# Experimental Design and Probability

## Sampling, observational studies, experiments

Robin Elahi

# Question

Consider statements made by journalists, lawyers, and scientists. What do they have in common? What sets the statements of scientists apart?

# A general process of investigation

1. Identify a question
2. Collect relevant data on the topic
3. Analyze the data
4. Form a conclusion

# A general process of investigation

1. Identify a question
2. Collect relevant data on the topic
3. Analyze the data
4. Form a conclusion

**Scientists quantify the uncertainty associated with a set of observations**

**Statistics helps us with steps 2-4, and the quantification of uncertainty**

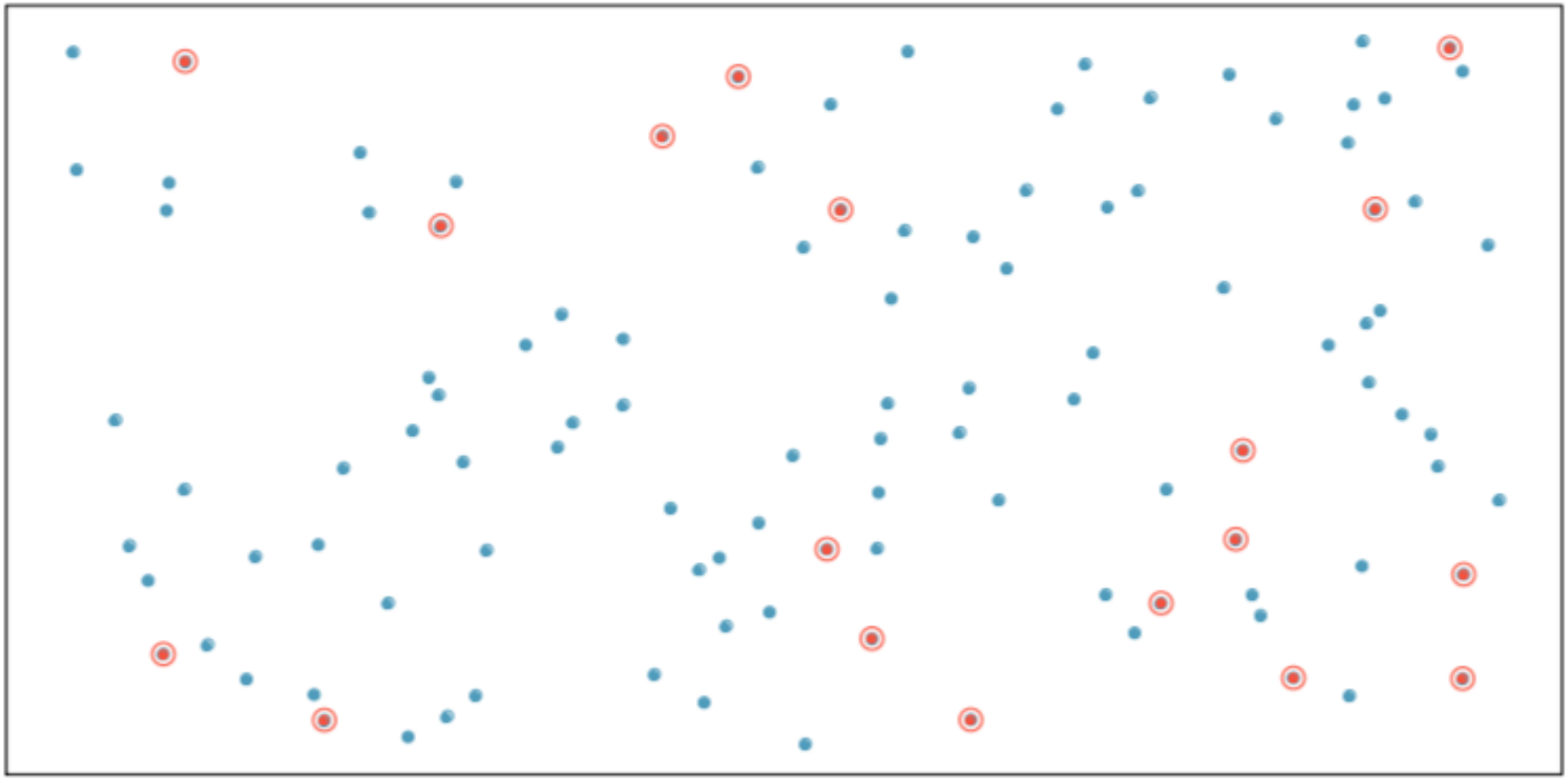
# Where does the uncertainty come from?

