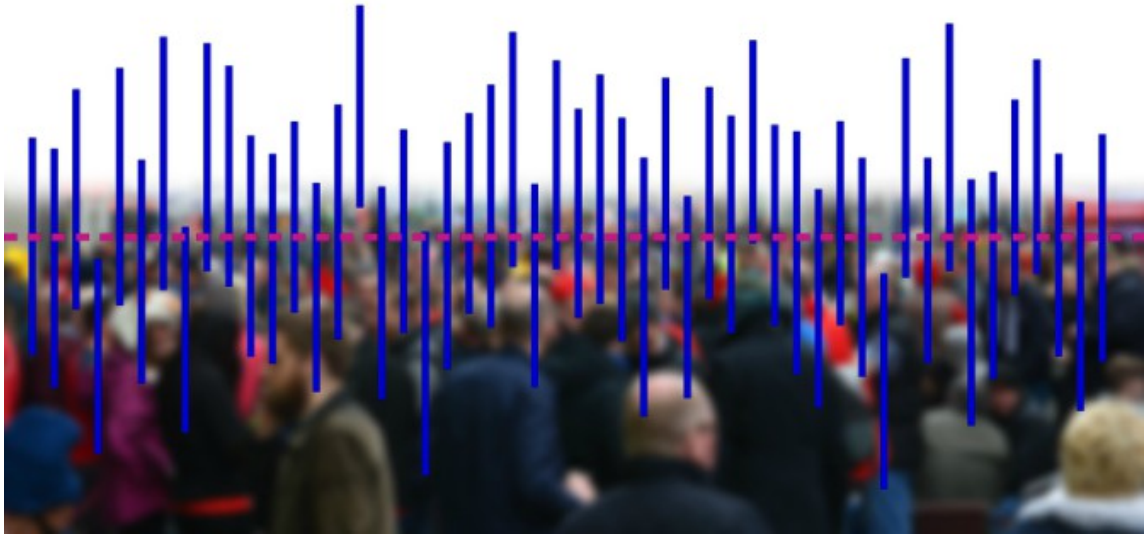
More data or better data? Using statistical decision theory to guide data collection



Big data has become an increasingly common topic of discussion. While the amount of available data and its role in the economy will continue to grow, we worry that the big data revolution will not live up to its promise if it is guided by the principle that bigger is always better. Data quality will limit the usefulness of big data.

Our research provides a clear framework for weighing the costs and benefits of allocating resources to acquiring more data as opposed to better data, for the purpose of inference about a population of interest. The objective may be to predict demand for a product at a store or criminal activity in a neighbourhood or vote shares in an election. If the only inferential problem arises from statistical imprecision, collecting more of the same kind of data is an obvious solution. However, collecting more data is not the solution if *identification problems* are a concern. Identification problems arise from data quality issues that do not diminish with sample size. Data quality may be impaired by selection of convenience samples, survey non-response, or inaccurate measurement. Confronting these problems, resources may be better spent collecting higher quality data rather than more of the same kind of data.

Higher quality data may cost substantially more per observation than lower quality data. We have seen this through our experience working with surveys of national probability samples of thousands of households, as opposed to surveys of so-called “internet access panels” that claim hundreds of thousands or even millions of members. Identification problems are not solved by just adding sample members. They can only be alleviated by collecting better data or by making assumptions that relate low-quality data to the objectives of research. To put it simply, would the Brexit and Trump election pollsters have made noticeably more accurate forecasts if they had merely surveyed more potential voters? We think not. Notably greater accuracy would have required some combination of better sampling schemes, higher response rates, and more informative measures of prospective voting decisions.

To make sample design a coherent subject of study, it is desirable to specify an explicit decision problem. We use the Wald (1950) framework of statistical decision theory to study allocation of a budget between two or more sampling processes for data collection. These processes all draw random samples from a population of interest and aim to collect data that are informative about the sample realisations of an outcome. But they differ in the cost of data collection and the quality of the data obtained. One may incur lower cost per sample member but yield lower data quality than another. Thus, increasing the allocation of budget to a low-cost process yields more data, whereas increasing the allocation to a high-cost process yields better data.

