ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

INTRODUCTION TO DATA

SAMPLING STRATEGIES AND BIAS

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Sampling Principles and Strategies

An important part of statistical research is identifying how data is collected so that it helps us answer our research questions.

- **Example:** We wish to know the prevalence of a particular gene in crows in California.
- **Population:** The entire set of things we wish to draw conclusions about.
  - In our example, it would be the set of all crows in California.

Frequently, it is too expensive, or even infeasible to collect data for every member of our population.

- **Sample:** A subset of the population for which data is collected.
  - In our example, it could be a random set of crows in California.

Sampling from a population

- **Example:** Over the last 5 years, what is the average time to complete a degree for UCSD undergrads?

- Here the population is all UCSD graduates in the last 5 years. Our sample, would be the subset of this population for which we collect data.

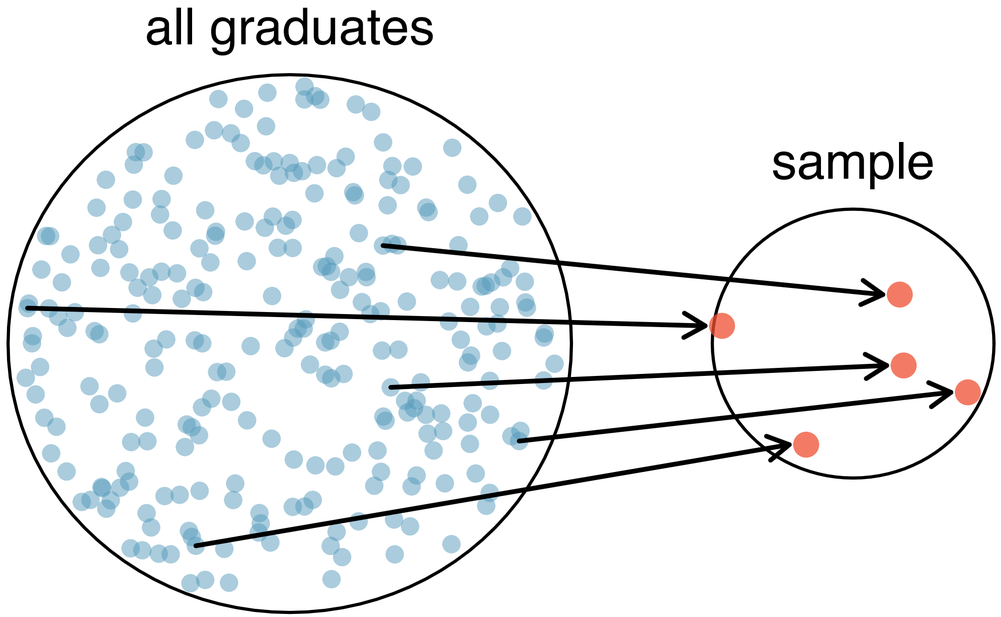

Figure from Open-intro Statistics textbook, Chapter 1

BIAS in a Sample

- You would like your sample to be representative of the population.
- In our example, we would want to avoid scenarios where graduates from one field are over-represented in the data.
- Or where graduates from a particular socio-economic status are over-represented...
- When selecting samples by hand you run the risk of introducing **bias** into the sample, even if it is unintentional.