- Observation error
  - Did you measure the weight of that whale perfectly?
- Process error
  - our model, or view, of the world is *never* right!
  - "All models are wrong, but some are useful" - George Box
- Random variation
  - nature is big and variable
  - can't sample everything
  - sampling nature is hard

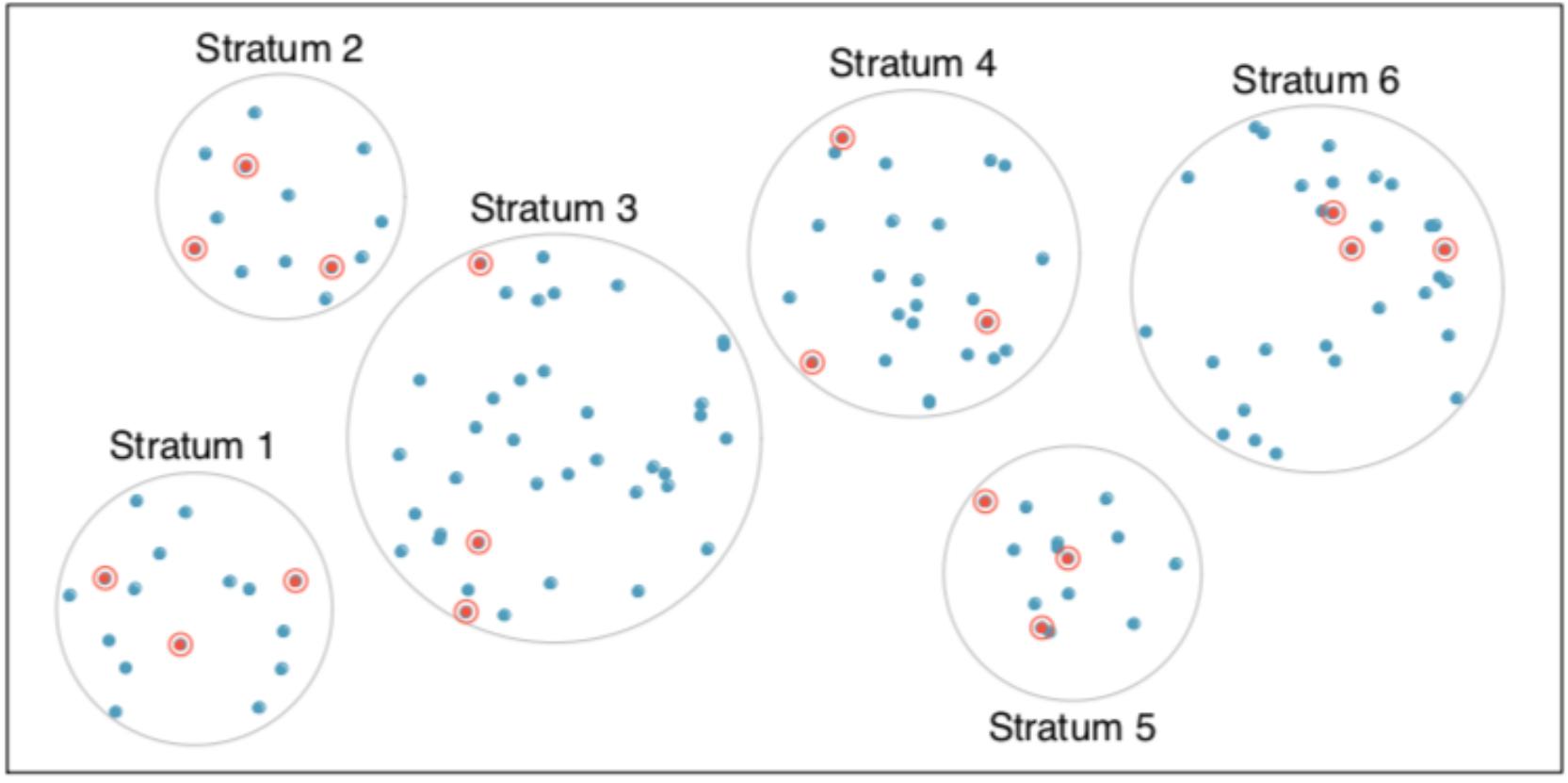
# Sampling and populations

- take a spoonful (*sample*) of the pot (*population*)
- your *exploratory data analysis* - not salty enough!
- if you conclude that the entire pot needs salt, that's an *inference*
- for the inference to be valid, the spoonful needs to be representative of the entire pot

