Module 3 Video 4: Prosector's Fallacy in Data Journalism

[00:00:00] The final fallacy that we're going to talk about today is the Prosecutor's afllacy. And this is called the Prosecutor's Fallacy, because it comes up a lot in court, but it's certainly not only in court, but in terms of data journalism, it's extraordinarily important that you understand some of the basics of probability and how probability works or you are very likely to accidentally get important data aspects of your story wrong.

[00:00:32] So we're just going to take an example of a woman in Britain named Sally Clark. Sally Clark was a lawyer. She was married to a lawyer. And she had a son in 1996. And he died very suddenly at home at eight weeks of age. They did a postmortem. And it wasn't conclusive, but essentially it looked like he died from natural causes. The next year sally Clark gave birth to another son and he'd also died very suddenly at home at about age of eight weeks. The postmortem on Harry was also inconclusive. It's extraordinarily complex to get a solid determination in many cases of infant death, but they wondered if it was death from shaking. They thought it was a little bit suspicious. And then they went back and they reexamined the deaths, the death of Christopher, Sally Clarke's first son. And she was put on trial and charged with the murder of both of her children.

[00:01:36] Now, what the prosecutor did, and this is why it's called the Prosecutor's Fallacy, is that the prosecutor got somebody to do the math to try and prove what the odds were for this happening as a way to get the jury to convict the jury or the judge to convict Sally Clark of murder because we had no conclusive evidence and there were no conclusive physical evidence. And so the mathematician that the prosecutor hired said, did some research and found that the odds of a sudden infant death in a healthy high income family in Britain was one in eight thousand five hundred. And so this relied on dividing infant death into three different categories. The first category was natural causes. The second category was murder. And the third category was Sudden Infant Death Syndrome. So because that's what Sally Clark's defense attorney was saying, that these were this was sudden infant death syndrome. She had not murdered these babies. So the prosecutor said the odds of a sudden infant death happening to a healthy income family is one in eight thousand five hundred. Then that mathematician said the probability of two infant deaths is the probability of one of those events multiplied by the probability of one of those events, because there were two. So the probability of two is the probability one times the probability of one. So if you multiply one in eight thousand five hundred by one in eight thousand five hundred, you get about one in 73 million. So this is what the prosecutor said. There was a one in 73 million chance that these deaths were Sudden Infant Death Syndrome and therefore they should convict Sally Clark. And in fact, they did convict Sally Clark based on this math. And then what happened was a whole bunch of other mathematicians stepped in and said that is not the correct math and we have to reopen this trial and think about this again. The reason that this is not correct, math is based on how probability works.

[00:04:03] So the chance of two sudden infant death syndrome is happening in the same household. Being one in 73 million. It's a correct nu number. It's a correct calculation, but it's a correct calculation that answers the incorrect question. The one in 73 million calculation answers the probability of the evidence. What were the probability that these deaths were SIDS? Given that she was innocent? So the probability of E. Given I so the probability of the evidence, given the fact that she was innocent, is what they actually calculated. So the one in 73 million. Is how likely it is for Sally Clark's, both Sally Clark's babies, to die from SIDS versus to die from other natural causes, it's not the probability of whether or not it was SIDS or murder. You can see this from the bar chart we have at the bottom. The chances of it being murder are yellow. The chances of it being SIDS are maroon. And the chance of it being all their natural causes are gray. And the prosecutor accidentally compared the chances of it being SIDS to the chances of it being any other natural cause. What we really need to do, of course. As the prosecutor or to use math to decide what the probability of this happening, is the probability of her innocence given to infant deaths of unknown cause. Go back to those slide minutes. I know this can be complicated. One in 73 million is the probability of E given I. So that's the probability of the evidence given that she is innocent.

[00:05:59] What we need is the probability of innocence, given the evidence. So the probability of innocence of Sally Clark, given that there are two infant deaths in her home of unknown causes.

So the probability of innocence is whether she's an innocent mother or a guilty mother. These two bars on the left in purple. We want to know whether Sally Clark is innocent or whether she's guilty. And if she's innocent there's two ways that those babies could die. They could have died of SIDS or they could have died of all or natural causes. If she's guilty, there's only one way that those babies could have died. And that's murder. So the probability for innocence, given the evidence, is the probability for innocence up here, divided by the combined probability that she's innocent and guilty. And here we use the bars differently, the probability of innocence is SIDS, divided by the combination of the probability that those death were SIDS plus murder. What is the probability if that is two in three? So the probability of Sally Clarke's innocence given to death of an unknown cause is two in three, not one in 73 million.

[00:07:21] So after the big group of mathematicians and statisticians stepped in, Sally Clark was released. But unfortunately, it was too late. She later essentially died of alcoholism. And so getting data wrong is very easy. And getting data wrong is very, very problematic.

[00:07:42] So we've gone through four key fallacies. And if you're going to use data in a data story, you need to double check your stories for these fallacies. The ecological fallacy. Answering a question, using data that's actually at a different unit of analysis or level. Causal fallacy, including or excluding variables changes the question. Simpson's paradox, answering a question that you think is important. Rather than letting the nuances of the data show you what's happening. And then the Prosecutors Fallacy, getting confused about the nature of conditional probability. All of these mistakes are very easy to make. They immediately impact the ethics and equity of your data story. So you need to either check them yourselves or get a friend or an expert or a colleague to check your data story for you.