TEST (2019) 28:345–348
https://doi.org/10.1007/s11749-019-00643-9

DISCUSSION

Comments on: Data science, big data and statistics

Abigael C. Nachtsheim[1] · **John Stufken**[1]

Published online: 8 April 2019

Mathematics Subject Classification 62 (Statistics)

We would first like to thank the authors for writing this thought-provoking article on such an important topic. This piece explores the intersection of data science and statistics in a world increasingly concerned with the analysis of massive data sets. The authors consider seven areas in which the increased presence of big data may alter and expand traditional statistical approaches; they give an overview of the emerging field of data science; and they provide two examples of statistical analyses driven by big data. Finally, they provide some insight into the future of statistics, imagining it as one piece of the multi-faceted and evolving field of data science.

The authors stress the role that big data has and will continue to have in the development of data science as a field of study, and in the ways that statistics as a discipline must adapt. We would like to note that the field of statistics already has a long history of evolution. The study of statistics dates at least to the late 18th century, when scholars like Laplace and Legendre applied mathematical statistics to problems in the field of astronomy. Through the 19th century, probability theory was further developed and applied to problems in the social sciences by mathematicians like Gauss and Poisson. Entering the 20th century and the modern era of statistics, scholars developed the analysis of variance and regression to study heredity (Stigler 1986). The pace of change only increased in the 20th century, moving beyond the realm of genetics to tackle problems from agriculture to manufacturing and from marketing to medicine. Thus, the field of statistics grew out of the desire to answer complicated questions in new and innovative ways. We see no reason to doubt that, as it always has, the field of statistics will continue to evolve in the presence of new problems, regardless of the form that they take.

This comment refers to the invited paper available at: https://doi.org/10.1007/s11749-019-00651-9.

✉ Abigael C. Nachtsheim
anachtsh@asu.edu

[1] School of Mathematical and Statistical Sciences, Arizona State University, Tempe, USA

As the authors note, many problems of import today revolve around the presence of large data sets. While big data is a much-discussed topic, we do not believe that it is a well-defined one. Big data may mean a large number of observations, or a large number of covariates, or both. It might mean covariates that outnumber observations, or observations that outnumber covariates. Further, the size of big data can itself fluctuate drastically, implying millions to some but trillions to others. Since the form of the data greatly affects the statistical questions we may pose and the methodologies we may employ, we believe that seeking to more clearly define what we mean when we talk about big data is a conversation worth having within the field of statistics.

Nevertheless, whatever form big data may take, we agree with the authors that it is important to consider how our field of statistics will continue to adjust in its presence. However, in our view any discussion of big data must include strong words of caution. The authors discuss some pitfalls of big data. They mention briefly the problem of selection bias and they give particular attention to the problem of multiple testing and false discovery rate. We believe, though, that additional warning could prove useful. We fear that it is too easy to lose sight of the downsides of big data when faced with the sheer quantity and availability of it. However, and especially as statisticians, we must stress data quality over quantity.

Meng (2018) gave a thorough argument against the overzealous use of large datasets. With his introduction of the data defect index, a measure of data quality, he quantifies the ways in which even very large amounts of observational data can be inferior to small datasets collected via a simple random sample. For instance, he shows that an observational dataset covering 50% of US voters, or 115 million observations, is no better than a simple random sample of size 400, when the observational dataset includes even an "extremely modest" sampling bias. Moreover, he shows that if the quality of the data is poor, the larger the population is the more severe the effect of the low-quality data. That is, if the number of US voters grew, then the effective sample size in terms of the simple random sample would be even less than 400. Therefore, Meng shows that counting on the size of big data to protect from its low quality is a losing strategy.

That is why we would like to stress the importance of high-quality, small data. In particular, let us not forget the powerful advantages of randomized, controlled studies. By their nature, studies of this type yield small but exceptionally high-quality data. Crucially, randomized, controlled studies allow researchers to draw conclusions about causal relations. With large, observational datasets on hand, practitioners often hope to build predictive models. However, with observational data—even when it is relatively high-quality—questions of causality must be approached with caution.

Ioannidis (2005) shows how observational studies can fail to produce beneficial results. After evaluating 49 studies, each with at least 1000 citations, from the recent medical literature, Ioannidis found that, in follow-up replications, 90% of experimental study results were confirmed, while only 20% of observational study results could be validated. Thus, experimental data is necessary to effectively advance scientific questions that hinge on questions of causality. Holland (1986) reminds us that "randomized experiments have transformed many branches of science, and the early proponents of such studies were the same statisticians who founded the modern era of our field."

As our field moves into yet another new era, we should not lose sight of the critical importance of high-quality, small data.

