Data

Recap

- Data science = solve scientific problems with data
- Can you identify the scientific problem?

Today's outline

- Research questions
- Population and sample
- Variables

Research questions

Research questions

- 相關性 (association)
- 因果關係 (causation)
- 預測 (prediction)

Association

- hours of study v.s. GPA
- medical treatment v.s. survival rate
- 心電圖 (electrocardiography) v.s. heart attack
- image v.s. object label
- etc.

# 48112

Causation

本頁刻意留白

# 48112

Association≠Causation

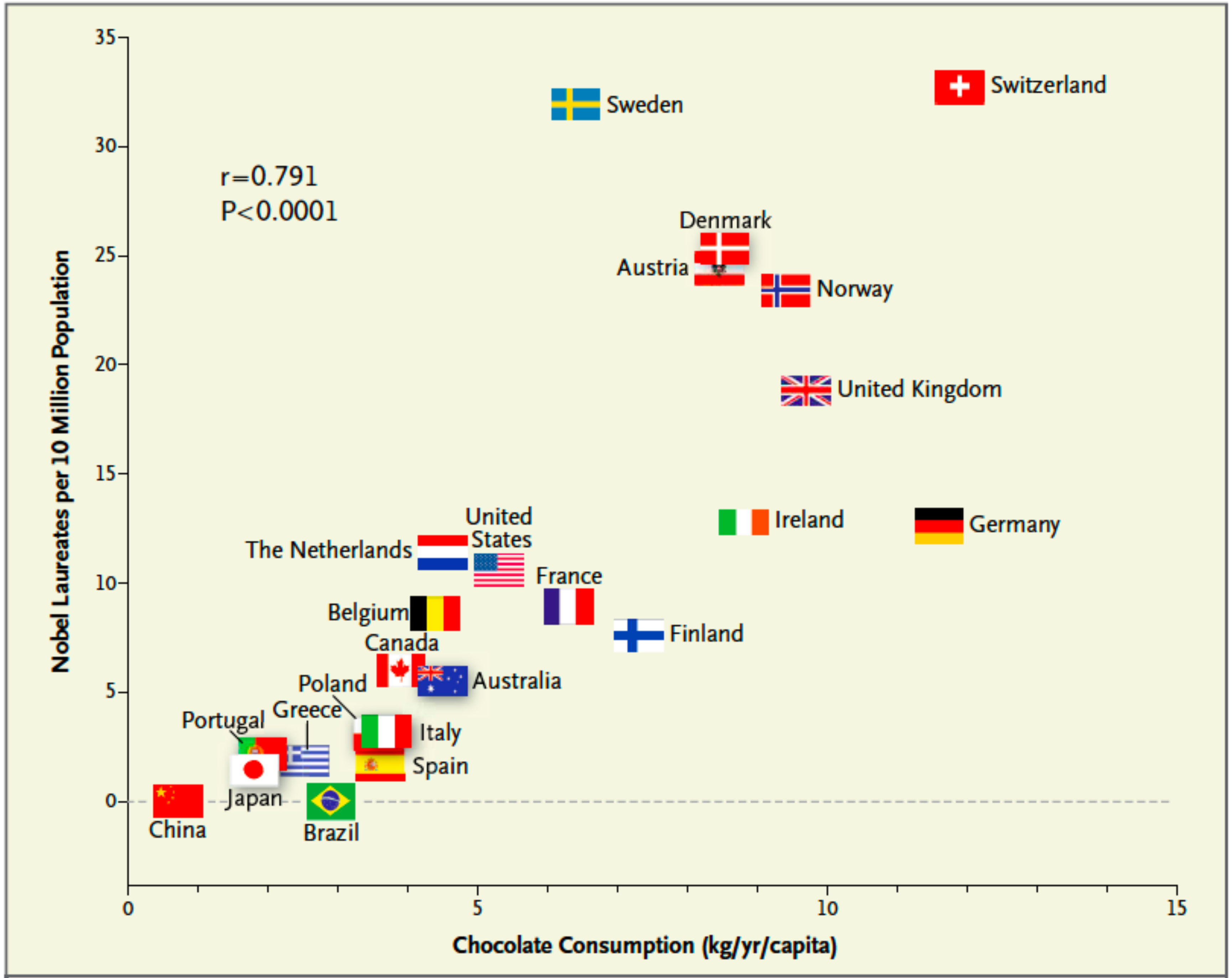
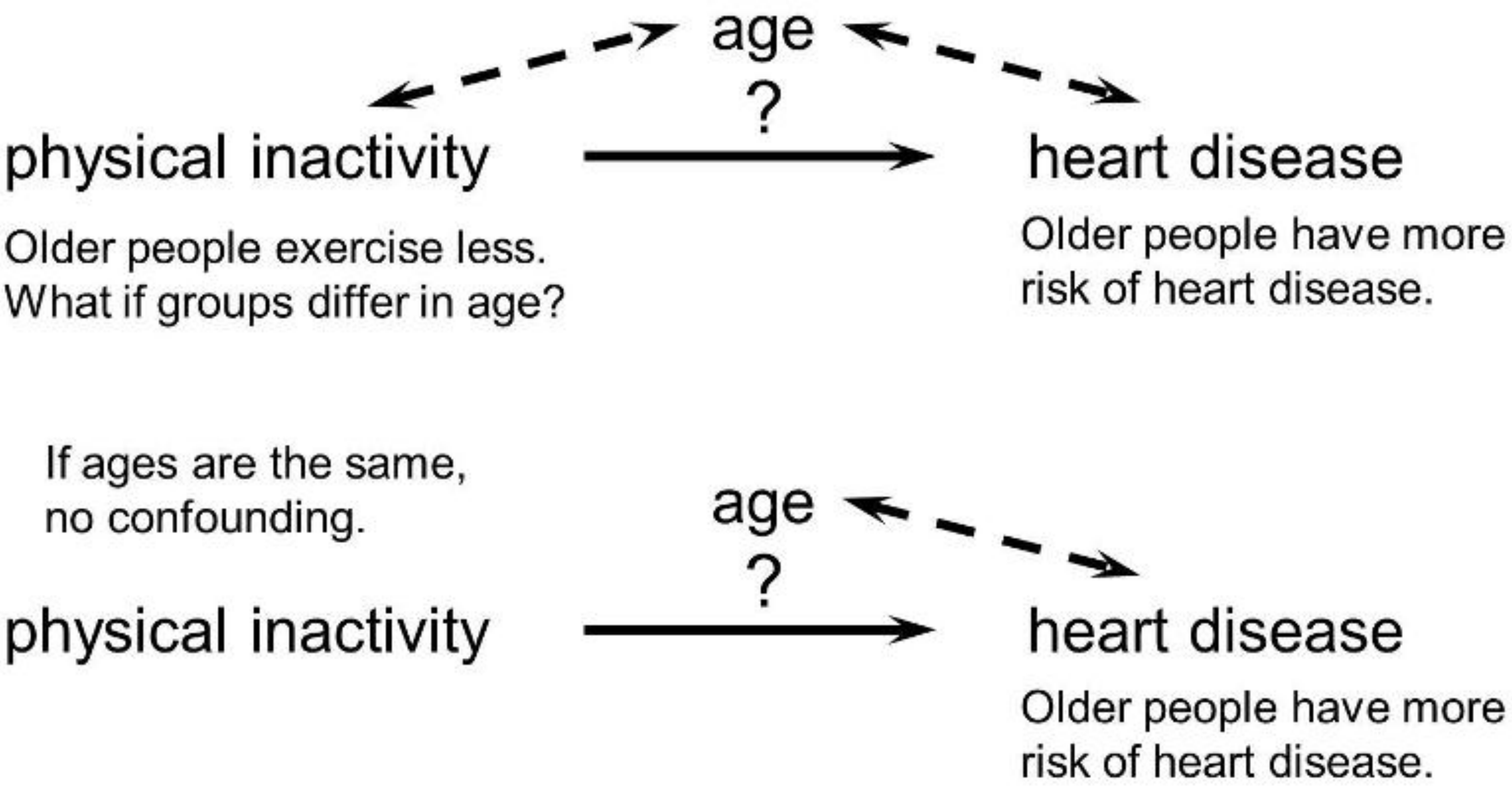


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Confounding



Older people exercise less.
What if groups differ in age?

