ASA
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics
NEWS
732 North Washington Street • Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org

STATISTICS EDUCATION, EVIDENCE-BASED DATA ANALYSIS PRACTICES NEEDED TO FIGHT REPRODUCIBILITY CRISIS IN SCIENCE

Lack of analytical skills means scientific findings are neither replicable nor reproducible

ALEXANDRIA VA, JUNE 16, 2015 – Dramatic increases in data science education coupled with robust evidence-based data analysis practices could stop the scientific research reproducibility and replication crisis before the issue permanently damages science's credibility, asserts Roger D. Peng in an article in the newly released issue of *Significance* magazine.

"Much the same way that epidemiologist John Snow helped end a London cholera epidemic by convincing officials to remove the handle of an infected water pump, we have an opportunity to attack the crisis of scientific reproducibility at its source," wrote Peng, who is associate professor of biostatistics at the Johns Hopkins Bloomberg School of Public Health.

In his article titled "The Reproducibility Crisis in Science"—published in the June issue of *Significance*, a statistics-focused, public-oriented magazine published jointly by the American Statistical Association (ASA) and Royal Statistical Society—Peng attributes the crisis to the explosion in the amount of data available to researchers and their comparative lack of analytical skills necessary to find meaning in the data.

"Data follow us everywhere, and analyzing them has become essential for all kinds of decision-making. Yet, while our ability to generate data has grown dramatically, our ability to understand them has not developed at the same rate," he wrote.

This analytics shortcoming has led to some significant "public failings of reproducibility," as Peng describes them, across a range of scientific disciplines, including cancer genomics, clinical medicine and economics.

Perhaps the most recent infamous example is a Duke University cancer research project in 2006 in which researchers published a paper claiming they had built an algorithm using genomic microarray data that predicted which cancer patients would respond to chemotherapy. A subsequent attempt to reproduce the results found a morass of poorly conducted data analyses with errors ranging from trivial and strange to devastating. The original study was retracted by *Nature Medicine* in 2011.

"The common thread between each of these public failings was the poor or questionable quality of the original analysis. The errors that were made showed a lack of judgement, training and quality control," wrote Peng.

Peng said to improve the quality of data analysis in science, stakeholders need to go beyond the call for reproducibility and increase the number of trained data analysts in the scientific community and identify statistical software and tools proven to improve study reproducibility and replicability. These latter items must be moderately robust to user error, noted Peng.

“If we could prevent problematic data analyses from being conducted, we could substantially reduce the burden on the [peer review] community of having to evaluate an increasingly heterogeneous and complex population of studies and research findings,” asserted Peng.

Unfortunately, most scientists receive basic to moderate training in data analysis, creating the potential for generating individuals with enough skill to perform data analysis, but without enough knowledge to prevent data mistakes.

To improve the global robustness of scientific data analysis, we must take a two-pronged approach and couple massive-scale education efforts with the identification of data-analytic strategies that are reproducible and replicable in the hands of basic or intermediate data analysts, explained Peng.

Peng said a fundamental component of scaling up data science education is performing empirical studies to identify statistical methods, analysis plans and software that lead to increased replicability and reproducibility by scientists.

“We call this approach ‘evidence-based data analysis,’” described Peng. “Just as evidence-based medicine applies the scientific method to the practice of medicine, evidence-based data analysis applies the scientific method to the practice of data analysis. Combining massive-scale education with evidence-based data analysis can allow us to quickly test data-analytic practices in a population most at risk for data analytics mistakes.”

***About* Significance:**

Significance explains how and why statistics contributes in many areas of life, science, government and commerce. Its articles are written by statisticians for anyone with an interest in the analysis and interpretation of data. The magazine’s writers and editors challenge myths, provide a unique perspective on the stories of the day and use statistics to answer society’s most difficult questions.

About the American Statistical Association:

The American Statistical Association is the world’s largest community of statisticians and the second-oldest continuously operating professional society in the United States. Its members serve in industry, government and academia in more than 90 countries, advancing research and promoting sound statistical practice to inform public policy and improve human welfare. For additional information about the American Statistical Association, please visit the ASA website at www.amstat.org.

###

For more information or to interview the article author:
Roger D. Peng
Office: (410) 955-2468
Email: rdpeng@gmail.com

Jeffrey A. Myers
Office: (703) 684-1221, Ext. 1865; Mobile: (540) 623-7777
Email: Jeffrey@amstat.org