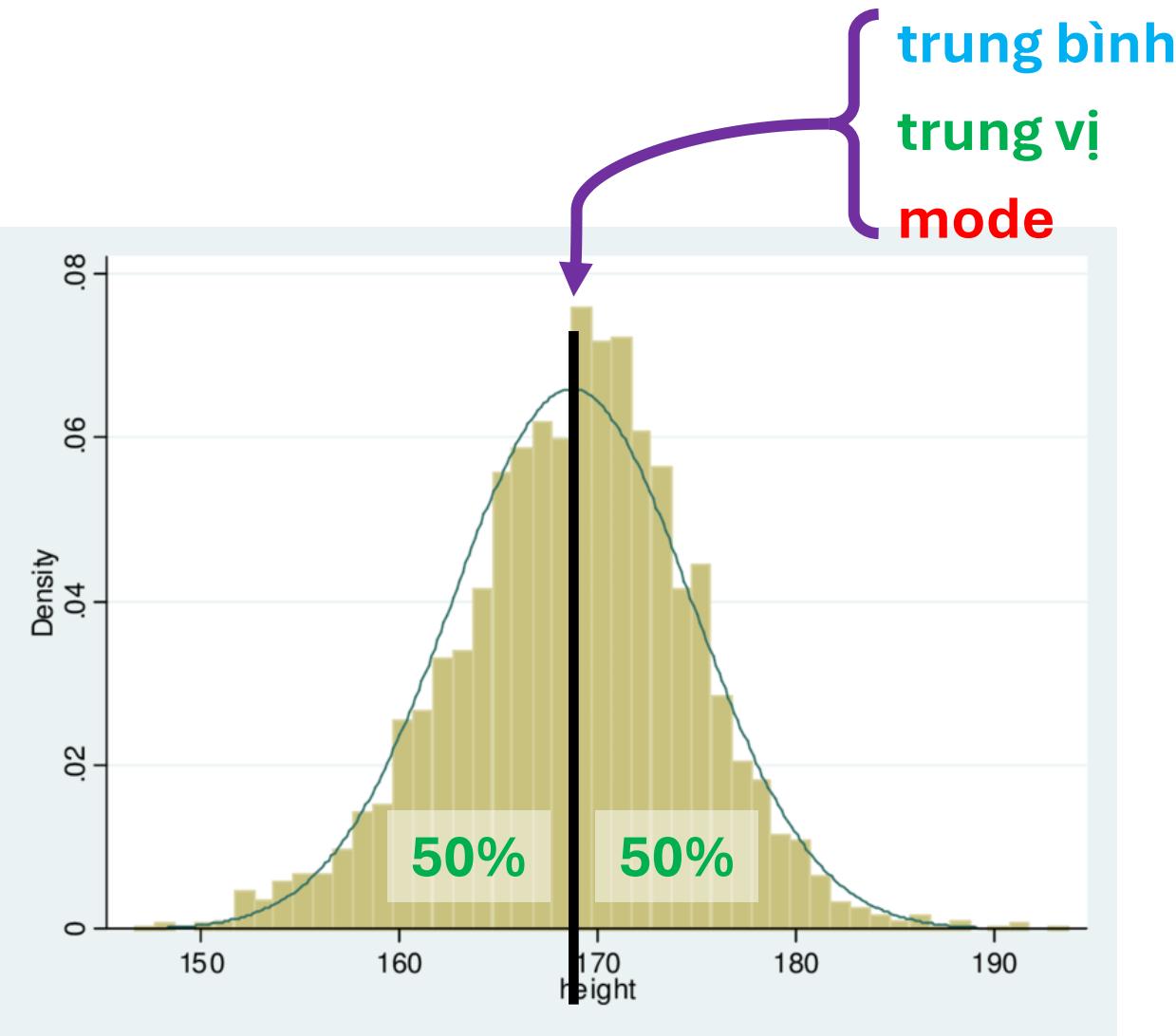
# Thống kê mô tả - Phân phối



# Thống kê mô tả - Phân phối



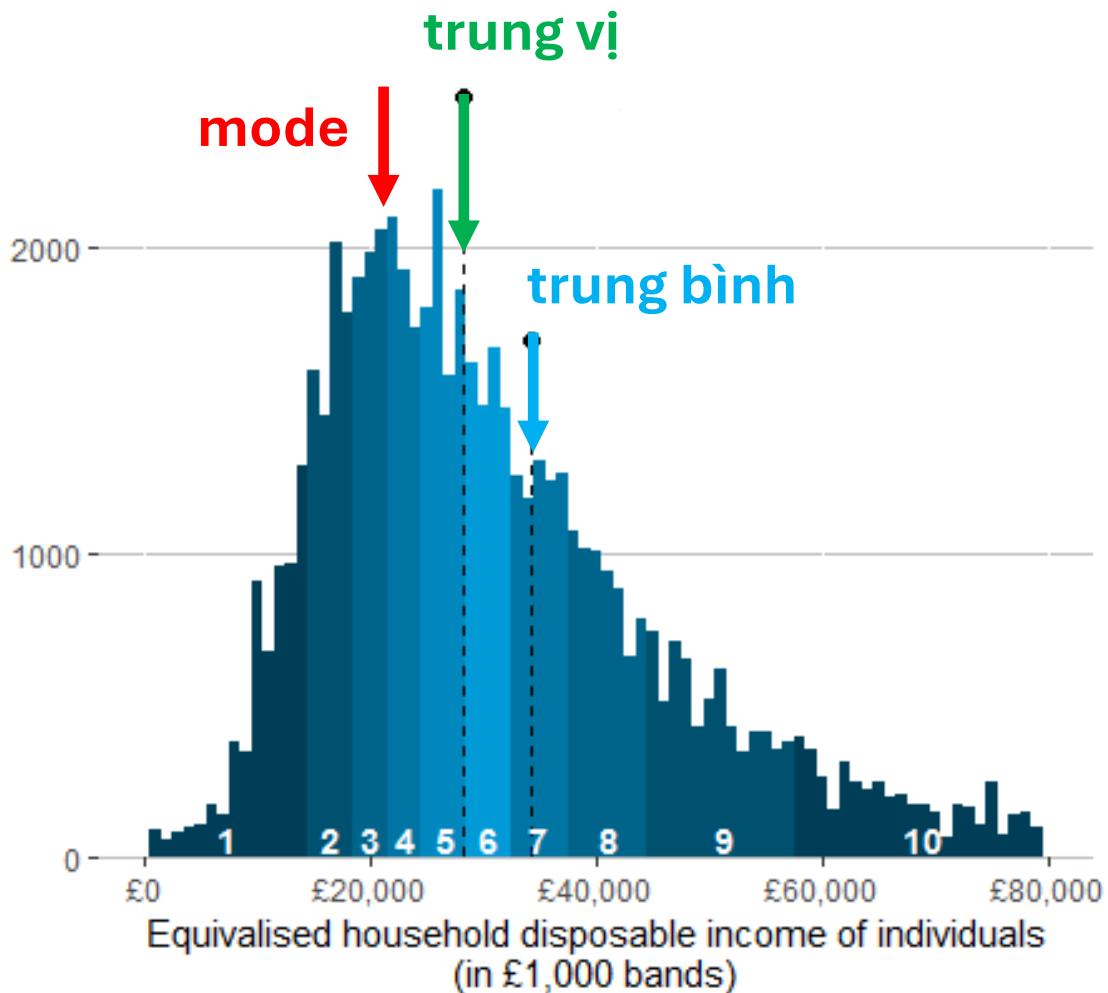
# Thống kê mô tả - Phân phối



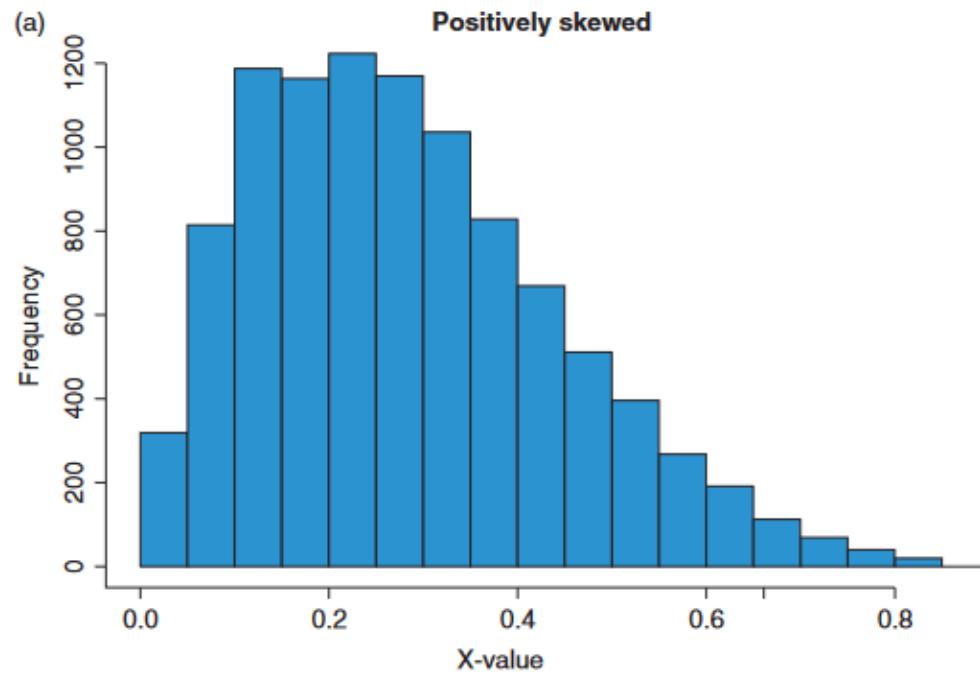
# Thống kê mô tả - Phân phối

Number of individuals (in thousands)

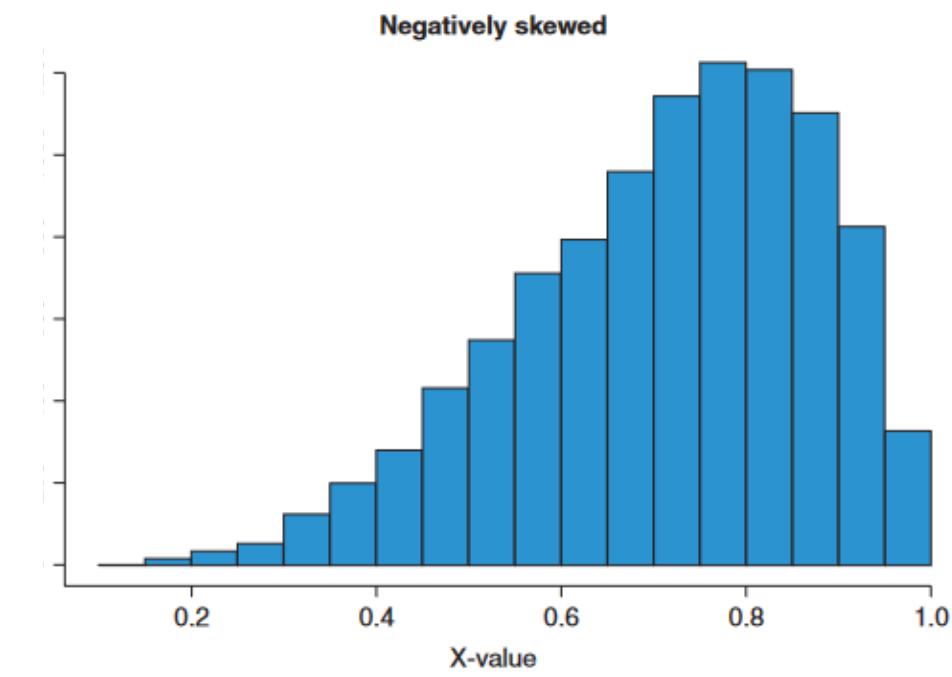
3000



# Thống kê mô tả - Phân phối



Lệch phải/ Lệch dương



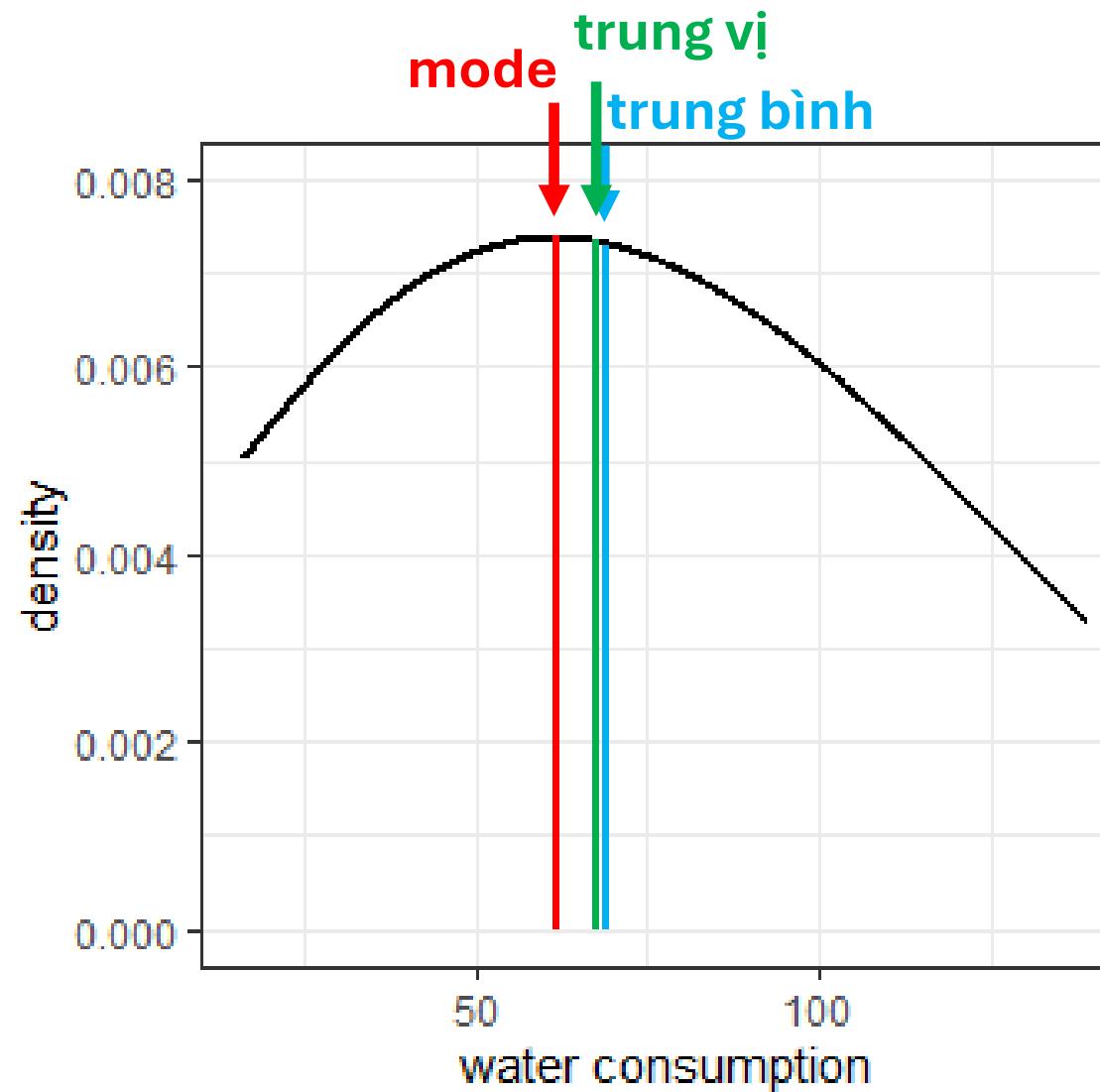
Lệch trái/ Lệch âm

# Thống kê mô tả - Biến liên tục

- csmptv: lượng nước cấp tiêu thụ trong 1 năm
- rwtank: có bể nước mưa
- iceqac2: thu nhập của gia đình

<b>id</b>	<b>csmptv</b>	<b>rwtank</b>	<b>iceqac2</b>
2714	16.23	yes	modest
1476	25	yes	average
2345	29.2	yes	precarious
780	30	no	average
1048	30.93	yes	modest
2375	33.46	no	average
2687	45.63	no	precarious
1405	52.95	yes	modest
431	56.99	no	average
781	66.36	yes	higher
2183	69	yes	modest
1757	71.25	yes	average
730	74.57	no	average
2334	86.06	no	modest
1403	100	no	higher
137	105	no	modest
2752	105.09	no	higher
655	122	yes	modest
2704	126	yes	higher
1432	139	no	modest

# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

trung bình = 69.24

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

<b>id</b>	<b>csmptv</b>	<b>rwtank</b>	<b>iceqac2</b>
137	105	no	modest
431	56.99	no	average
655	122	yes	modest
730	74.57	no	average
780	30	no	average
781	66.36	yes	higher
1048	30.93	yes	modest
1403	100	no	higher
1405	52.95	yes	modest
1432	139	no	modest
1476	25	yes	average
1757	71.25	yes	average
2183	69	yes	modest
2334	86.06	no	modest
2345	29.2	yes	precarious
2375	33.46	no	average
2687	45.63	no	precarious
2704	126	yes	higher
2714	16.23	yes	modest
2752	105.09	no	higher

# Thống kê mô tả - Biến liên tục

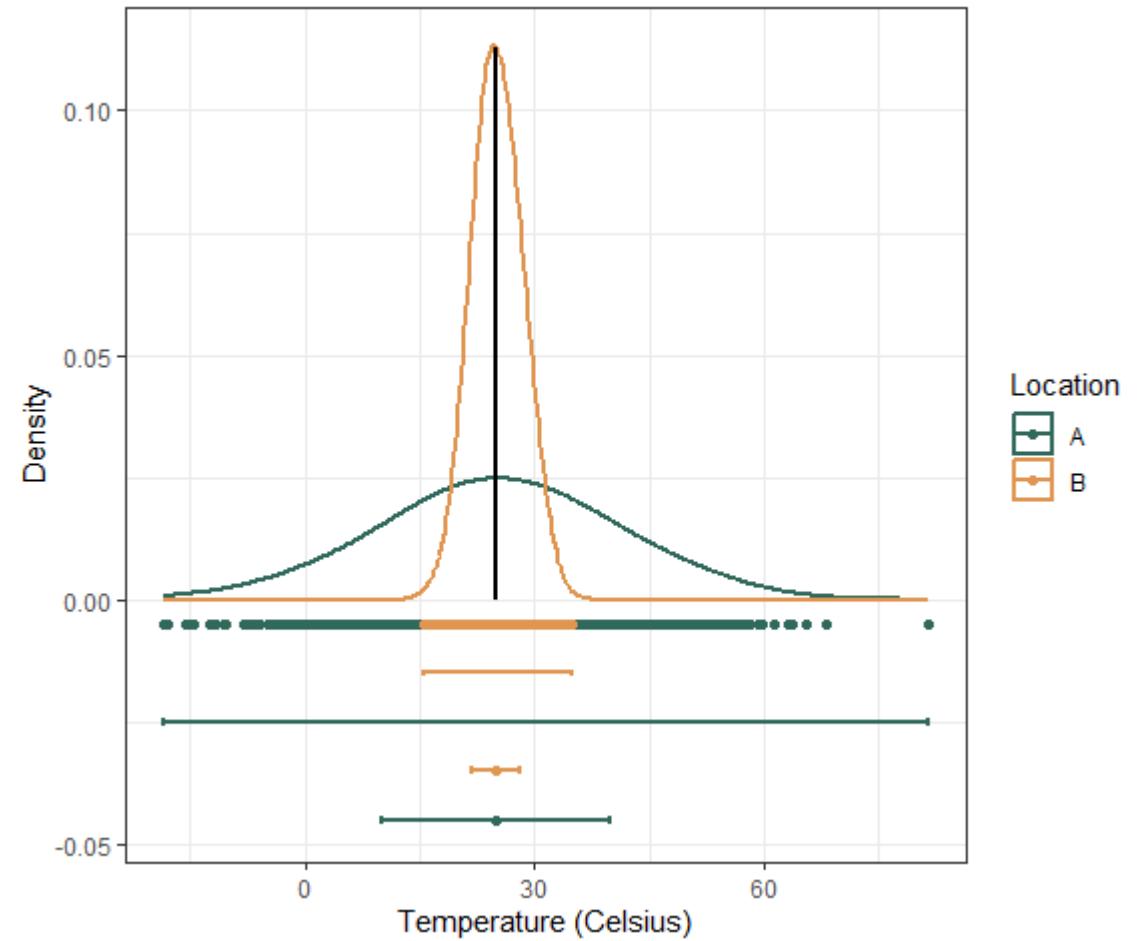
<b>id</b>	<b>csmpty</b>	<b>rwtank</b>	<b>iceqac2</b>
2714	16.23	yes	modest
1476	25	yes	average
2345	29.2	yes	precarious
780	30	no	average
1048	30.93	yes	modest
2375	33.46	no	average
2687	45.63	no	precarious
1405	52.95	yes	modest
431	56.99	no	average
781	66.36	yes	higher
2183	69	yes	modest
1757	71.25	yes	average
730	74.57	no	average
2334	86.06	no	modest
1403	100	no	higher
137	105	no	modest
2752	105.09	no	higher
655	122	yes	modest
2704	126	yes	higher
1432	139	no	modest

# Thống kê mô tả - Biến liên tục

trung vị = 67.68

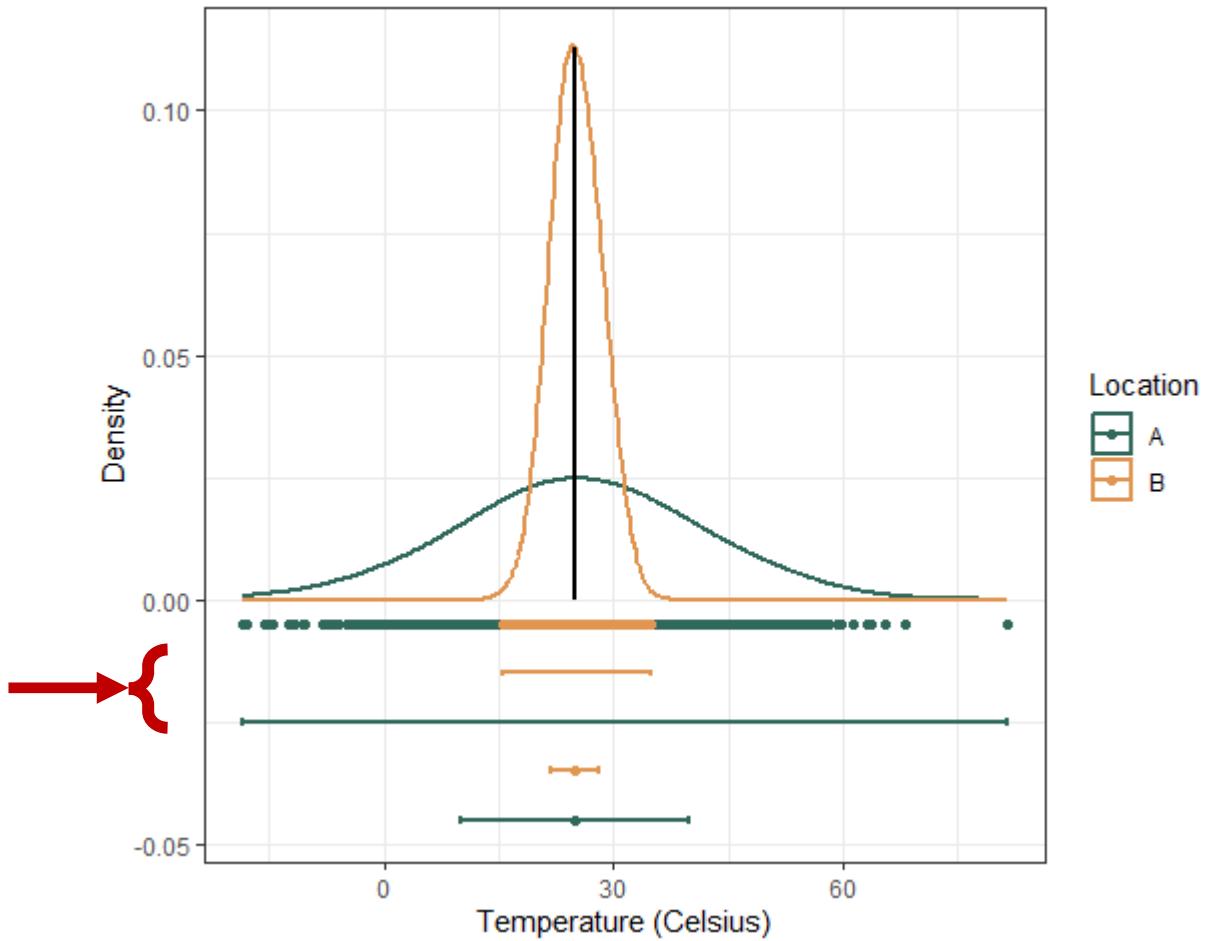
<b>id</b>	<b>csmpty</b>	<b>rwtank</b>	<b>iceqac2</b>
2714	16.23	yes	modest
1476	25	yes	average
2345	29.2	yes	precarious
780	30	no	average
1048	30.93	yes	modest
2375	33.46	no	average
2687	45.63	no	precarious
1405	52.95	yes	modest
431	56.99	no	average
781	66.36	yes	higher
2183	69	yes	modest
1757	71.25	yes	average
730	74.57	no	average
2334	86.06	no	modest
1403	100	no	higher
137	105	no	modest
2752	105.09	no	higher
655	122	yes	modest
2704	126	yes	higher
1432	139	no	modest