As trained statisticians, we have a unique skill set that can help us accomplish at least two important tasks: (1) avoid the blind use of poor-quality data (as Meng did); and (2) effectively make use of big data when appropriate. In the two case studies included in this article, the authors illustrated the important ways in which statistical tools can be applied powerfully to tackle big data problems. Therefore, we disagree that statisticians' skills are not well suited to the world of big data. In fact, as Donoho (2017) points out, the term "statistics" was coined in the context of census data. That is, the modern notion of statistics grew out of the need to analyze and make sense of very large datasets. Further, just as we need to be careful to avoid elevating quantity over quality in data collection, we should not lose sight of the importance of the quality of data analysis in the face of big data. As Donoho argues, we are increasingly constrained in our analyses by algorithms that can handle gigantic quantities of data. He writes that with these constraints, "one inevitably tends to adopt inferential approaches which would have been considered rudimentary or even inappropriate in olden times. Such coping …holds us back from data analysis strategies that we would otherwise eagerly pursue." We argue that allowing constraints imposed by the size of the dataset to weaken the sophistication of the statistical analyses is short-sighted. Further, it is not always necessary.

Rather than using the large dataset to carry out the analysis, statisticians can make use of subsampling schemes to select a subset of the full data and conduct the statistical analysis using this subdata alone. A statistician may approach subdata selection in a variety of ways. One straightforward approach is by taking a simple random sample from the full data. However, it can be shown that this approach is not always the best (Ma et al. 2015; Wang et al. 2018). Instead, selecting the subsample judiciously can give better results. For instance, a statistician may choose to select the subsample to optimize a specified criterion, with the goal of improving properties of the resulting model estimators. Several related methods have been proposed, including leverage-based subsampling methods and the Information-Based Optimal Subdata Selection method (IBOSS) (Ma et al. 2015; Wang et al. 2018). Under D-optimality, IBOSS selects observations from the full data to be included in the subdata in order to maximize the determinant of the information matrix for the model parameters, given a fixed subdata size. This strategy gives several advantages over both simple random sampling and over other methods that seek to optimize a given criterion. First, parameter estimates obtained from the subdata selected via IBOSS are unbiased. Second, unlike the leverage-based methods, the IBOSS subdata selection algorithm is computationally efficient, running in linear time. Finally, depending on the distribution of the covariates, the variance of the parameter estimates that result from the IBOSS subdata converge to zero as the size of the full data converges to infinity, with the size of the subdata fixed (Wang et al. 2018).

Thus, using strategic subdata selection methods like IBOSS, statisticians can select the most informative observations from the full data, resulting in a subdata set that is smaller but still provides inferences or predictions consistent with those that would be obtained from the full data. The reduction in size then allows statisticians to carry out appropriate analyses, even if those are complicated or computationally demand-

ing. We note that the case studies presented in this article offer a useful example of statistical analyses that are sophisticated, potentially computationally expensive, but produce compelling and worthwhile results. Both the BS and DIA datasets consisted of observations numbering in the low millions. Had these datasets instead been much larger, the full analyses performed might have become infeasible. However, that should not be a reason to compromise on the quality of the statistical analysis. Instead, using a subdata selection method like IBOSS would have allowed the researchers to still perform a thorough and careful exploration of the data.

Therefore, we believe that the advances made over many decades of work in the field of statistics should not be ignored or downplayed in the presence of newer and bigger data. Allowing the excitement of vast new quantities and types of data to compromise on the quality of the data itself, or the analysis of the data, is not useful. Rather, we must find innovative ways to both apply and continue to build upon the tools of our field to meet new challenges. As the authors stated, continued cross-collaboration and discussion about the future of the fields of statistics and data science are crucial to confronting this undertaking effectively. Thus, we would again like to thank the authors for writing this thoughtful piece and for encouraging reflection and discussion. Continuing to talk and work together collaboratively will help us ensure a future with better ideas and higher quality, exciting areas of research.

Acknowledgements The work by JS was partially supported by NSF grant DMS-1811363.

References

Donoho D (2017) 50 years of data science. J Comput Graph Stat 26(4):745–766

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960

Ioannidis JP (2005) Contradicted and Initially stronger effects in highly cited clinical research. JAMA 294(2):218–228

Ma P, Mahoney M, Yu B (2015) A statistical perspective on algorithmic leveraging. J Mach Learn Res 16:861–911

Meng XL (2018) Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. Ann Appl Stat 12(2):685–726

Stigler SM (1986) The history of statistics: the measurement of uncertainty before 1900. Harvard University Press, Cambridge

Wang H, Yang M, Stufken J (2018) Information-based optimal subdata selection for big data linear regression. J Am Stat Assoc 1–13. https://doi.org/10.1080/01621459.2017.1408468