# Introduction to data

# Data matrix

*variable*

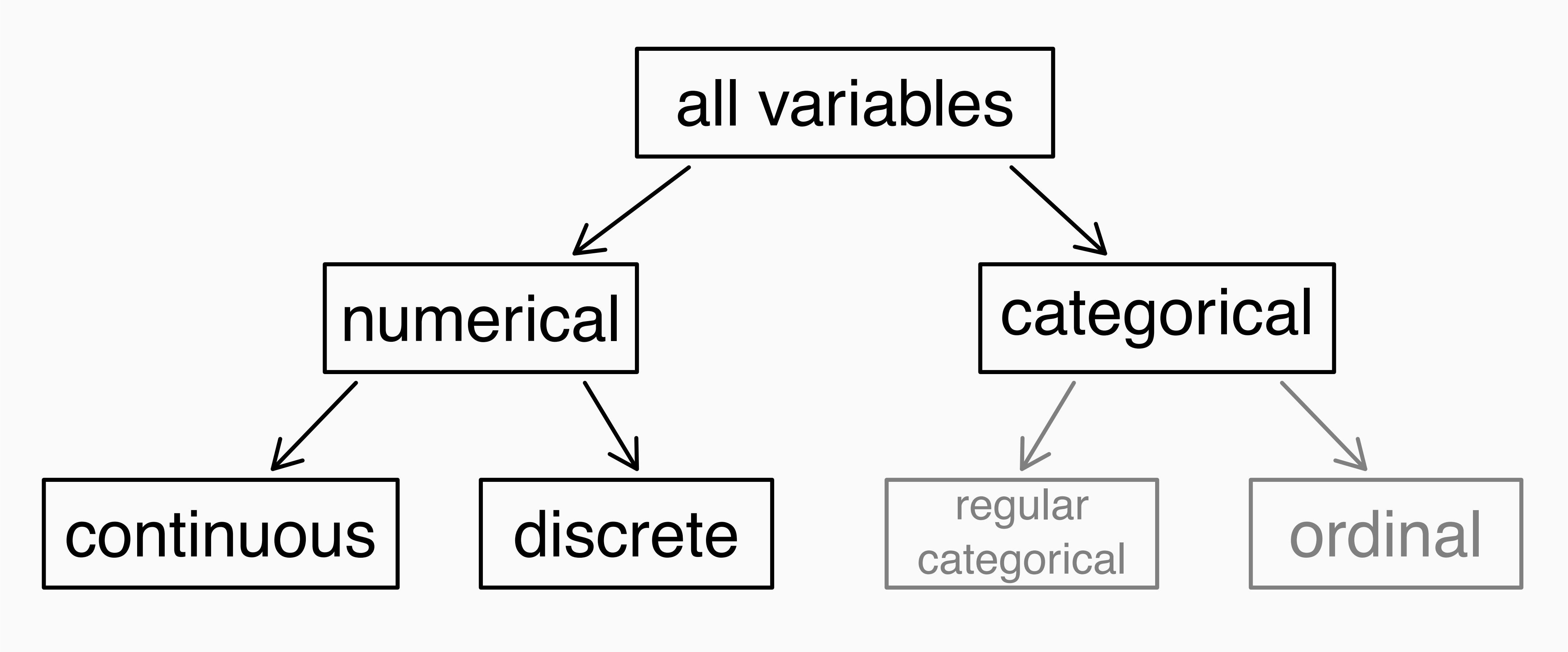
↓

| 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*

# Types of variables



all variables

numerical

categorical

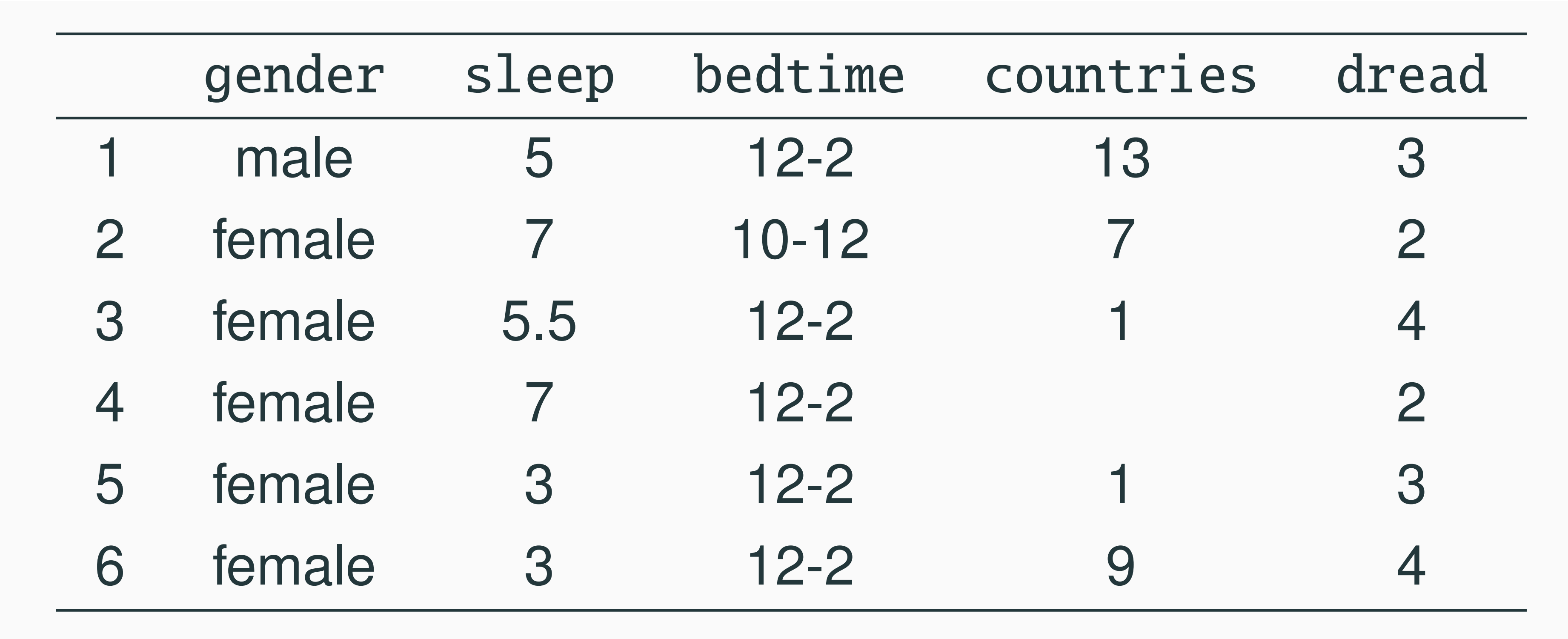
continuous

discrete

regular categorical

ordinal

# Types of variables (cont.)

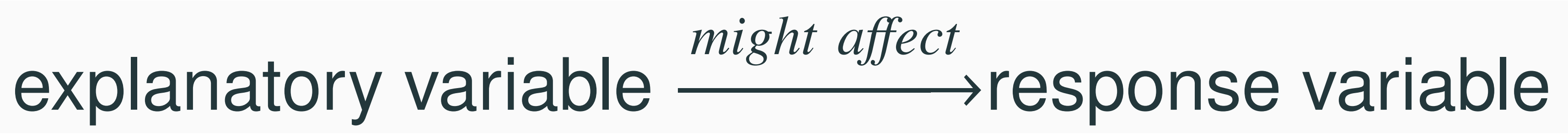


| | gender | sleep | bedtime | countries | dread |
|---|---|---|---|---|---|
| 1 | male | 5 | 12-2 | 13 | 3 |
| 2 | female | 7 | 10-12 | 7 | 2 |
| 3 | female | 5.5 | 12-2 | 1 | 4 |
| 4 | female | 7 | 12-2 | | 2 |
| 5 | female | 3 | 12-2 | 1 | 3 |
| 6 | female | 3 | 12-2 | 9 | 4 |

- 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:



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

- 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.