# Thống kê mô tả - Biến liên tục

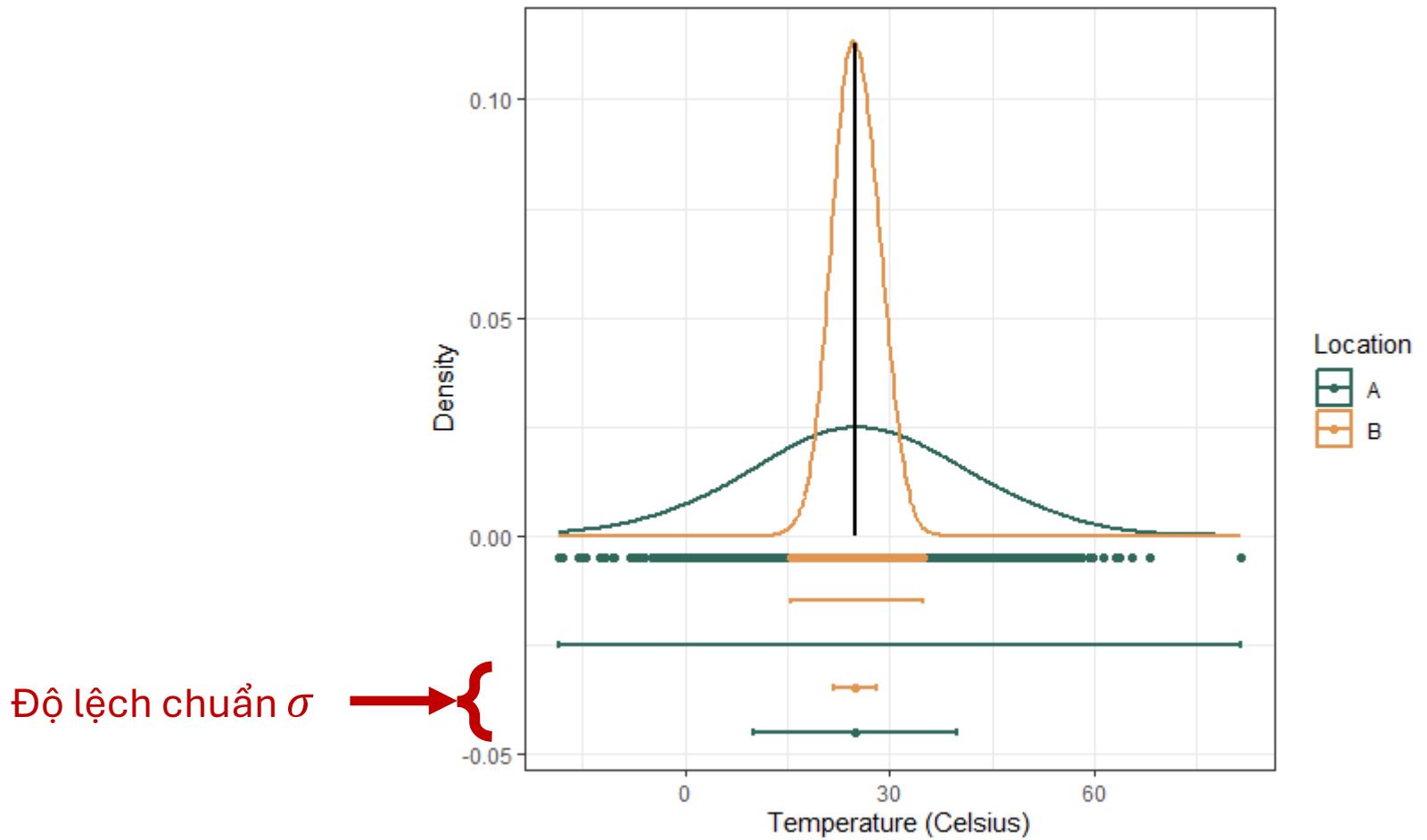


# Thống kê mô tả - Biến liên tục

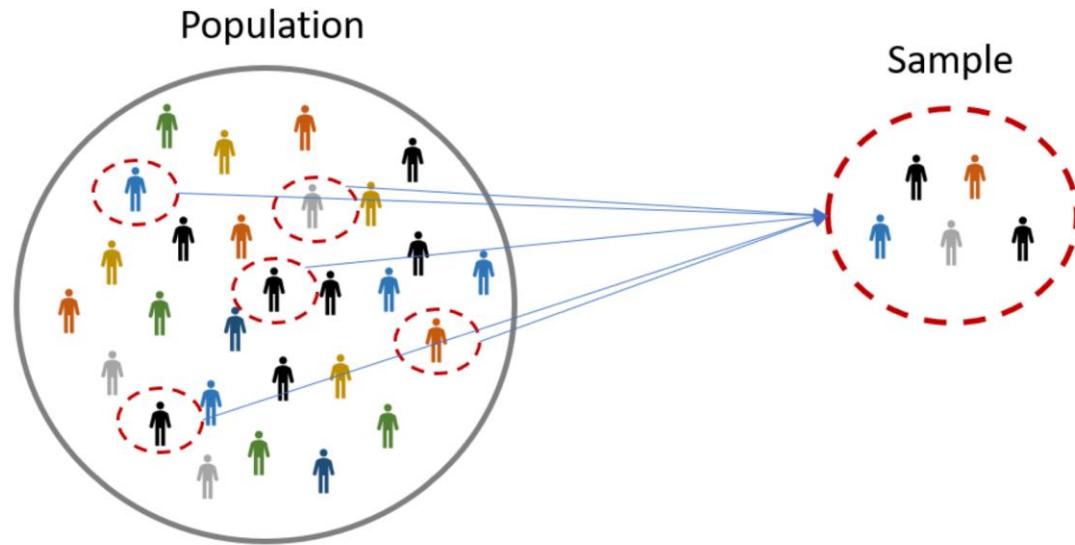
Khoảng biến thiên  
 $\min(x) \rightarrow \max(x)$



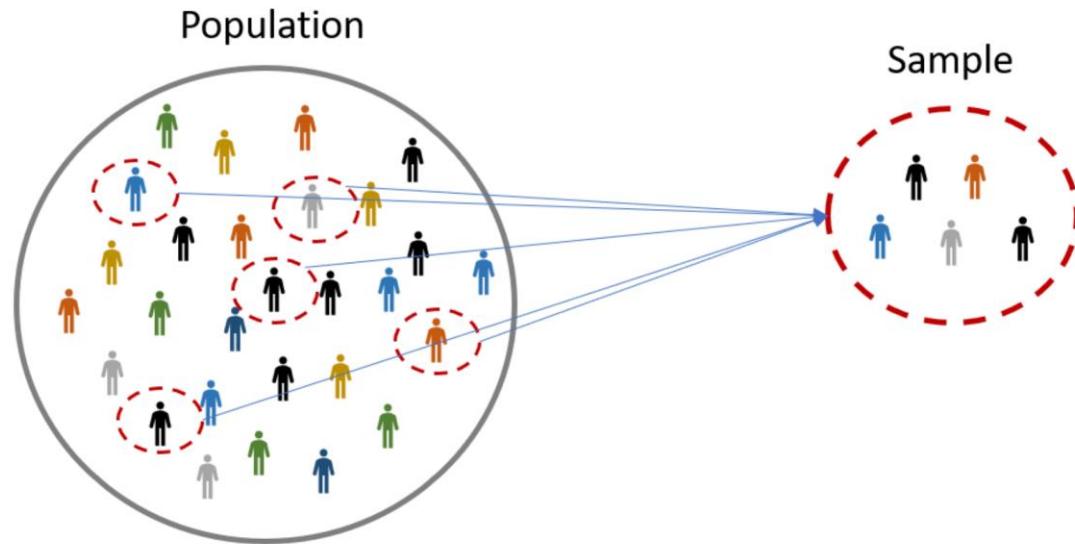
# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục



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

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

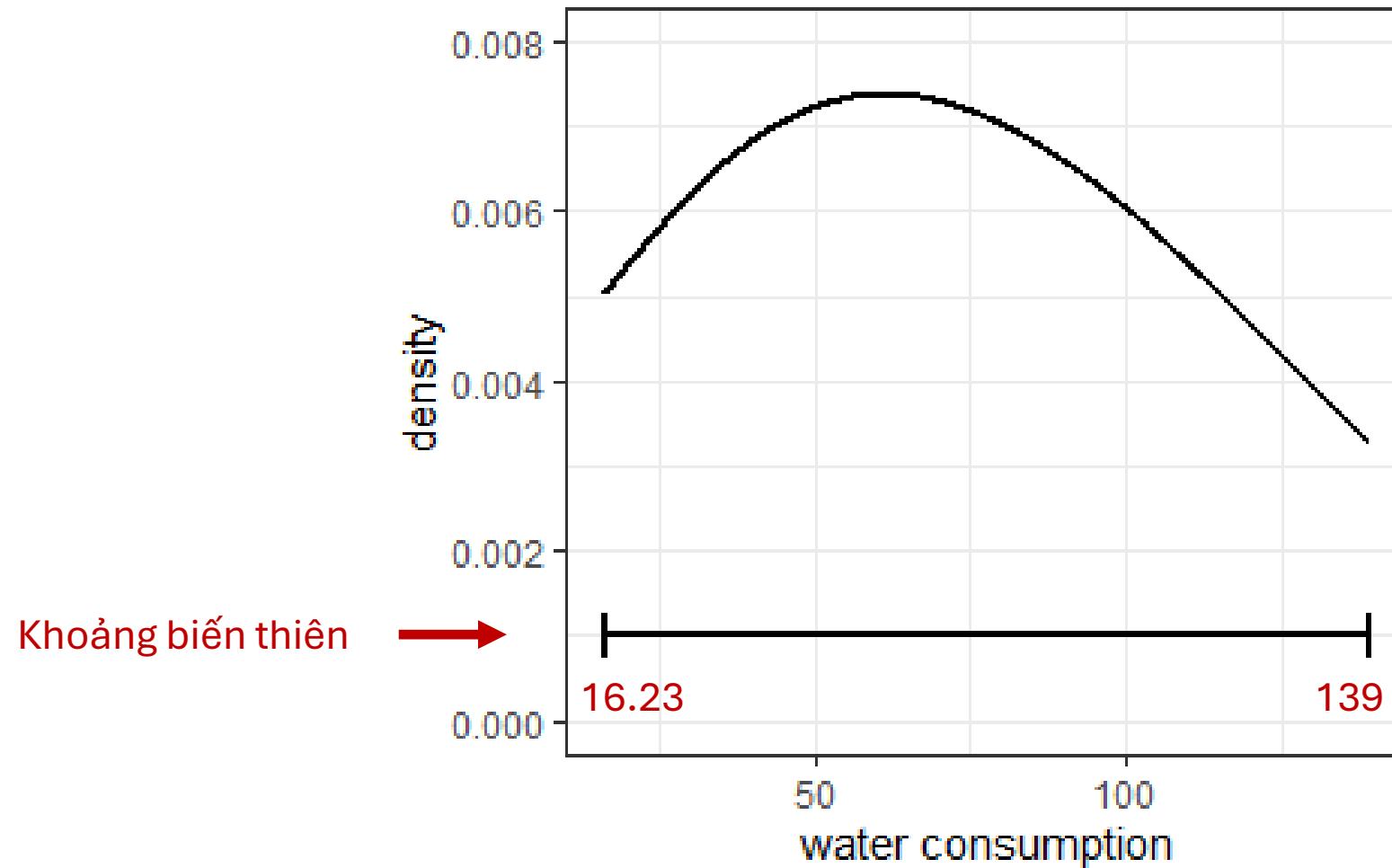
$\sigma$	Độ lệch chuẩn của quần thể
$\sigma^2$	Phương sai của quần thể
$X_i$	Giá trị của đối tượng $i$
$\bar{X}$	Trung bình của quần thể
$N$	Số lượng đối tượng trong quần thể
$s$	Độ lệch chuẩn của mẫu
$s^2$	Phương sai của mẫu
$x_i$	Giá trị của đối tượng $i$
$\bar{x}$	Trung bình của mẫu
$n$	Số lượng đối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

# Thống kê mô tả - Biến liên tục

- csmptv: lượng nước cấp tiêu thụ trong 1 năm
- rwtank: có bể nước mưa
- iceqac2: thu nhập của gia đình

<b>id</b>	<b>csmptv</b>	<b>rwtank</b>	<b>iceqac2</b>
137	105	no	modest
431	56.99	no	average
655	122	yes	modest
730	74.57	no	average
780	30	no	average
781	66.36	yes	higher
1048	30.93	yes	modest
1403	100	no	higher
1405	52.95	yes	modest
1432	139	no	modest
1476	25	yes	average
1757	71.25	yes	average
2183	69	yes	modest
2334	86.06	no	modest
2345	29.2	yes	precarious
2375	33.46	no	average
2687	45.63	no	precarious
2704	126	yes	higher
2714	16.23	yes	modest
2752	105.09	no	higher

# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

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

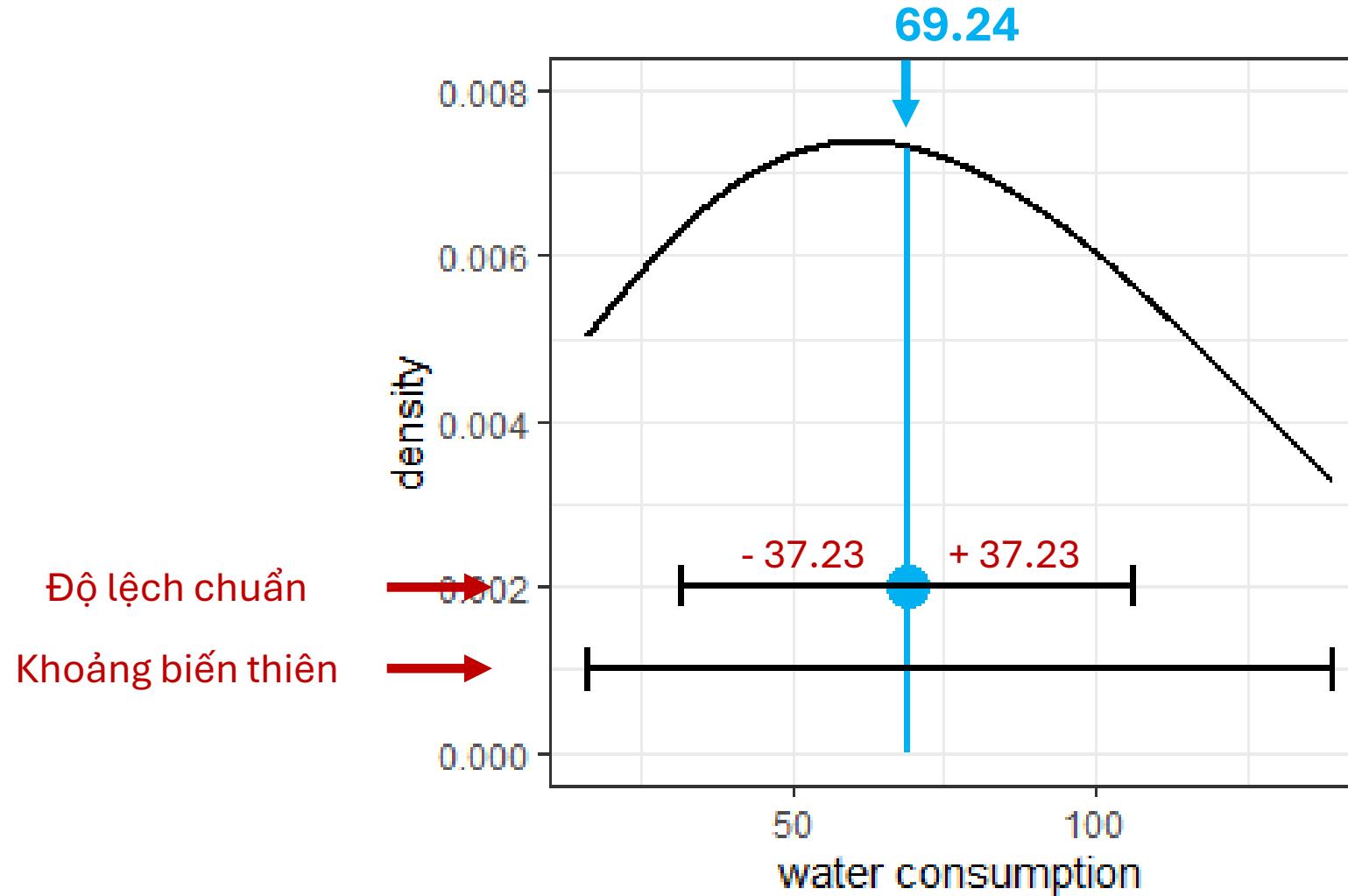
$$\frac{(105 - 69.24)^2 + (56.99 - 69.24)^2 + \dots + (105.09 - 69.24)^2}{(20 - 1)}$$

$$s^2 = 1386.14$$

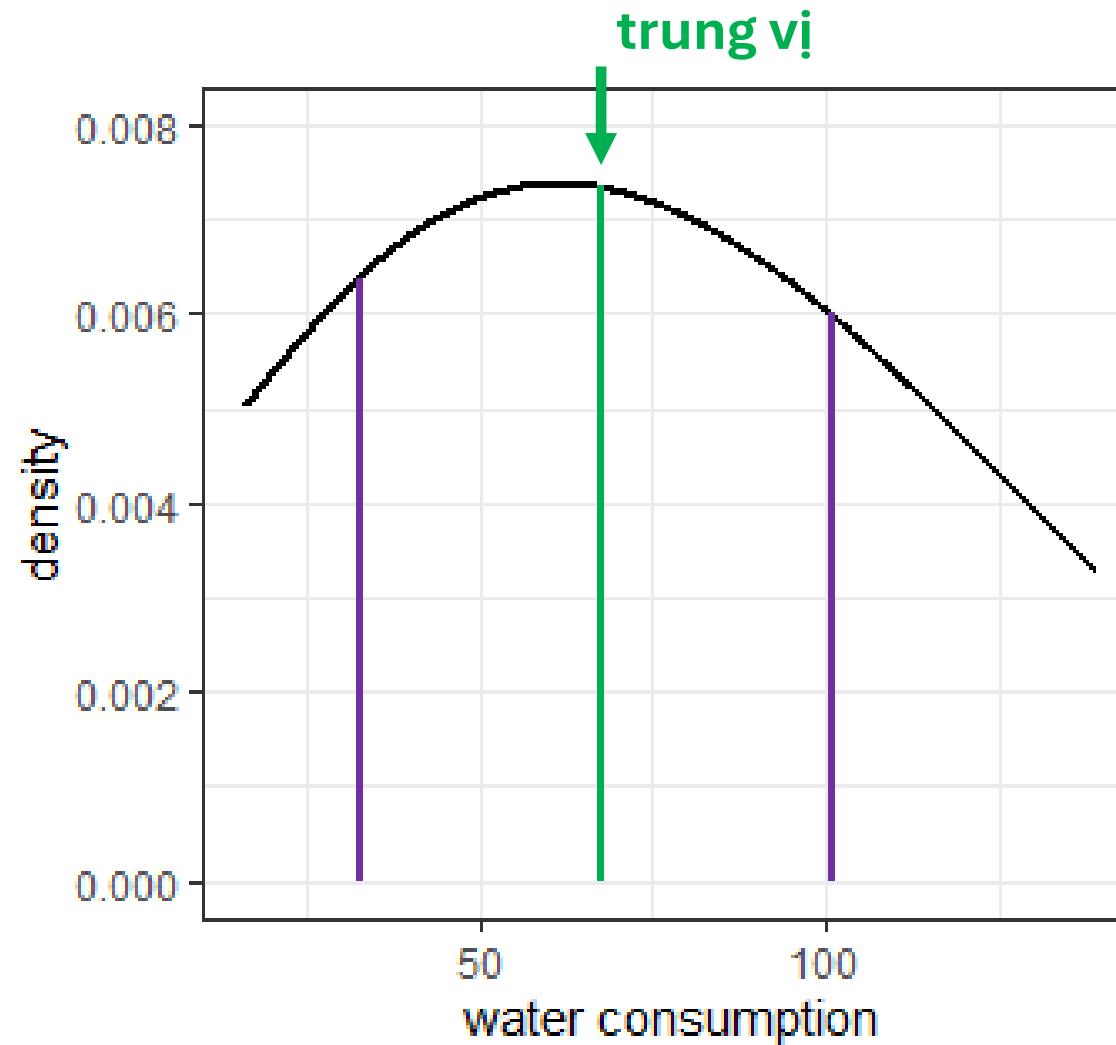
$$s = 37.23$$

id	csmptv	rwtank	iceqac2
137	105	no	modest
431	56.99	no	average
655	122	yes	modest
730	74.57	no	average
780	30	no	average
781	66.36	yes	higher
1048	30.93	yes	modest
1403	100	no	higher
1405	52.95	yes	modest
1432	139	no	modest
1476	25	yes	average
1757	71.25	yes	average
2183	69	yes	modest
2334	86.06	no	modest
2345	29.2	yes	precarious
2375	33.46	no	average
2687	45.63	no	precarious
2704	126	yes	higher
2714	16.23	yes	modest
2752	105.09	no	higher

