Can we really trust our statistics? - A short essay on the Simpson's Paradox

Written by Charlie Tomlinson

It is widely accepted that in our world, "the facts don't lie". The phrase is used all throughout your life, from education, to occupation, to healthcare and to politics, and statistics at their core are both the means we use to justify many of these facts, and often times, the facts themselves.

But what if the facts did lie?

As quoted by the US National centre for educational statistics - and used as an example by the educational YouTube channel "minutephysics", The overall statistics and facts point to Wisconsin having a superior education system to Texas.

At first glance, saying Wisconsin's education system being better than that of Texas looks to be a fairly obvious assumption, with consistent higher standardised test results on average across a range of years.

YEAR	WISCONSIN RESULT	TEXAS RESULT
2009	157	150
2011	159	153
2015	159	156

The average reader would see these statistics, the raw facts that we base our opinions and decisions on, and would come to the conclusion that Texas should be more like Wisconsin in terms of their teaching methods, as Wisconsin clearly achieves better results.

However, if we delve a bit deeper into the statistics available, things start to become a bit less clear. This is where the Simpson's paradox comes into play.

We've already established that Wisconsin clearly achieves better results on average across the years, but if we split the average results by race and look at them in a separate table, we see a bit of a change.

RACE/ETHNICITY	WISCONSIN RESULT	TEXAS RESULT
Black	120	137
Hispanic	138	145
White	166	169

This is where confusion starts to arise. A person not aware of the Simpson's Paradox would see this statistics as falsified or completely irrelevant to the first lot. All Texas ethnic groups scored higher than all of that of Wisconsin's, but still score lower on average? How could this be the case?

The reason for this is simple, but often overlooked. In both cases, white students score higher than that of Hispanic students, and likewise with Hispanic students to black students. In this case, Proportionally, the number of white students to Hispanic and black students in Wisconsin is so different to the proportion of that in Texas (Being that Texas has a far higher percentage of black and Hispanic students), that the percentage average of the cohort on the whole becomes skewed, and favours the Wisconsin results.

This is known as the Simpson's Paradox, and is a phenomena where trend that occurs in groups can be completely removed or entirely reversed by combining these groups together.

Now that you are aware of this, and have analysed both tables of results yourself, I'm sure that your stances on the teaching or educational ability of both states have changed since the top of the page to the bottom of the page, but what about when this occurs in real life?

In our day to day lives, statistics are thrown at us every day in forms of charts and percentages, and play a huge role in forming our opinions, especially political ones, which is why we have to ask ourselves, **Do our facts really lie to us?**

This simple but often unnoticeable phenomena is actually incredibly dangerous, and can be used as a form of manipulation, either knowingly or unknowingly, without actually doing anything at all.

If I were to have made a claim at the beginning of this essay that Wisconsin had a superior educational system than that of Texas, with my cold hard proof of that first table, my chances of pushing your opinion toward agreeing with mine would be greatly improved. Why wouldn't you believe me after all? I've made a claim and then shown you seemingly obviously definitive evidence from a reputable government source to back me up.

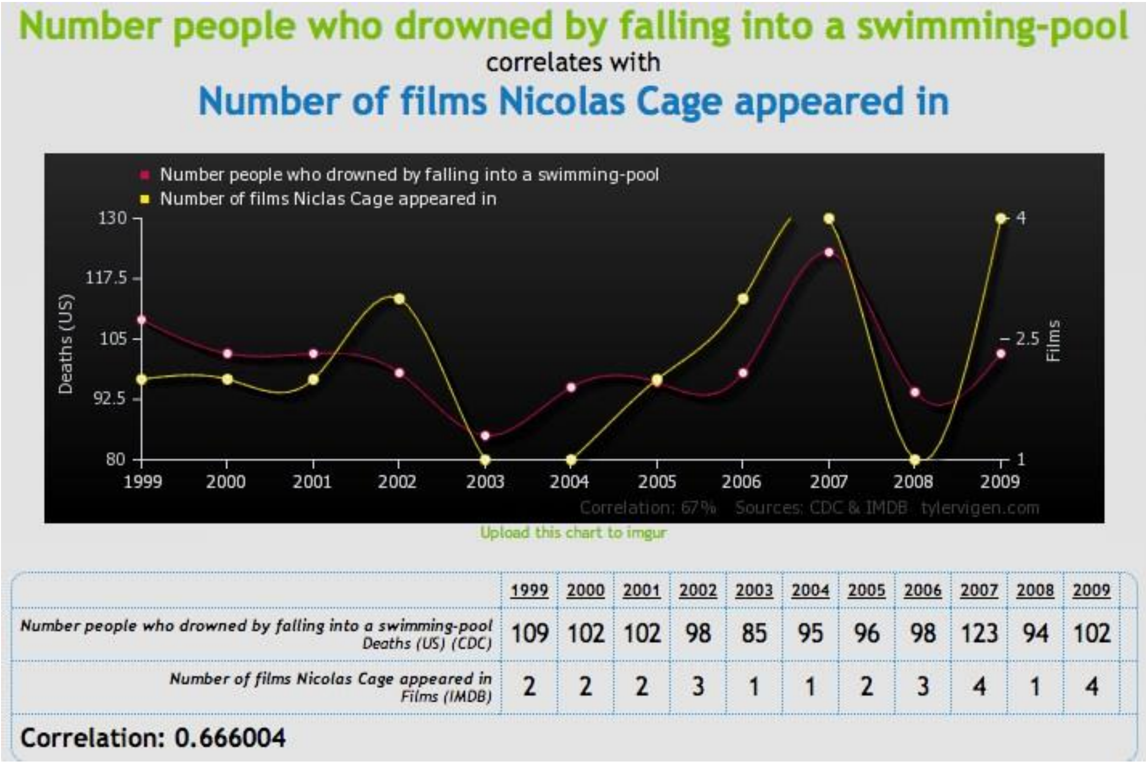
By leaving out that second table, which completely reverses the trend, you would have most likely placed your faith in the educational system of Wisconsin, and based on the evidence you had available to you, this would normally have been the correct assumption to make.