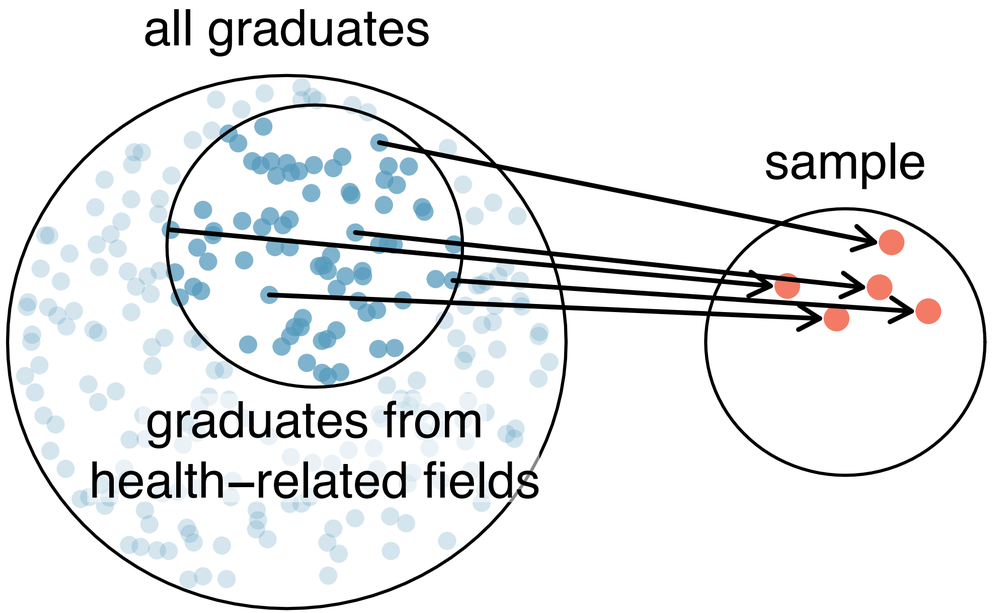

Figure from Open-intro Statistics textbook, Chapter 1

Simple Random Sampling

- Simple random sampling helps remove bias.
- Each case in the population has the same probability of being included in the sample, and the samples are picked independently of each other.

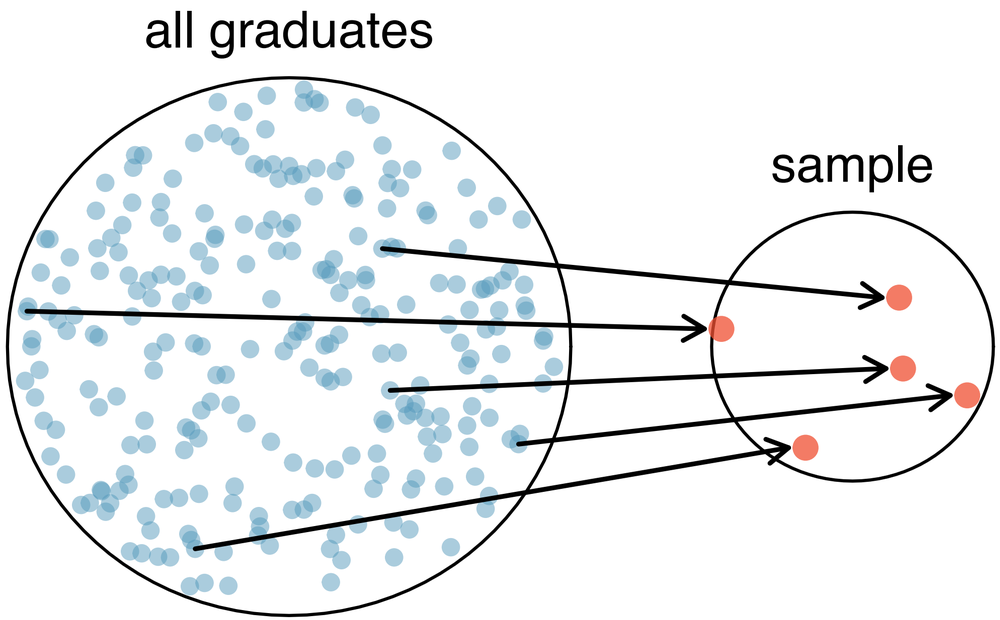

Figure from Open-intro Statistics textbook, Chapter 1

Non-response bias

- In our example, if we are able to obtain data from a random sample of UCSD's graduation records, our sample would not be biased.
- However, if we send out a survey to a random selection of UCSD graduates, bias can creep back into our sample.
- If, say, only 40% of our randomly selected graduates respond to the survey, it isn't clear whether our respondents are representative of the population.