Older people have more
risk of heart disease.

If ages are the same,
no confounding.

Older people have more
risk of heart disease.

Confounding (cont.)

- 味精有害健康嗎?
- Exclude confounding factors proper design of experiments

Prediction

- Predict the future behavior of a new observation based on the other variables (features), e.g.,
  - genes v.s. disease
  - 機台狀態 v.s. defect
  - weather prediction
  - 烘豆溫度曲線 v.s. 咖啡豆品質

Association vs prediction

- Association ⇒ prediction

- To Explain or to Predict?

Population and sample

Population

- 某個研究問題的所有研究對象稱為母體, e.g.,
  - 2020選舉勝負: 所有合格選民
  - image object detection/recognition: 所有images
  - 心電圖 v.s. 心臟病: 所有病人的心電圖
  - weather prediction: 空間中所有氣象變數

Sample

- 母體的任意子集合稱為樣本
- 理想中的樣本: 利用樣本得到的結果可**推廣**至母體結果 (generalization)
  - 如何收集到理想的樣本？

Sample

- Random sample
- Nonrandom sample
- Sample of interest

Sampling bias

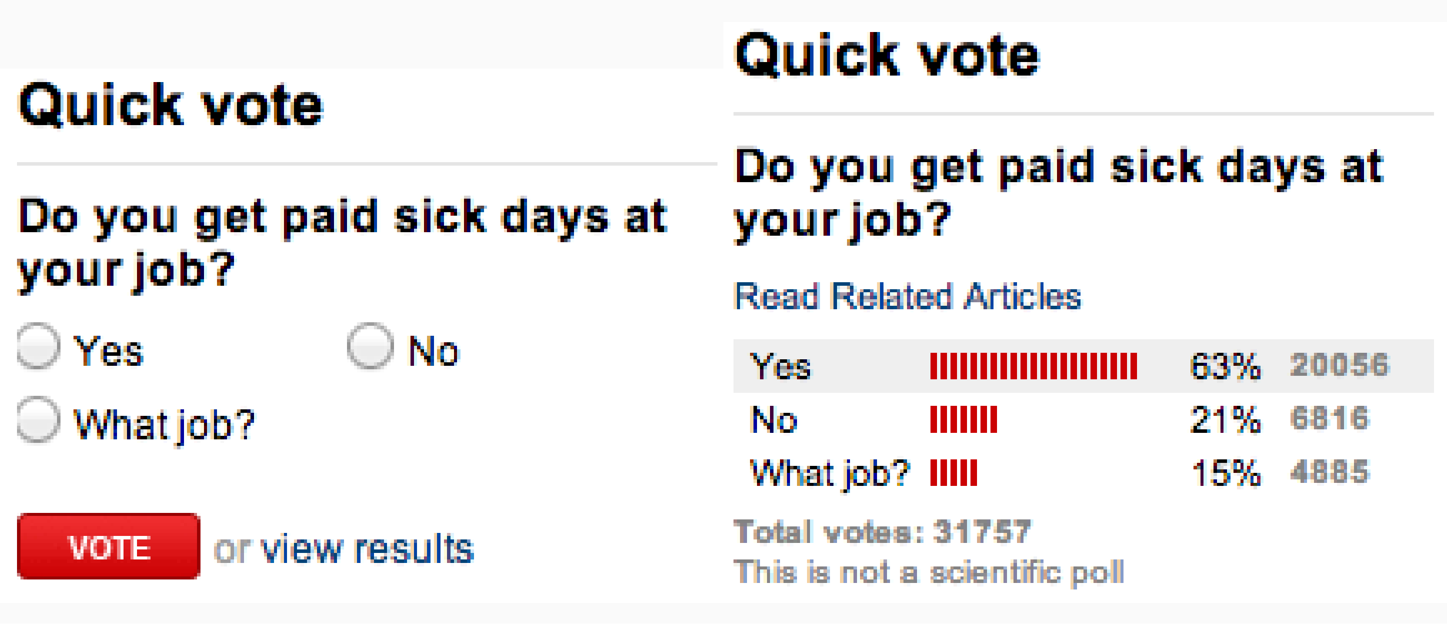
- Non–response: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.

