Threats to Validity

(Also available in Pyret)

Students consider possible threats to the validity of their analysis.

Lesson Goals	Students will be able to… • Define several types of Threats to Validity • Identify those threats by reading the description of an analysis • Identify those threats in their own analysis
Student-facing Lesson Goals	• Let's identify issues that could affect our data analysis.
Prerequisites	• Introduction to Data Science
Materials	***This lesson is unplugged*** *and does not require a computer.* • PDF of all Handouts and Page • Lesson Slides • Printable Lesson Plan (a PDF of this web page)
Supplemental Materials	• *Optional Project: Threats to Validity [rubric]* • Additional Printable Pages for Scaffolding and Practice • Project: When Data Science Goes Bad
Supplemental Resources:	Poster Set of Data Fallacies to Avoid

Glossary

threats to validity :: factors that can undermine the conclusion of a study

Threats to Validity *20 minutes*

Overview

Students are introduced to the concept of *validity*, and a number of possible threats that might make an analysis invalid.

Launch

Let's say a survey says that people prefer cats to dogs…

As good Data Scientists, the staff at the animal shelter are constantly gathering data about their animals, their volunteers, and the people who come to visit. But just because they have data doesn't mean the conclusions they draw from it are correct! For example: suppose they surveyed 1,000 cat-owners and found that 95% of them thought cats were the best pet. Could they really claim that people generally prefer cats to dogs?

Have students share back what they think. The issue here is that cat-owners are not a representative sample of the population, so the claim is invalid.

There's more to data analysis than simply collecting data and crunching numbers. In the example of the cat-owning survey, the claim that "people prefer cats to dogs" is **invalid** because the data itself wasn't representative of the whole population (of course cat-owners are partial to cats!). This is just one example of what are called ***Threats to Validity***.

There are several major threats to validity you should be on guard against:

(1) **Selection bias** - Data was gathered from a biased sample of the population. This is the problem with surveying *cat owners* to find out which animal is most loved!

(2) **Bias in the study design** - Data was gathered using a "loaded" question like "Since annual vet care comes to about $300 for dogs and only about half of that for cats, would you say that owning a cat is less of a burden than owning a dog?" This could easily lead to a misrepresentation of people's true opinions.

(3) **Poor choice of summary data** - Even if the selection is unbiased, sometimes outliers are so extreme that they make the mean completely useless at best - and misleading at worst.

(4) **Confounding variables** - A study might find that cat owners are more likely to use public transportation than dog owners. But it's not that owning a cat means you drive less: people who live in big cities are more likely to use public transportation, *and* also more likely to own cats. More examples of confounding variables can be found in the correlations lesson: Correlation Does Not Imply Causation!.

This is just a small list of different threats to validity. There are plenty more!

Investigate

- On Identifying Threats to Validity and Identifying Threats to Validity (2), you'll find four different claims backed by four different datasets.
- Each one of those claims suffers from a serious threat to validity. Can you figure out what those threats are?
- *Optional:* Respond to Selection Bias or Biased Study?

Optional Project: When Data Science Goes Bad

In this Project: When Data Science Goes Bad, students pretend to be terrible data scientists who develop and support claims based on faulty sampling techniques (selection bias, bias in the study design, poor choice of summary data, and confounding variables). This is a fun opportunity for your students to demonstrate their understanding of the impact of various threats to validity.

Synthesize

Give students time to discuss and share back.

Life is messy, and there are *always* threats to validity. Data Science is about doing the best you can to minimize those threats, and to be up front about what they are whenever you publish a finding. When you do your own analysis, make sure you include a discussion of the threats to validity!

Fake News! *20 minutes*

Overview

Students are asked to consider the ways in which statistics are misused in popular culture, and become critical consumers of some statistical claims. Finally, they are given the opportunity to misuse their *own* statistics, to better understand how someone might distort data for their own ends.

Launch

Students have already seen a number of ways that statistics can be misused:

1. Using the mean instead of the median with heavily-skewed data
2. Using the wrong language when describing a Linear Regression
3. Using a correlation to imply causation

There are other ways to mislead the audience as well:

1. **Intentionally using the wrong chart** - suppose the census asks for data from different groups of people, and gets *none* from one group. That would be very suspicious! That group would show up as an empty space on bar chart, making the absence visible. A pie chart, however, would hide that absence completely - making it less likely that anyone would even notice that group had been "erased"!
2. **Changing the scale of a chart** - Changing the y-axis of a scatter plot can make the slope of the regression line seem smaller: "look, that line is basically flat anyway!"

With all the news being shared through newspapers, television, radio, and social media, it's important to be critical consumers of information!

Investigate

- On Fake News, you'll find some deliberately misleading claims made by slimy Data Scientists. **Why shouldn't these claims should be trusted?**
- Once you've finished, consider your own dataset and analysis: what misleading claims could someone make about your work? Turn to Lies, Darned Lies, and Statistics, and **come up with four misleading claims based on data or displays from your work.**
- Trade papers with another group, and see if you can figure out why each other's claims are not to be trusted!

Synthesize

Have students share back their "lies". Was anyone able to stump the other group?

Additional Exercises

- Identifying Threats to Validity (3)
- *Optional Project:* Threats to Validity *[rubric]*