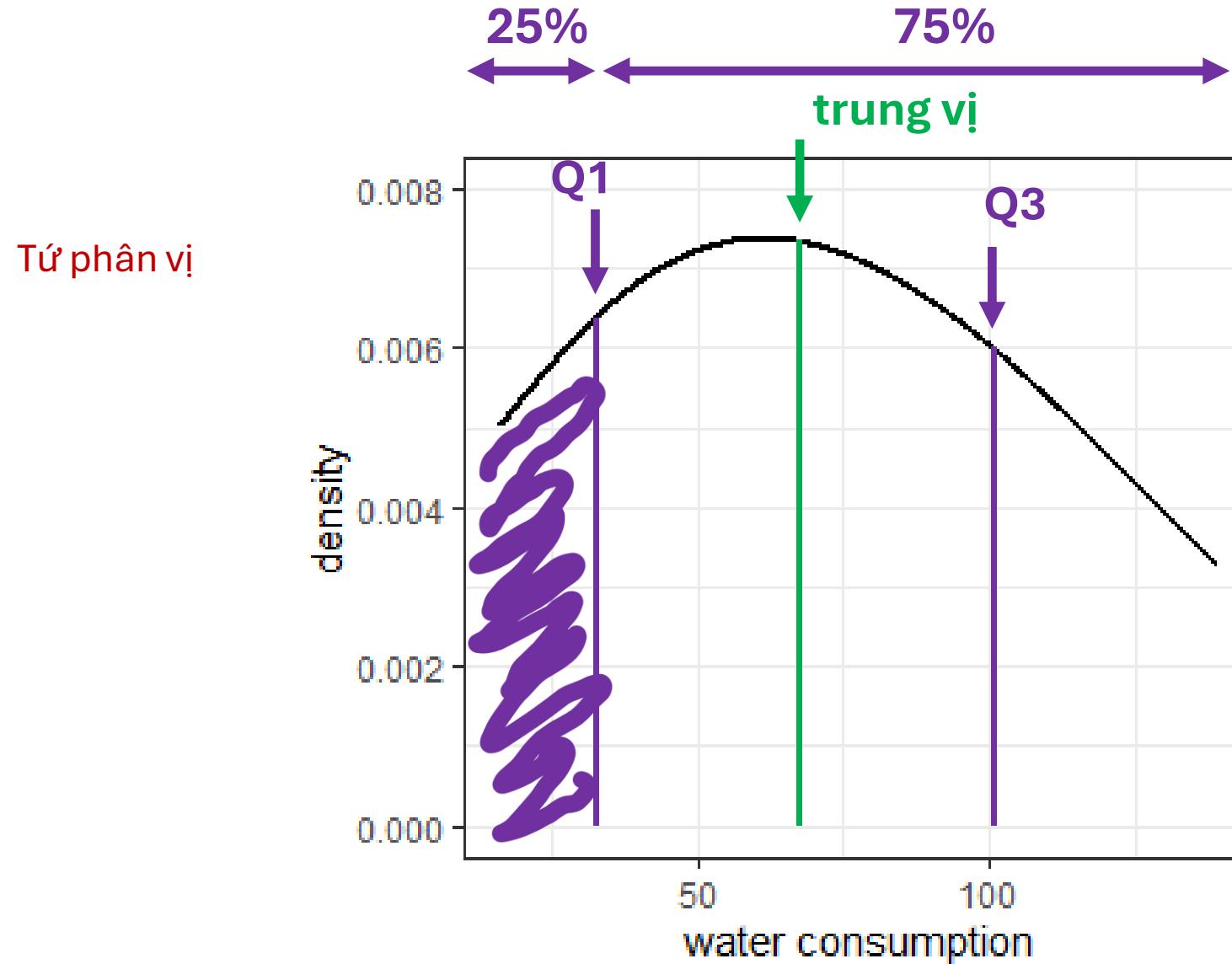
# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục



# Thống kê mô tả - Biến liên tục

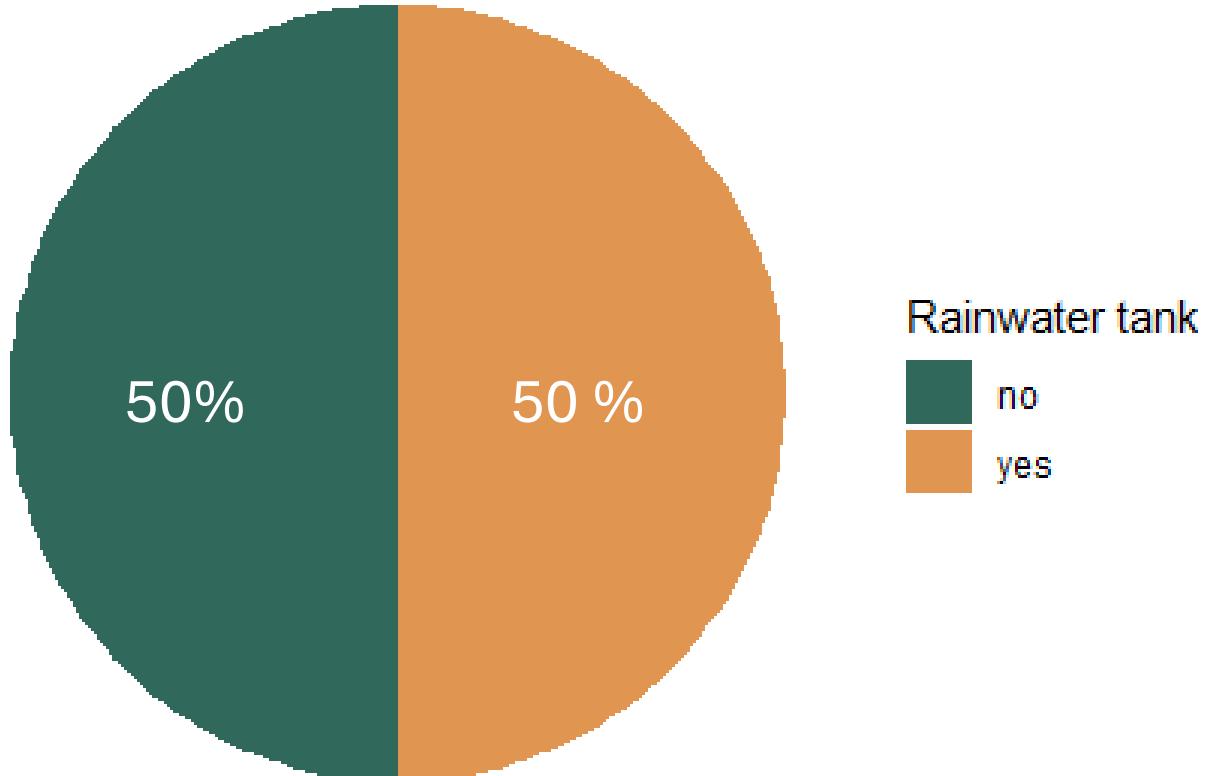
tứ phân vị 1 = Q1 = 32.83

trung vị = Q2 = 67.68

tứ phân vị 3 = Q3 = 101.25

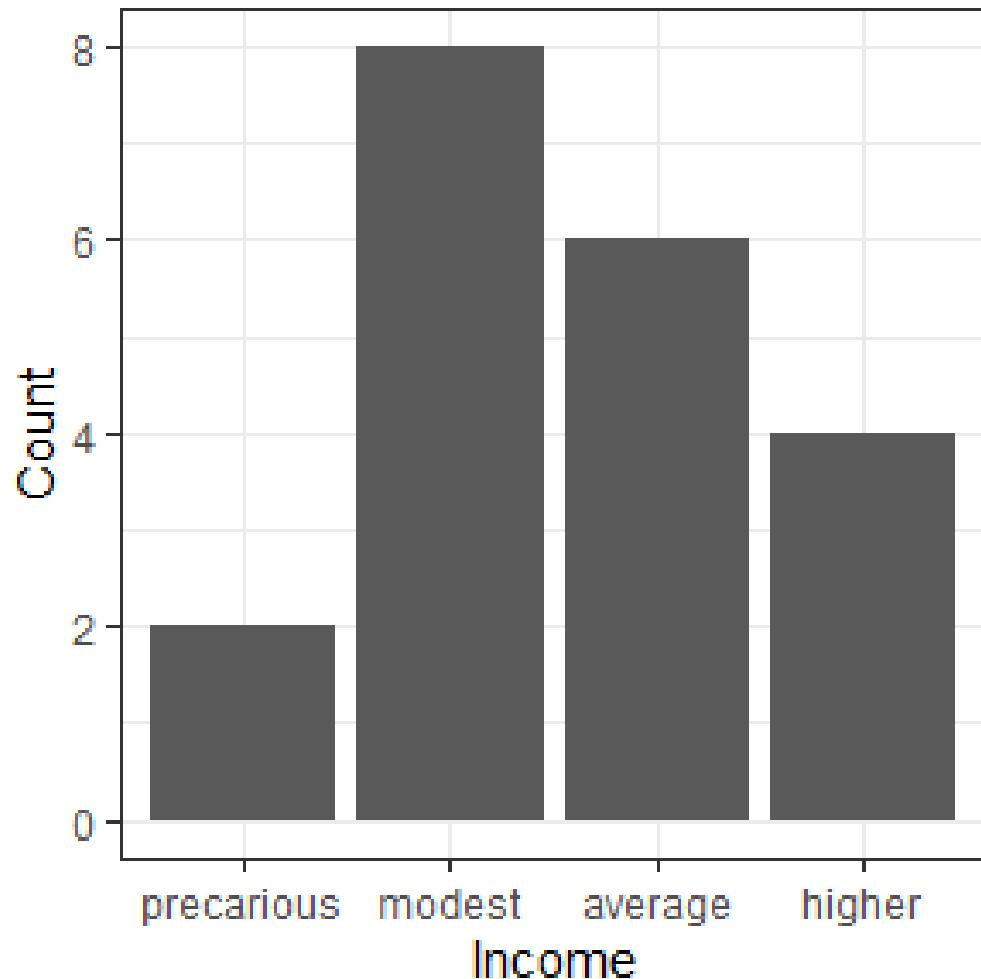
<b>id</b>	<b>csmpty</b>	<b>rwtank</b>	<b>iceqac2</b>	<b>hhs_tot</b>	<b>cfdiwq</b>	<b>livara</b>
2714	16.23	yes	modest	1	confident	80
1476	25	yes	average	2	confident	100
2345	29.2	yes	precarious	3	confident	160
780	30	no	average	1	rather confident	70
1048	30.93	yes	modest	3	suspicious	110
2375	33.46	no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

# Thống kê mô tả - Biến rời rạc



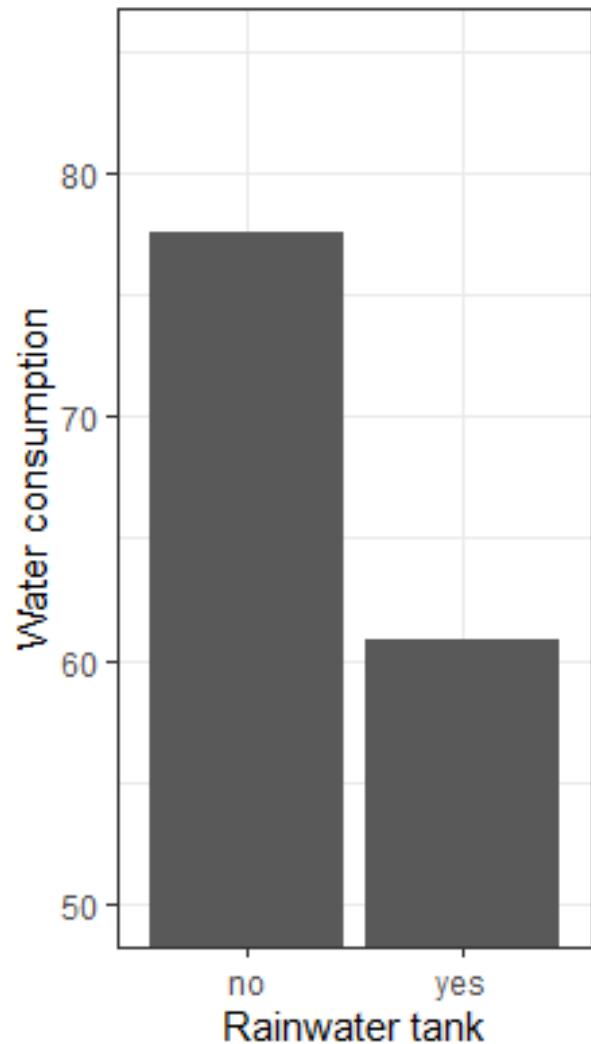
<b>Rainwater tank</b>	<b>Count</b>	<b>%</b>
No	10	50
Yes	10	50
<b>Total</b>	<b>20</b>	<b>100</b>

# Thống kê mô tả - Biến rời rạc

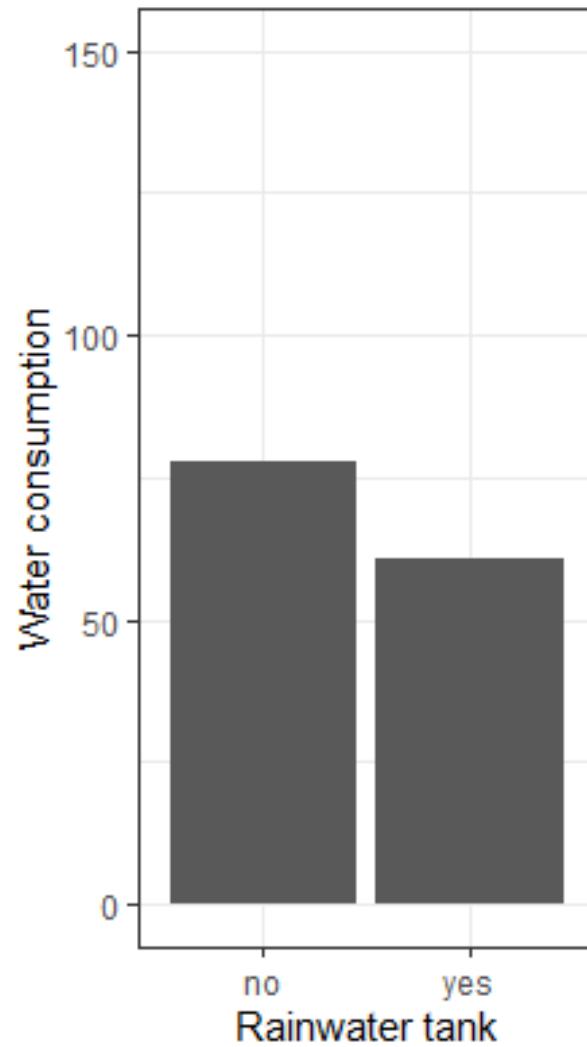
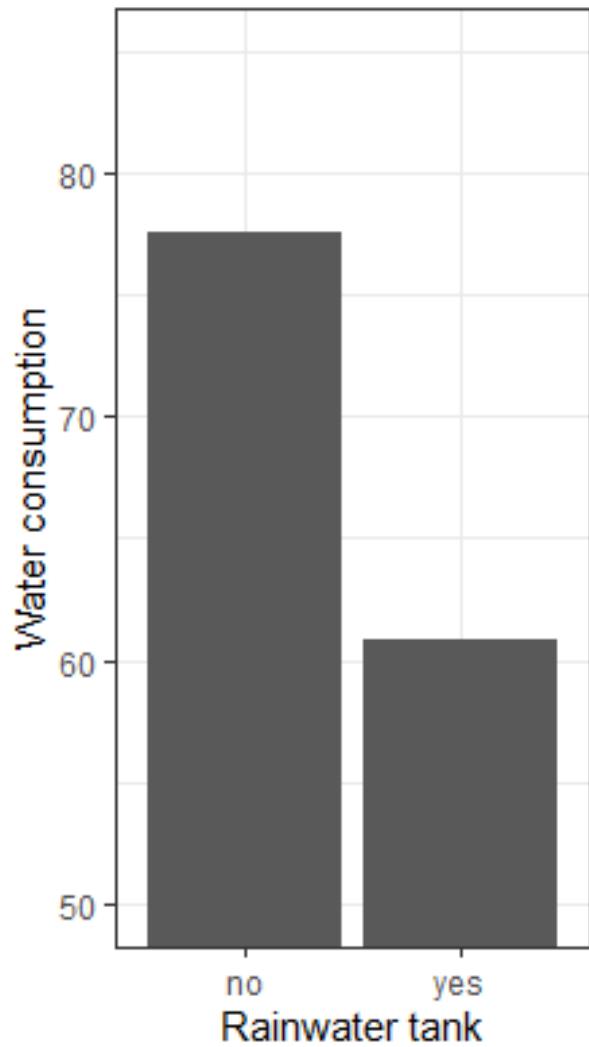


Income	Count	%
precarious	2	10
modest	8	40
average	6	30
higher	4	20
<b>Total</b>	<b>20</b>	<b>100</b>

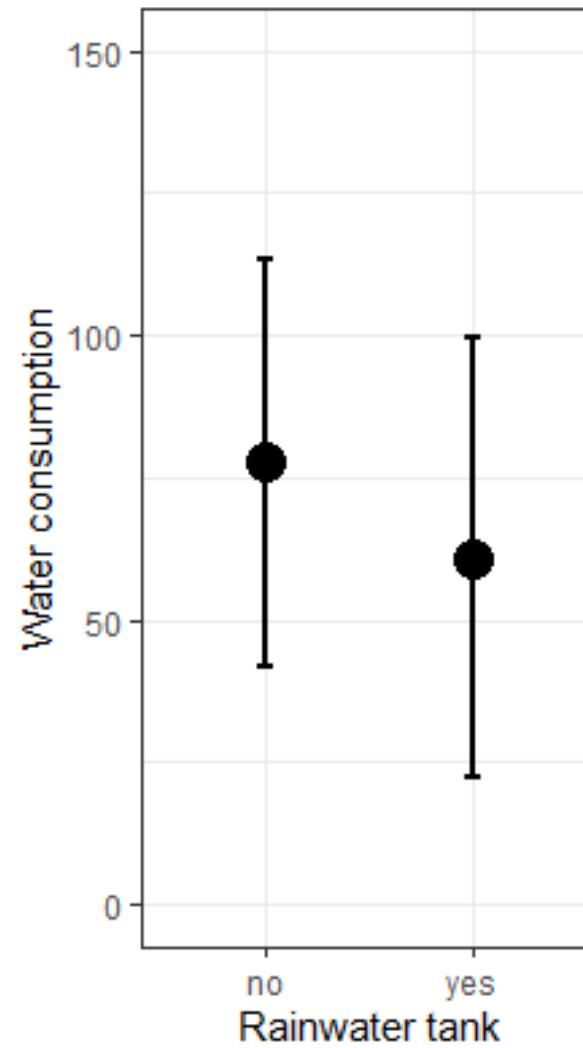
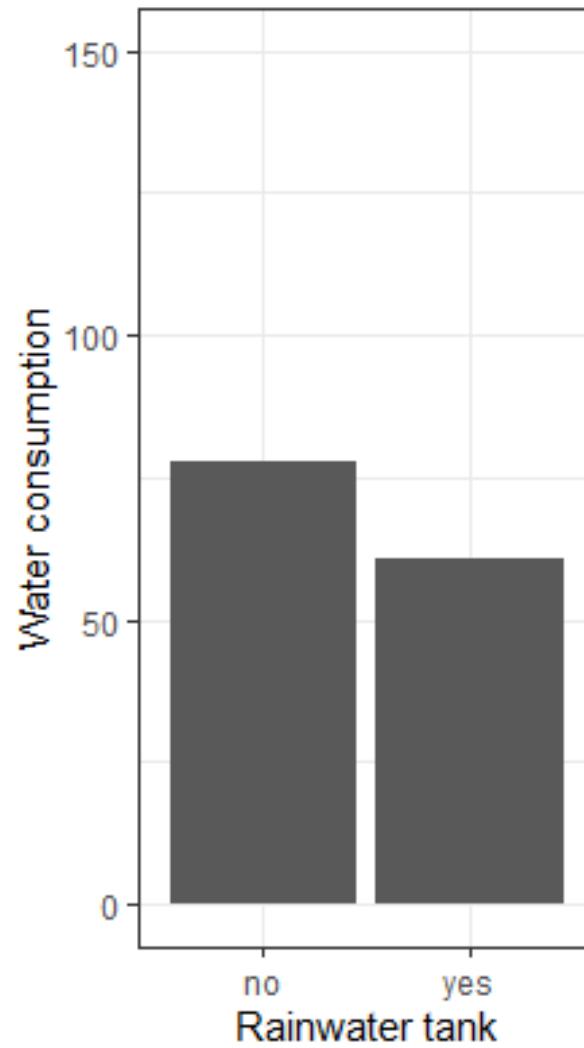
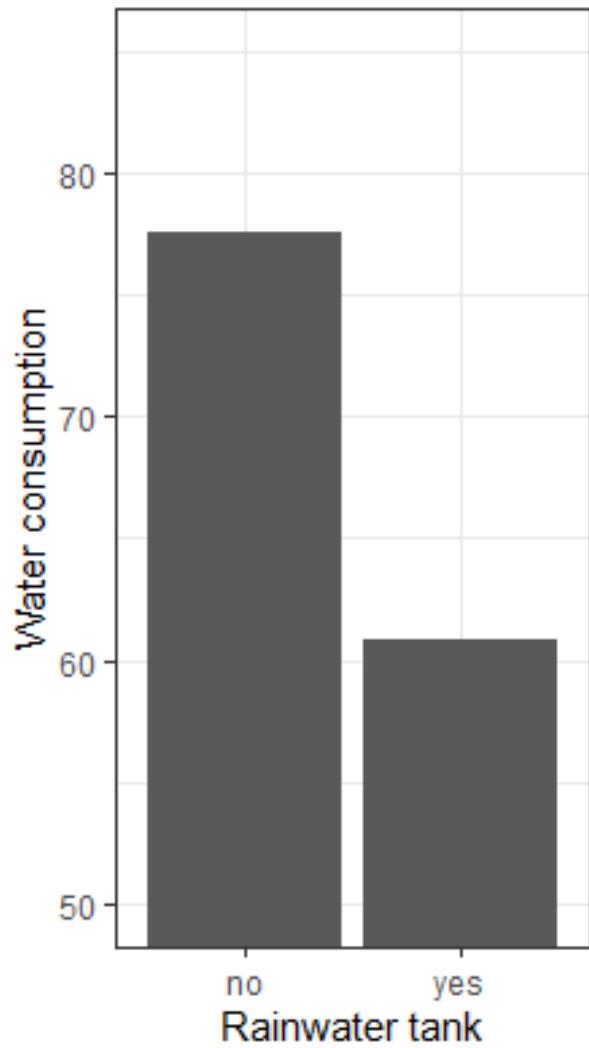
# Thống kê mô tả - Cho từng nhóm