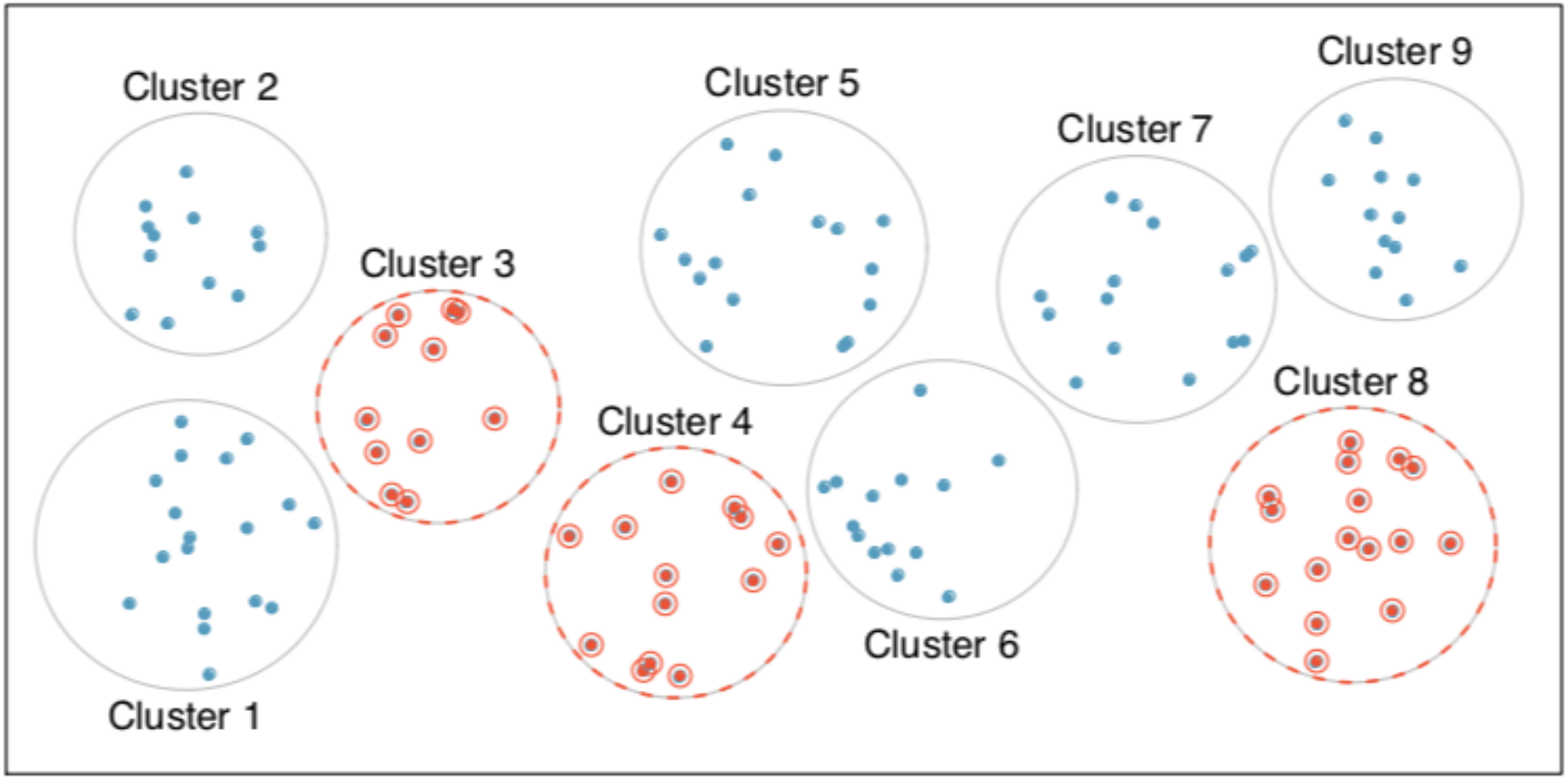
# Simple random sample



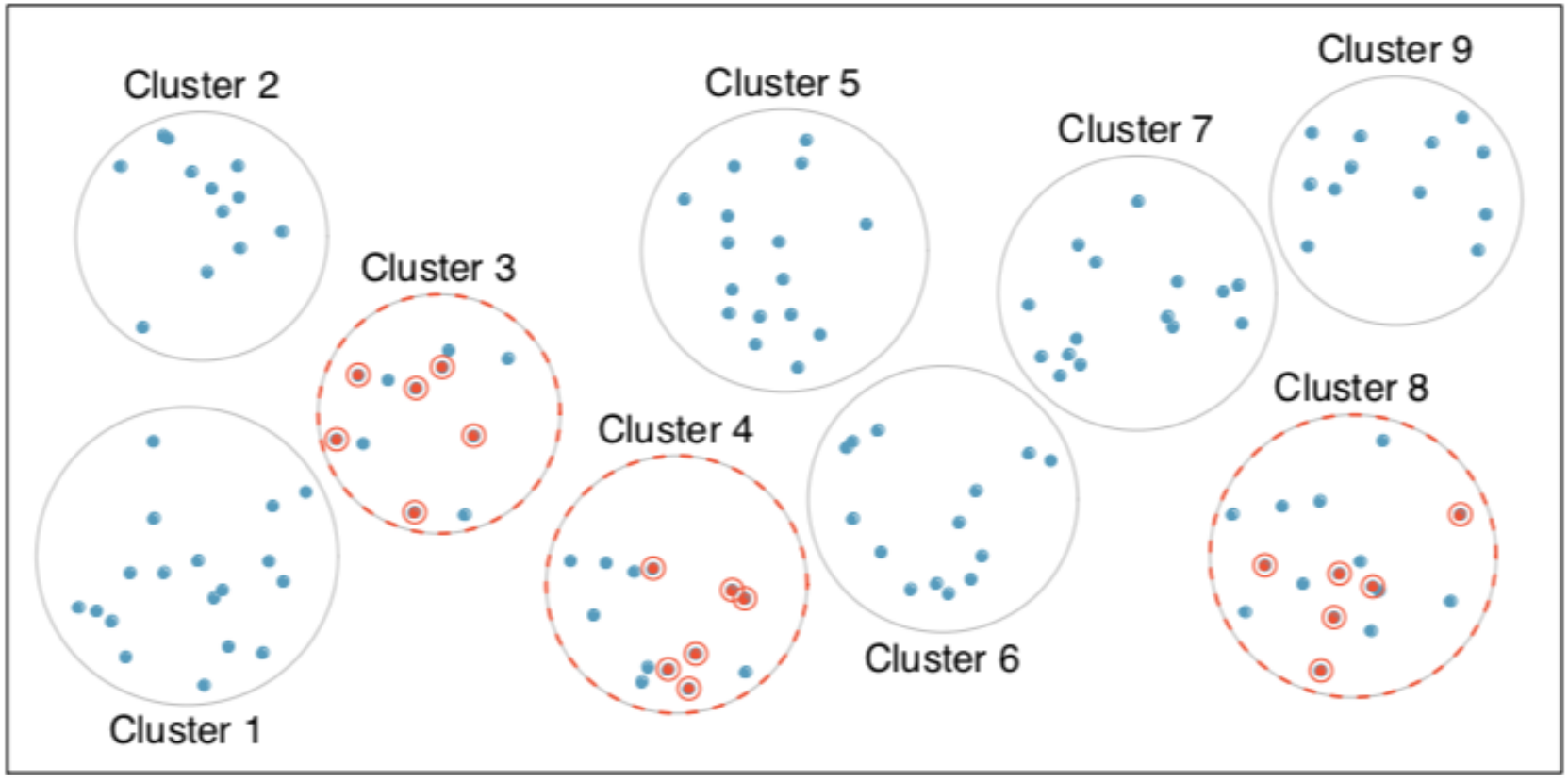
# Stratify to control for a variable



# Cluster to make sampling feasible



# Random samples in clusters



# Question

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Which approach would likely be the **least** effective?

(a) Simple random sampling

(b) Stratified sampling, where each stratum is a neighborhood

(c) Cluster sampling, where each cluster is a neighborhood

# Question

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Which approach would likely be the **least** effective?

