How statistics can be misleading

Statistics are persuasive. So much so that people, organizations, and whole countries base some of their most important decisions on organized data.

But there's a problem with that. Any **set** of statistics might have something **lurking** inside it, something that can turn the results completely **upside down**.

For example, imagine you need to choose between two hospitals for an elderly **relative**'s surgery. **Out of** each hospital's last 1000 patient's, 900 survived at Hospital A, while only 800 survived at Hospital B. So it looks like Hospital A is the **better** choice.

But before you **make** your decision, remember that not all patients arrive at the hospital with the same level of health. And if we divide each hospital's last 1000 patients into those who arrived in good health and those who arrived in poor health, the picture starts to look very different.

Hospital A had only 100 patients who arrived in poor health, of which 30 survived. But Hospital B had 400, and they were able to save 210. So Hospital B is the better choice for patients who arrive at hospital in poor health, with a survival **rate** of 52.5%.

And what if your relative's health is good when she arrives at the hospital? Strangely enough, Hospital B is still the better choice, with a survival rate of over 98%. So how can Hospital A have a better **overall** survival rate if Hospital B has better survival rates for patients in each of the two groups?

What we've **stumbled** upon is a case of Simpson's paradox, where the same set of data can appear to show opposite **trends** depending on how it's grouped. This often **occurs** when aggregated data hides a conditional variable, sometimes known as a **lurking** variable, which is a hidden additional factor that significantly influences results. Here, the hidden factor is the relative proportion of patients who arrive in good or poor health.

Simpson's paradox isn't just a hypothetical scenario. It **pops up** from time to time in the real world, sometimes in important contexts.

One study in the UK appeared to show that smokers had a higher survival rate than non-smokers over a twenty-year time period. That is, until dividing the participants by age group showed that the non-smokers were significantly older on **average**, and **thus**, more **likely** to die during the **trial** period, precisely because they were living longer in general. Here, the age groups are the lurking variable, and are vital to correctly interpret the data.

In another example, an analysis of Florida's death penalty cases seemed to reveal no racial disparity in sentencing between black and white **defendants convicted** of murder. But dividing the cases by the race of the victim told a different story.

In **either** situation, black defendants were more likely to be **sentenced** to death. The **slightly** higher overall sentencing rate for white defendants was due to the fact that cases with white victims were more likely to **elicit** a death sentence than cases where the victim was black, and most murders occurred between people of the same race.

So how do we avoid **falling for** the paradox? Unfortunately, there's no **one-size-fits-all** answer. Data can be grouped and divided in any number of ways, and overall numbers may sometimes give a more **accurate** picture than data divided into **misleading** or arbitrary categories.

All we can do is carefully study the **actual** situations the statistics describe and consider whether lurking variables may be present. Otherwise, we leave ourselves vulnerable to those who would use data to manipulate others and promote their own **agendas**.

Mark Liddell – TED-Ed