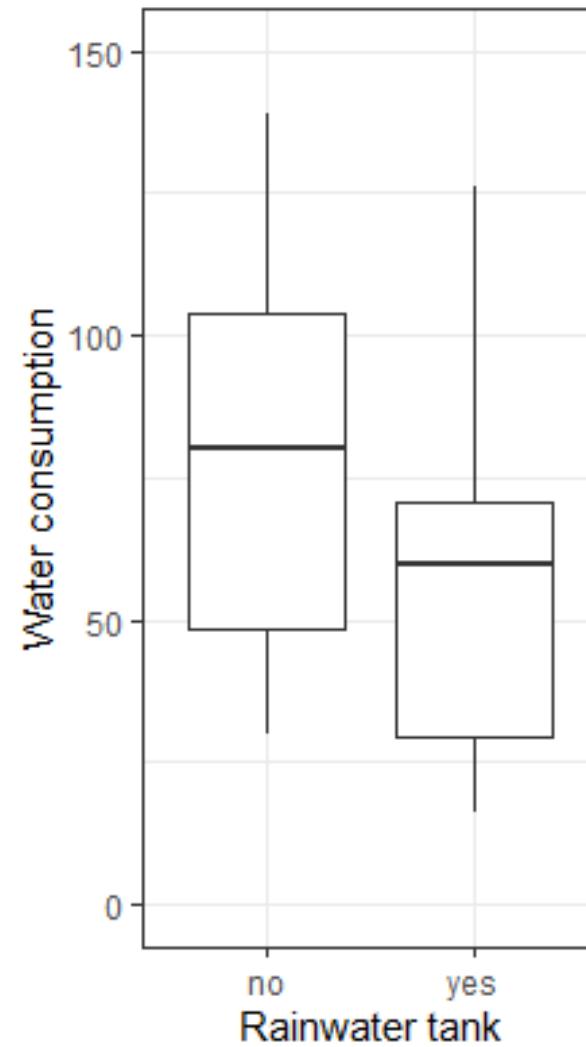
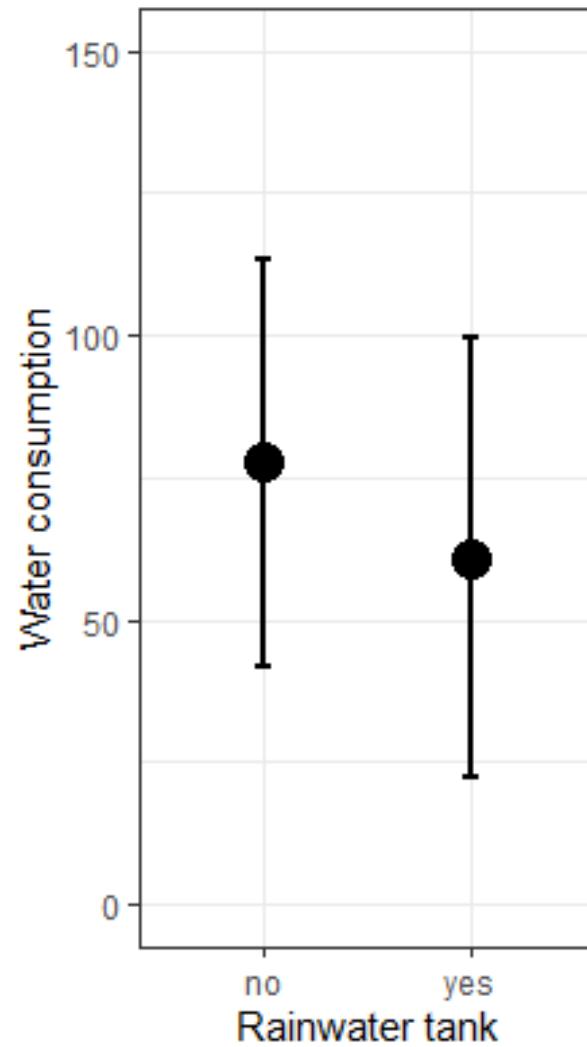
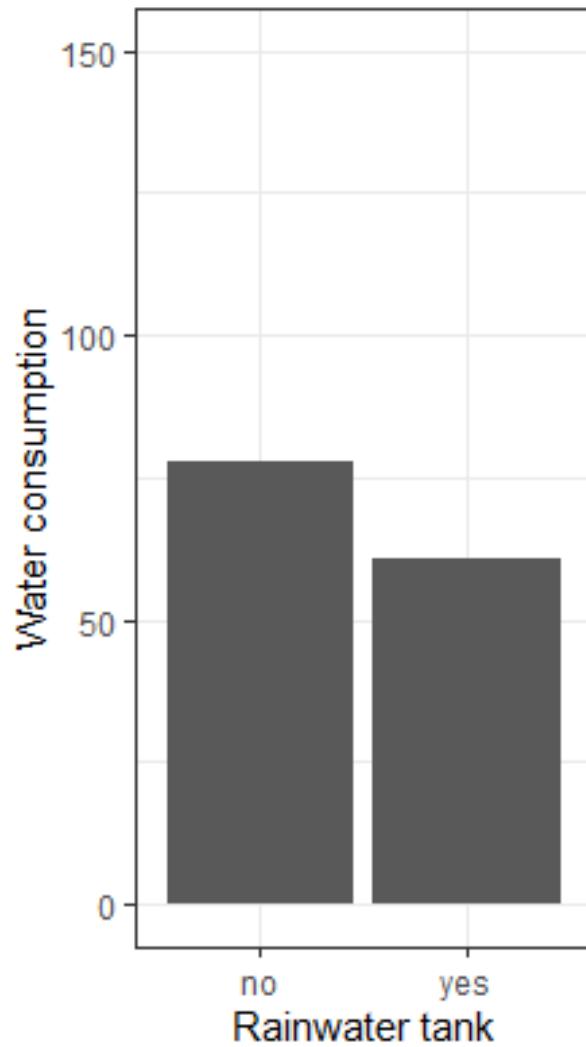
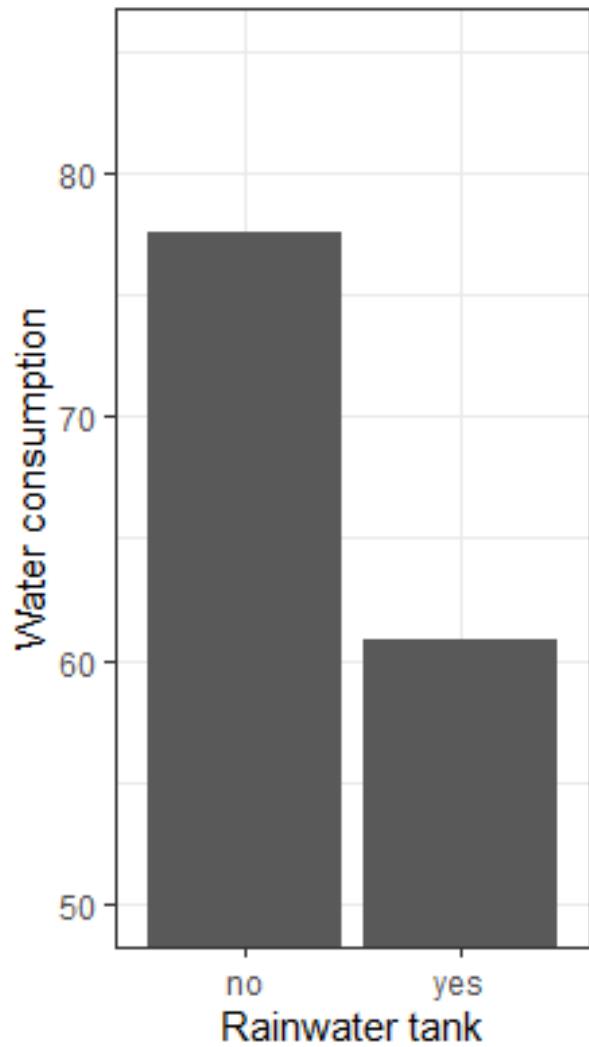
# Thống kê mô tả - Cho từng nhóm



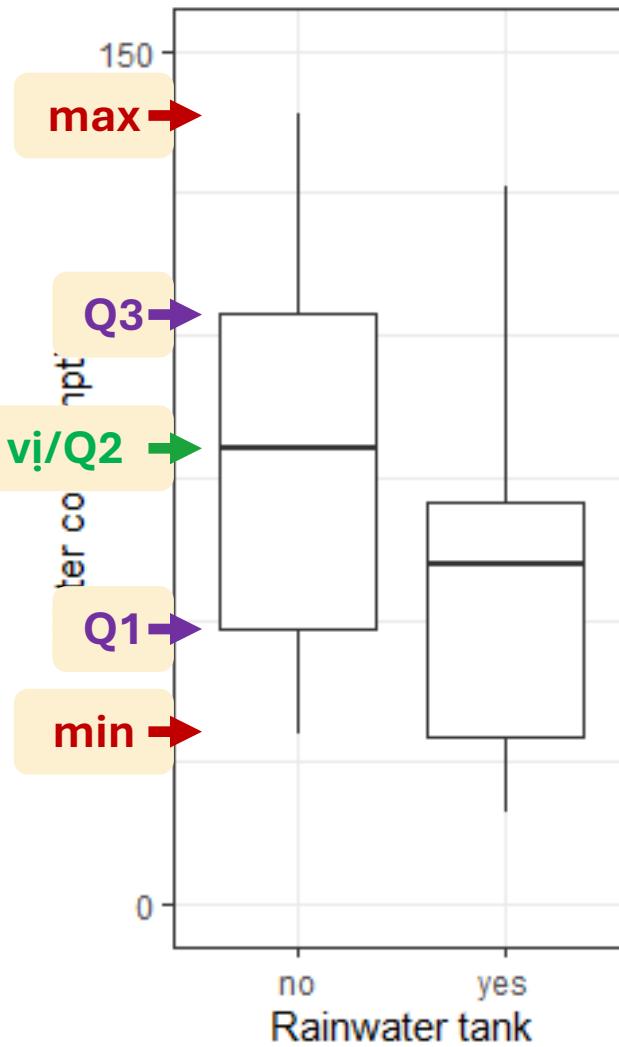
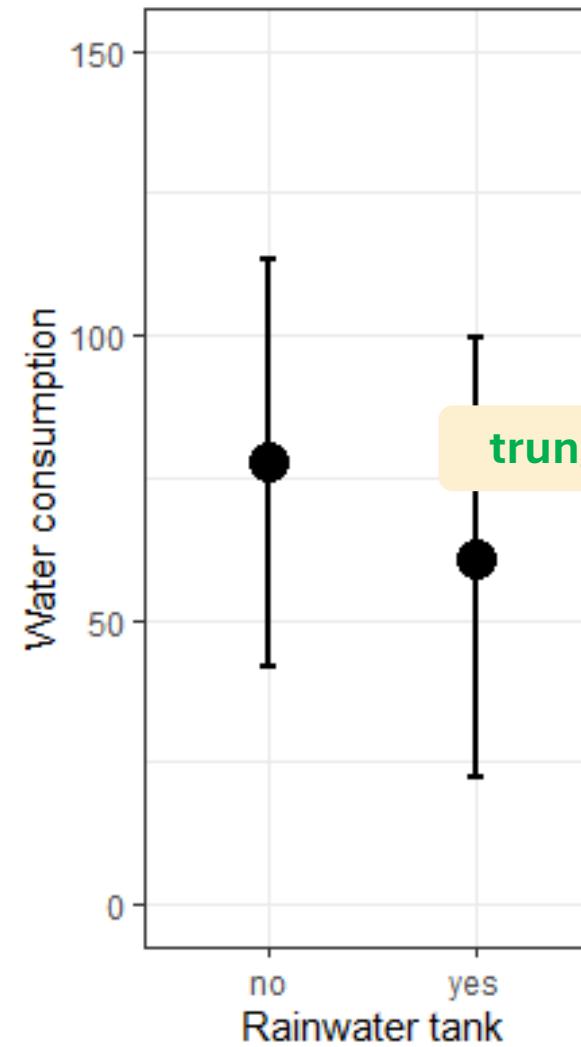
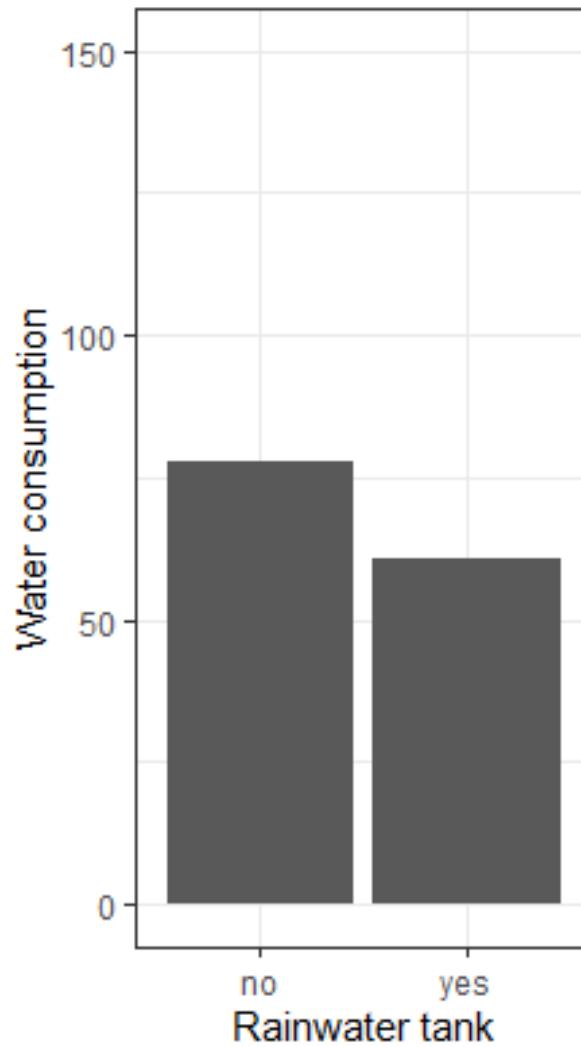
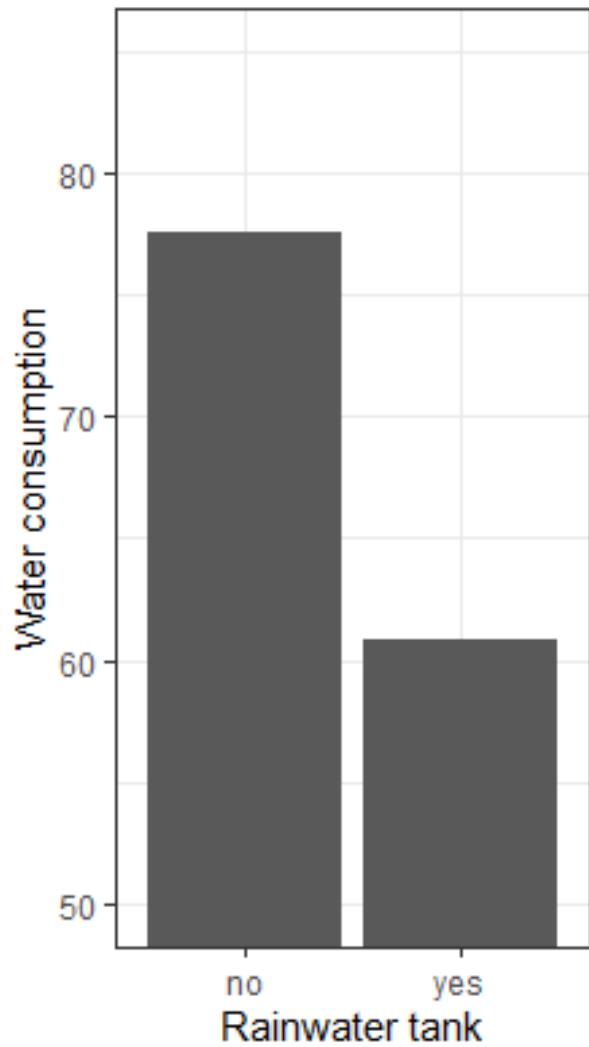
# Thống kê mô tả - Cho từng nhóm



# Thống kê mô tả - Cho từng nhóm

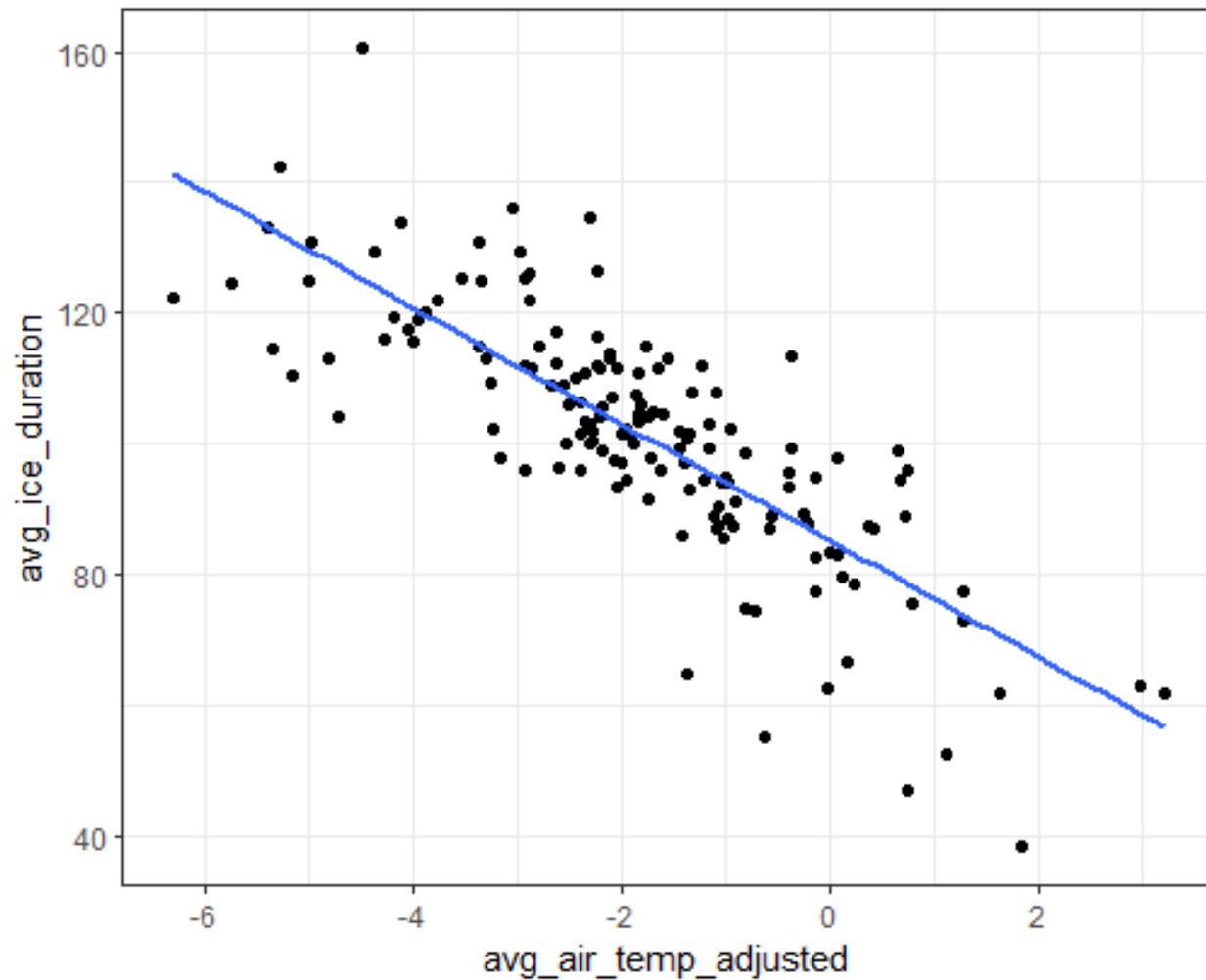


# Thống kê mô tả - Cho từng nhóm



# Thống kê mô tả - Tương quan

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

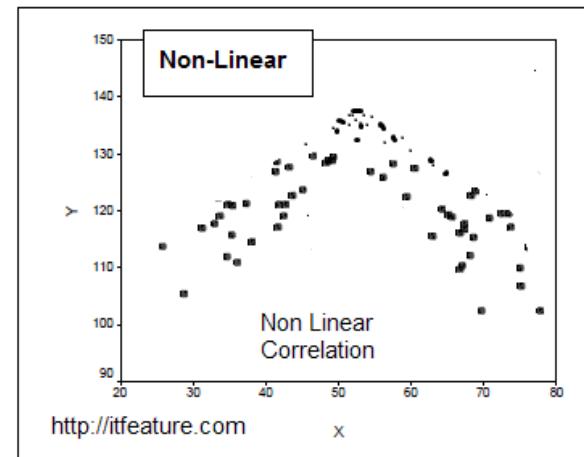
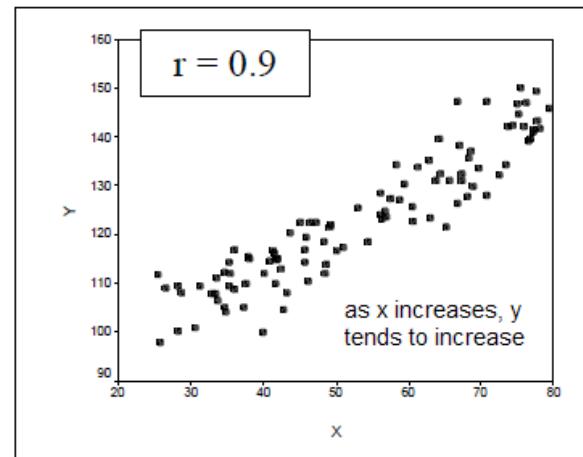
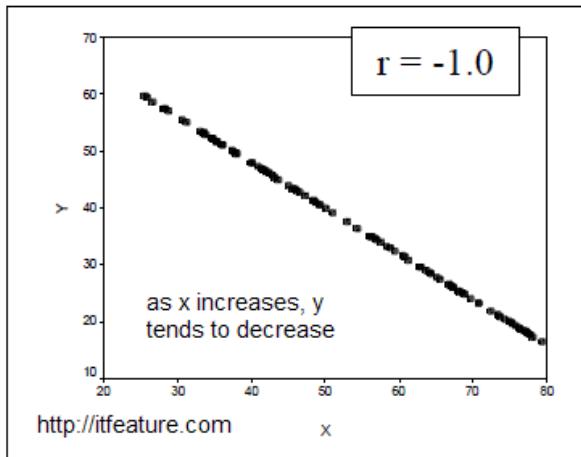
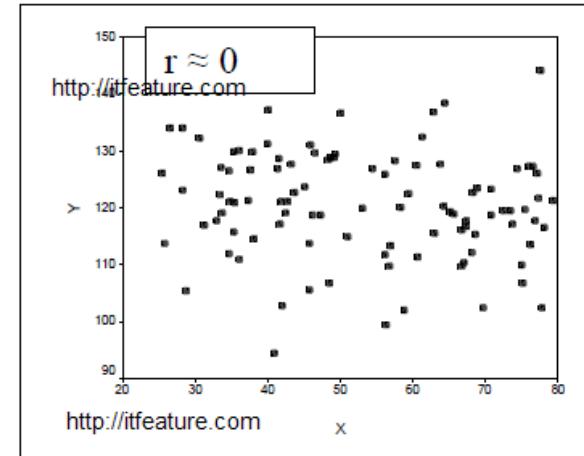
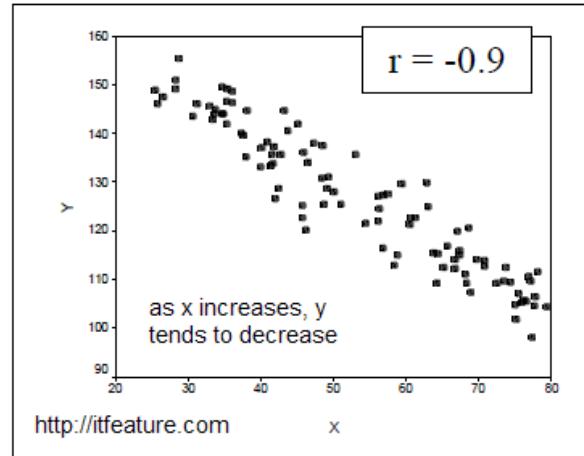
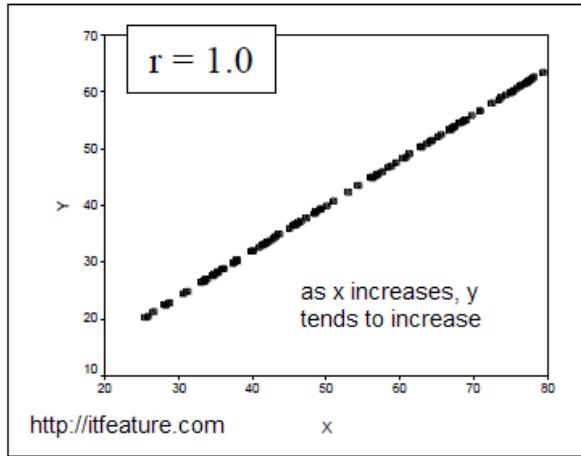