(a) Simple random sampling

(b) Stratified sampling, where each stratum is a neighborhood

**(c) Cluster sampling, where each cluster is a neighborhood**

# Sampling biases

**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

**Voluntary response**: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population

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

# Question

A school is considering to no longer allow high school students to park at school after two accidents. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys, 1,200 are returned. Of these 1,200, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.
II. Overall, the school district has strong support from parents to move forward with the policy approval.
III. It is possible that majority of the parents of high school students disagree with the policy change.
IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

# Question

A school is considering to no longer allow high school students to park at school after two accidents. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys, 1,200 are returned. Of these 1,200, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.
II. Overall, the school district has strong support from parents to move forward with the policy approval.
III. It is possible that majority of the parents of high school students disagree with the policy change.
IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II **(c) I and III** (d) III and IV (e) Only IV

# Are response biases relevant in marine science?

Traditionally, no.

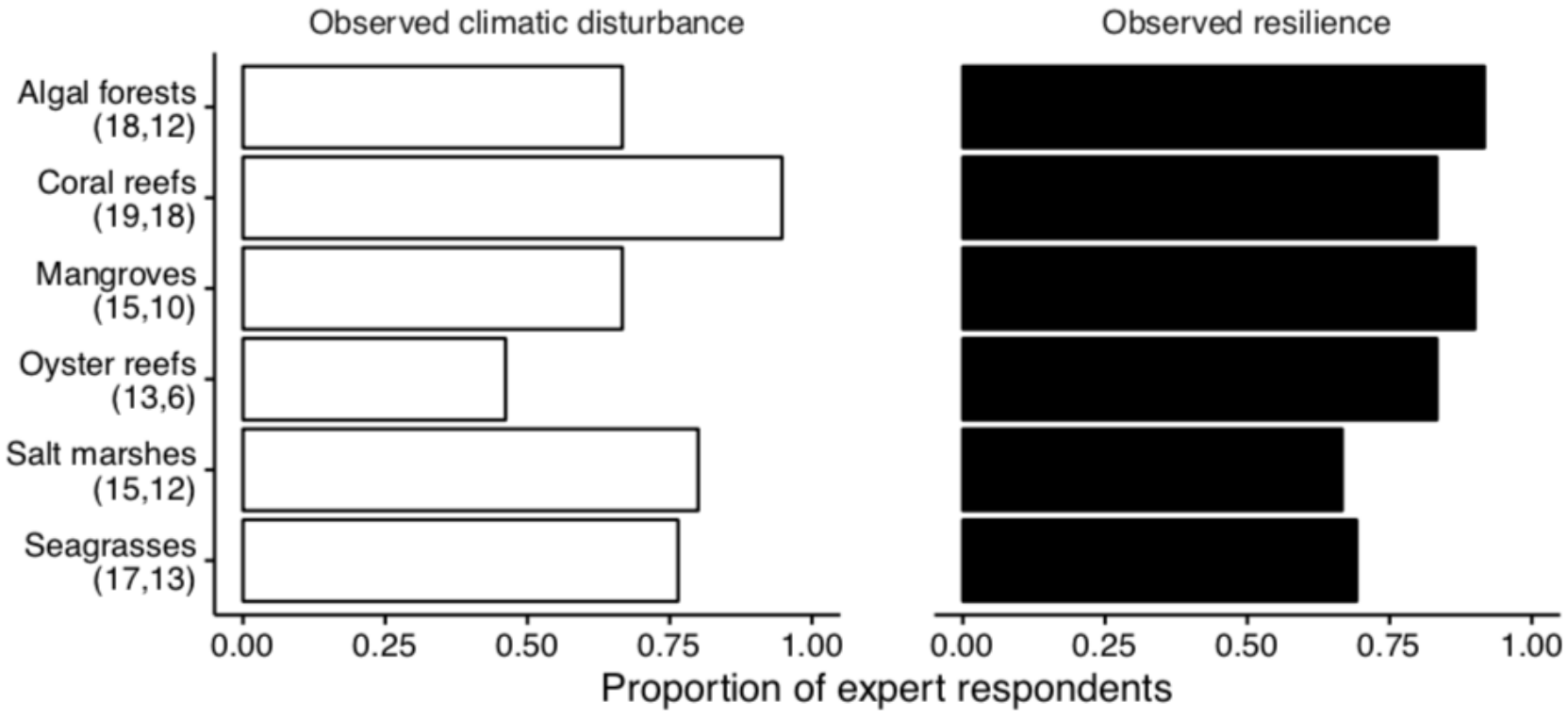
But social-ecological science is a burgeoning field. We can learn plenty from talking to people - both scientists and resource users (e.g., fishers).

