Module 3 Video 1: Common Mistakes in Data Analysis

[00:00:00] Hi, it's Heather. And we're now in the data analysis section of our course on how to embed equity and ethics into your data journalism project. And today we're going to talk about the foremost common mistakes that people make in working with data analysis in data journalism.

[00:00:18] So the foremost common mistakes that people make in working with data for data journalism are the ecological fallacy, the causal fallacy, Simpson's paradox and the prosecutor's fallacy. And making those mistakes has deep and important real world consequences for ethics and equity. Often, the primary theme that runs through all four of these mistakes is that the data does not answer the question that's being asked. Often we get an answer from data, either data analysis that we do ourselves or data analysis that we get from someone else. And it often answers a very similar question to the core question in our data journalism project. And because it answers a very similar question and we really want an answer. Emotionally, we allow ourselves to believe that it is the answer we are looking for, but it is not the answer that we're looking for. And this really, really matters, because when we missed when we make these mistakes, people go to jail. People die. Harmful policies get made. And everyday citizens are given incorrect information. All things that are problematic for data journalism from an ethics and equity point of view.

[00:01:37] Why do we care? In journalism data design has a really strong set of emerging best practices and principles, and we need to catch up a little bit in terms of data analysis and helping make sure that we don't accidentally lie with our data.

[00:01:55] So is smoking cigarettes good for your health? Let's look at this data. The X axis, the flat, straight, horizontal axis is the average cigarette consumption by country. And the horizontal the up and down access is the life expectancy by country. And this data is real. Each one of these dots represents a country. And we do some statistical analysis on this. And the statistical analysis is correct. There's no problem here with the math. And we find out that there's a very strong relationship between average cigarette consumption by country and life expectancy for country. And it's actually a positive relationship. So what this means is that smoking cigarettes is going to add for smoking four cigarettes a year is going to add 10 years to your life.

[00:02:44] Is that what this data says? No, it is not. The problem isn't with the data. The problem isn't with the math. The problem is with the title. The title is an example of an ecological fallacy. Smoking is good for your health is not what this data is telling us. This data is telling us that countries with higher average cigarette consumption have longer life expectancies.

[00:03:13] That isn't so surprising because often countries with more cigarette consumption are countries where people have more money. In countries where people have more money are countries where people have a higher life expectancy. And the reason that this is an example of the ecological fallacy is because we are taking data that's at the country level. So each one of these dots represents a country. So this data is at the country level. And our temptation was to make an interpretation about individuals that the positive relationship between cigarette consumptions and life expectancies could be applied to individuals, even though the data is about countries.

[00:03:57] Now, this probably would have caught your attention and not getting gotten published the journalism article, because it goes against something that we pretty much have lived experience about that we know is not true. We would not publish an article that says smoking cigarettes is good for your health. However, lots of times in data journalism, we accidentally do publish stories that make an ecological fallacy mistake. One of the reasons for this is that as journalists, we usually want to tell stories about people. And the easiest and most commonly accessible data is usually data that is about groups of people, whether it's aggregated into communities or zip codes or states. If you have data that's been aggregated to any level, that's the level that you need to tell your story about. If you have a zip code data, you cannot tell data up, cannot tell data stories about individual people. It's very frustrating. But let this be an example that you are doing the equivalent of saying that smoking is good for your health. It's about the unit of analysis. I like to think of it as ecological fallacy, frogs, lily pads and ponds. Do you have data about frogs? Do you have data about lily pads or do you have data about ponds, whatever level you have data about? Is this the story that you can tell? And please avoid the temptation to tell stories about individuals when you have data about zip codes or states.