G Model
JOI-692; No. of Pages 6
ARTICLE IN PRESS
Journal of Informetrics xxx (2016) xxx–xxx
Contents lists available at ScienceDirect
Journal of Informetrics
journal homepage: www.elsevier.com/locate/joi

Correspondence

The imaginarium of statistical inference when data are the population: Comments to Williams and Bornmann

1. Introduction

We thank the Editor for providing us with the opportunity to comment on the suggestions put forward by Williams and Bornmann (2016; WB). We will focus on Section 3.1 of their paper, which concerns the use of inferential (frequentist) statistics in analyses of population data. In the article, the authors do acknowledge that not all agree with their conjectures and provide a few references to such antagonists, yet WB do not present the essence of these objections, nor do they provide any detailed discussion of the inherent problems in their proposal. In this comment we will address a few of these fundamental problems albeit in a restricted manner, we refer to Western and Jackman (1994), Berk, Western and Weiss (1995), Wagner and Gill (2005) and Winkler (2009) amongst others for more elaborate discussions. Also, we have previously addressed critical issues concerning the use of statistical inference in scientometrics (Schneider, 2013, 2015).

The main argument put forward by WB is that it is both desirable and possible to "make [statistical] inferences about the underlying process that generated" citation scores in settings where we have collected all available data for one or several research institutions or other similar units of analysis in scientometric settings. In such situations the data available for analysis exhaust the population of interest, no additional data can be collected, not even in principle. Berk, Western and Weiss (1995) call such data sets "apparent populations" and they are common in the social sciences, also in scientometrics where we often exhaustively collect publication and citation data from international databases. In many scientometric settings such databases constitute the finite universe we can sample from.

WB base their reasoning on frequentist notions of inference. They argue that if it is somehow possible to "repeat the citation process over and over", it would also be possible to determine whether potential differences in citation scores are "due to luck . . . [or] are too large to attribute to chance alone." WB are certainly not alone in such conjectures, see for example Fox (2016, p. 9).

WB speculate that the complex behavioural/social processes leading to references in manuscripts, publication choices and eventually accrued citations in citation databases, are prone to "chance" outcomes along the way. Therefore "uncertainty" is seemingly linked to such processes as multiple possibilities in the same conditions in theory could lead to different, "chance-like", outcomes, one path is chosen, albeit others were also likely. The key questions would therefore be if such "uncertainty" is manifest in data from historical social processes, as well as in situations where data are an apparent population, and if so, whether we can model such "uncertainty", and how best to proceed with it?

We cannot go into a deeper discussion about the potential meanings of "uncertainty" and "randomness", these are subtle issues in both probability theory and philosophy of statistics. We are sympathetic, however, to the idea of uncertainty in a set of final citation scores, but contrary to WB, and in agreement with Berk, Western and Weiss (1995), we do not think that uncertainty can be addressed as a frequentist phenomenon of random sampling variation when data constitutes an apparent population. As we will elaborate below, data from apparent populations, at least in the social sciences, are not replicable, neither physically nor as a meaningful thought experiment. In our view, this fatally undermines the frequentist assumptions for statistical inference.

References given and articles published are manifest, and the compiled citation scores are an empirical fact. The data we have collected are the result of complex social and historical processes which we have very little knowledge about. WB ask us to imagine a counterfactual historical "citation process"; in fact, they want us to imagine infinite numbers of counterfactual processes where we again and again conceivably re-run the historical processes eventually leading to perceived variations in citation scores. The implicit claim is that such an imaginable infinite sampling process will provide us with a realistic sampling distribution from which meaningful standard errors and confidence intervals can be derived. We disagree with these claims and the analogies put forward by WB to support them for simple epistemic and empirical reasons: Frequentist theory cannot provide a reasonable, realistic or meaningful foundation for statistical inference in a context where repeated sampling makes no sense, neither empirically nor theoretically.

http://dx.doi.org/10.1016/j.joi.2016.09.011
1751-1577/

What then? If we dismiss statistical inference when we have an apparent population, we implicitly acknowledge that the processes leading to the citation scores are taken to be deterministic, at least when it comes to "sampling error". But there are alternatives to quantifying uncertainty, Berk, Western and Weiss (1995) suggest a Bayesian approach for situations when data are not replicable which we are sympathetic of, but we will not address it further in this comment.

Below we will elaborate some of our views, but first we recapitulate what frequentist statistical inference is all about as we fear that many of us tend to forget its underpinnings needed for proper interpretations.