# Thống kê mô tả - Tương quan

- Tương quan Pearson's  $r$

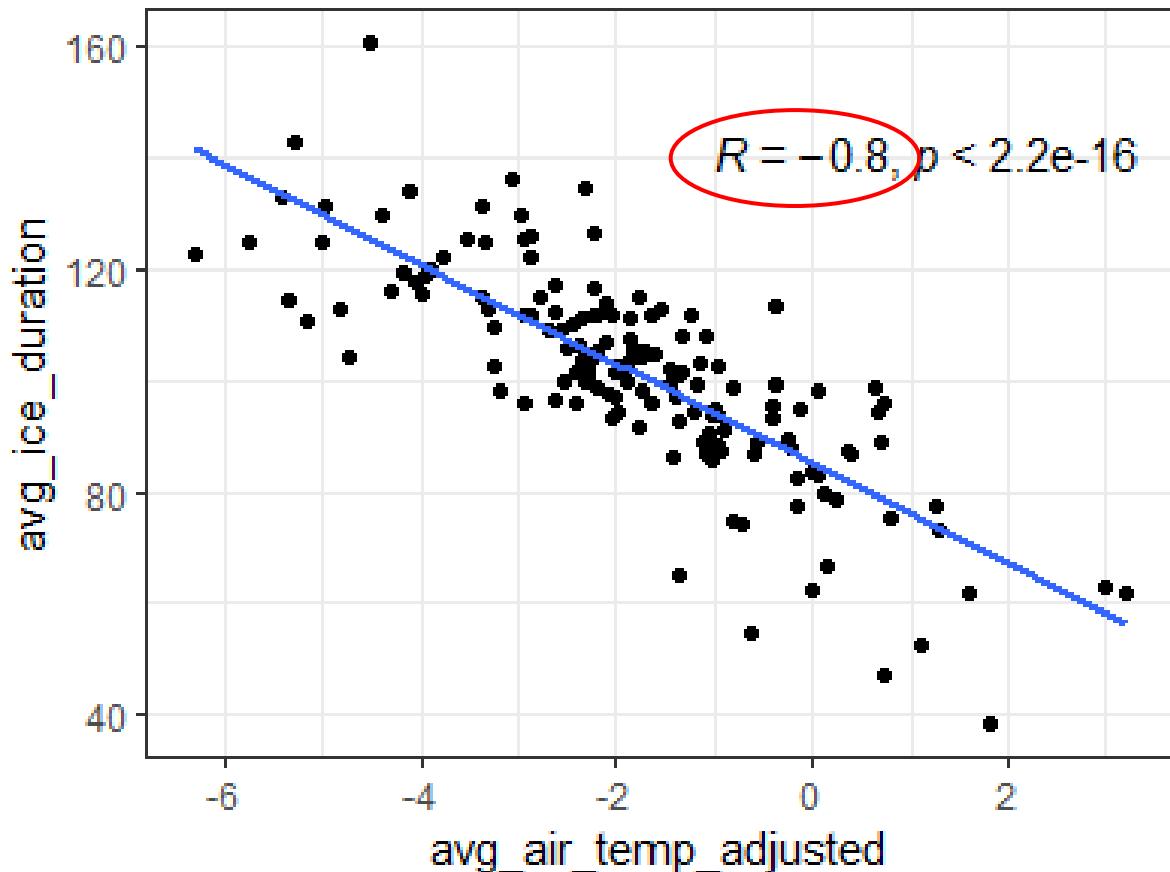
$$r = \frac{N \sum xy - \sum x \sum y}{\sqrt{\left[ N \sum x^2 - (\sum x)^2 \right] \left[ N \sum y^2 - (\sum y)^2 \right]}}$$

# Thống kê mô tả - Tương quan

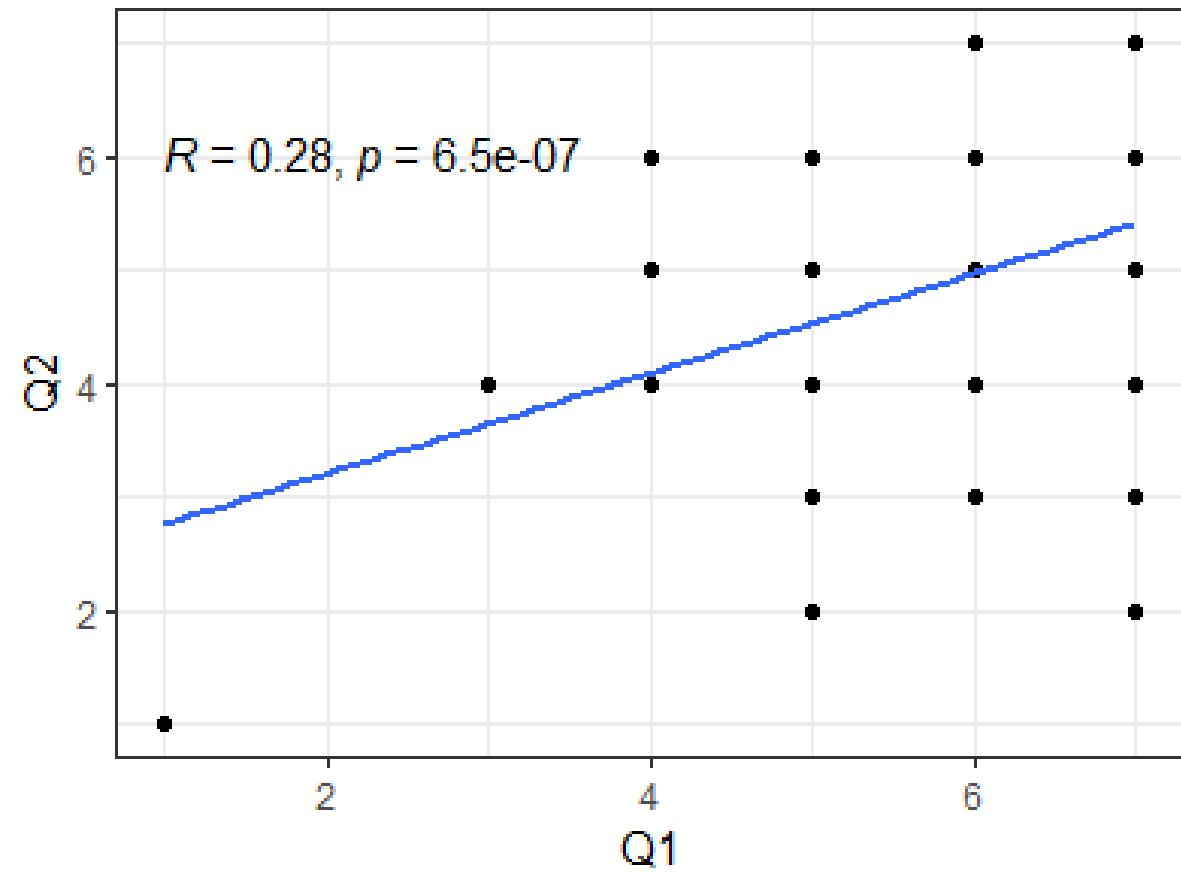


# Thống kê mô tả - Tương quan

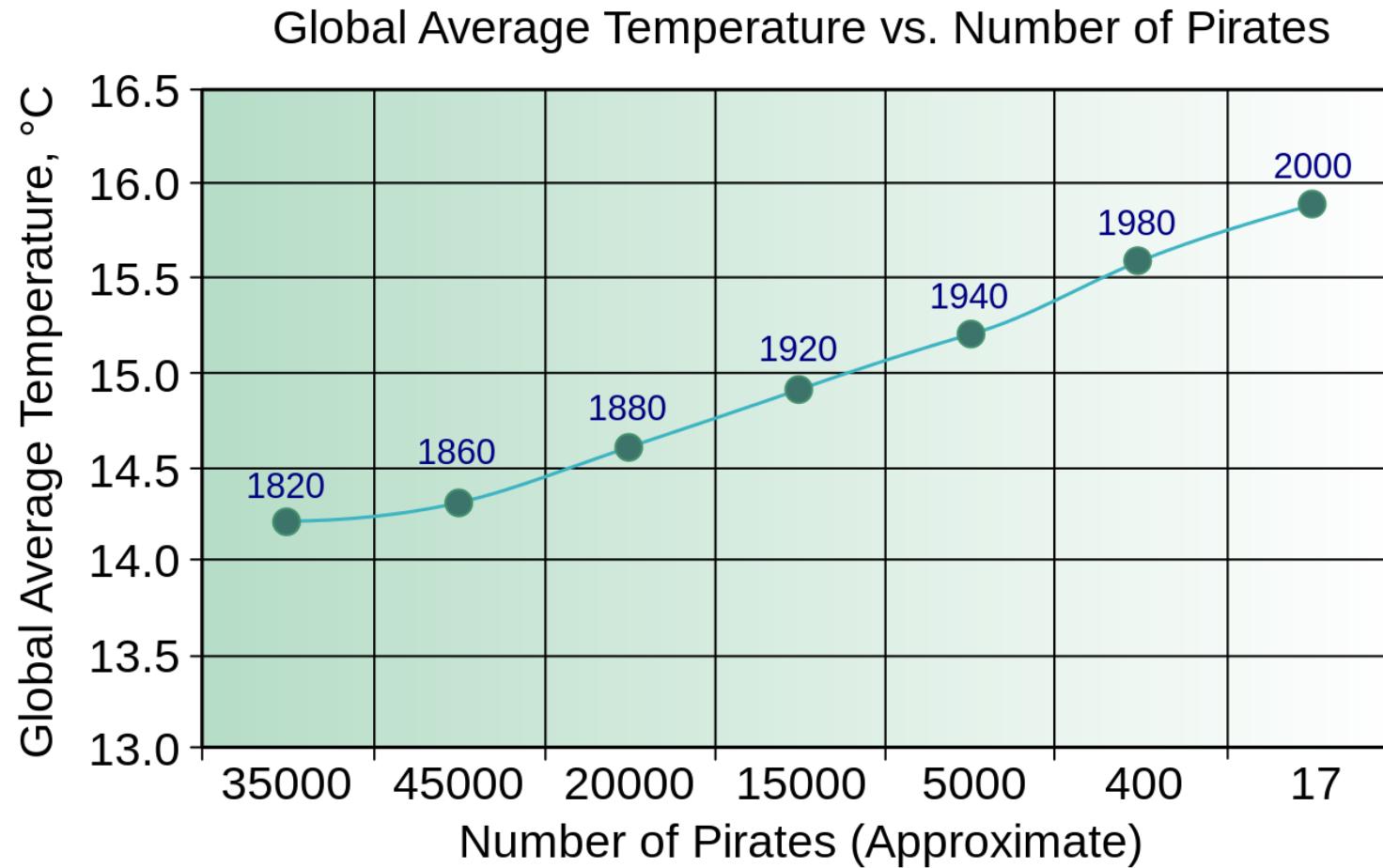
Pearson



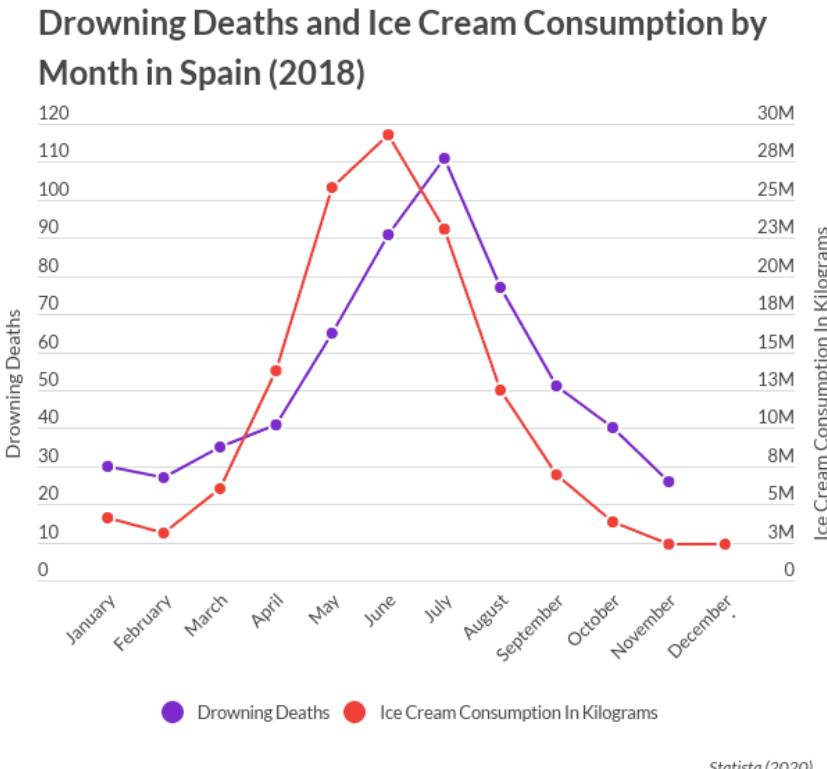
Spearman



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



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

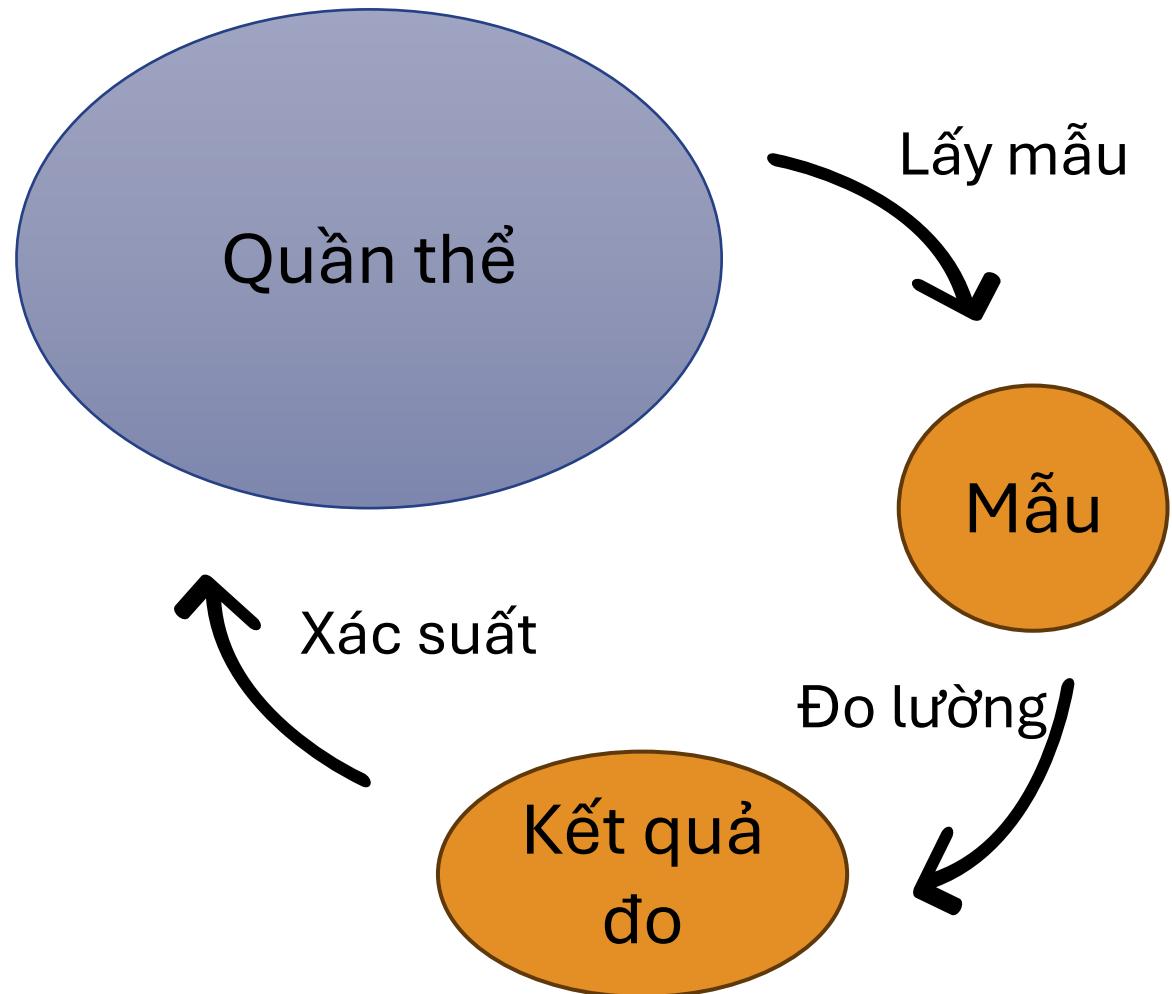


# Các bước phân tích

- 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

- 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

# 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 trống là đúng
- Giá trị  $p < 0.05$ : 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 – Ví dụ

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

# Thống kê suy luận – Ví dụ

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

Kiểm định t

$$H_0: \mu_1 = \mu_2$$

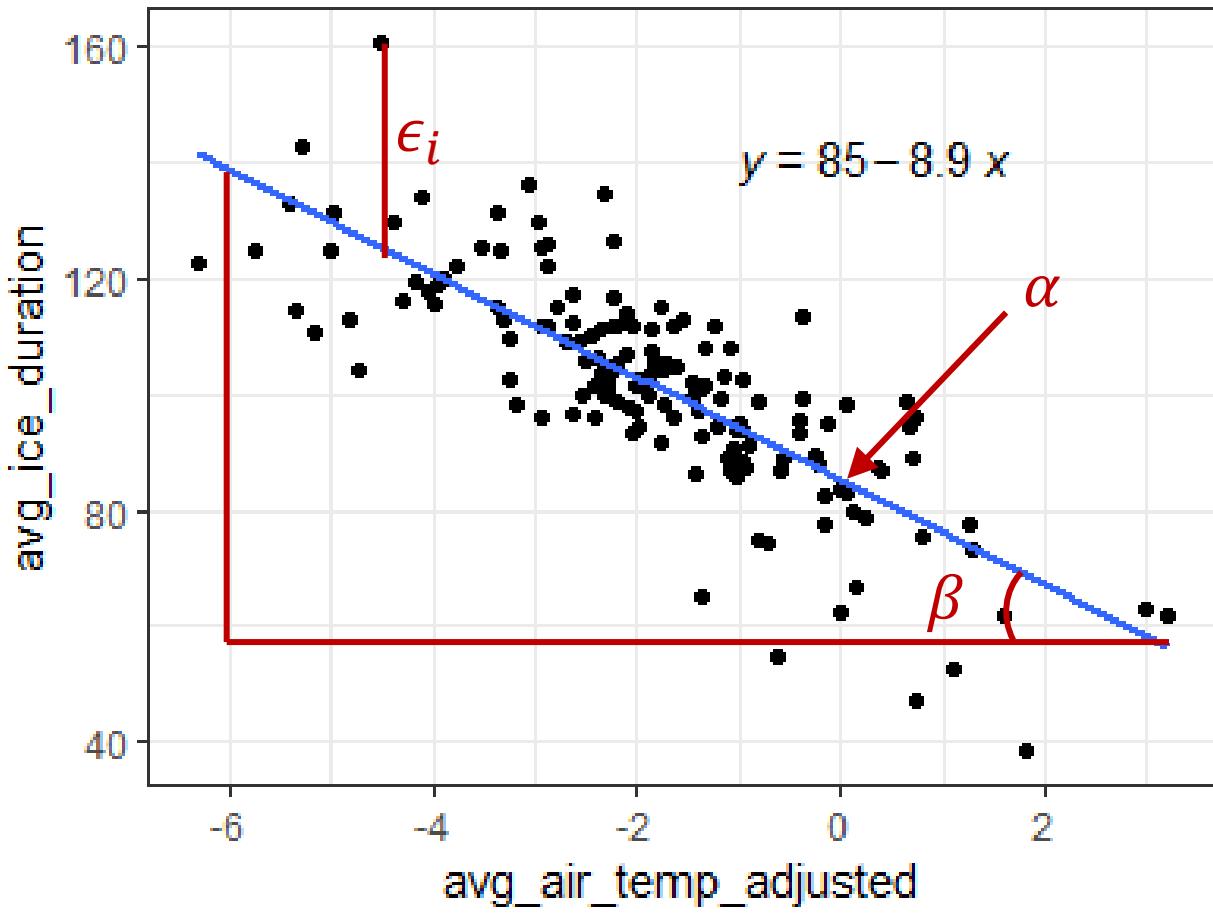
$$H_a: \mu_1 \neq \mu_2$$

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

# 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)
- ...

# 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**

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

$$\alpha = 45.31 \qquad \qquad \beta = 0.11$$

# Hồi quy tuyến tính đơn giản

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

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

$$\alpha = 45.31 \quad \beta = 0.11$$

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

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

$$R^2 = 0.037$$

# 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

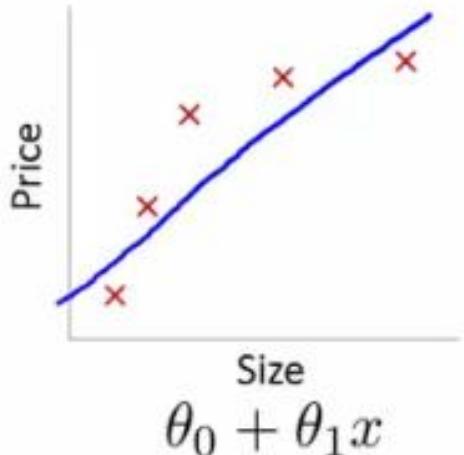
Ví dụ: Giải thích sự biến thiên của **lượng nước tiêu thụ** bởi **diện tích nhà ở và sử dụng nước mưa**

*csmptv*

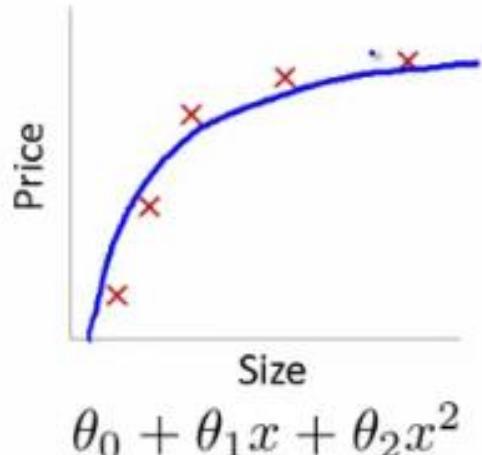
$$= \alpha + \beta_1 * livara + \beta_2 * livara^2 + \beta_3 * rwtank + \beta_4 * livara * rwtank + \epsilon_i$$

?