- Convenience sample: Individuals who are easily accessible are more likely to be included in the sample.

# 48112

Sampling bias (cont.)

- Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue.



Quick vote

Do you get paid sick days at your job?

○ Yes ○ No

○ What job?

Quick vote

Do you get paid sick days at your job?

Read Related Articles

Yes	63%	20056
No	21%	6816
What job?	15%	4885

Total votes: 31757

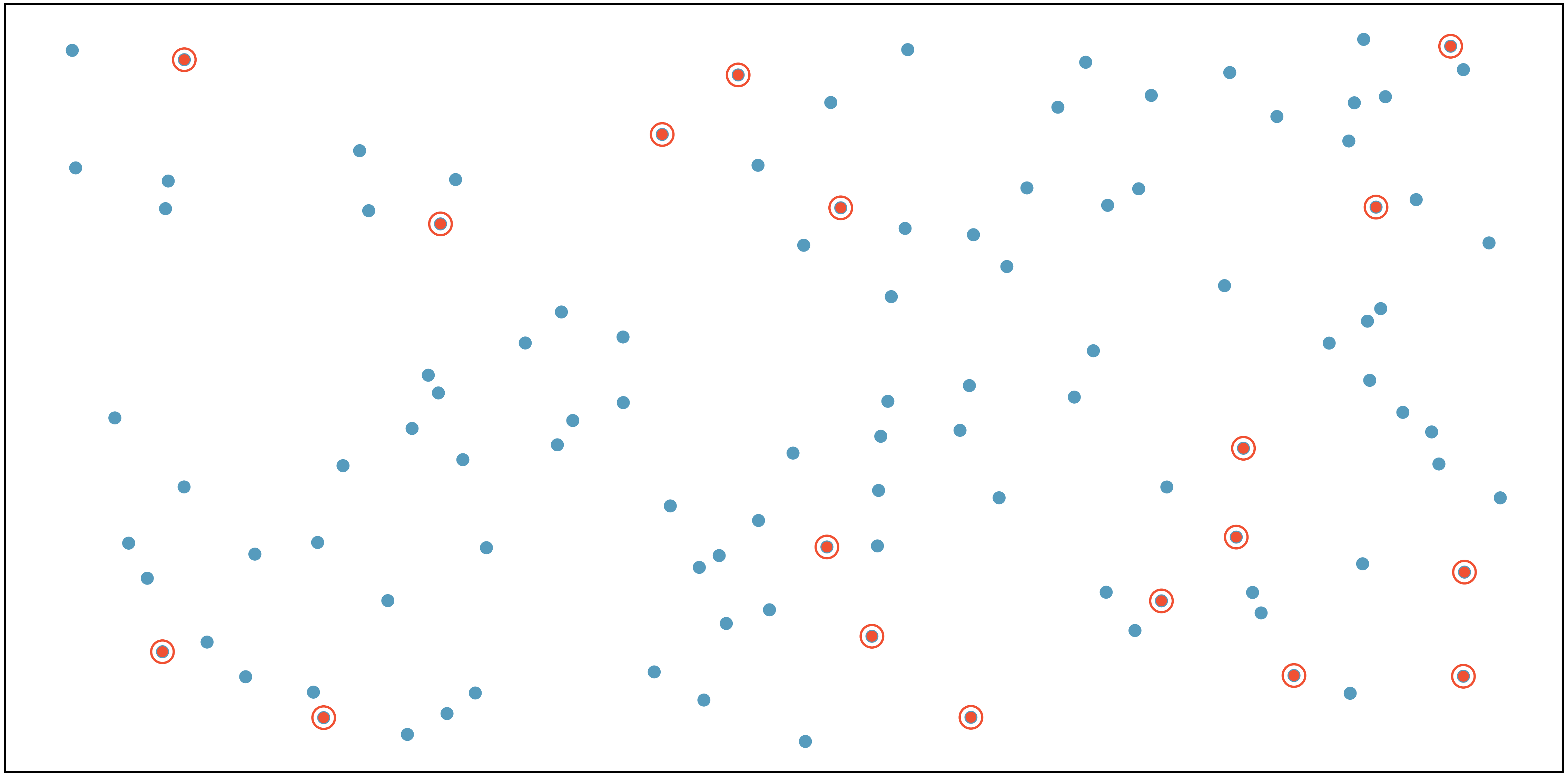
This is not a scientific poll

Random samples

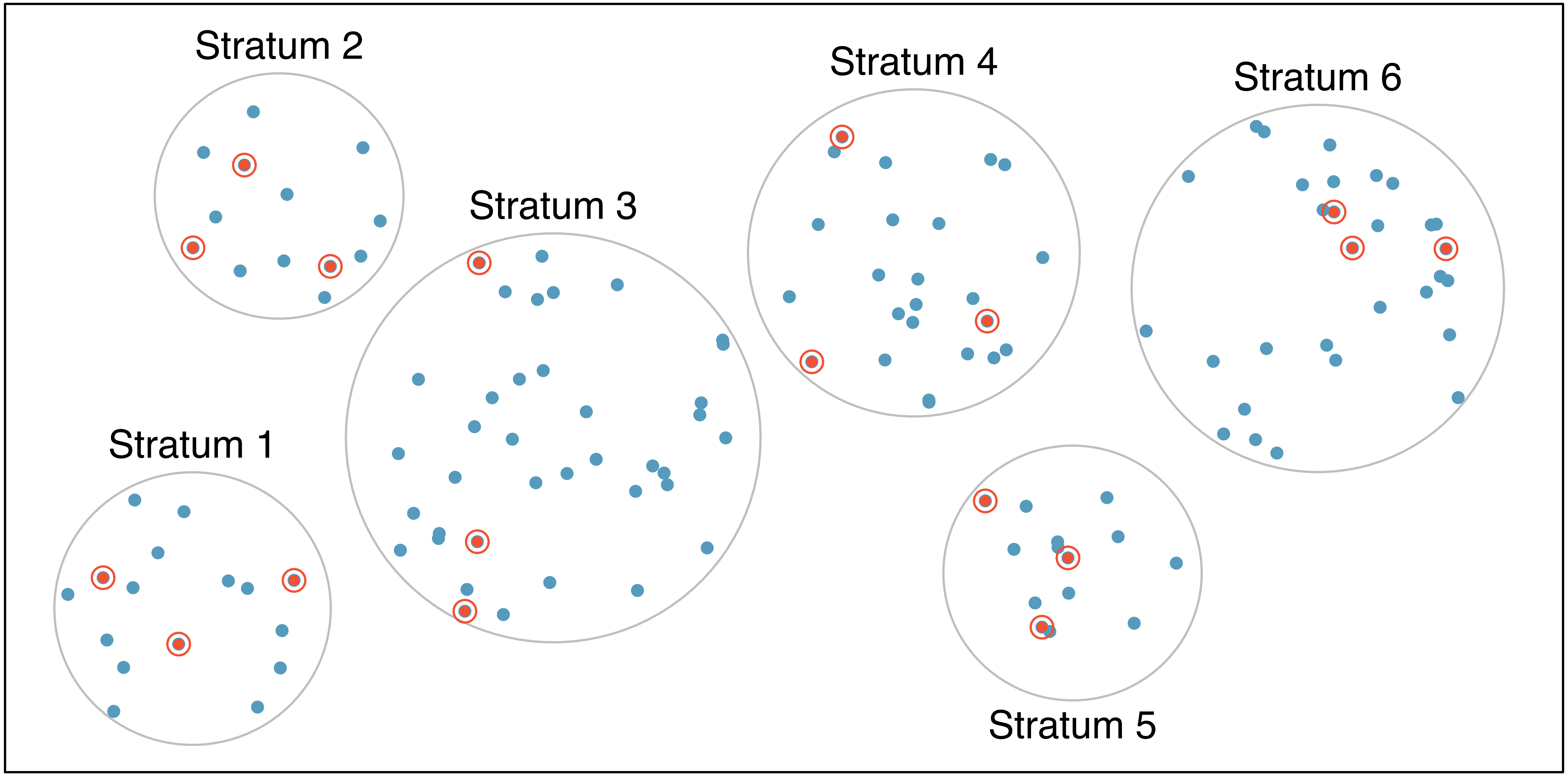
- The only way to collect a representative sample
  - simple random sample
  - stratified random sample
  - 抽樣理論 (統計所)