Now think to yourself, how often could this happen to you in day to day life, especially when it comes to forming your opinions? How often are you presented with accurate and truthful information, so base your opinion on that information?

We assume that this is the correct thing to do, because why wouldn't it be? It's common sense 101 to form our opinions based on fact. But what if there was other facts? What if there was other factors? What if we split up these broad statistics into their individual percentages by category?

This paradox is a shocking example of the more well known theory that "correlation does not equal causation". Factors may seem to link together so cleanly and so closely that they have to be true, and among those interested in statistics (or simply aware of these facts), has become a topic of comedy.

An example of an absurd case of this effect is shown in the graph below.



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool Deaths (US) (CDC)	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in Films (IMDB)	2	2	2	3	1	1	2	3	4	1	4

Correlation: 0.666004

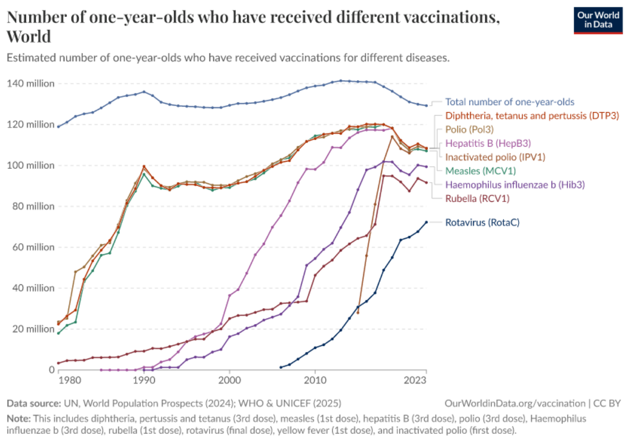
Obviously, Nicholas Cage has no impact on the number of people who drowned by falling into a swimming pool, but if we were to take the labels off the red and yellow lines and refer to them as "X' and "Y", many would assume them to be related as the correlation in the graph is so similar.

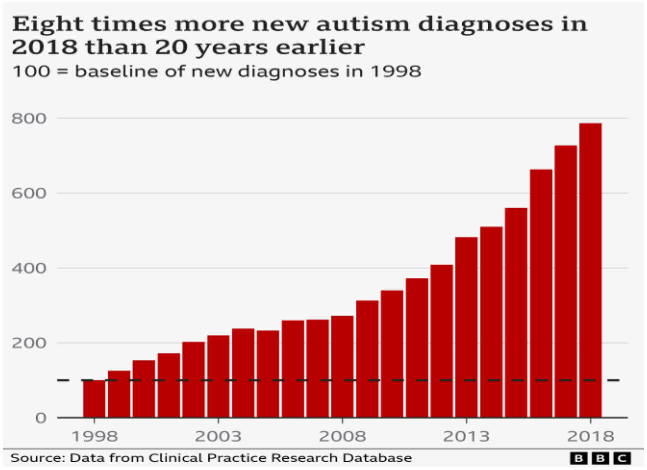
This is a widely known example of how statistics can be misinterpreted and become so misleading.

This particular graph results in no consequences as the vast majority of people can make the link that Nicholas Cage isn't running around pushing people into swimming pools to drown them, and that maybe, it's just a coincidence.

But when we put together some information that is less obviously not connected to each other, it can be a bit more difficult to distinguish if causation does in fact equal causality.

The graphs below show the number of children who have received vaccinations in the past 43 years (as of 2023) and the number of autism diagnoses in the past 20 years (as of 2018).





Based off similar graphs and statistics, a widespread and extremely dangerous misconception that vaccinations cause autism has plagued much of the western world, where this information is so freely accessible.

With this information being so easy to find, many see these graphs and make what they believe to be an accurate assumption, without being aware of the other factors within medicine, but since they are both in the medical field, many make the link between the two, and assume that correlation results in causality.