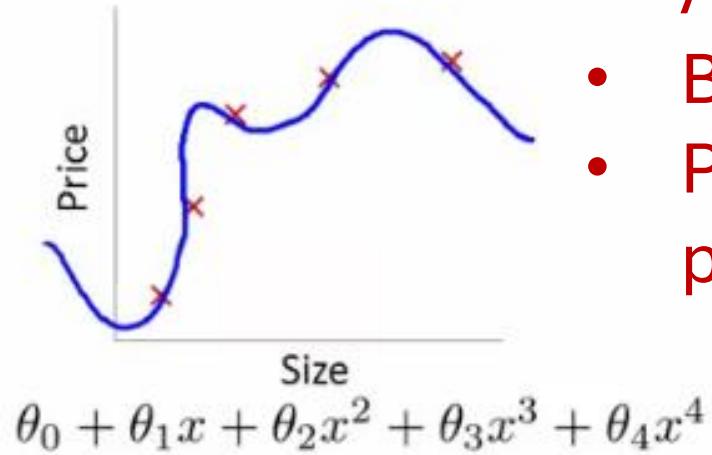
# Quá khớp hoặc thiếu khớp với số liệu



High bias  
(underfit)



"Just right"

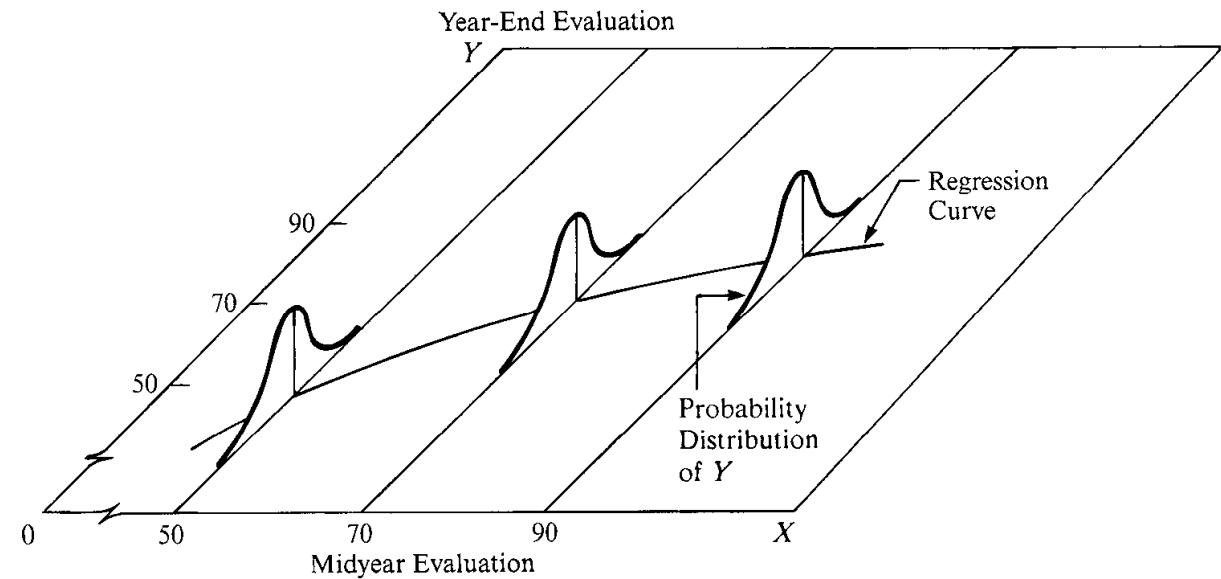


High variance  
(overfit)

- Adjusted R<sup>2</sup>
- AIC
- BIC
- Predictive power

# Hồi quy tuyến tính – Các giả định

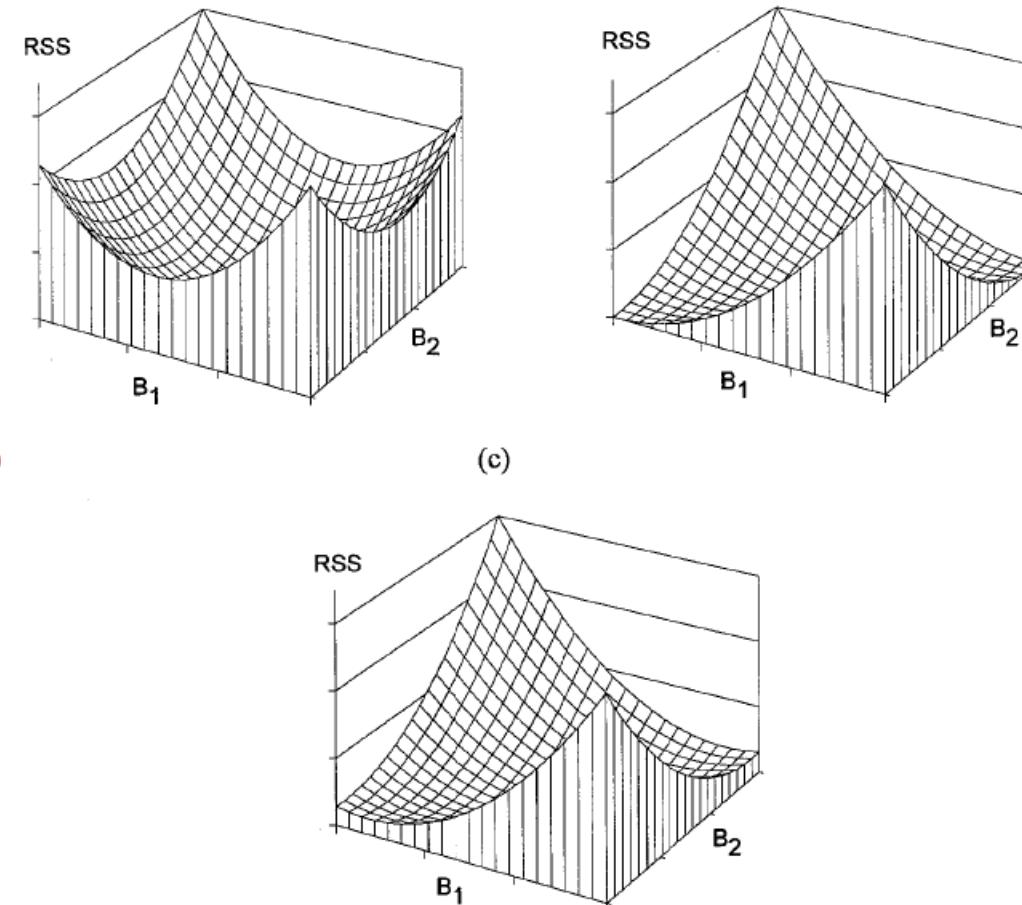
- $Y$  là biến liên tục
- Mối liên hệ tuyến tính giữa  $Y$  với các tham số khảo sát
- Các giá trị  $Y$  độc lập với nhau
- Các sai số ngẫu nhiên tuân theo phân phối chuẩn có cùng phương sai và trung bình = 0



# 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)

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

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

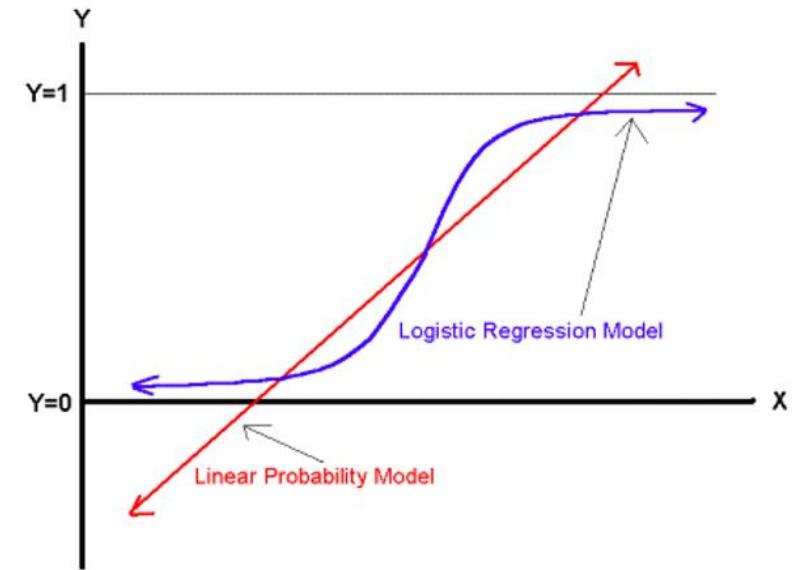
Khi  $Y$  là biến rời rạc?

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

Định danh: Multinomial logistic regression

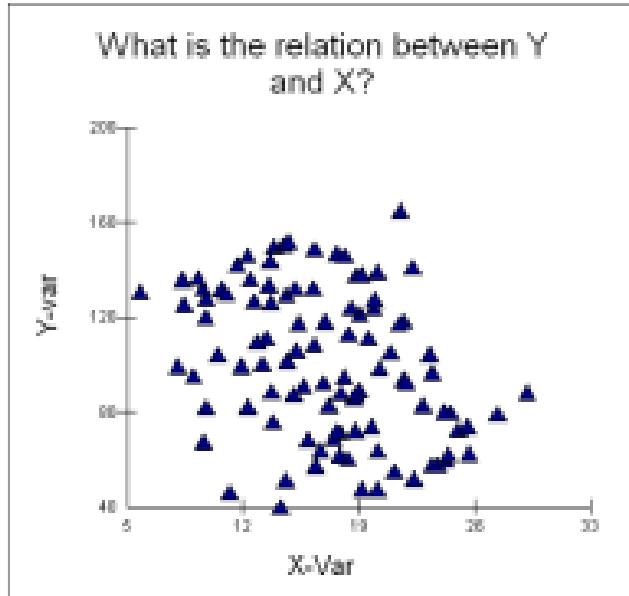
Thứ bậc: Cumulative logistic regression

Biến đếm: Poisson regression



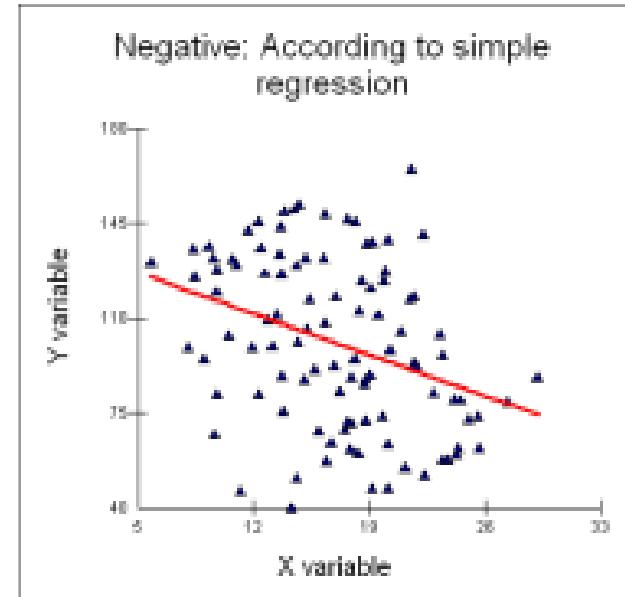
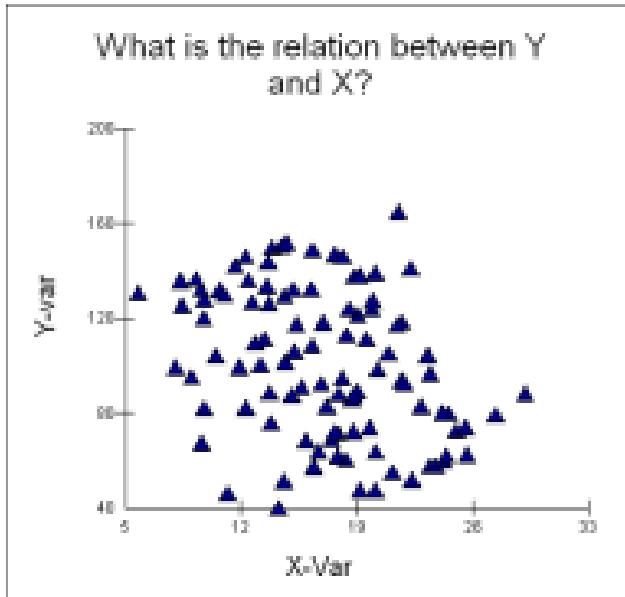
# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

## Multilevel model/Mixed effect model



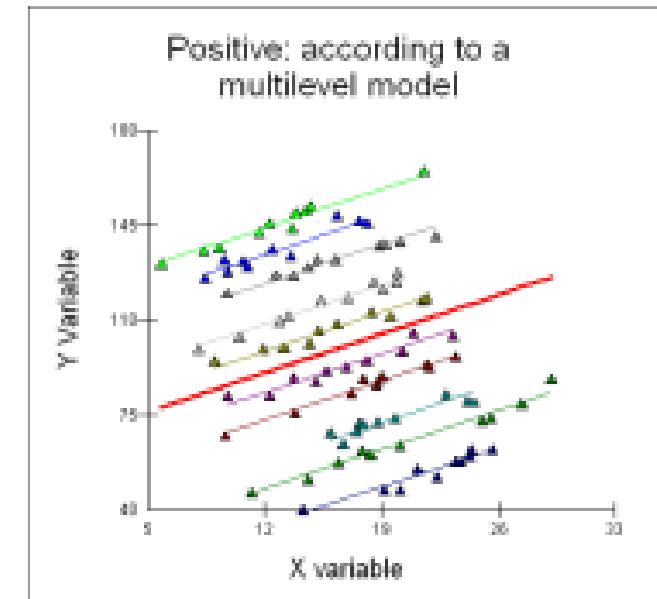
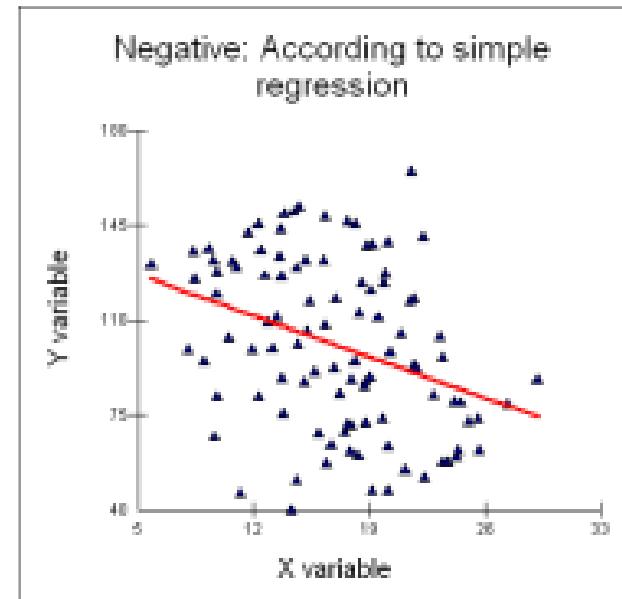
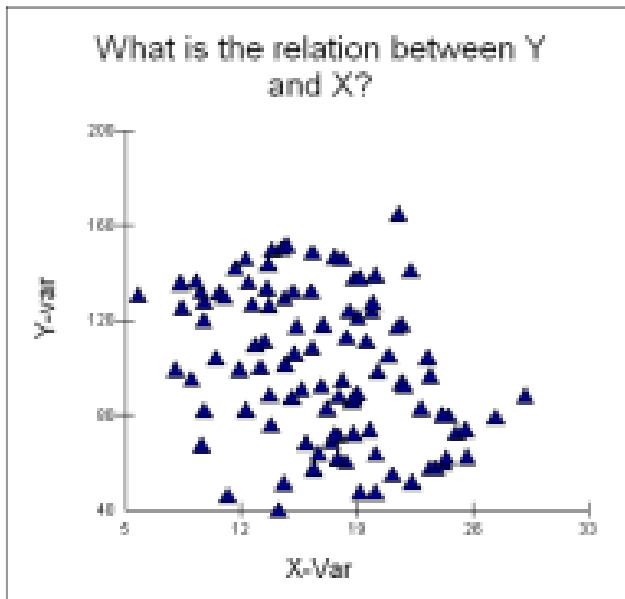
# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

## Multilevel model/Mixed effect model



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

## Multilevel model/Mixed effect model

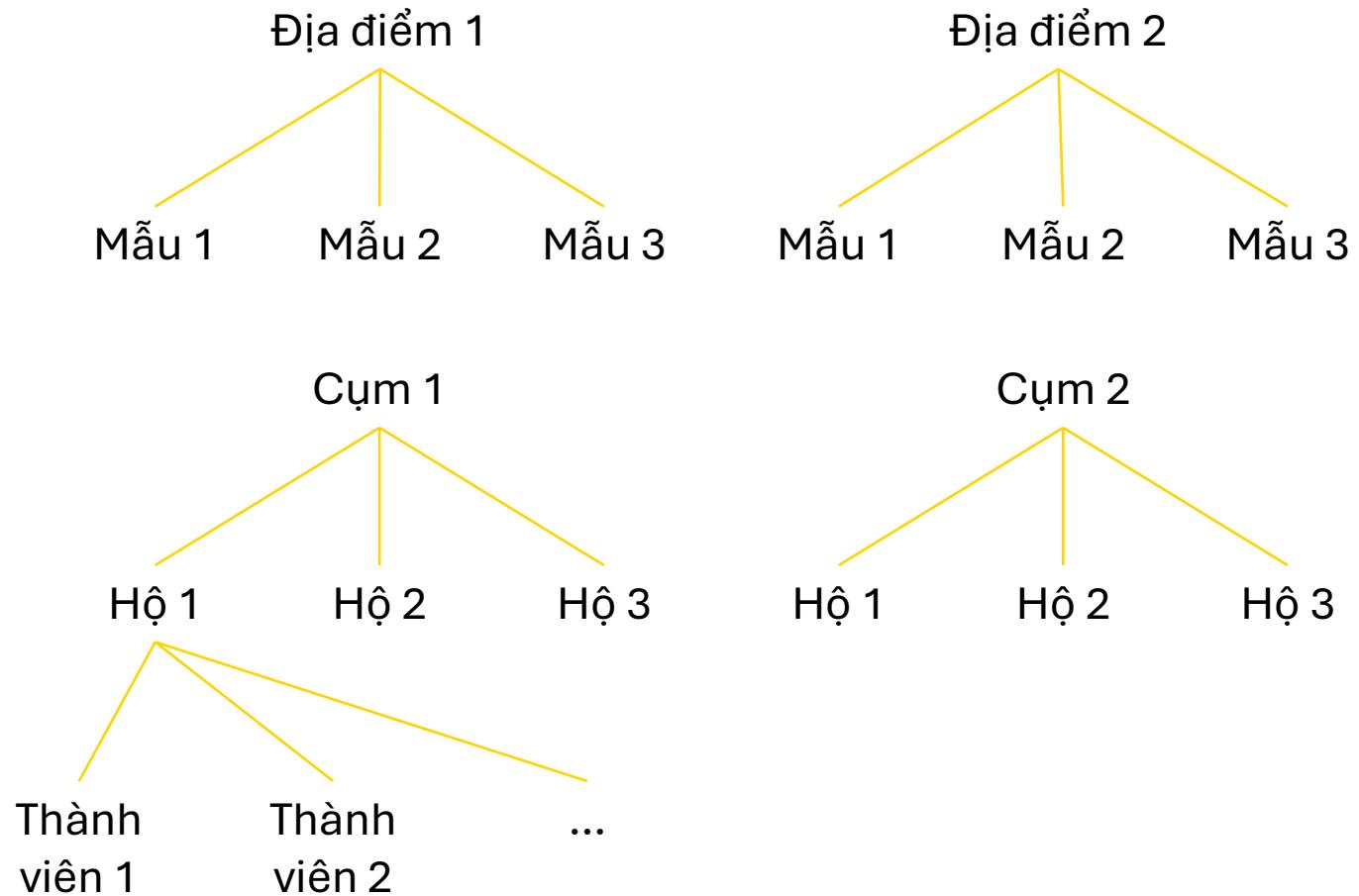


# Mô hình đa cấp/Mô hình ảnh hưởng hỗn hợp

## Multilevel model/Mixed effect model

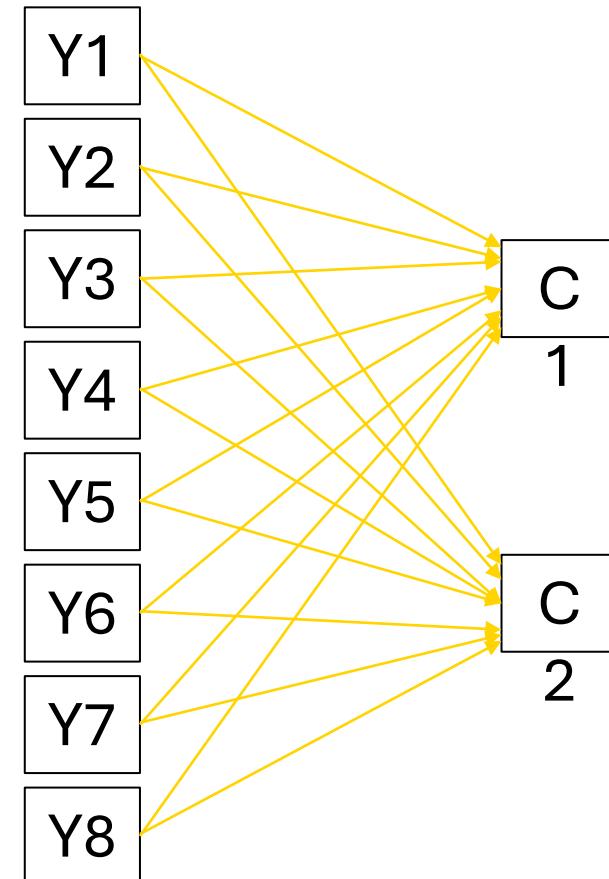


Image by [Chelsea Parlett-Pelleriti](#)



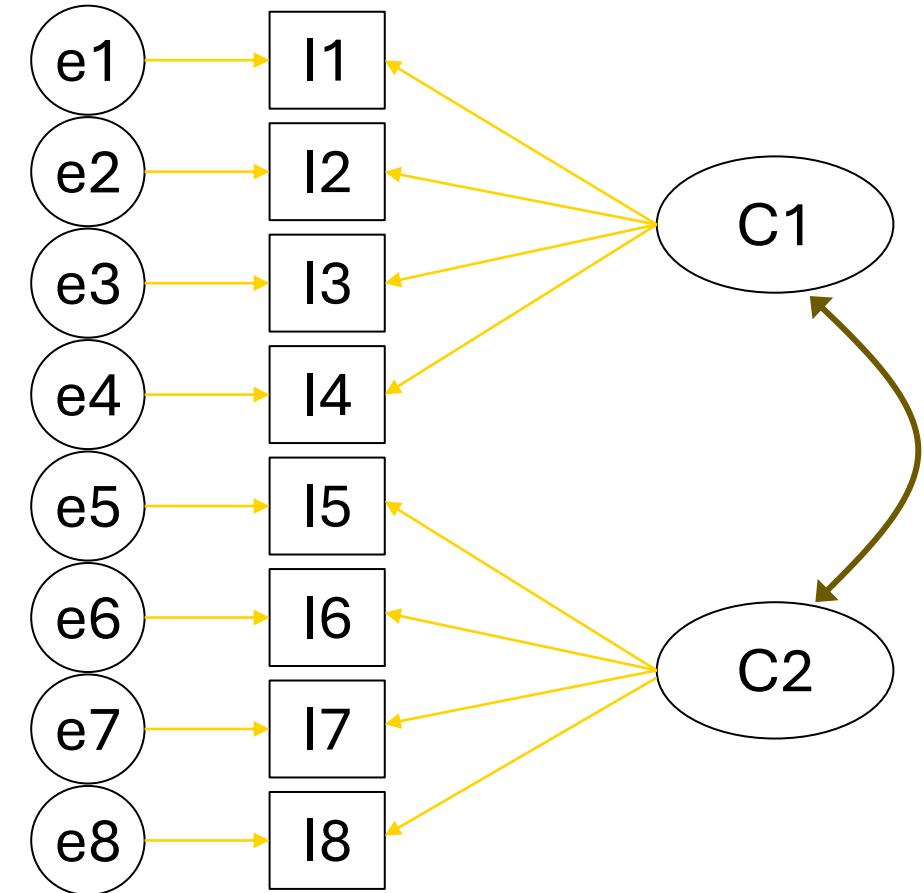
# Phân tích thành phần chính (PCA - Principal component analysis)

- 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)



