Article from

Predictive Analytics and Futurism

December 2019

Big Data—You've Rocked My World!

By Dorothy Andrews

We can no longer be sure of what we believe thanks to big data. The advent of big data has caused researchers to question the soundness of age-old sampling techniques and scientific methods,[1] the chain of custody and stability of big data,[2] the necessity for its neatness[3] and the need for causation over correlation.[4] The following is my appreciation for these revelations.

In statistics, a metric is "unbiased" if its mathematical expectation based on a sample equals its population equivalent. If this relationship holds, then there is no need to calculate the metric on the entire population in order to make inferences about the population. The sample metric will do. Historically, it has been cheaper to run statistical tests on samples rather than on entire populations, as Cukier and Mayer-Schoenberger[5] would agree. Depending on the experiment, however, samples are still preferred to running experiments on entire populations. Fast forward to the age of big data and big data analytics and we are seeing many analyses being performed on populations. In these cases, there is no need to question "unbiasedness" in the results because the sample is the population, N = All.[6] One would think the scientific community would be more excited than they are about being able to use populations over samples but that does not seem to be the case. Big data is "destabilizing" their models and systems,[7] forcing them to develop new approaches to solving problems before there are willing or able to do so.

> Many statistical texts exist that prescribe techniques for handling "messy data."

There other considerations that raise concerns over the future applicability of results even from population sized datasets. A major concern is that "real life" unlikely reflects all the variations the future may hold and, in those cases, simulating data to anticipate future scenarios is an often employed technique. This poses risks, which are significant, for interpreting the results, but such discussion is beyond the scope of this writing.

Plantin et al. point to a compromise in the "chain of custody" of data, which they call the "control zone."[8] They express concern for the integrity of data collected by individuals lacking "traditional scientific credentials"[9] and affiliations with respected institutions. (One must wonder how they feel about Nate Silver basing his election predictions, in part, on Yahoo polls.) Despite their views sounding a bit elitist, it is important to apply sound data collection and governance controls to minimize errors and biases in results to ensure results are stable over time.

The well-known 80/20 rule as applied to model building means about 80 percent of the effort in building a model is spent on cleaning and scrubbing the data and about 20 percent of the effort is spent building model code and results, and validating the results. Many statistical texts exist that prescribe techniques for handling "messy data." For example, when data is missing in a field in a record, one technique is to estimate the missing value with the mean, median or mode of existing values of the field to prevent having to exclude the record from the calculation of a metric. Cukier and Mayer-Schoenberger[10] highlight Google Flu Trends (GFT)[11] as an example where messy data was tolerated because the dataset was big. The GFT was a flu-tracking system, grounded in big data, that was meant to predict influenza outbreaks. Its predictions outnumbered those of the Centers for Disease Control and Prevention (CDC) for all the wrong reasons. The researchers point to Google's willingness to tolerate some "messiness"[12] in their big data because of their view that volume trumps messiness in detecting patterns in data. In effect, Google was saying all the messiness will be lost in the decimal places.

All Google really did was swap causation for correlation, according to the researchers. There is a lot of that going around, according to Barrowman,[13] and not for the betterment of analytics. He talks about one of my heroes in statistics, Ronald A. Fisher, father of modern-day statistics and father-in-law to my other statistical hero, George E.P. Box. Barrowman points out Fisher was skeptical of the data and the research done linking smoking to lung cancer,

believing such linkage was spurious at best. To be clear, Fisher questioned the correlational basis being advanced to support the link, not the possibility of the linkage. Fisher spent a significant portion of his career ferreting out "spurious correlations," creating a "cottage industry,"[14] according to Barrowman. Tyler Vigen has profited from this industry, writing books and creating websites on the subject. This writer employs his website[15] when presenting on predictive modeling issues involving correlations. It is particularly amusing to find a spurious correlation relevant to the audience, like how the number of lawyers in Iowa is positively corelated with the number of days of sunshine in the state. It is a spurious relationship, but highly amusing.

Barrowman[16] provides insights to some excellent tools useful in explaining causal relationships, such as path analysis, structural equation modeling, counterfactual analysis, instrumental variables analysis and directed acyclic graphs that no data scientist should live without. His discussion on selection bias is particularly relevant. In modeling, it is just as important to understand the data that is excluded from a sample as it is to understand the data that is included. Statistically significant is an aberration of the data included in the modeling dataset. This aberration is directly related to selection bias.

O'Neil[17] has made it her mission to become an evangelist of big data skepticism and West[18] is indeed a disciple. He sees the need for the data scientists-in-training to attend Sunday services to soak in the gospel of data skepticism preached from the pulpit by Her Holiness O'Neil. West[19] is critical of upcoming data scientists being too focused on techniques and not paying enough attention to the social and ethical implications of the results of their analyses. I am sold on O'Neil's teachings, too. In *On Being a Data Skeptic*, O'Neil defines a skeptic as "someone who maintains a consistently inquisitive attitude toward facts, opinions, or (especially) beliefs stated as facts. A skeptic asks questions when confronted with a claim that has been taken for granted."[20] Further, she exclaims a truly outstanding data scientist knows how to put "science" in the phrase data science. Many in academia feel the art of designing balanced and unbiased modeling datasets is being lost because data scientists are blinded by their quest for statistical significance. The result of this blindness is modelers who forget about what can go wrong with their models and where they can fail. The model does not speak the truth if it is founded on a poorly designed dataset.

