STAT 113
Sampling, Randomization and Confounding

Colin Reimer Dawson

Oberlin College

August 31 and Sept 5, 2017

Cases and Variables Warmup

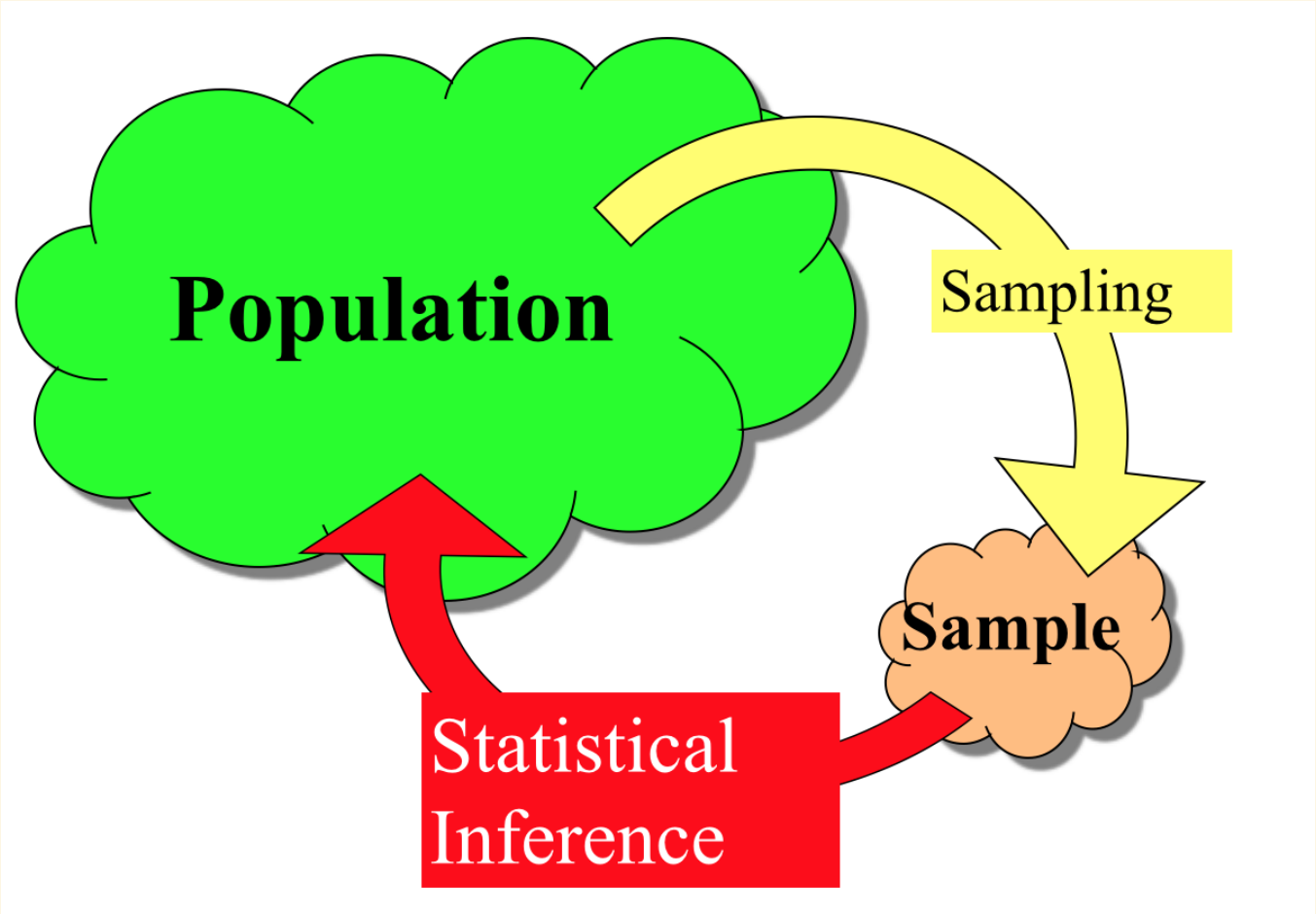
The handout has a graph depicting some information about monuments depicting Confederate leaders and soldiers involved in the U.S Civil War in the 1860s. (Source: Southern Poverty Law Center)