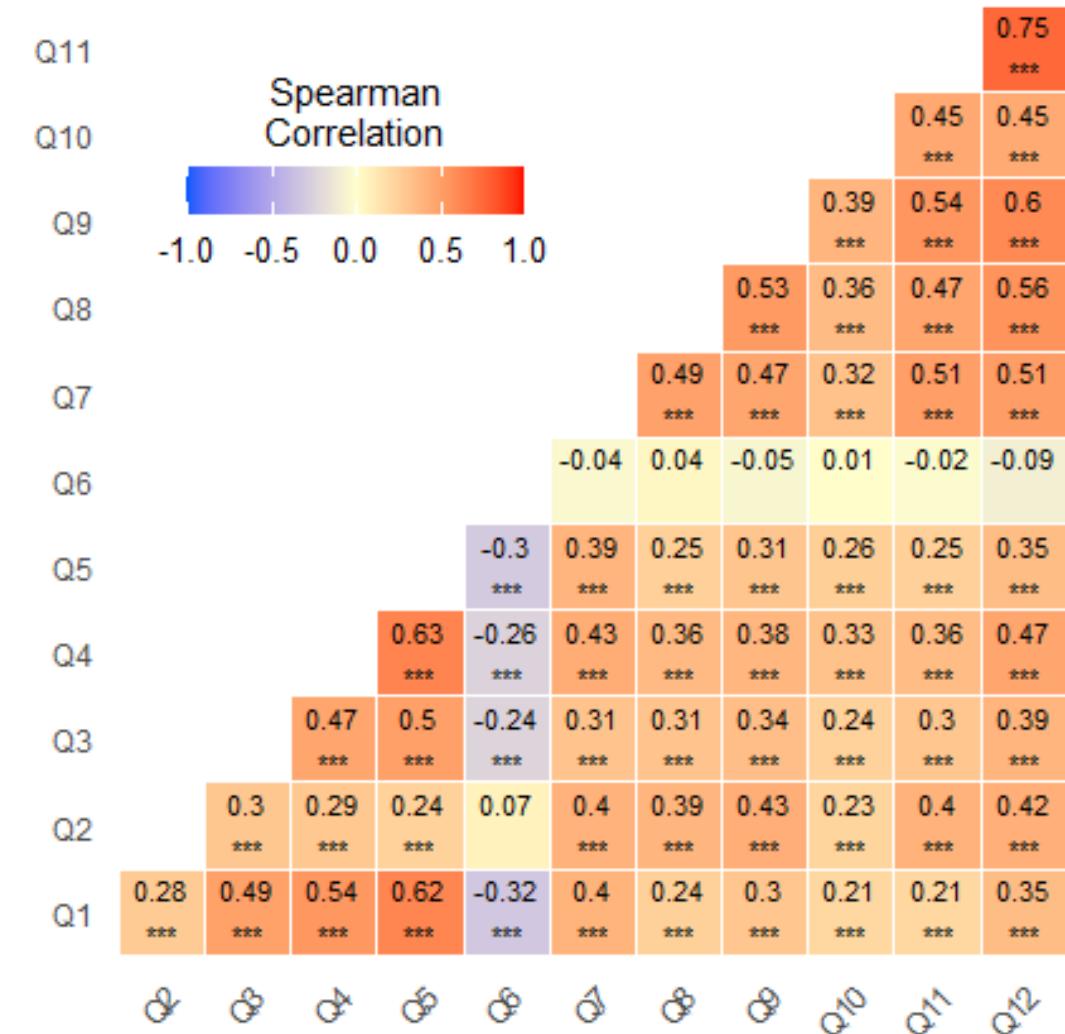
# Phân tích nhân tố (Factor analysis)

- 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)



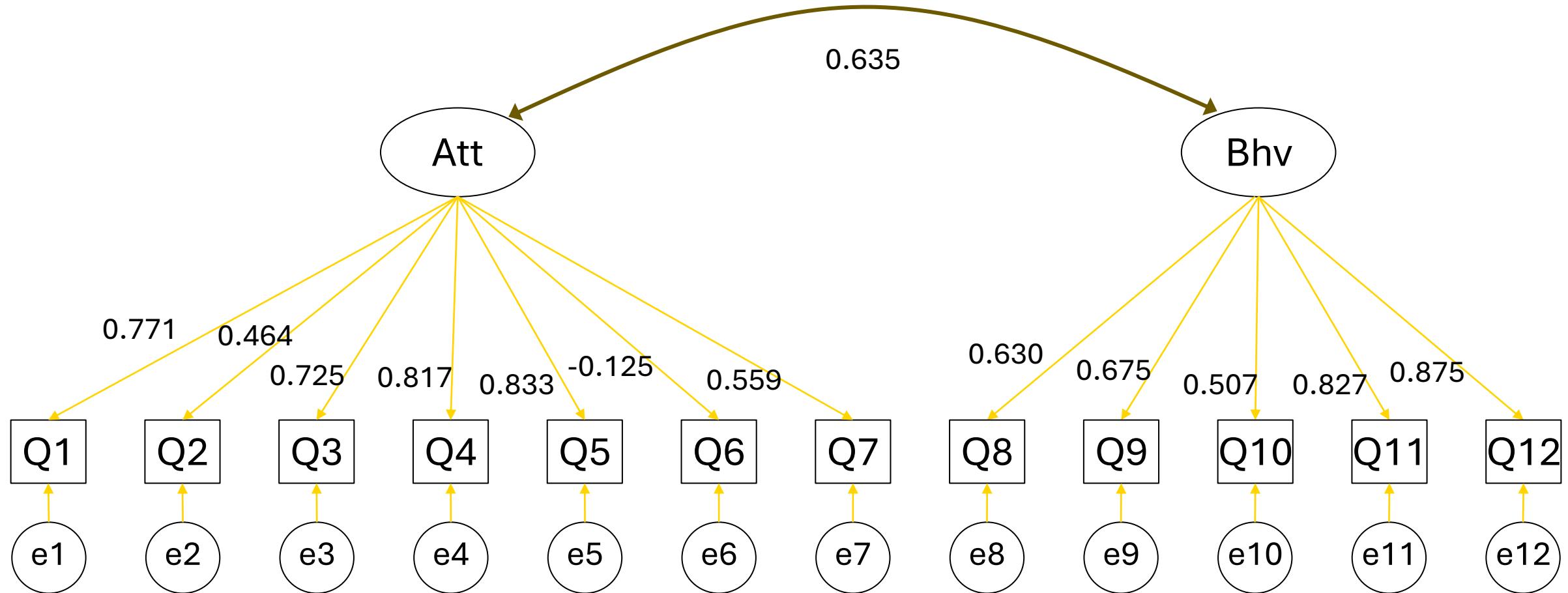
# Phân tích nhân tố khẳng định (CFA)

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.



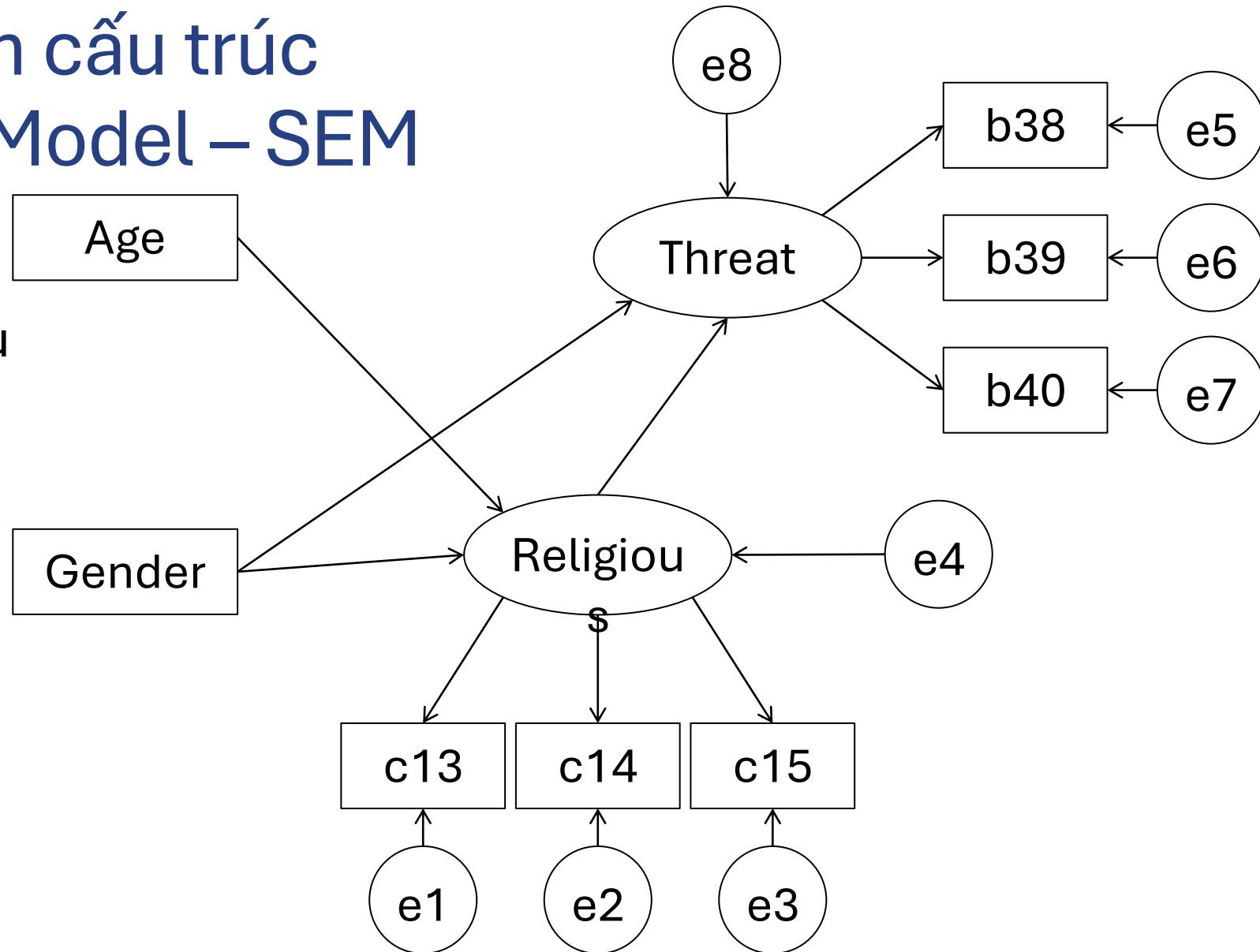
12

# Phân tích nhân tố khẳng định (CFA)



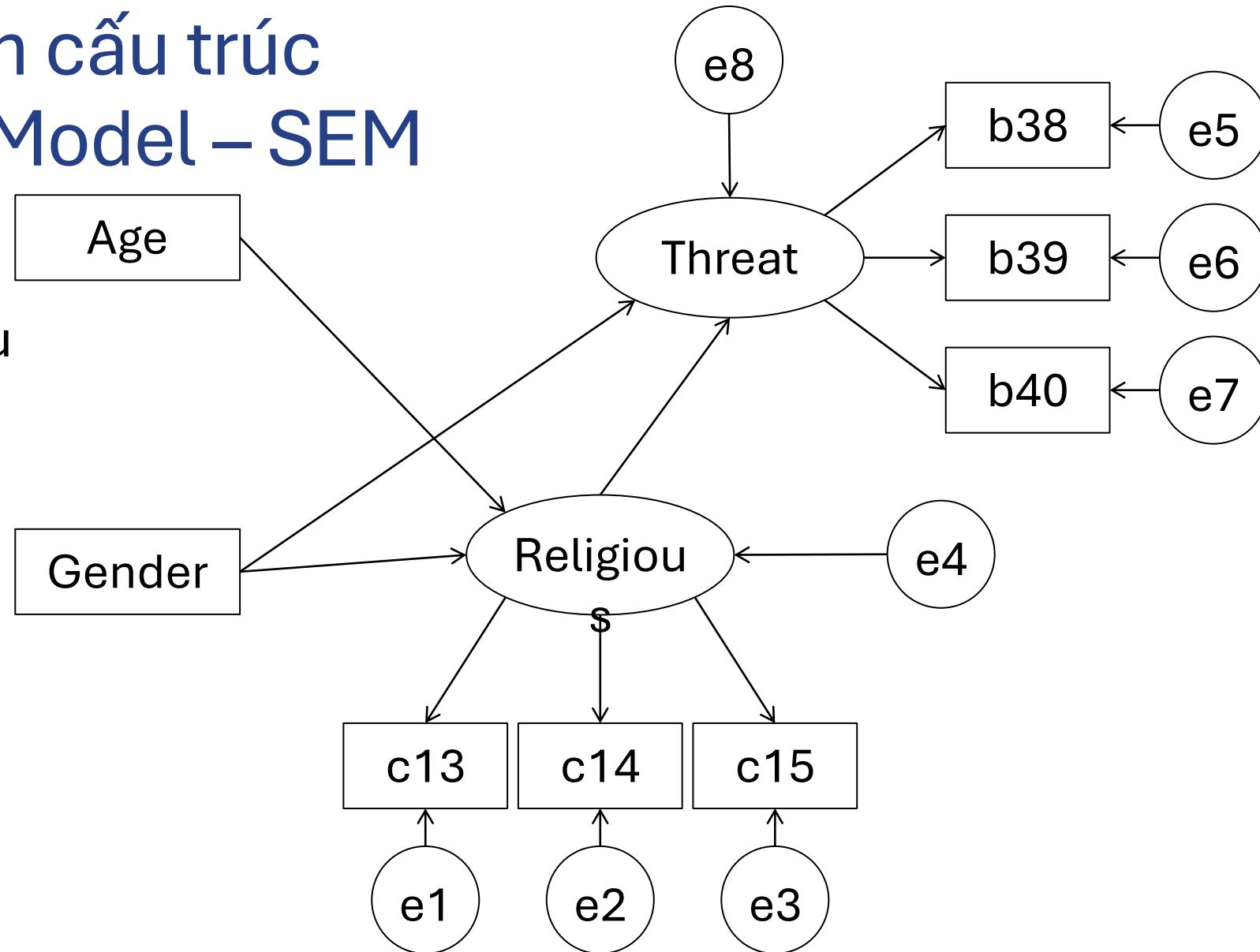
# Mô hình phương trình cấu trúc Structural Equation Model – SEM

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



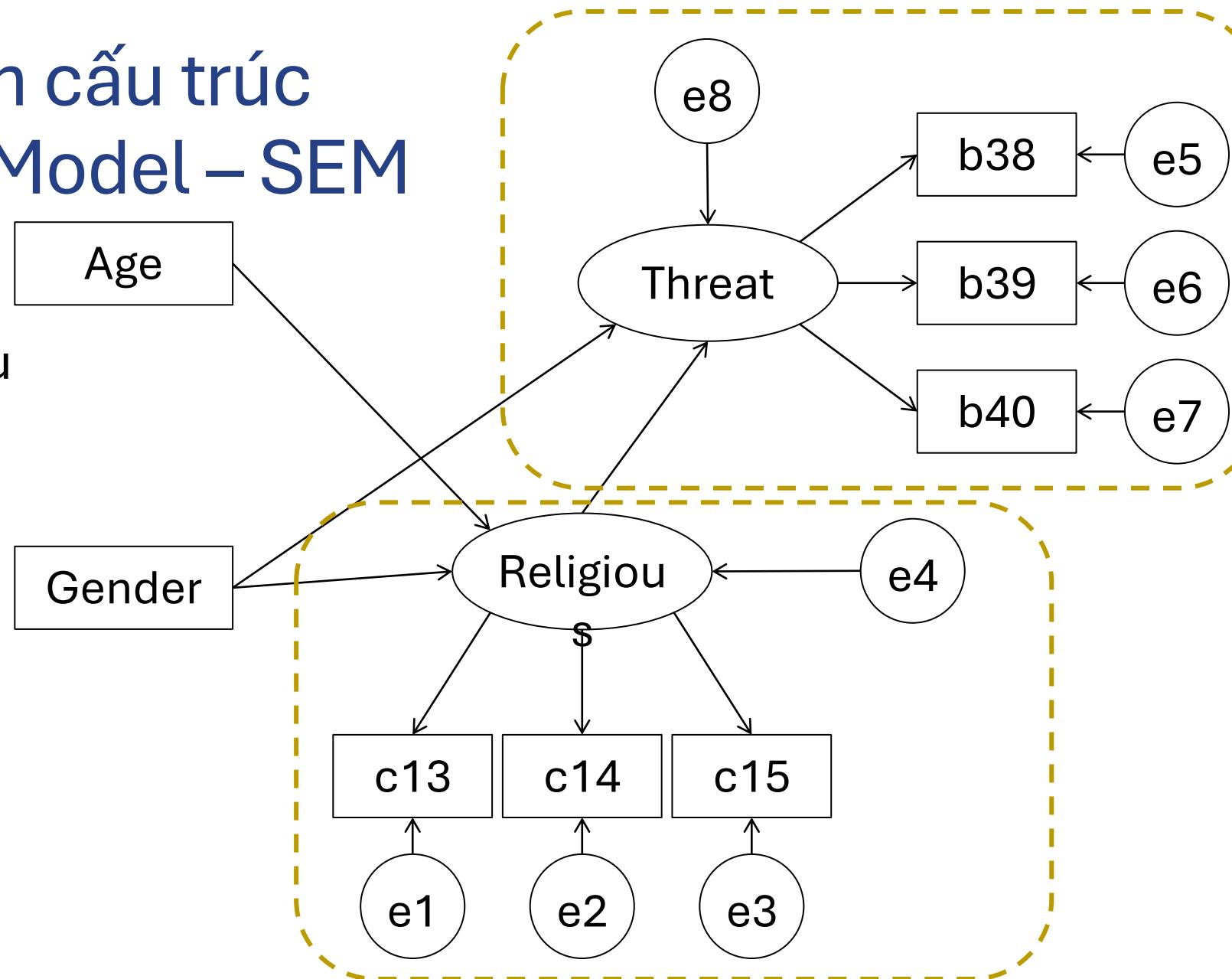
# Mô hình phương trình cấu trúc Structural Equation Model – SEM

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



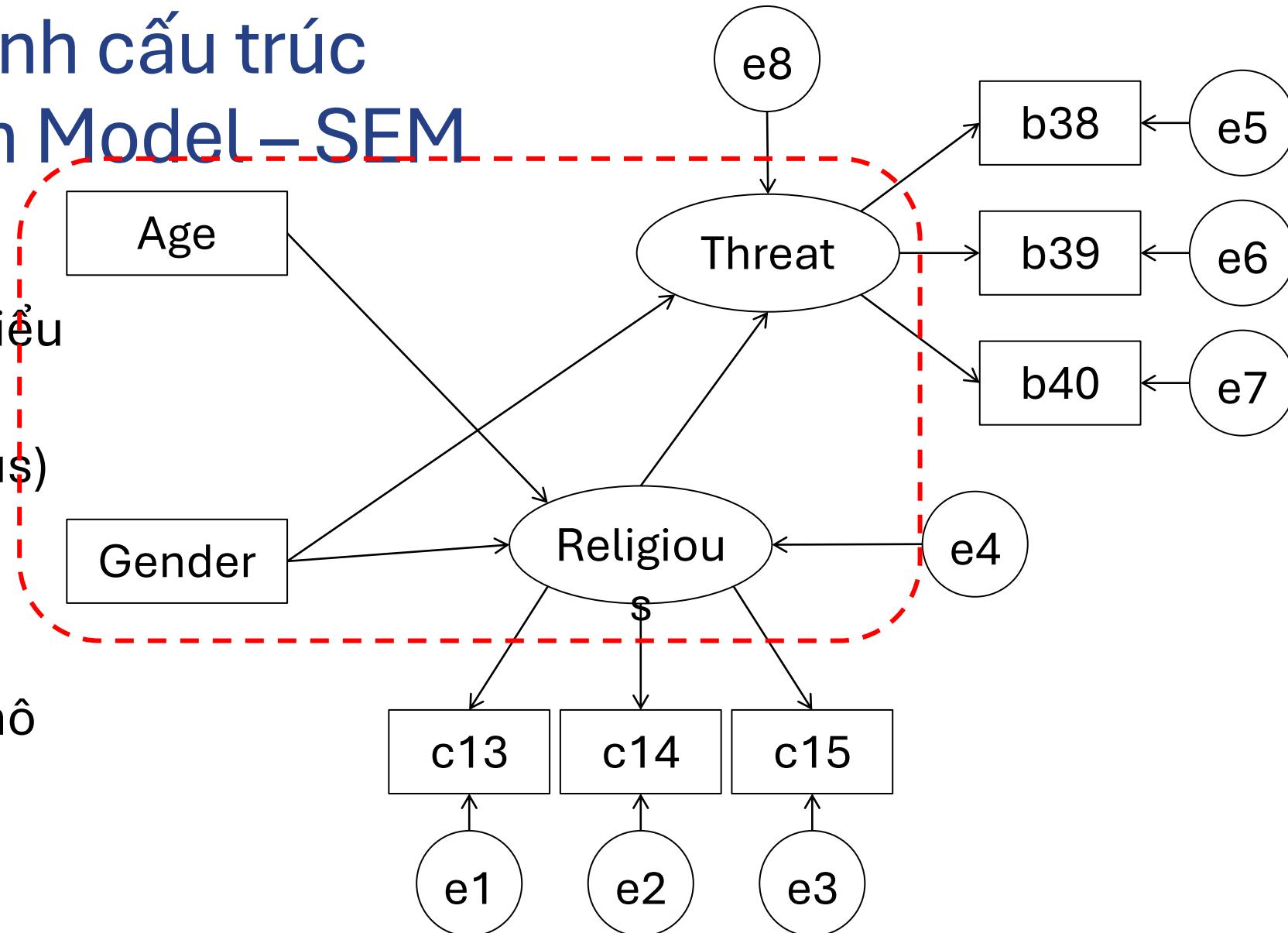
# Mô hình phương trình cấu trúc Structural Equation Model – SEM

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



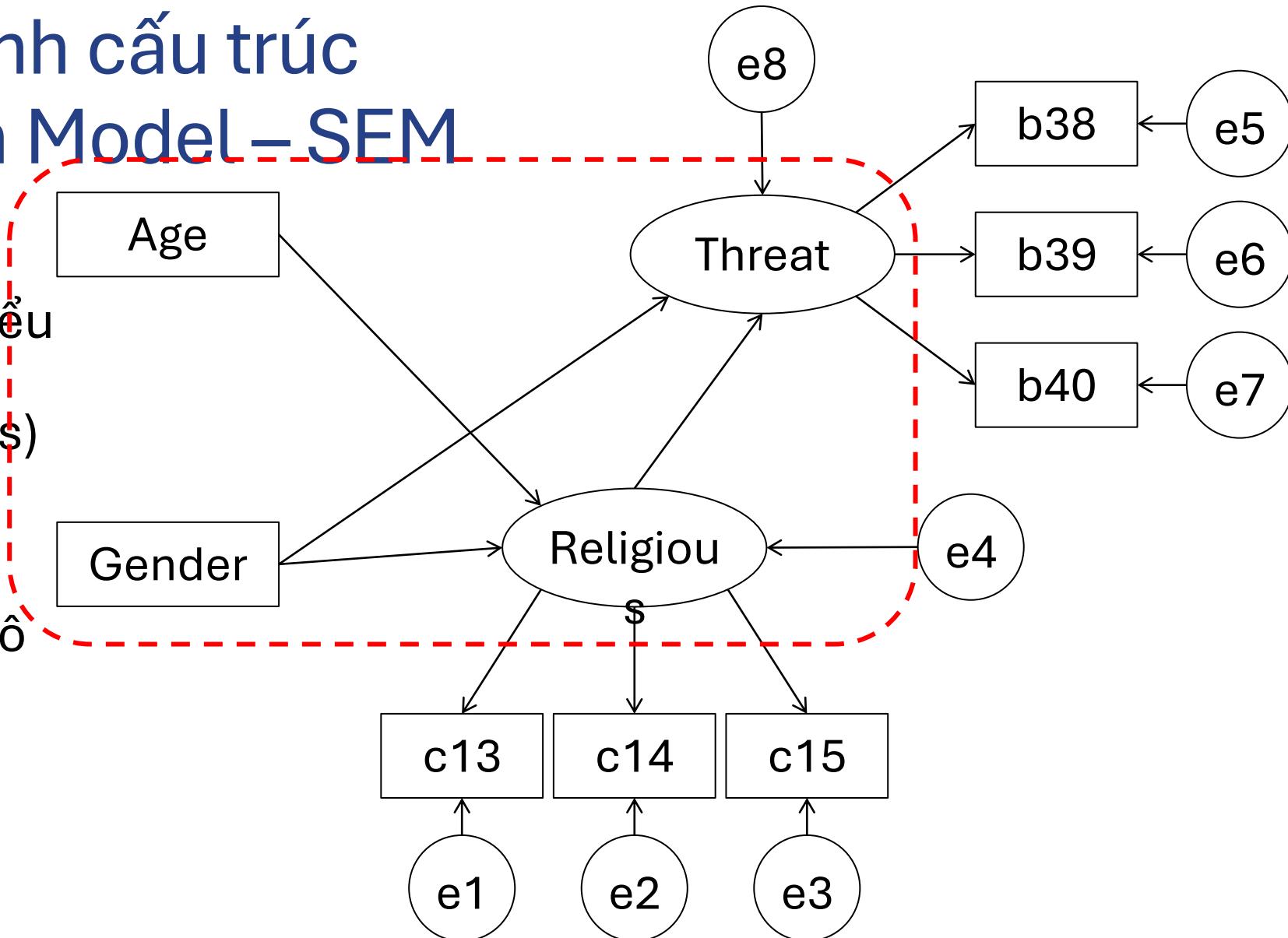
# Mô hình phương trình cấu trúc Structural Equation Model - SEM