O'Neil[21] gives some significant insights regarding data blindness, which she has termed "The Measurement Addiction." This addiction problem creates four hinderances to skepticism. The first hinderance is an addiction to metrics since they are grounded in mathematics, which is perceived as hard, objective, logical, axiomatic and trustworthy. Non-skeptics are unlikely to question the appropriateness of metrics used to assess a model

because of this perception. The second hinderance is an over focus on numbers and not enough on behavior. It is important not to confuse correlation with causation. Causation is the root of all behavior, not a p-value. The third hinderance is incorrectly framing the problem. It is important to have your model peer reviewed as a check that the correct mathematics has been applied to the problem to minimize model risk. Modeling assumptions should be kept at a minimum to prevent biasing the range of results. Finally, the fourth hinderance is ignoring perverse incentives. Models naturally beg for gaming because they cannot account for all possible contingencies modeled phenomenon respond to. This is an area O'Neil says is sorely ignored by modelers. The models most susceptible to gaming are those that heavily utilize proxy variables and assumptions. Proxy assumptions are often used to model missing data. Where data is missing, it is worth the effort to have data corrected at its source before modeling. Campbell's law summarizes the impact of proxies quite poignantly. It states, "The more any [proxy] quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."[22] This statement generalizes with the removal of the word "social." The message to actuaries, especially those working as or with data scientists, is it is important to identify and resolve weakness in data, big or small, to increase its value and reliability to analytical modeling.

The last bit of gospel O'Neil would say is critical to recognize is the wider cultural consequences of models. The Problem of Externalities is the modeler's problem, according to O'Neil. In her view, modelers have a responsibility to ensure the external effects of their models are benign, that the positive effects outweigh the negative effects, or be subject to the heavy hand of government regulation. ■

Dorothy L. Andrews, ASA, MAAA, CSPA, is the chief behavioral data scientist for Insurance Strategies Consulting LLC. She can be reached at *dorothylandrews@msn.com*.

ENDNOTES

1 Cukier, Kenneth, and Viktor Mayer-Schoenberger. 2013. The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs* 92, no. 3:28–40; Tufekci, Zeynep. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. Presented at the Eighth International Conference on Weblogs and Social Media (ICWSM), June 2–4, 2014, Ann Arbor, Michigan.

2 Plantin, Jean-Christophe, Carl Lagoze, Paul Edwards and Christian Sandvig. 2017. Big Data is Not About Size: When Data Transforms Scholarship. In Ouvrir, Partager, Réutiliser: Regards critiques sur les données numériques, 128–48. Paris: Éditions de la Maison des Sciences de l'Homme.

3 Cukier and Mayer-Schoenberger, The Rise of Big Data.

4 Barrowman, Nick. 2014. Correlation, Causation, and Confusion. *The New Atlantis* Summer/Fall:23–44; Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, no. 6176:1203–05.

5 Cukier and Mayer-Schoenberger, The Rise of Big Data.

6 Ibid., 30.

7 Plantin, et al., Big Data is Not About Size, 14.

8 Ibid., 6.

9 Ibid., 10

10 Cukier and Mayer-Schoenberger, The Rise of Big Data.

11 Lazer, et al., The Parable of Google Flu.

12 Cukier and Mayer-Schoenberger, The Rise of Big Data, 33

13 Barrowman, Correlation, Causation, and Confusion.

14 Barrowman, Correlation, Causation, and Confusion, 25.

15 *www.spuriouscorrelations.com*.

16 Barrowman, Correlation, Causation, and Confusion.

17 O'Neil, Cathy. The Rise of Big Data, Big Brother. *Mathbabe: Exploring and Venting About Quantitative Issues* (blog). May 2, 2013. *https://mathbabe.org/2013/05/02/the-rise-of-big-data-big-brother*.

18 West, Jevin. Calling B.S. in the Age of Data Science Euphoria. Presented at National Academies of Sciences, Engineering, and Medicine's Integrating Ethics and Privacy Concerns into Data Science Education meeting and webcast Dec. 8, 2017, Washington, D.C. *https://vimeo.com/250857594?ref=em-share*.

19 Ibid.

20 O'Neil, Cathy. 2014. *On Being a Data Skeptic*. Sebastopol, California: O'Reilly Media, Inc., 1.

21 Ibid.

22 Campbell, Donald. 1979. Assessing the Impact of Planned Social Change. *Evaluation and Program Planning* 2, no. 1:49. *https://doi.org/10.1016/0149-7189(79)90048-X*.