1. Identify what the cases are (what does one dot represent?)
2. Identify all of the variables that the graph depicts (what do we know about each dot?) and how their value is conveyed.
3. Sketch the first few rows of the data table that might have been used to create this graph.
4. What features of the graph stand out?
5. Does this graph help resolve the following question: "To what extent do Confederate monuments represent southern heritage, honoring those who died, and to what extent were they meant to intimidate Black citizens in the south?"

Sampling

Summarizing the Gettysburg Address

Sampling and Inference: The "Big Picture"



Population, Samples, and Inference

Population: All potential cases that we are interested in saying something about

Sample: The set of cases we actually have data for (a subset of the population)

Statistical Inference: Using sample data to obtain information about the population

For inference to be effective, samples ought to be **representative** of the population.

Simple Random Sampling

- To guard against sampling bias, we typically want to collect a **random** sample. Versions...
  - Simple Random Sampling $\leftarrow$ (Our focus)
  - Stratified Sampling
  - Cluster Sampling
  - Systematic Sampling

Feasibility of Random Sampling

It is often not feasible to get a truly random sample. Options:

- Reduce the scope of your population, and limit generalization accordingly
- Collect a non-random sample, avoid as many sources of bias as possible

Not all Non-Random Samples are Created Equal

You want to estimate the average hours per week that Oberlin students spend studying. None of the following is random; which would you go with?

(a) Go to Mudd and ask people there

(b) Email every student and use all responses

(c) Require all students in a statistics class to respond

(d) Go to the gym and ask everyone going in

(e) Go to The Local and ask everyone

Not all Non-Random Samples are Created Equal

None of the above are representative in every way, but some are more obviously non-representative *for the variable we care about*.

"Don't Ask Don't Tell"

2010 CBS/NYT polls (when DADT was being reconsidered):
"Do you favor or oppose homosexuals gay men and lesbians serving openly in the military?"

	Favor	Oppose
"homosexuals"	44%	42%
"gay men and lesbians"	58%	28%

Non-Sampling Bias

Other sources of bias not due to sampling procedure

- Question wording
- Non-response bias
- Context

Confounding Variables

Does a packed arena help or hurt the home team in basketball?