# 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.

# Causation

# Association≠Causation

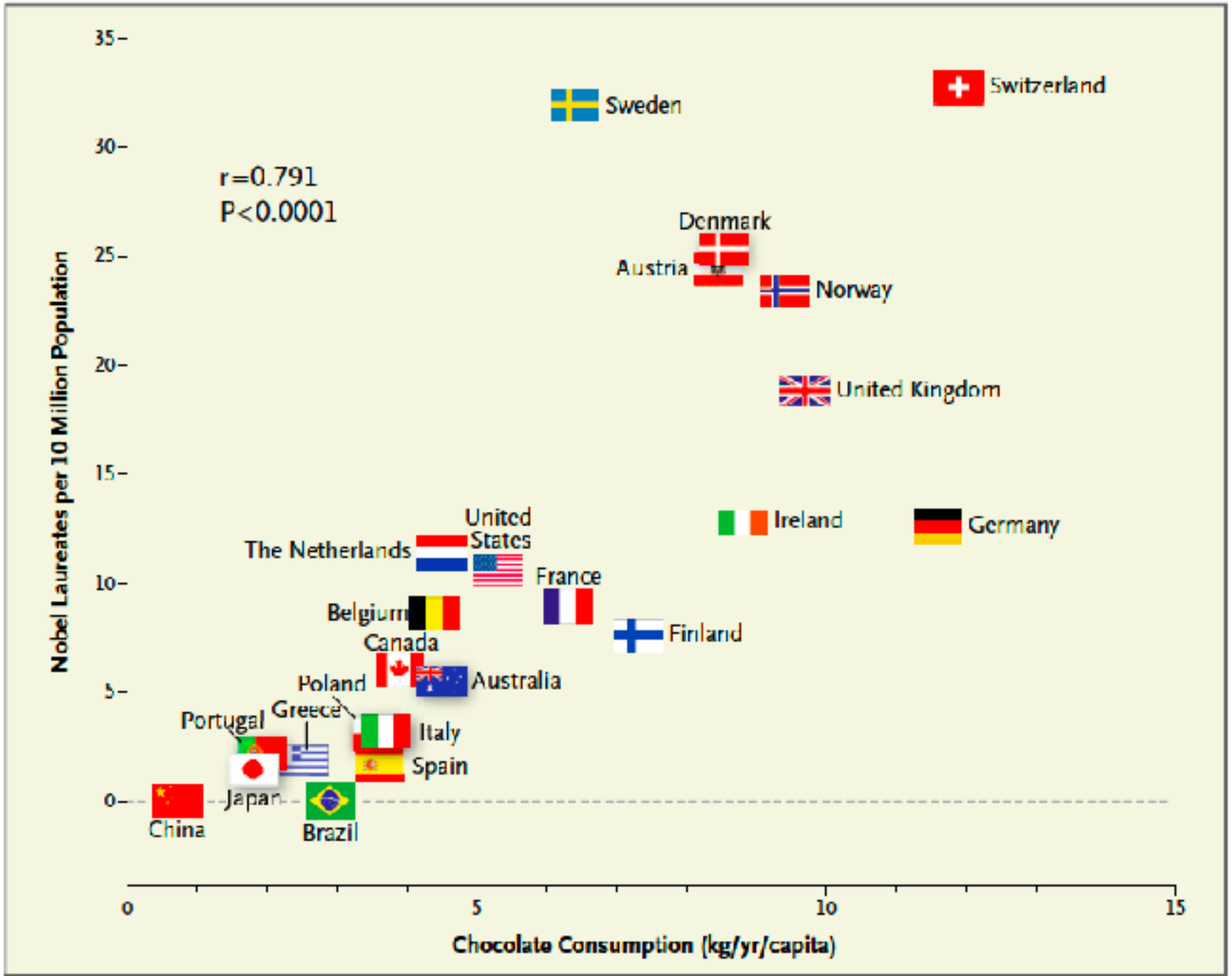
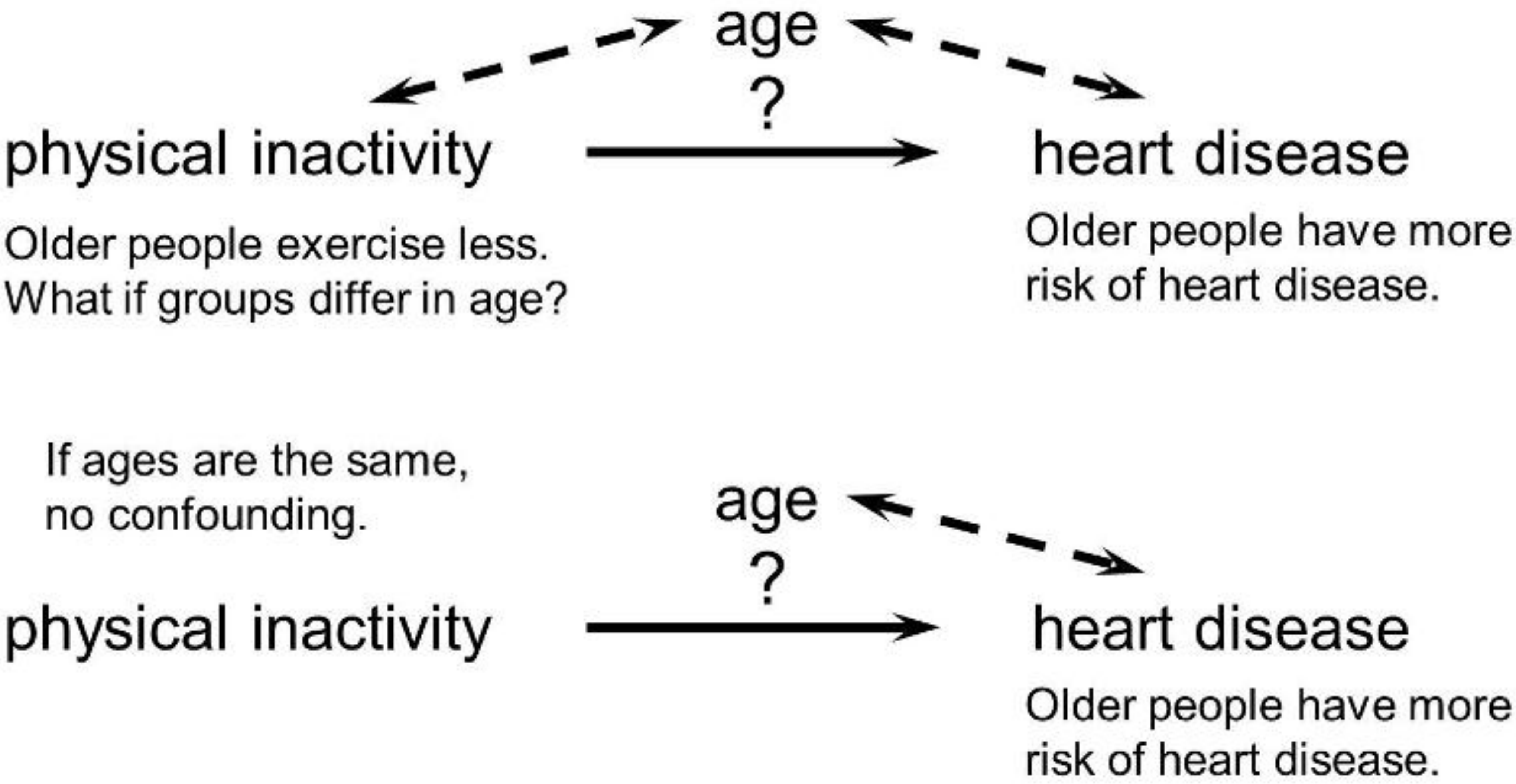