- 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)



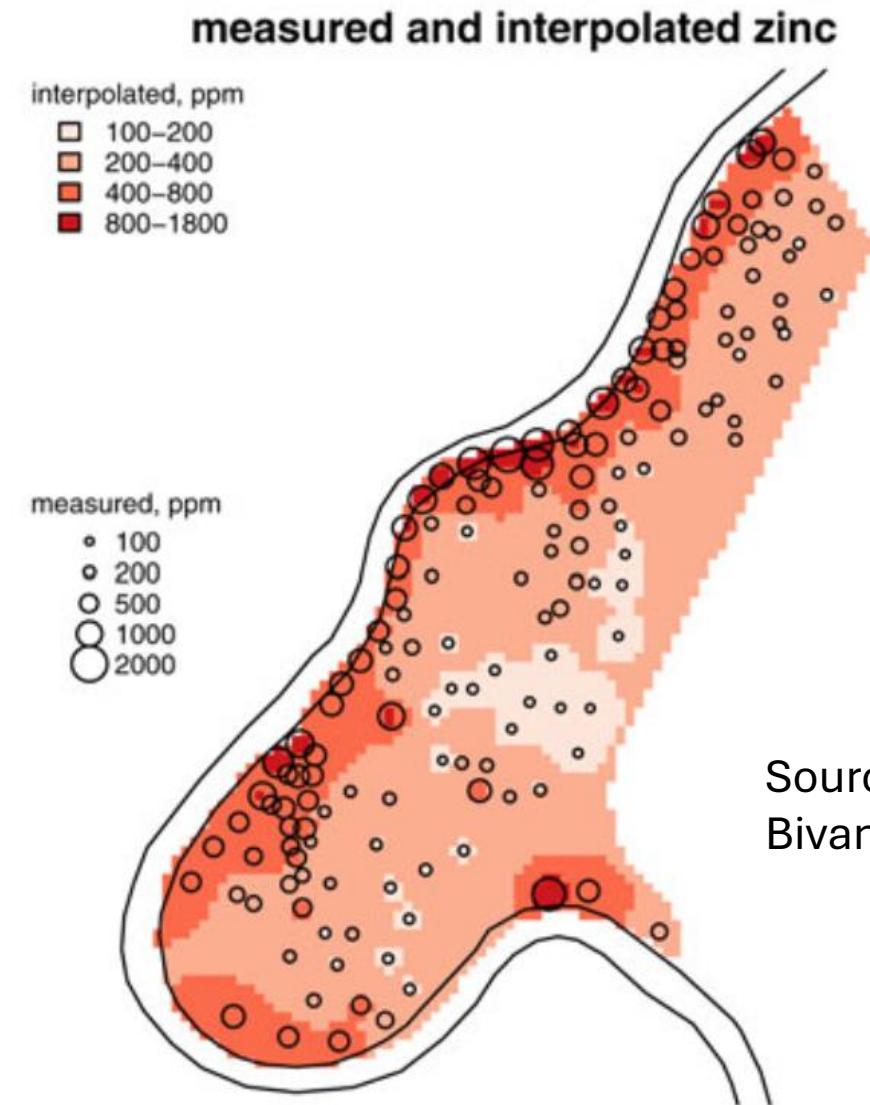
# Mô hình phương trình cấu trúc Structural Equation Model - SEM

- 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

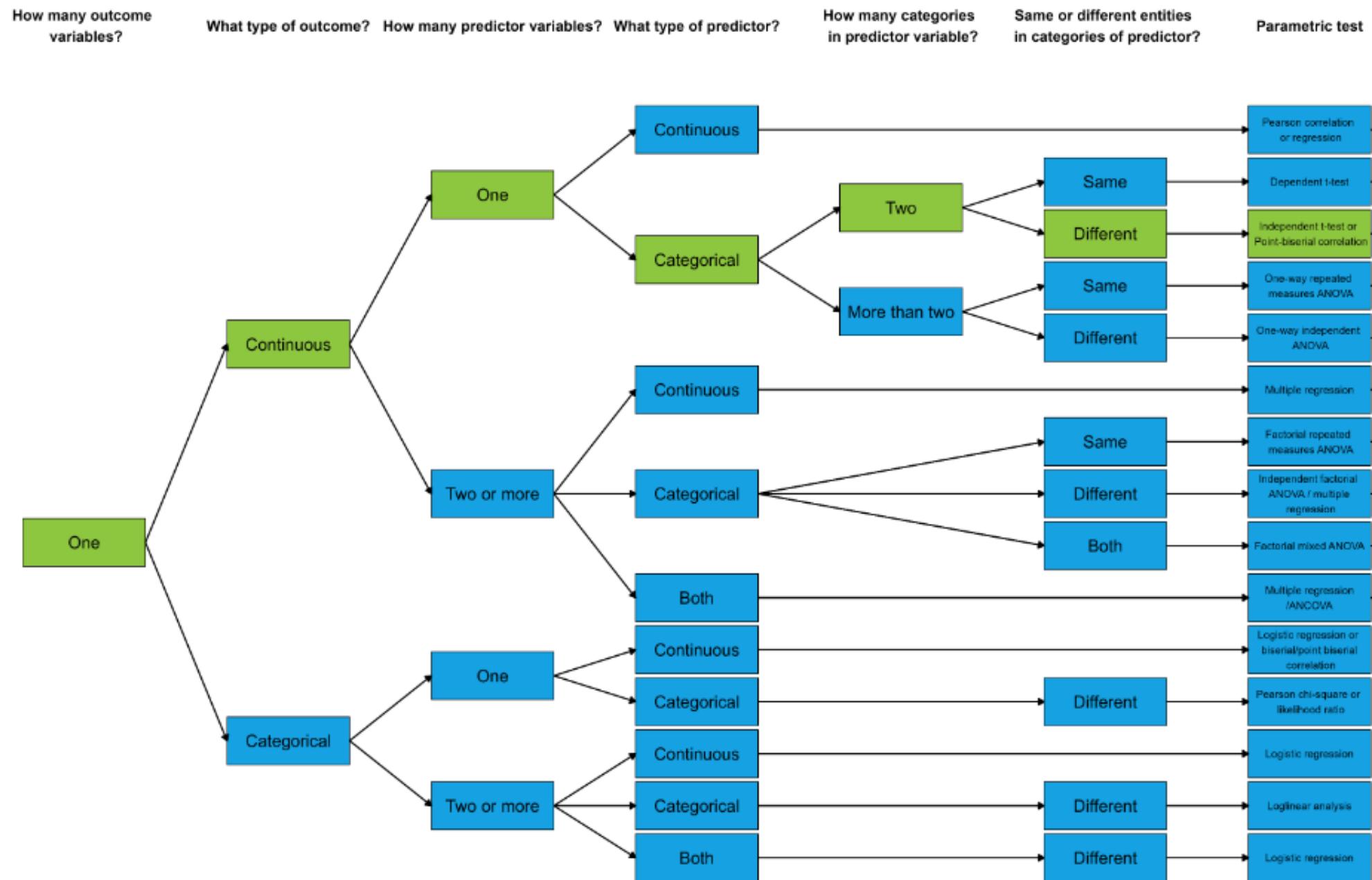
- Tương quan không gian
- GIS



Source:  
Bivand (2013)

# Lựa chọn phương pháp

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



# Công bố

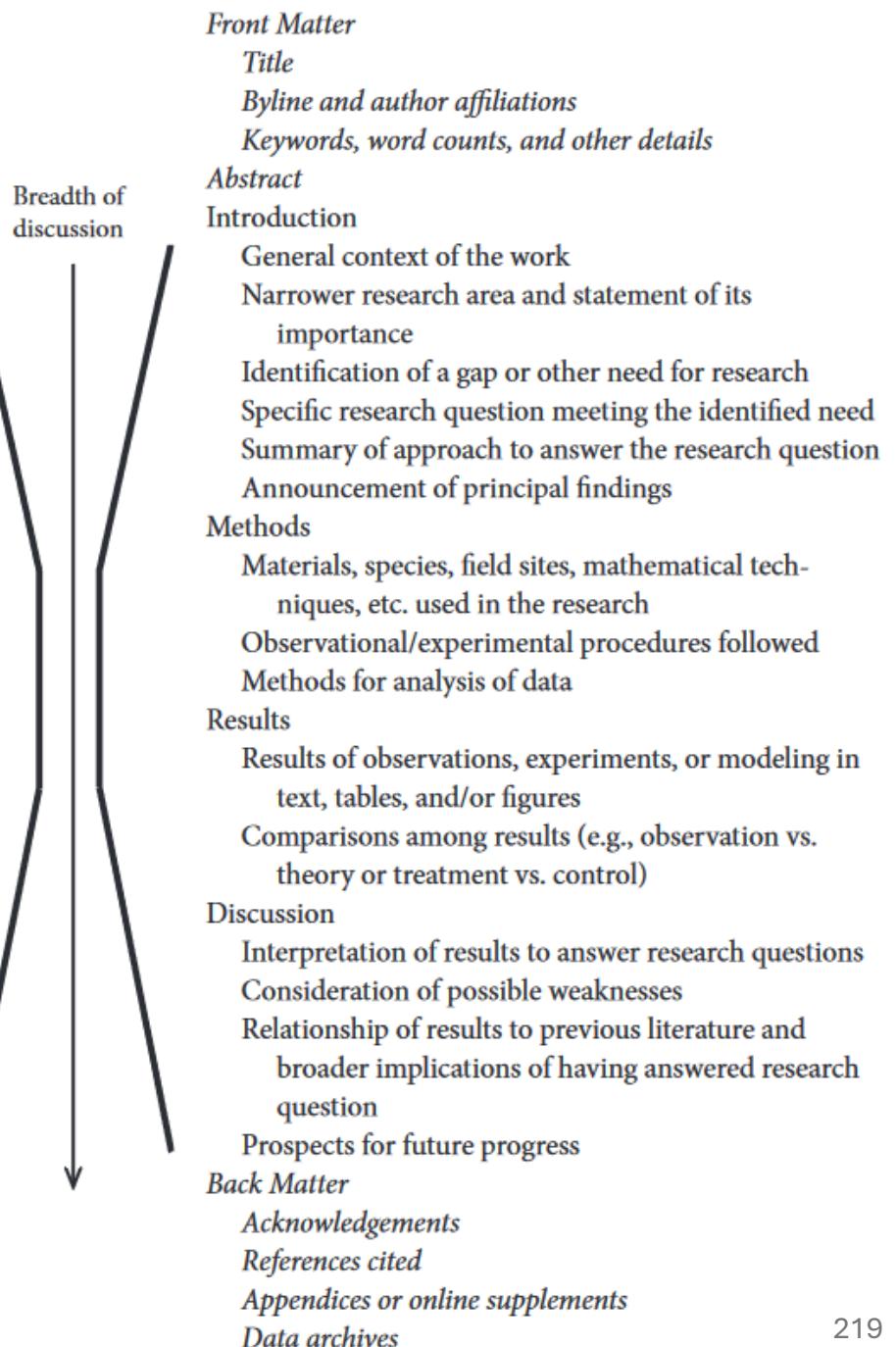
# Các loại tài liệu

- Sách chuyên khảo
- Sách tổng hợp/chương sách
- Bài báo (có/không qua phản biện đồng cấp)
- Bài hội nghị
- Bản thảo tiền xuất bản (preprint)
- Luận án/luận văn
- Báo cáo tư vấn
- Văn bản luật

# Bài báo

- Cấu trúc một bài báo - **IMRaD**
  - Abstract
  - Introduction
  - Methods
  - Results
  - Discussion
  - (Conclusion)

“It is impossible to be a good writer without being a good reader first”



# Đặt vấn đề - Introduction

- Bối cảnh chung của nghiên cứu
- Thu hẹp lại thành chủ đề nghiên cứu và tại sao nó quan trọng
- Xác định khoảng trống trong các nghiên cứu trước đây và tại sao nghiên cứu hiện tại là cần thiết
- Nêu cụ thể câu hỏi nghiên cứu/giả thuyết nghiên cứu
- (Tóm tắt định hướng hoặc phương pháp nghiên cứu)
- (Giới thiệu kết quả chính nhất)

# Phương pháp - Methods

- Nêu cụ thể và chi tiết vật liệu, địa điểm lấy mẫu được sử dụng
- Quy trình thu thập mẫu, hoặc thực nghiệm
- Phương pháp phân tích dữ liệu

! Nếu quá dài có thể để một phần chi tiết xuống phần phụ lục

# Kết quả - Results

- Kết quả của bài báo
- So sánh kết quả giữa các lần thực nghiệm hoặc các nhóm đối tượng
- Hình và bảng cần xếp theo trật tự kết quả đưa ra
- Đưa ra cả các kết quả âm tính

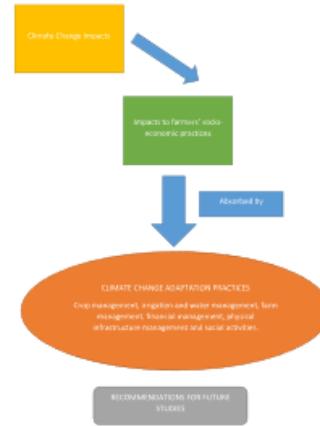
# Thảo luận – Thảo luận

- Phân tích kết quả thu được để trả lời cho câu hỏi nghiên cứu
- Xem xét các điểm yếu có thể
- Liên hệ đến các nghiên cứu trước đây
- Những ý nghĩa ứng dụng mở rộng của kết quả nghiên cứu hiện tại cho lĩnh vực/ngành
- Định hướng cho các nghiên cứu sau

# Tóm tắt - Abstract

- Tóm tắt:
  - Câu hỏi nghiên cứu
  - Phương pháp sử dụng
  - Kết quả chính
  - Kết luận chính thu được từ kết quả
- Nên viết cuối cùng
- Không nên
  - Quá dài (một số tạp chí giới hạn ở 100 từ)
  - Chưa tài liệu tham khảo
  - Có từ viết tắt
  - Có hình hay bảng
- Tóm tắt bằng hình ảnh (Graphical abstract)

## GRAPHICAL ABSTRACT



## ABSTRACT

Climate change in Asia is affecting farmers' daily routines. Much of the focus surrounding climate change has targeted the economic and environmental repercussions on farming. Few systematic reviews have been carried out on the social impacts of climate change among farmers in Asia. The present article set out to analyse the existing literature on Asian farmers' adaptation practices towards the impacts of climate change. Guided by the PRISMA Statement (Preferred Reporting Items for Systematic reviews and Meta-Analyses) review method, a systematic review of the Scopus and Web of Science databases identified 38 related studies. Further review of these articles resulted in six main themes – crop management, irrigation and water management, farm management, financial management, physical infrastructure management and social activities. These six themes further produced a total of 35 sub-themes. Several recommendations are highlighted related to conducting more qualitative studies, to have specific and a standard systematic review method for guide research synthesis in context of climate change adaptation and to practice complimentary searching techniques such as citation tracking, reference searching, snowballing and contacting experts.

© 2018 Elsevier B.V. All rights reserved.

# Định dạng văn bản, tài liệu tham khảo

- Đọc và theo chặt chẽ quy định của tạp chí mình hướng tới
- Sử dụng phần mềm hỗ trợ trích dẫn (Zotero, Mendeley, Endnote) để đảm bảo tính chính xác, thống nhất của tài liệu tham khảo
- Zotero có hỗ trợ định dạng tài liệu tham khảo của rất nhiều tạp chí

# Đạo văn

- Sử dụng lại ý tưởng, đoạn viết, hình ảnh của người khác trong tài liệu của mình mà không trích dẫn đúng chuẩn
- Trích dẫn nguyên văn không dùng dấu “...”
- Tự đạo văn?
- Sử dụng hình ảnh khung lý thuyết, biểu đồ quy trình từ sách giáo khoa, bài báo khác?
- Dịch nguyên văn từ tài liệu đã xuất bản từ Tiếng anh sang tiếng Việt hoặc ngược lại?

# Tạp chí

- Các công cụ lựa chọn tạp chí

[https://www.myendnoteweb.com/EndNoteWeb.html?func=journalRecom  
mend&](https://www.myendnoteweb.com/EndNoteWeb.html?func=journalRecommend&)

<https://journalsuggester.springer.com/>

<https://journalfinder.elsevier.com/>

<https://journalfinder.wiley.com/>

- Xếp hạng tạp chí

- Được “indexed” bởi Web of Science, Scopus

<https://mjl.clarivate.com/search-results>

<https://www.scopus.com/sources>

- <https://www.scimagojr.com/>

# Thư gửi tạp chí (cover letter)

- Văn phong lịch sự (formal)
- Bắt đầu bằng việc giới thiệu tên bài báo, nhóm tác giả, và thể loại bài báo (review, original research)
- Mô tả lý do thực hiện nghiên cứu và kết luận chính
- Giải thích tại sao bài báo này phù hợp với tạp chí mình đang gửi đến

# Quá trình phản biện

- Biên tập (editor) và phản biện đồng cấp (reviewer)
- Các kết quả có thể nhận
  - Chấp nhận đăng như hiện tại
  - Sửa chữa nhỏ
  - Sửa chữa lớn và sẽ qua phản biện lại
  - Từ chối nhưng mời nộp lại sau khi sửa
  - Từ chối

# Trả lời phản biện

- **KHÔNG** trả lời phản biện ngay lập tức
- Trả lời tất cả các câu hỏi/nhận xét của phản biện và gửi lại cho tạp chí/chủ biên cùng bản thảo đã sửa với chú thích vị trí sửa

-·Why·the·reclassification·of·water·use·(lines130-132)?·What·purpose·does·it·serve·in·the·analysis?¶

Alternative·water·use·in·our·study·was·obtained·by·asking·the·participants·to·indicate·which·source·of·water·(besides·piped·water)·they·used·for·which·purposes.·With·4·water·sources·and·12·purposes,·the·raw·data·contain·48·binary·variables.·To·reduce·the·dimensions·of·the·predictor·matrix,·and·to·obtain·meaningful·interpretation·for·alternative·water·sources,·we·have·created·a·nominal·variable·with·four·levels·(no·use,·only·indoor,·only·outdoor,**both·indoor·and·outdoor**)·for·each·type·of·water.·We·have·included·this·explanation·in·the·revised·text·at·line·137-143.¶

- Khi không đồng ý với phản biện giải thích lý do một cách lịch sự, tôn trọng
- Nên cố gắng sửa theo nhận xét hợp lý của phản biện kể cả khi quyết định đổi tạp chí

# Nỗi sợ trang giấy trống (Fear of blank paper)

- Viết trong lúc đọc!

# Nỗi sợ trang giấy trống (Fear of blank paper)

- Viết trong lúc đọc!
- Bộ từ (wordstack)

Figure 7.1. My wordstack for this chapter.

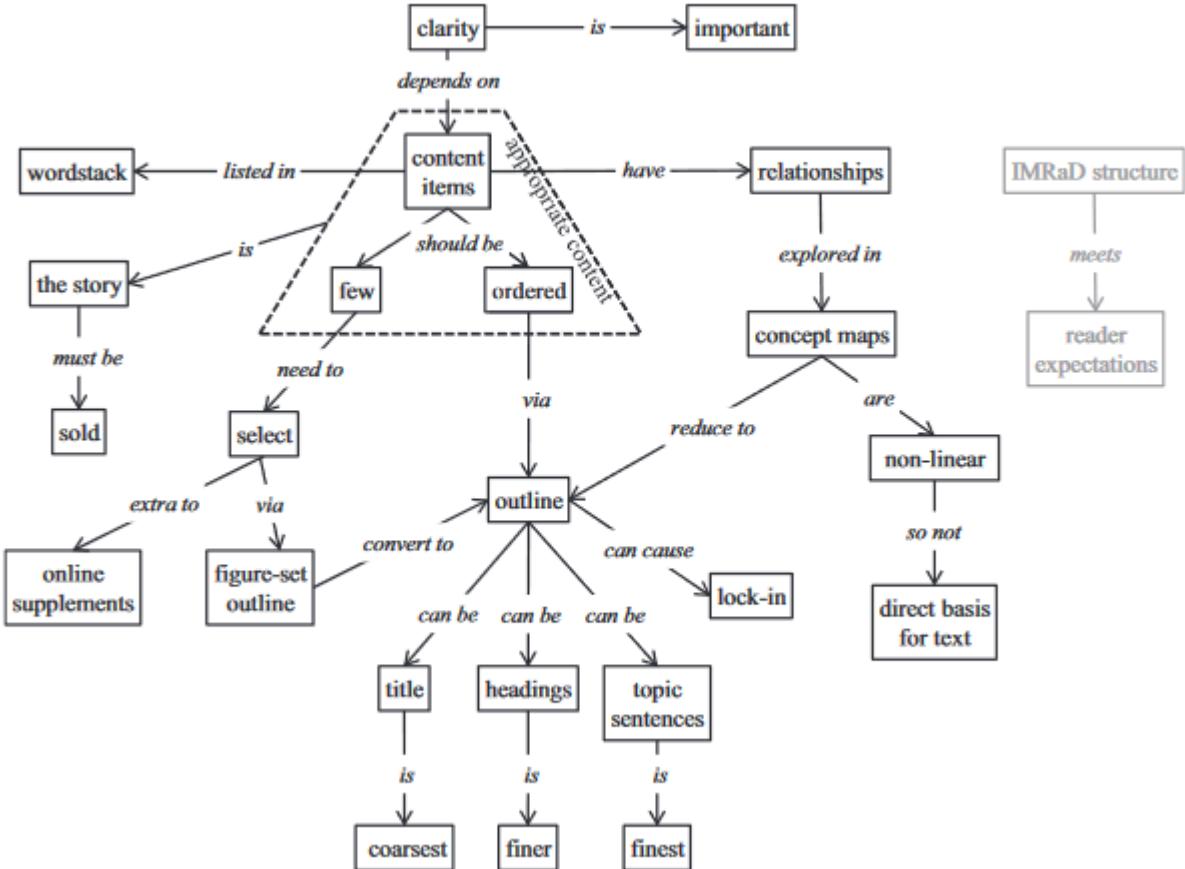
cohesive story . . . “thesis”  
?titles - shortest summary of story  
outline  
story about how I thought I didn’t do outlines  
when to outline – when story is ready; vs. to find the story  
head and subheads as coarse outline  
topic sentences as detailed outline  
what goes in and doesn’t, and what order  
*not everything you did*  
*not in the order you did it*  
concept map: non-linear  
intermediate step to outline  
despite HTML, basic form still linear  
wordstack/idea pile  
*first step*  
*accumulate pre-writing*  
?Cahill – “pitch”  
fail to consider story: leads to overlong, poorly organized MS  
IMRaD structure  
online supplements  
“retroactive storytelling”  
avoid lock-in  
simple clear direction  
selling the story  
*not “I was interested in”*  
*not “no studies have examined”*  
*not “increase our understanding of”*  
contrast “writing backwards” (Magnusson 1996)  
figure shuffling

# Nỗi sợ trang giấy trống (Fear of blank paper)

- Viết trong lúc đọc!

- Bộ từ (wordstack)

- Bản đồ ý tưởng



# Các lưu ý khác

- Tập trung vào thông điệp chính/kết luận chính của bài báo
- Cấu trúc song song
- Ghi nhận rõ ràng trong nội dung các hạn chế của nghiên cứu
- Đối với văn phong, ưu tiên rõ nghĩa dễ hiểu
- Luôn tự đọc và sửa bài báo trước
- Nhờ người không trong nhóm nghiên cứu đọc và sửa