Published on *American Council on Science and Health* (https://www.acsh.org)

How Big Data Has Created a Big Crisis in Science

By ACSH Staff — December 14, 2018

Credit: Storyblocks [1]

By Kai Zhang [2], *University of North Carolina at Chapel Hill [3]*

There's an increasing concern among scholars that, in many areas of science, famous published results tend to be impossible to reproduce.

This crisis can be severe. For example, in 2011, Bayer HealthCare reviewed 67 in-house projects [4] and found that they could replicate less than 25 percent. Furthermore, over two-thirds of the projects had major inconsistencies. More recently [5], in November, an investigation of 28 major psychology papers found that only half could be replicated.

Similar findings are reported across other fields, including medicine [6] and economics [7]. These striking results put the credibility of all scientists in deep trouble.

What is causing this big problem? There are many contributing factors. As a statistician, I see huge issues with the way science is done in the era of big data. The reproducibility crisis is driven in part by invalid statistical analyses that are from data-driven hypotheses – the opposite of how things are traditionally done.

Scientific method

In a classical experiment, the statistician and scientist first together frame a hypothesis. Then scientists conduct experiments to collect data, which are subsequently analyzed by statisticians.

A famous example of this process is the "lady tasting tea" story. [8] Back in the 1920s, at a party of academics, a woman claimed to be able to tell the difference in flavor if the tea or milk was added first in a cup. Statistician Ronald Fisher doubted that she had any such talent. He hypothesized that, out of eight cups of tea, prepared such that four cups had milk added first and the other four

cups had tea added first, the number of correct guesses would follow a probability model called the hypergeometric distribution [9].

Such an experiment was done with eight cups of tea sent to the lady in a random order – and, according to legend, she categorized all eight correctly. This was strong evidence against Fisher's hypothesis. The chances that the lady had achieved all correct answers through random guessing was an extremely low 1.4 percent.

That process – hypothesize, then gather data, then analyze – is rare in the big data era. Today's technology can collect huge amounts of data [10], on the order of 2.5 exabytes a day.

While this is a good thing, science often develops at a much slower speed, and so researchers may not know how to dictate the right hypothesis in the analysis of data. For example, scientists can now collect tens of thousands of gene expressions from people, but it is very hard to decide whether one should include or exclude a particular gene in the hypothesis. In this case, it is appealing to form the hypothesis based on the data. While such hypotheses may appear compelling, conventional inferences from these hypotheses are generally invalid. This is because, in contrast to the "lady tasting tea" process, the order of building the hypothesis and seeing the data has reversed.

Data problems

Why can this reversion cause a big problem? Let's consider a big data version of the tea lady — a "100 ladies tasting tea" example.

Suppose there are 100 ladies who cannot tell the difference between the tea, but take a guess after tasting all eight cups. There's actually a 75.6 percent chance that at least one lady would luckily guess all of the orders correctly.

Now, if a scientist saw some lady with a surprising outcome of all correct cups and ran a statistical analysis for her with the same hypergeometric distribution above, then he might conclude that this lady had the ability to tell the difference between each cup. But this result isn't reproducible. If the same lady did the experiment again she would very likely sort the cups wrongly – not getting as lucky as her first time – since she couldn't really tell the difference between them.

This small example illustrates how scientists can "luckily" see interesting but spurious signals from a dataset. They may formulate hypotheses after these signals, then use the same dataset to draw the conclusions, claiming these signals are real. It may be a while before they discover that their conclusions are not reproducible. This problem is particularly common in big data analysis [11] due to the large size of data, just by chance some spurious signals may "luckily" occur.

What' worse, this process may allow scientists to manipulate the data [12] to produce the most publishable result. Statisticians joke [13] about such a practice: "If we torture data hard enough, they will tell you something." However, is this "something" valid and reproducible? Probably not.

Stronger analyses

How can scientists avoid the above problem and achieve reproducible results in big data analysis? The answer is simple: Be more careful.

If scientists want reproducible results from data-driven hypotheses, then they need to carefully take the data-driven process into account in the analysis. Statisticians need to design new procedures that provide valid inferences. There are a few already underway [14].

Statistics is about the optimal way to extract information from data. By this nature, it is a field that evolves with the evolution of data. The problems of the big data era are just one example of such evolution. I think that scientists should embrace these changes, as they will lead to opportunities to develop of novel statistical techniques, which will in turn provide valid and interesting scientific discoveries.

***

Kai Zhang [2], Associate Professor of Statistics and Operations Research, *University of North Carolina at Chapel Hill* [3]

Source URL: https://www.acsh.org/news/2018/12/14/how-big-data-has-created-big-crisis-science-13668

Links

[1] https://www.storyblocks.com/stock-image/businessperson-studying-electronic-data-in-digital-tablet-bq1ekjwp7obj6gt4d17
[2] https://theconversation.com/profiles/kai-zhang-549921
[3] http://theconversation.com/institutions/university-of-north-carolina-at-chapel-hill-1353
[4] https://doi.org/10.1038/nrd3439-c1
[5] https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/
[6] https://doi.org/10.1371/journal.pmed.0020124
[7] https://doi.org/10.1111/ecoj.12461
[8] https://io9.gizmodo.com/how-a-tea-party-turned-into-a-scientific-legend-1706697488
[9] http://mathworld.wolfram.com/HypergeometricDistribution.html
[10] ftp://public.dhe.ibm.com/software/data/sw-library/bda/zone/index.html
[11] https://doi.org/10.1109/TIT.2017.2700202
[12] https://doi.org/10.1371/journal.pone.0005738
[13] http://books.wwnorton.com/books/How-to-Lie-with-Statistics/
[14] https://projecteuclid.org/euclid.aos/1369836961
[15] http://theconversation.com
[16] https://theconversation.com/how-big-data-has-created-a-big-crisis-in-science-102835