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

# Confounding



age

?

physical inactivity

heart disease

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.

age

?

physical inactivity

heart disease

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, e.g.,
    - gene 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 means prediction

# Population

- 所有研究對象稱為母體, e.g.,
    - 2018台中市長選舉勝負: 台中市合格選民
    - image object detection/recognition: 所有images
    - 心電圖 v.s. 心臟病: 所有病人的心電圖

# Sample

- 母體的任意子集合稱為樣本 (以管窺天)
  - random sample
  - nonrandom sample
  - sample of interest

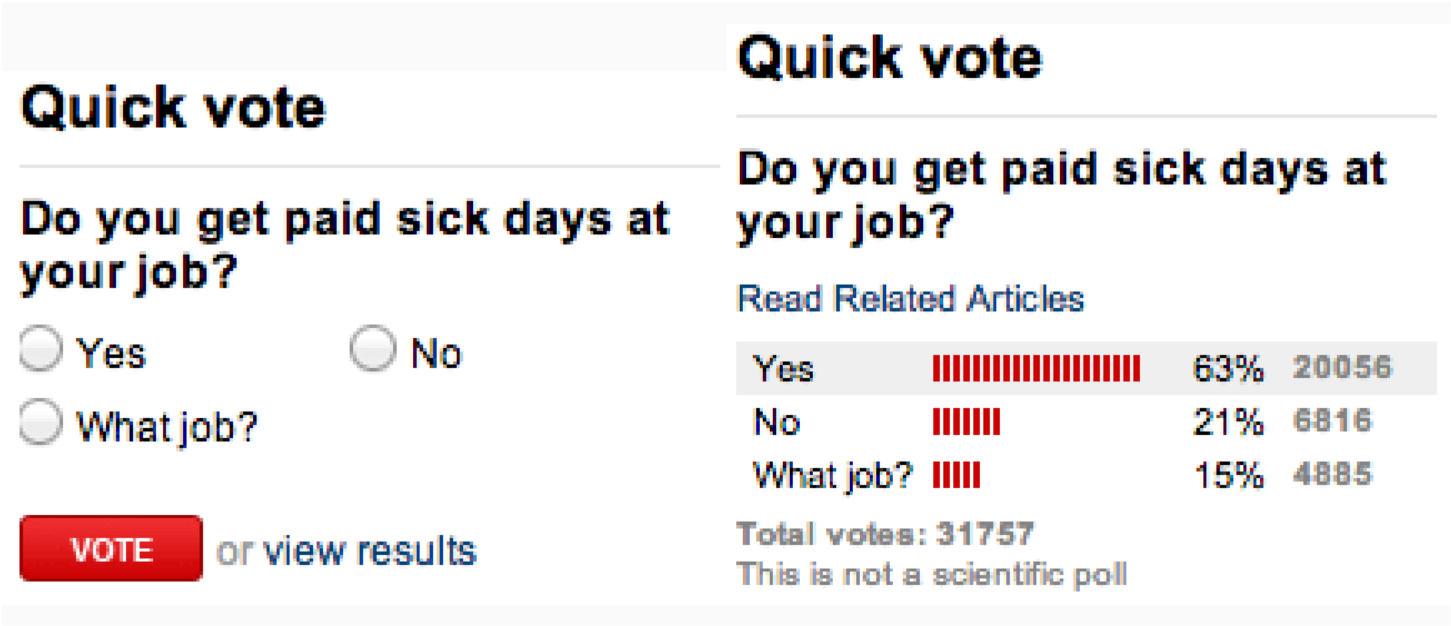
# Sampling bias

- 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.

# 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?

VOTE or view results

Quick vote

Do you get paid sick days at your job?

