Module 3 Video 3: Simpson’s Paradox in Data Journalism

[00:00:00] The third is something you've probably heard of, this is called Simpson's paradox, and this is essentially where with groups you divide the data by ,dramatically changes the results. And this is a real world example of this. We were working in a community that was having a youth mental health crisis, and we were asked to help figure out where the mental health crisis was coming from. Where was the risk? This is an important story for a community that is watching its young people suffer. So if you take the dataset and you divide it by race, whether the young person is black or white. You find out that the black youth are at greater risk for a mental health crisis in this community. I should point out that this data is not generalizable to society as a whole.

[00:01:01] If you take the very same data set and you divide it by sex, male youth and female youth, we find out that it's the male youth that are the most at risk. If you take the same dataset and you divide it by whether or not the in-person is currently living in poverty, we find out that is the young people not living in poverty that are the most at risk. Again, this is not generalizable. This is in this community only. However, if you take the data and you look at all those aspects of a young person at the same time. So instead of just breaking out the data by white or black, male or female, or in poverty or not in poverty, if we do a more complicated three way analysis, we find out that it's the white males living in poverty who are actually the highest at risk.

[00:02:00] When we looked at it, one by one, it was the black children, the male children and the children not living in poverty. And then when we make the groups a different way, we find out that it's white males living in poverty. And what this is called is Simpson's paradox. And it's called a paradox. Instead of a fallacy, because all of these answers are actually correct. So if we want to know whether black children or white children are at risk, it's black children. That is the correct answer is not a fallacy. It's not a mistake. The correct answer is black children. If we want to use a gender lens and find out if it's male children or female children at risk. It's male children. That's a correct answer. And if we want to know whether it's children living in poverty or not living in poverty who are at risk, it's children not living in poverty. That is also a correct answer. And if we want to look at children with all of these social identities simultaneously. We get a very different answer, we find out that it's white males living in poverty.

[00:03:07] So all four of these answers are correct, just like in our very first video, whether the average classroom size is three or four. All of these answers are correct. So it's very important for you as a data journalist to understand that the way that you choose to group or slice and dice your data has a very strong influence over the results that you will get. So your assumptions going into the data analysis, or your priorities going into the data analysis are absolutely going to affect the results coming out of the data analysis. And that's called Simpson's paradox.