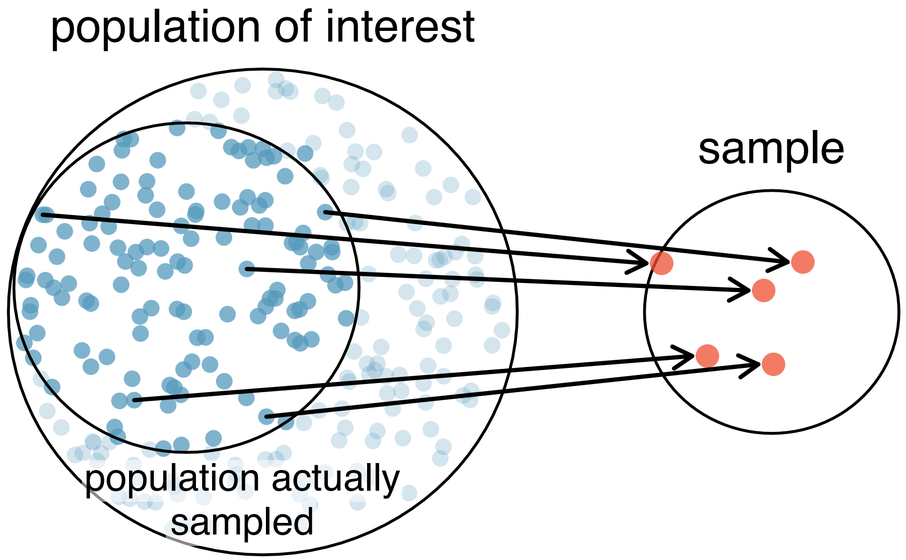

Figure from Open-intro Statistics textbook, Chapter 1

Other forms of Bias: Convenience Sample

- A convenience sample occurs when individuals that are easier to reach are more likely to be included in the sample.

- In our example, if we advertise our survey of UCSD graduate students on social media, this will represent only the graduates who are on social media.

- It may be difficult to discern the relevant characteristics of the sub-population associated with convenience sampling.

- **Example:** Movie/product reviews that can be found on various websites represent only the subset of the population that took the effort of providing a review. Do they tend to be more positive? More negative? Are they representative?

Experiments and Experiment Design

- In the graduation rate example we just saw, no treatment was applied or withheld. Thus, the previous example was an **observational study**.

- In contrast, studies where researchers assign treatments to cases are called **experiments**. When the assignment is random *(e.g., by flipping a coin)*, the experiment is called a **randomized experiment**.

Experiments and Experiment Design

Four principles of experiment design

- Controlling: when assigning treatments to cases, researchers must do their best to control other differences in the groups.
- Randomization: researchers randomize patients into treatment groups to account for variables that cannot be controlled (or that are even unknown).
- Replication: having more cases allows for more accurate estimation of the effect of the explanatory variable on the response. (One may even replicate an entire study to verify an earlier finding)
- Blocking: subdividing individuals into blocks on the basis of a variable, other than the treatment, that may affect the response, and then randomly assigning cases within each block to treatment groups. (may not always be necessary)

Bias in Human Experiments

Example

- Researchers design a *randomized experiment* to draw causal conclusions about the effect of a drug.
- Volunteers were randomly assigned to either the **treatment group**, which received the drug, or the **control group** which did not.

What could go wrong?

- Volunteers in the treatment group, may anticipate that the drug will help them. On the other hand, volunteers in the control group do not (and may even fear that not receiving any treatment could harm them).
- In this experiment there are possibly **two effects**:
  - the one of interest is the effectiveness of the drug
  - the second is an emotional effect (difficult to quantify).

Reducing Bias in Human Experiments

Example (cont'd)

- To circumvent this problem, we can keep the patients uninformed about their treatment status. Such a study is said to be **blind**. Here one gives fake treatments to patients in the control group (so they can't identify which group they're in).
- A fake treatment is called a **placebo**, and is key to making a study truly blind.
- Sometimes a placebo results in a slight but real improvement in patients $\Longrightarrow$ **placebo effect**.
- Another issue: Doctors and researchers can accidentally bias a study, e.g., by inadvertently giving patients in the treatment group more attention/care.
- To circumvent this problem, most modern studies employ a **double-blind** setup where researchers who interact with patients are also unaware of who is or is not receiving the treatment.