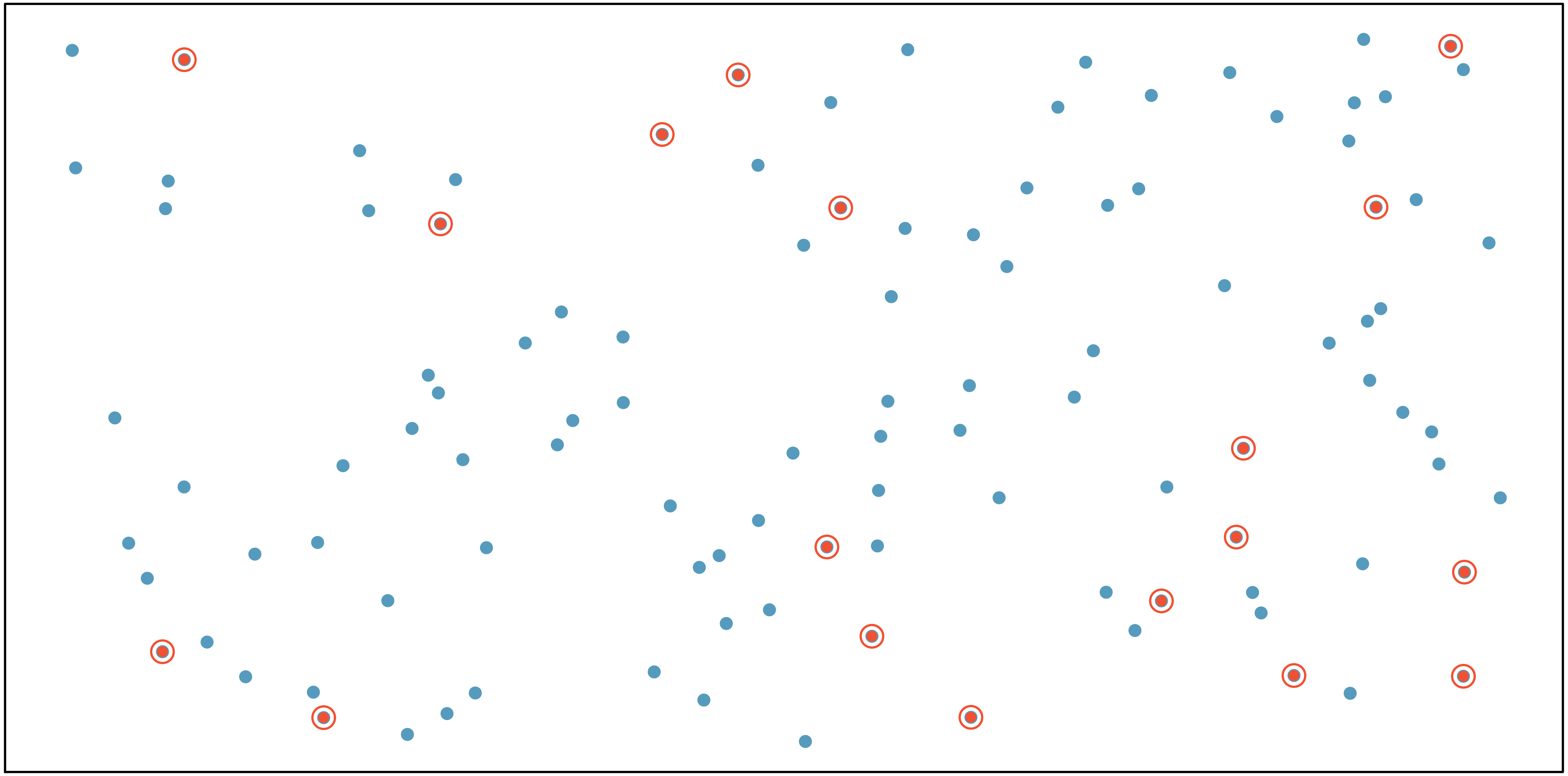
Read Related Articles

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

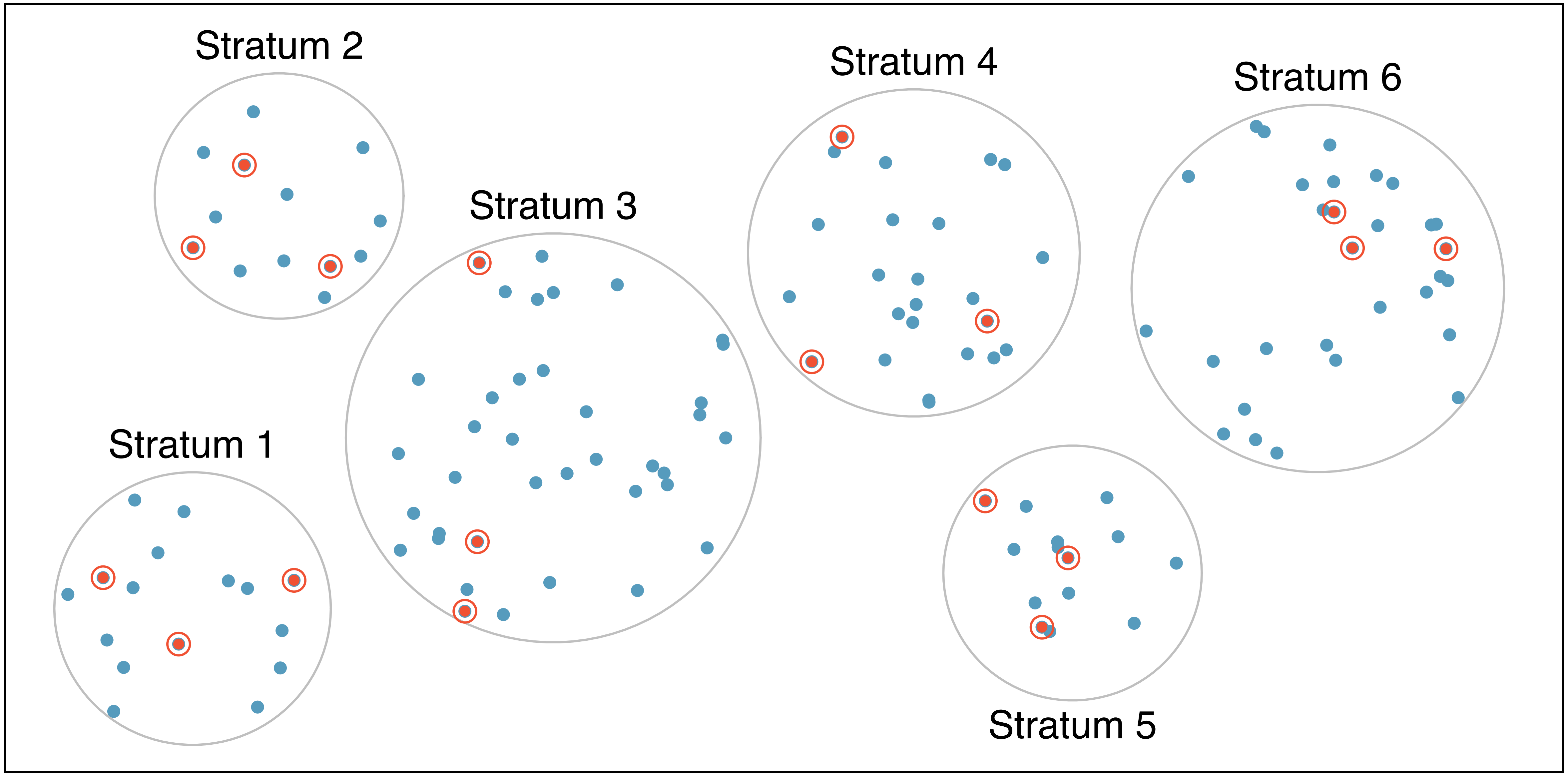
Total votes: 31757

This is not a scientific poll

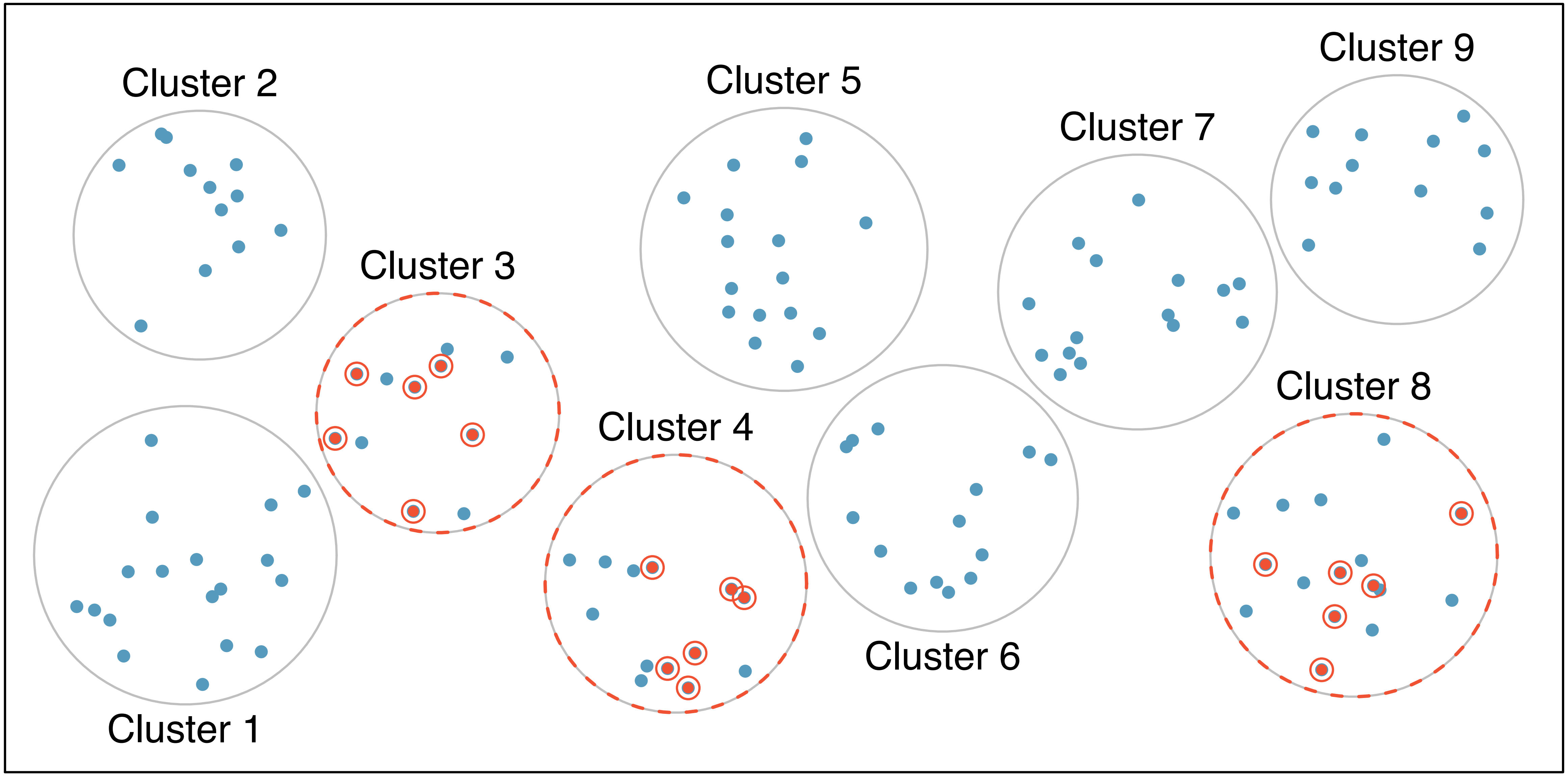
# Simple random sample



# Stratified sample



# Multistage sample



# Experiment

- compare treatments
- control variables (reduce confounding variables)
- fractional design
- bayesian optimization
- etc.

# Readings

- Chapter 2 of our first reference
- Chapter 1 of Introductory Statistics with Randomization and Simulation

# Homework #1

1. Plot the function $y = \sin(2\pi x)$ with Numpy and Matplotlib.

2. Generate 100 random points $x_1, \ldots, x_{100}$ in [0,1] with $y_i = \sin(2\pi x_i)$. Fit the data pairs $(x_i, y_i)$ by radial basis functions (provided by Scipy) and plot the fitted function.

3. Generate another 100 equispaced points in [1/4,3/4] and repeat 2 again.