Race and the Death Penalty

Data from 1981 Florida Homocide Convictions

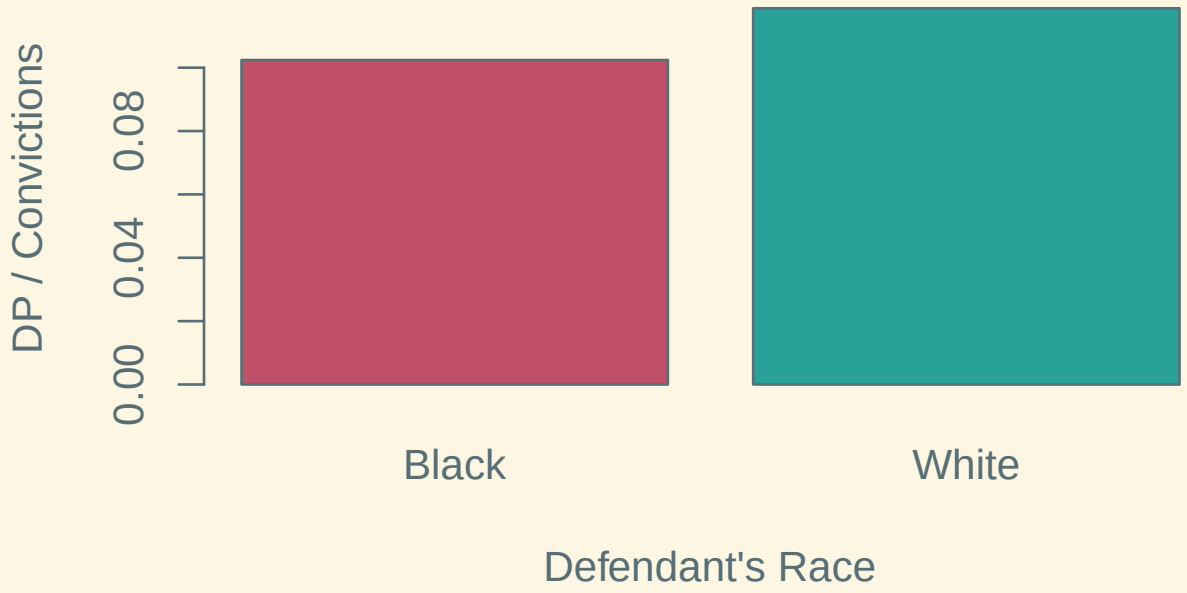


Figure: Proportions of Death Sentences out of Total Convictions, by Defendant's Race. Source: Agresti (2002)

Race and the Death Penalty

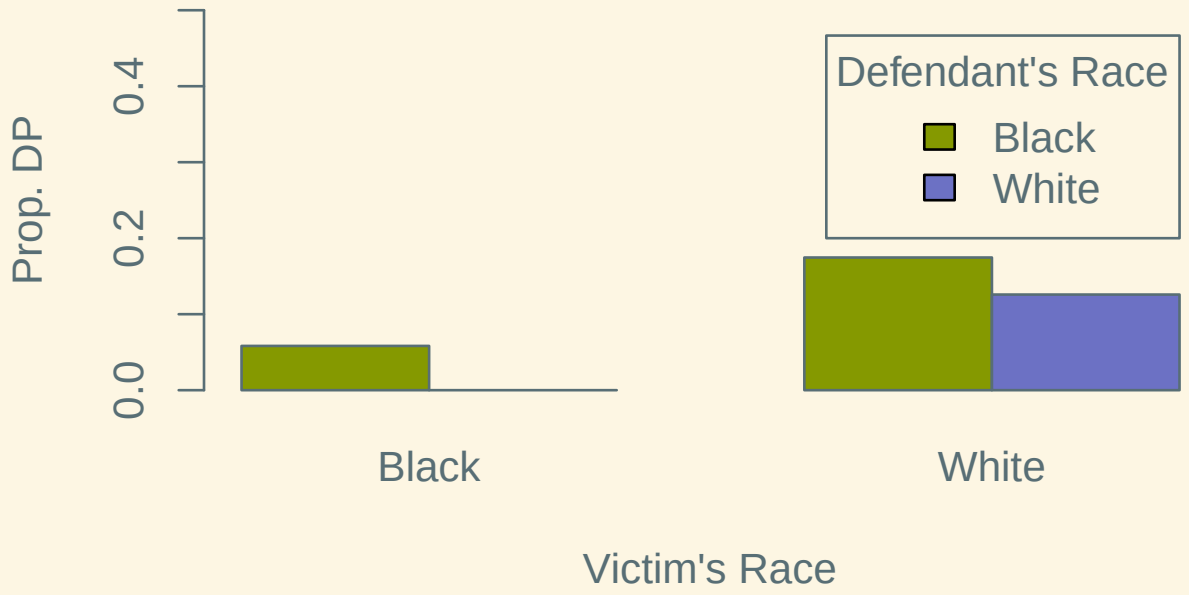
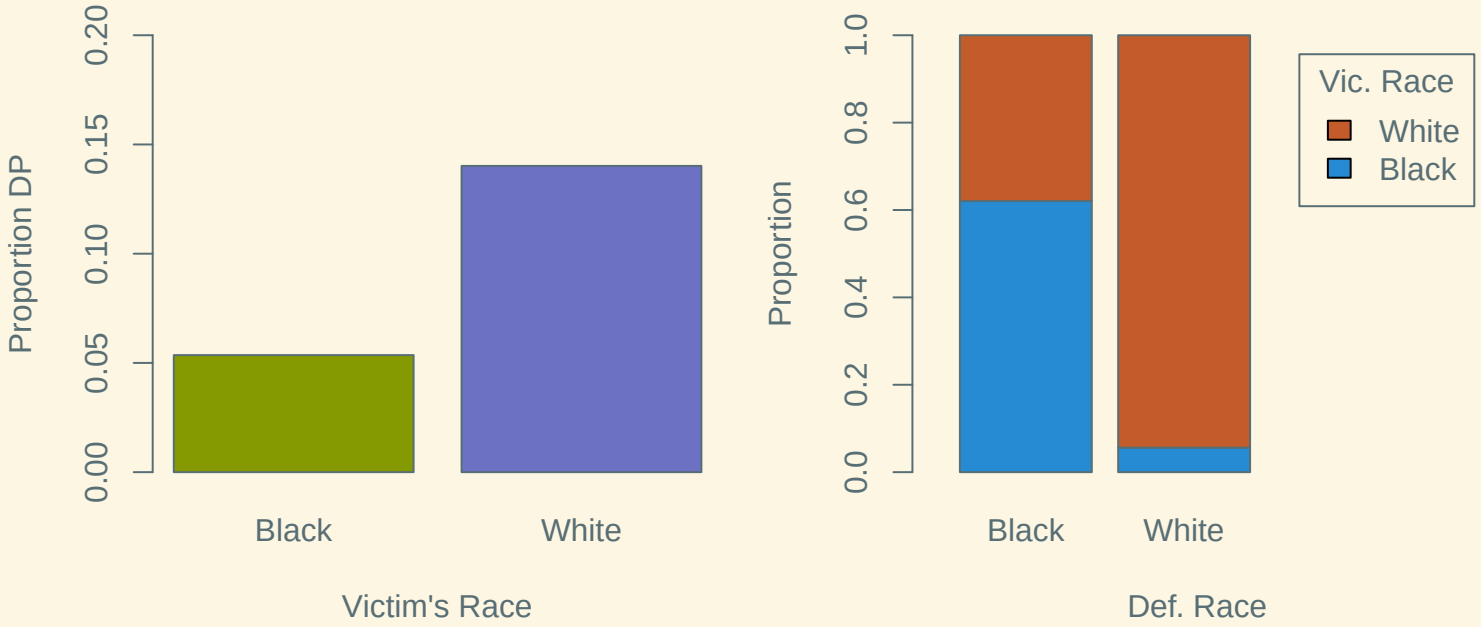