2. Frequentist statistical inference

Statistical inference is based upon and bounded by theory. Frequentist and Bayesian theories are the best known and they differ considerably when it comes to modelling uncertainty. WB are committed to frequentist statistical inference, in fact, most social scientists have been for the past century. We presume that its widespread use to many is a testament to its all-encompassing feasibility; but if so, we disagree, in our view frequentist inference is restricted to specific settings.

Standard errors, *p* values and confidence intervals can be produced *en bloc* for almost any data set. It is deceptively easy to do so. But if there is a desire to move beyond description of the observed data, credible assumptions are required in order for the inferences to give meaning. Frequentists, like WB, must provide credible information beyond the data set about *how* the data were generated. And the data generation procedure must be stochastic and, at least in theory, the stochastic procedure must also be identical, independent and infinite for the frequentist derivatives such as sampling distributions and standard errors to provide any useful meaning.

Frequentist inference addresses uncertainty in data *introduced* by a stochastic data generation mechanism that produced the data. Other uncertainties or biases are not addressed. Eligible data generation mechanisms have *known* probability procedures and include probability sampling from a well-defined population, random assignments in experiments, or their equivalents produced by some well-described natural or perhaps social processes (for simplicity we will call these procedures random sampling). In the literature these frameworks are often characterized as either design-based or model-based inference. The first refers to empirical randomness introduced by the random sampling design, and the latter framework refers to epistemic randomness via imposition of distribution assumptions (Koch & Gillings, 2004).

Random sampling means that sample units are drawn independently, and each unit in the population has an equal chance to be drawn at each stage. Frequentist statistical inference depends on the concept of a sampling distribution. If we have a random sample from a well-defined population, then it is possible to construct a sampling distribution and subsequently derive meaningful standard errors, and the logic and mathematics of frequentist inference play through, at least in principle.

Frequentist inference also assumes an infinite number of independent and identically drawn random samples which are needed in order to establish the sampling distribution; i.e. the well-known frequentist "thought experiment" exemplified by WB's coin toss example. If such a sampling procedure is impossible to implement physically, or to imagine as plausible, then frequentist interpretations based on images of the "long run" become a fiction. The concept of a proper sampling distribution simply does not apply. Let us assume that we have created a data set without using a random sampling procedure and that we, at least in theory, could repeat this study a number of times. The subsequent results might well vary but most likely not in a manner consistent with a specifiable sampling distribution due to the lack of proper random sampling. Therefore, if a valid sampling distribution cannot be constructed, there can be no meaningful interpretable standard error, and any statistical inference that is attempted will not convey any useful information. Renowned statistician David Freedman suggests that absent random sampling, standard errors are "best viewed as a de minimis error estimate: if this were – contrary to the fact – a simple random sample, the uncertainty due to randomness would be something like the SE" (Freedman, 2004, p. 989). It is simply not obvious what the inferences refer to.

Statistical inference, whether frequentist or Bayesian, is no free lunch. We are deeply concerned with "assume-and-proceed" statistics where fundamental requirements needed for statistical inference to give meaning mindlessly are assumed to be present everywhere and always. Our main point of view is that assumptions, conditions and models required for frequentist-based inference to make sense must be reasonable, theoretically justified and empirical feasible.

We do not think WB's arguments hold up to such requirements, and neither do similar views by other esteemed colleagues who also endorse statistical inference from non-stochastic settings (e.g. Abramo, D'Angelo & Grilli, 2015; Dieks & Chang, 1976; Fairclough & Thelwall, 2015). We think it matters because violations of random sampling assumptions such as independence will transfer to the sampling distributions, and subsequently the standard errors, resulting in bias. Obviously, standard errors will not tell us whether they are biased, to what extent, or if they are totally meaningless. We have to probe such issues with credible assumptions and empirical feasibility.

Given its particular focus and its required assumptions, we think that frequentist statistical inference has severe limitations of application in the social sciences. One such inappropriate setting is when our data set is an apparent population exactly as the one suggested by WB.

3. When data *are* the population

Making frequentist statistical inferences from apparent populations in social science settings raise some important conceptual challenges. First, the data are not generated by some known random sampling procedure, they are a grab of what is available, perhaps restricted by a specific time period. No other data can be sampled for such a time period. Second, since

Please cite this article in press as: Schneider, J.W. The imaginarium of statistical inference when data are the population: Comments to Williams and Bornmann. *Journal of Informetrics* (2016), http://dx.doi.org/10.1016/j.joi.2016.09.011