Date originally posted: 2018-01-22

Permalink: http://blogs.lse.ac.uk/businessreview/2018/01/22/more-data-or-better-data-using-statistical-decision-theory-to-guide-data-collection/

Blog homepage: http://blogs.lse.ac.uk/businessreview/

Our case study of survey non-response is particularly instructive. We study minimax-regret sample design for prediction of a real-valued outcome under square loss; that is, design which minimises maximum mean square error. The analysis imposes no assumptions that restrict the unobserved outcomes. Hence, the decision maker must cope with both statistical imprecision and identification problems.

The need to specify the decision criterion and the loss function are both the strength and the vulnerability of applying statistical decision theory to sample design. The strength of the theory is that it requires one to take an explicit stand on the decision problem to be addressed and delivers specific conclusions about what constitutes a good sample design. The vulnerability is that findings obtained for the specified decision problem may not satisfy persons who would choose a different specification. Some may view the dependence of findings on the specification to be a deficiency, but we think it a virtue. Statistical decision theory faces up to the reality that one cannot pose and study a well-defined optimization problem without taking a stand on what one wants to optimize.

Survey researchers who want to minimize the maximum mean square error of estimates should be concerned with both bias and variance, as recommended in the literature on total survey error. However, the focus has been on variance, as explained by Groves and Lyberg (2010): "The total survey error format forces attention to both variance and bias terms. . . . . Most statistical attention to surveys is on the variance terms—largely, we suspect, because that is where statistical estimation tools are best found" (p. 868).

Our research provides tools to directly assess both bias and variance. It formally shows the conditions under which reductions in maximum mean square error will be more efficiently obtained from an increased response rate than from increased sample size. We find that the threshold beyond which one should choose better over bigger data may be reached long before the sample numbers in the thousands, much less the hundreds of thousands.

Our findings make the case for better data over bigger data. Long ago, Cochran, Mosteller, and Tukey (1954) reached a similar conclusion in their report assessing the statistical methodology of the Kinsey study of male sexual behaviour. They wrote (p. 282): "Very much greater expenditure of time and money is warranted to obtain an interview from one refusal than to obtain an interview from a new subject." Unfortunately, their exploratory work was not followed up subsequently.

We believe that the proper role of statistical decision theory to guide data collection has been neglected for far too long. Our research develops tractable methods for using statistical decision theory in a setting where there is a concern with both statistical imprecision and partial identification. We hope that our paper will encourage increased use of statistical decision theory to inform data collection more generally, including collection of big data.

♣♣♣

Notes:

- *This blog post is based on the authors' paper More Data or Better Data? A Statistical Decision Problem, The Review of Economic Studies, October 2017.*
- *The post gives the views of its authors, not the position of LSE Business Review or the London School of Economics.*
- *Featured image credit: Population statistics, by geralt, under a CC0 licence*
- *When you leave a comment, you're agreeing to our Comment Policy.*

Jeff Dominitz is an independent economic consultant with recent clients in a variety of public sector and private sector organisations. He is also adjunct staff at the RAND Corporation, where he was previously Senior Economist. He holds a Ph.D. in Economics from the University of Wisconsin. He held faculty positions at Carnegie Mellon University, the University of Southern California, the California Institute of Technology, and the University of Michigan. He co-founded the consulting firm Resolution Economics LLC in Los Angeles. He previously worked as Director of Statistics for the Philadelphia Eagles and Football Research Manager for the Washington Redskins of the National Football League.

Date originally posted: 2018-01-22

Permalink: http://blogs.lse.ac.uk/businessreview/2018/01/22/more-data-or-better-data-using-statistical-decision-theory-to-guide-data-collection/

Blog homepage: http://blogs.lse.ac.uk/businessreview/

Charles F. Manski has been Board of Trustees Professor in Economics at Northwestern University since 1997. He previously was a faculty member at the University of Wisconsin-Madison (1983-98), the Hebrew University of Jerusalem (1979-83), and Carnegie Mellon University (1973-80). He holds a Ph.D. in economics from MIT. Manski's research spans econometrics, judgement and decision, and analysis of public policy. He is an elected Fellow of the American Academy of Arts and Sciences, the Econometric Society, the American Statistical Association, and the American Association for the Advancement of Science, Distinguished Fellow of the American Economic Association, and Corresponding Fellow of the British Academy.

Date originally posted: 2018-01-22

Permalink: http://blogs.lse.ac.uk/businessreview/2018/01/22/more-data-or-better-data-using-statistical-decision-theory-to-guide-data-collection/

Blog homepage: http://blogs.lse.ac.uk/businessreview/