Figure: Proportions of Death Sentences out of Total Convictions, by both Victim's and Defendant's Race. Source: Agresti (2002)

Black defendants are sentenced to death at higher rates with both white and black victims. But combined, white defendants have a (slightly) higher rate. Why the seeming contradiction?

Race and the Death Penalty



The DP was applied much more often for white victims, and most homocides involved same-race individuals. The tendency to be involved in cases with white victims was a slightly bigger disadvantage as a defendant than being white was an advantage.

Confounding Variables

A **confounding variable** has a relationship with *both* the explanatory and response variable, making it difficult or impossible to interpret the relationship between the two.

- If we view defendant's race as explanatory variable and sentencing outcome as response, then victim's race is a **confounding variable**.
- When the confound is ignored, we get a **spurious association**.

Simpson's Paradox

When controlling for the confound results in a reversal of the direction of association, this is an instance of **Simpson's Paradox**.

- Ignoring victim's race flips the direction of the association between defendant's race and sentencing outcome.

Example: Cursive Handwriting and SAT Scores

Handout

Observational vs. Experimental Designs

In an **experimental design**, the researchers control which cases are assigned to which levels of the explanatory variable(s). Otherwise (if cases have "naturally occurring" values of the e.v.), the study is **observational**.

A "gold standard" procedure for an experiment is **random assignment** of cases to levels of the explanatory variable(s). This ensures that there is no systematic relationship between the explanatory variable and any would-be confounding variables.