# 48112

Simple random sample



Stratified sample

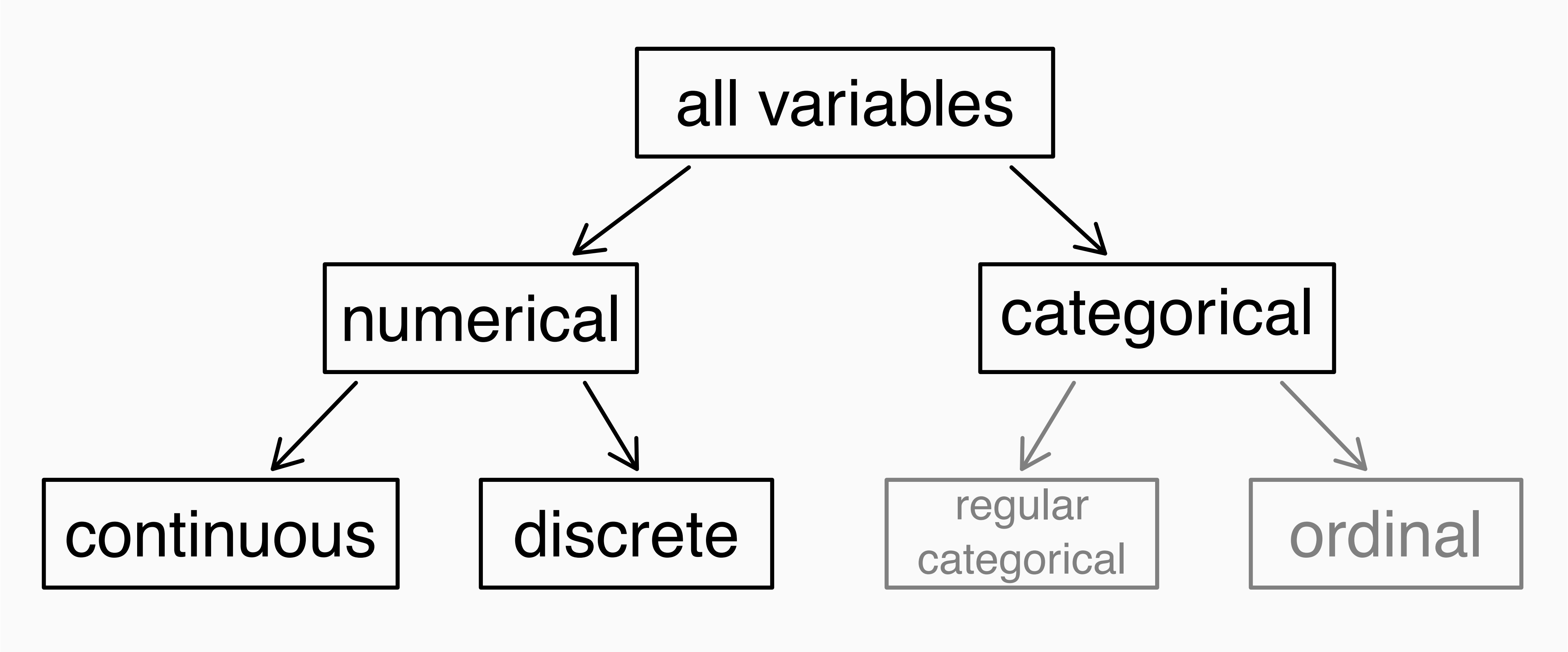


Experiment

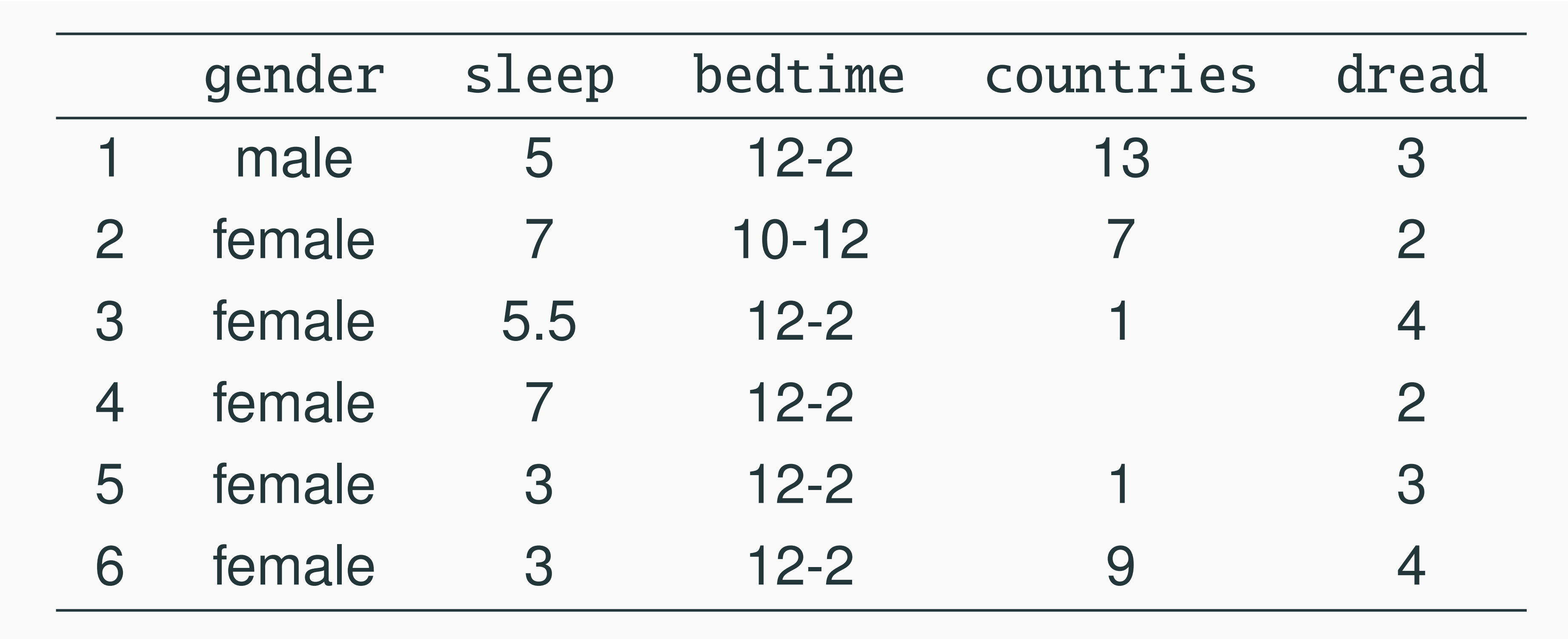
- Block design: reduce confounding variables by randomized experiments
- Factorial design
- Optimal design
- 實驗設計

Variables

Types of variables



Types of variables (cont.)



	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: categorical
- sleep: numerical, continuous
- bedtime: categorical, ordinal

Explanatory and response variables

- To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

$$\text{explanatory variable} \xrightarrow{might\ affect} \text{response variable}$$

- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables.

Data matrix (wide table)

variable

↓

Stu.	gender	intro_extra	⋯	dread
1	male	extravert	⋯	3
2	female	extravert	⋯	2
3	female	introvert	⋯	4
4	female	extravert	⋯	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	⋯	3

← *observation*

Data matrix (wide table)

- An observation contains several variables
- In a random sample, all the observations are collected randomly
    - i.e. all the variables are random variables

# 48112

Wide table vs long table

Long table

Person	Variable	Value
Bob	Age	32
Bob	Weight	168
Bob	Height	180
Alice	Age	24
Alice	Weight	150
Alice	Height	175
Steve	Age	64
Steve	Weight	144
Steve	Height	165

Wide table

Person	Age	Weight	Height
Bob	32	168	180
Alice	24	150	175
Steve	64	144	165