An example: We asked ecologists whether they have personally observed resilience to climate disturbances.
(O'Leary et al. 2017 Bioscience)

# We surveyed experts on six coastal ecosystems

# Experts often observed ecological resilience

300 surveys sent to experts; 97 responses.



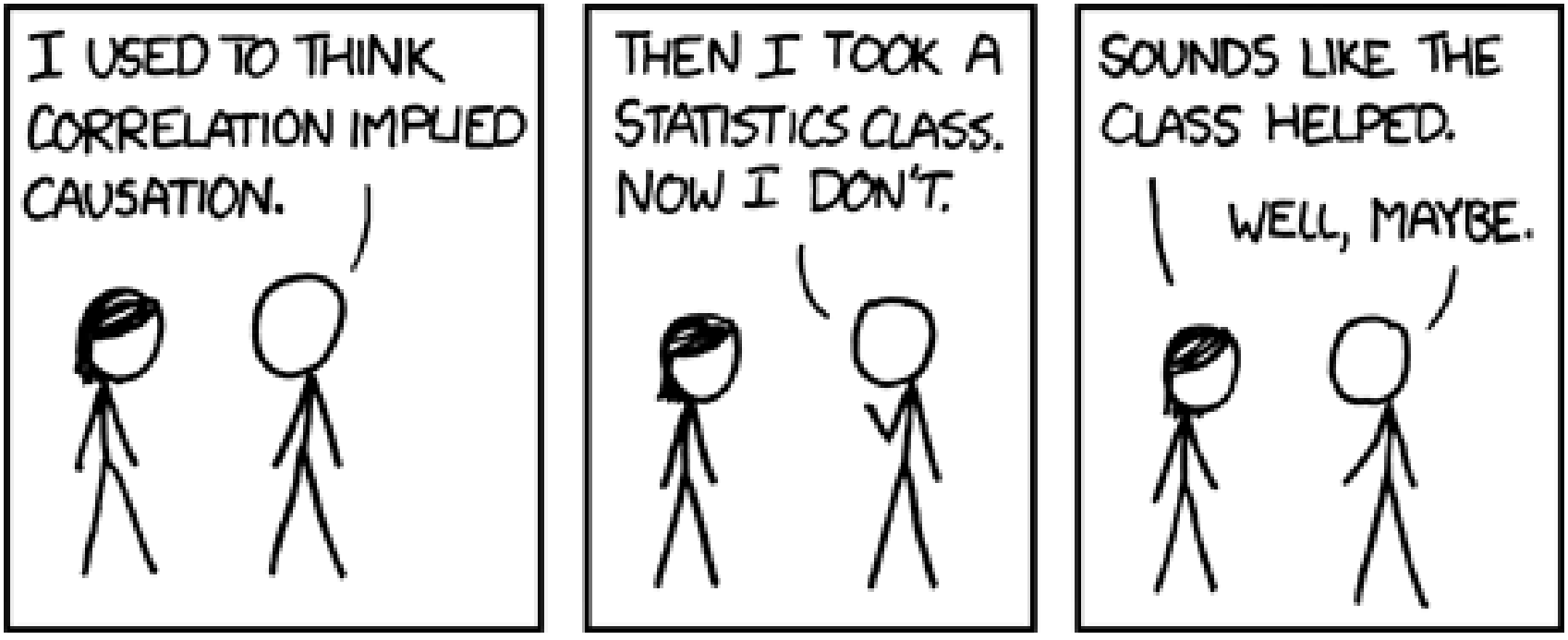
# Observational studies and experiments

# Observational studies and experiments

**Observational study**: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.

**Experiment**: Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.

# Correlation does not imply causation




https://xkcd.com/552/

# Question

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed.

What type of study is this?

What is the conclusion of the study?

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

# Question

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed.

What type of study is this? **Observational**

What is the conclusion of the study? **There is an association between increased stress & muscle cramps.**

Can this study be used to conclude a causal relationship between increased stress and muscle cramps? **No. Muscle cramps might also be due to increased caffeine consumption or sleeping less – these are potential confounding variables.**

