










Threats to Validity

Students consider possible threats to the validity of their analysis.

Prerequisites	Linear Regression																				
Relevant Standards	Select one or more standards from the menu on the left (⌘-click on Mac, Ctrl-click elsewhere).																				
<div>OK</div> <div>K12CS</div> <div>CSTA</div> <div>NGSS</div> <div>CC-Math</div>																					
Lesson Goals	Students will be able to... <ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• Let’s identify issues that could affect our data analysis.																				
Materials	<ul style="list-style-type: none">• Lesson Slides (Google Slides)• Computer for each student (or pair), with access to the internet• Student workbook, and something to write with																				
Preparation	<ul style="list-style-type: none">• Make sure all materials have been gathered• Decide how students will be grouped in pairs																				
Supplemental Resources																					
Language Table	<table><tr><th>Types</th><th>Functions</th><th>Values</th></tr><tr><td>Number</td><td>num-sqrt, num-sqr, mean, median, modes</td><td>4, -1.2, 2/3</td></tr><tr><td>String</td><td>string-repeat, string-contains</td><td>"hello", "91"</td></tr><tr><td>Boolean</td><td>==, <, <=, >=, string-equal</td><td>true, false</td></tr><tr><td>Image</td><td>triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized, histogram, scatter-plot, lr-plot</td><td></td></tr><tr><td>Table</td><td>count, .row-n, .order-by, .filter, .build-column</td><td></td></tr></table>			Types	Functions	Values	Number	num-sqrt, num-sqr, mean, median, modes	4, -1.2, 2/3	String	string-repeat, string-contains	"hello", "91"	Boolean	==, <, <=, >=, string-equal	true, false	Image	triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized, histogram, scatter-plot, lr-plot	  	Table	count, .row-n, .order-by, .filter, .build-column	
Types	Functions	Values																			
Number	num-sqrt, num-sqr, mean, median, modes	4, -1.2, 2/3																			
String	string-repeat, string-contains	"hello", "91"																			
Boolean	==, <, <=, >=, string-equal	true, false																			
Image	triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized, histogram, scatter-plot, lr-plot	  																			
Table	count, .row-n, .order-by, .filter, .build-column																				

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

As good Data Scientists, the staff at the animal shelter is 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, non-representative sample of the population. This is the problem with surveying cat owners to find out which animal is most loved. *Remember that, in general, randomness is the key to obtaining unbiased samples!*
2. **Bias in the study design** - Suppose you survey a random sample of pet owners that includes representative numbers of both cat and dog owners. But you ask them 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** - Even if the selection is unbiased, sometimes outliers are so extreme that they shift the results of our analysis (such as the mean) in ways that don't represent the population as a whole. For example, if the shelter happened to house a 100-year-old tortoise, and summarized its animals' ages with the mean, this would inflate our perception of what age is typical.
4. **Sample error** - Even if the selection is unbiased and has a large enough sample size, sometimes outliers are so extreme that they shift the results of our analysis in ways that don't represent the population as a whole.
5. **Confounding variables** - The gathered data does not take into account other factors that might influence a relationship. For example, a study might conclude that cat owners are more environmentally conscious: they're more likely to use public transportation than dog owners. The confounding variable here could be urban versus rural dwelling: people who live in big cities are more likely to use public transportation and also more likely to own cats.

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

Investigate

On [Identifying Threats to Validity \(Page 84\)](#) and [Identifying Threats to Validity \(Page 85\)](#), 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?

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

You've already seen a number of ways that statistics can be misused:

1. Intentionally using the wrong chart
2. Changing the scale of a chart
3. Using the mean instead of the median with heavily-skewed data
4. Using the wrong language when describing a Linear Regression
5. Using a correlation to imply causation

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! \(Page 86\)](#), you'll find some deliberately misleading claims made by slimy Data Scientists. Can you figure out *why these claims should not 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 \(Page 87\)](#), 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?

Your Analysis

flexible

Overview

Students repeat the previous activity, this time applying it to their own dataset and interpreting their own results. **Note:** this activity can be done briefly as a homework assignment, but we recommend giving students an *additional class period* to work on this.

Launch

In every analysis, there are always threats to validity. It's important to always be upfront about what those threats are, so that anyone who reads your analysis can make their own decision.

Investigate

- Students should fill in the [Findings](#) portion of their Research Paper, discussing threats to validity and drawing conclusions from their linear regression results.
-

Additional Exercises:

- [Identifying Threats to Validity \(Part 1\)](#)
- [Identifying Threats to Validity \(Part 2\)](#)
- [Identifying Threats to Validity \(Part 3\)](#)
- Project: [Project: Threats to Validity](#)