But it is not a total coincidence that these two graphs rise with each other so exponentially, unlike the one with Nicholas Cage and swimming pools, which are utterly and completely unrelated and total coincidences. In fact the reason for these graphs to rise side by side is the advances in medicine which affect them both, with more people getting vaccinated as we have a greater number of vaccines to a greater number of diseases, and more people getting autism diagnoses based off our greater understanding of mental health and psychological problems in medicine.

Without being aware of these external factors, it would be, for many people, safe to assume that the two are in fact related, which in the modern world, results in plenty of medical ignorance, as once people have made their mind up, it can be very difficult to convince them otherwise, especially when they have “evidence” which seemingly proves their point. This confusion between correlation and causation has become a plague in many countries, especially the USA, where even those in high positions of power, (some even in medicine!), cannot distinguish between the two.

This resulted in massive issues especially during the 2019-(formally)2023 crisis of Covid-19, where millions of mostly westerners refused to get vaccinated due to the misconception that these vaccines would harm them in other ways, or were being used as a form of government control.

This highlights how a lack of understanding of these statistical effects can be a truly dangerous situation, harming not only the person with the misunderstandings, but everyone around them as well.

Moving on from effects based on correlation and causation, the Simpson's effect has similar issues to the ones just described, especially in medical applications.

Now pretend that you are a patient at a hospital with kidney stones, you are in absolute agony, you can't even think straight, and you have to decide which treatment you want to go with. Take a look at this graph below, depicting the success as a percentage of 2 kidney stones treatments and make your decision.

Success of treatment for:	**TREATMENT A**	**TREATMENT B**
Small kidney stones	93%	87%
Large kidney stones	73%	69%

I'm assuming that in either case, you'd obviously pick the first of the treatments, it's clearly more effective in both scenarios, right?

This time, I'll include the numbers. Keep in mind what's been discussed previously and re-evaluate your decision.

Success of treatment for:	**TREATMENT A**	**TREATMENT B**
Small kidney stones	93% (81/87)	87% (234/270)
Large kidney stones	73% (192/263)	69% (55/80)

Have you changed your mind?

You're either sticking by your guns and agreeing with the percentages, or you've taken the time to do the math. (Or you're sitting in a third camp of changing your mind regardless, because it would be a bit pointless to include the same outcome twice).

By including the numbers and re-evaluating the percentages, we get a different idea of which treatment would be more effective.

Success of treatment for:	**TREATMENT A**	**TREATMENT B**
Combined kidney stones	78% (273/350)	83% (289/350)

I'm sure that as it is the focus of the discussion, you anticipated something like this might happen, but because of the Simpson's paradox, by now grouping the two categories of large and small kidney stones together, we get a completely new set of percentages.

In our hypothetical scenario, you're sat in a hospital in far too much pain to effectively contemplate this, which is most likely why the doctors wouldn't show it to you, and you're probably sat there now, without the agony of kidney stones, still unaware of which treatment would be more effective.

If you decided A? You are wrong.

If you decided B? You are still wrong.

Well, maybe you are. It's situational. Contrary to what you may have first thought, this information doesn't actually accurately reflect the effectiveness of these treatments, but more so shows us that the size of the kidney stone has a greater influence on the outcome that the treatment which was picked.

A larger proportion of people treated by method A had larger kidney stones than method B, and the same goes for method B, with the number of smaller kidney stones removed being greater than the number removed with method A.

If these two groups were equal in proportion, we would not have data which follows the Simpson's paradox.

You may be thinking now, "well why don't we just make all of the groups the same size in order to avoid the effects of this paradox", which in theory, would remedy this issue, but in the grand scheme of things, when it comes to statistics and studies, no two groups are going to have the same proportions and numbers unless we force them to, as seen earlier with the Educational statistics I discussed.

In a broad conclusion, all of statistics is incredibly situational and every study contains external factors and variables which we cannot factor into our results, be it numbers, percentages or graphs. There are countless paradoxes or theories which we can put to almost any result we get, but I believe that more effectively than any other, the Simpson's theory teaches us that we must delve deeper into any information we have, and ironically, not take our facts as facts. So to answer our earlier question that we must ask ourselves, "do our facts really lie to us?", Of course not. They're just numbers.

Our facts don't lie to us, they are merely just circumstantial. The takeaway we can pull from this is that we absolutely cannot make our conclusions based off our statistics without further analysis, and that we should never take "facts as facts".

The dangerous nature of this is that you can find numbers to justify almost every point you could make, if you just negate the underlying context. This can be obvious in some cases, such as with Nicholas Cage and swimming pools, or extremely subtle

This is prominent in politics, economics, marketing, and every aspect of your life which involves people telling you things, and in the fitting words of the controversial Rev. Thomas Malthus, we often "prostitute our understanding to your interests". Fitting in this context, as in our abundance of numbers, we will probably pick the ones which justify the ideas we already have.