PERSPECTIVES

Invited article
Original version
DOI: http://dx.doi.org/10.1590/S0034-759020190609

PLUS ÇA CHANGE, PLUS C'EST LA MÊME CHOSE

In 1572, a single data point, Tycho's Supernova, showed that contrary to the accepted paradigm at the time, the heavens did indeed change (Wootton , 2015). Less than 40 years later, in 1610, Galileo Galilei published his sensational findings in *Sidereus Nuncius*, a short treatise that demonstrated the existence of stars not seen by the naked eye and revealed the nature of the Milky Way (Galilei, 1610). Since then, data analysis has been central to scientific research and examples of its use in solving important and difficult problems have multiplied. At the age of 24, Carl Friedrich Gauss (1809) used least squares to correctly predict Ceres' position in 1801 after it emerged from behind the Sun's glare. A simple spatial analysis identified the source of the Broad Street cholera outbreak in London in 1854 (Snow, 1855). Between 1856 and 1863, careful estimation of frequencies allowed Gregor Mendel (1866) to determine the basic rules of heredity of physical traits in plants. In the late 1940's, a large retrospective study led by Richard Doll and A. Bradford Hill (1950) demonstrated the strong link between smoking and lung cancer.

The steady development and refinement of new statistical methods after 1880 by Galton, Student, Fisher, and others allowed for a broader range of applications of data analysis methods in industry and business. Ideas and methods developed and popularized by Shewhart, Deming, and others made statistical quality control an integral part of the industrial manufacturing process. This also incorporated the use of modern experiment design after the war. Inexpensive computing power, automated data collection, and the development of some general and flexible data analysis software—especially *R*, which is both comprehensive and free—greatly expanded the spectrum of applications. Consequentially, the era of "Big Data" was born. If much of what is published in the press is to be believed, Big Data will solve critical problems in areas as disparate as medical diagnostics, credit evaluation, weather forecast, and facial recognition. We will have much better products and services along with a much deeper understanding of physical and cultural processes as a result of ever larger data sets and the modern computer's intensive methods.

While making predictions is a difficult business, I am sure the key issues we will face in applying statistical analysis methods to business problems in the era of Big Data will continue to be the same ones we have been dealing with for decades. Data analysis still is about information, insights, and conclusions. Data is used to tell a story; analytical thinking is required in the evaluation of the relationship between data and conclusions. In regard to that evaluation, the three big challenges, in order of importance, have been and will continue to be overstating statistical significance; lack of reproducibility; and the avoidance of providing exact answers, but to the wrong questions.

Statistical significance is viewed by many as the gold standard. Significance, simply put, is the probability that an observed set of observations is the result of random fluctuation. If the calculated value is small—usually less than 5%—one is led to believe that some factors must exist

FLAVIO BARTMANN[1]
fcb2122@columbia.edu
ORCID: 0000-0002-9308-3049

[1]Columbia University, School of International and Public Affairs, New York, NY, United States of America

that explain the non-random behavior. Behind the calculation of the *p*-value (the usual measure of significance), there is always a complex process of study design, data collection, and data analysis, including model selection. The level of significance claimed is only accurate if all the elements of the process are done well and if the model chosen provides a reasonable description of reality. However, this is rarely the case. The situation is particularly dire in the business context where the models used can be very complex. These intricate models are typically regression models with several explanatory variables where the relationships with the response are assumed to be linear and the calculation of the estimated coefficients' significance is frequently meaningless. Poor design and flawed data collection only add to the troubles. The problem is so serious that a large movement to demote or even eliminate p-values has gathered much support among statisticians (McShane, Gal, Gelman, Robert, and Tackett, 2019).

The second big problem with data analyses is that many—perhaps a majority—are not reproducible (Ioannidis, 2005). The most common cause is one or more flaws in the design of the investigation, be it an experiment or a survey. Flaws in the data collection process, use of inappropriate methods, and fraud are also common occurrences that invalidate studies. In scientific research, studies can be re-done and some degree of self-correction can be achieved. The problem is much more serious within industry environments where business decisions are frequently data-driven and time sensitive, making replication very rare. A poorly carried out market or feasibility study can have costly consequences. Remember Apple's Newton platform?

The third problem might be the most serious and the most difficult to address. John Tukey (1962) used to say that "an approximate answer to the right question is worth a great deal more than a precise answer to the wrong question." The classical story, told numerous times by statisticians everywhere (or "data scientists" in modern parlance), but even more frequently in the corridors of the Department of Statistics at Columbia University, is about the reinforcement of British planes used in the bombings of Germany late in the Second World War. A large proportion of the planes were lost due to the German anti-aircraft fire and accordingly the decision was made to bolster them with armor. The most important concern to address for those on this project was which location was the best place to put the additional armor. The returning planes were carefully examined, and a proposal was made to put the extra armor in the areas that had received the most damage by fire. Interestingly, it would have proved to be a fatal mistake if this decision had been carried out. The correct variable that should have been reviewed was not regarding the planes that could be examined, but about the planes that had not returned. Fortunately, Abraham Wald cleverly suggested that perhaps areas where the returning planes were unscathed were those that should have been reinforced.

Data sets, even those with terabytes of records, are merely the raw material of knowledge. Today, almost everything can be monitored and measured but the key challenge continues to be our ability to use and analyze these data sets and to make sense of them in order to tell the real story.

REFERENCES

Doll, R., & Bradford Hill, A. (1950) Smoking and Carcinoma of the Lung. *British Medical Journal*, *2*(4682), 739–-748. doi:10.1136/bmj.2.4682.739

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hamburg: Friedrich Perthes and I.H. Besser.

Galilei, G. (1610). *Sidereus Nuncius*. Thomam Baglionum: Venice.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

McShane, B. B., Gal, D., Gelman, A., Robert C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245. doi:10.1080/00031305.2018.1527253

Mendel, G. (1866). Versuche über Plflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865 (pp. 3-47). *Abhandlungen*.

Snow, J. (1855). *On the Mode of communication of cholera*. London, UK: John Churchill.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33*, 1-67. doi:10.1214/aoms/1177704711

Wootton, D. (2015). *The invention of science: A new history of the scientific revolution*. London, UK: Allen Lane.

  © RAE | São Paulo | **59(6)** | **November-December 2019** | 433-434 ISSN 0034-7590; eISSN 2178-938X