# Four principles of experimental design

**Control**: Compare treatment of interest to a control group.

**Randomize**: Randomly assign subjects to treatments, and randomly sample from the population whenever possible.

**Replicate**: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.

**Block**: If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups.

# Does increased stress cause muscle cramps?

**Control**

**Randomize**

**Replicate**

**Block**

# Does increased stress cause muscle cramps?

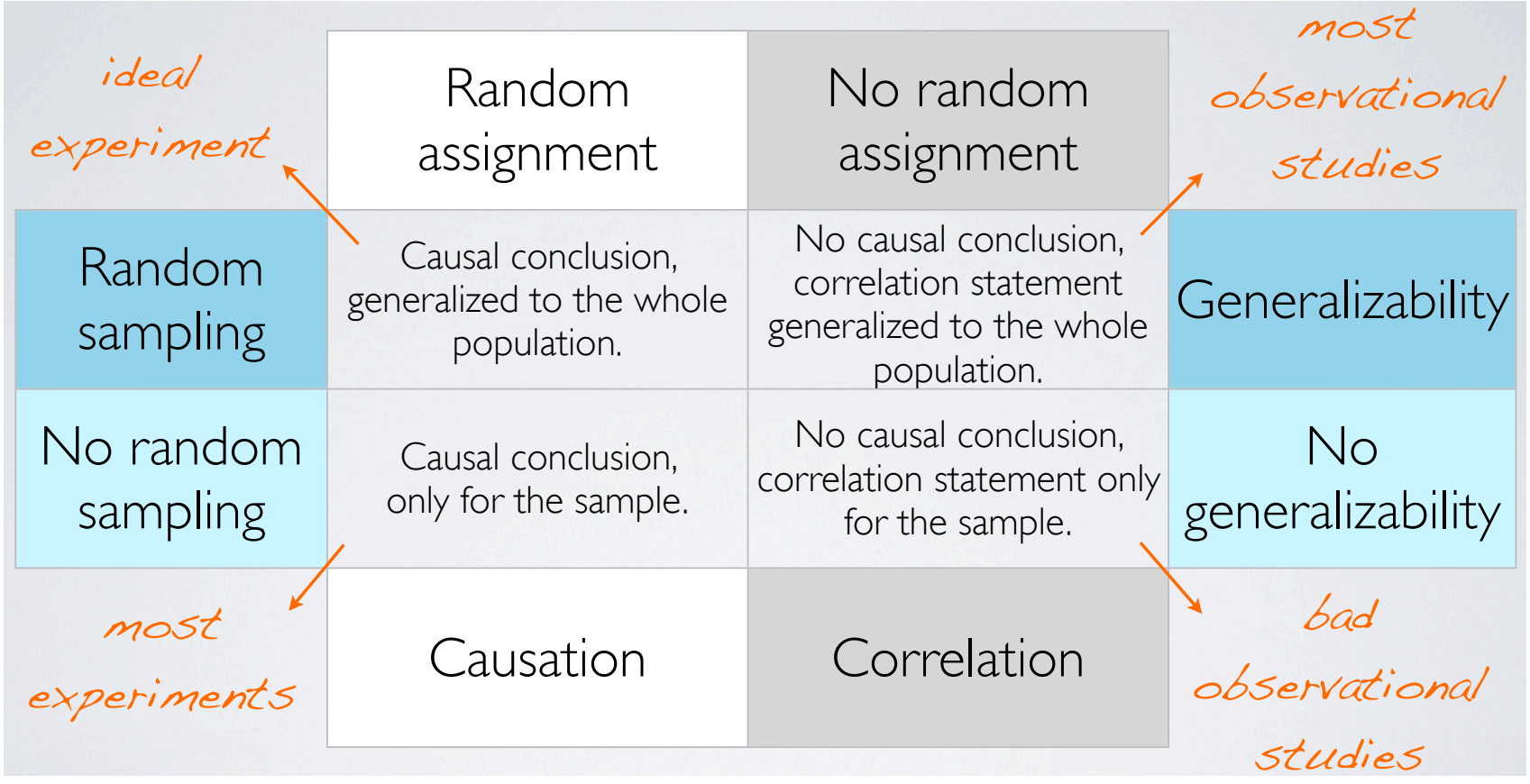
**Control**
Treatment: increased stress
Control: no or baseline stress

**Randomize**

**Replicate**

**Block**
It is suspected that the effect of stress might be different on younger and older people: *block* for age.

# Random sampling helps generalizability, random assignment helps causality



# Summary

1. Use a sample to make inferences about the population
2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
3. Sampling schemes can suffer from a variety of biases
4. Experiments use random assignment to treatment groups, observational studies do not
5. Four principles of experimental design: randomize, control, block, replicate
6. Random sampling helps generalizability, random assignment helps causality