G Model
JOI-604; No. of Pages 8

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

Sampling issues in bibliometric analysis

Richard Williams[a,*,1], Lutz Bornmann[b]

[a] Department of Sociology, 810 Flanner Hall, University of Notre Dame, Notre Dame, IN 46556, USA
[b] Division for Science and Innovation Studies Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539, Munich, Germany

ARTICLE INFO

Article history:
Received 9 January 2014
Received in revised form
16 November 2015
Accepted 16 November 2015
Available online xxx

Keywords:
Bibliometrics
Sampling
Population
Power analysis
Percentiles

ABSTRACT

Bibliometricians face several issues when drawing and analyzing samples of citation records for their research. Drawing samples that are too small may make it difficult or impossible for studies to achieve their goals, while drawing samples that are too large may drain resources that could be better used for other purposes. This paper considers three common situations and offers advice for dealing with each. First, an entire population of records is available for an institution. We argue that, even though all records have been collected, the use of inferential statistics, significance testing, and confidence intervals is both common and desirable. Second, because of limited resources or other factors, a sample of records needs to be drawn. We demonstrate how power analyses can be used to determine in advance how large the sample needs to be to achieve the study's goals. Third, the sample size may already be determined, either because the data have already been collected or because resources are limited. We show how power analyses can again be used to determine how large effects need to be in order to find effects that are statistically significant. Such information can then help bibliometricians to develop reasonable expectations as to what their analysis can accomplish. While we focus on issues of interest to bibliometricians, our recommendations and procedures can easily be adapted for other fields of study.

1. Introduction

Statistical significance tests and/or confidence intervals (CIs) are frequently used with bibliometric data. For example, Opthof and Leydesdorff (2010) compared leading scientists (professors) at the Academic Medical Center of the University of Amsterdam using the Kruskal–Wallis test. Statistical significance tests are strongly connected to questions of sampling, since these tests are usually applied to the analysis of samples in order to obtain information about an underlying population (Levy & Lemeshow, 2008). In bibliometrics, several papers have been published which deal with the use of significance tests and effect sizes (e.g., Bornmann & Williams, 2013; Schneider, 2012, 2013), but the literature on sampling of populations is scarce. In one of the rare papers, Bornmann and Mutz (2013) argue for clusters in a two-stage sampling design ("cluster sampling"), in which, first, one single cluster is randomly selected from a set of clusters (e.g., consecutive publication years, in which an institution have published) and second, all the bibliometric data (publications and corresponding citation metrics) is gathered (census) for the selected cluster. Then, this cluster sample can be statistically analyzed.

* Corresponding author. Tel.: +1 5743601017.
E-mail addresses: Richard.A.Williams.5@ND.Edu (R. Williams), bornmann@gv.mpg.de (L. Bornmann).

1 Web Page: http://www3.nd.edu/~rwilliam/

http://dx.doi.org/10.1016/j.joi.2015.11.004
1751-1577/

This paper deals with issues around samples and populations in bibliometrics. In many institutional evaluations, bibliometricians have complete publication and citation records for all the papers of an institution. These are sometimes referred to as "apparent populations;" Berk, Western, and Weiss (1995b) give as examples of apparent populations all states in the United States or all nations in the developing world. We argue that, even though all records have been collected, the use of inferential statistics and significance testing is both common and desirable. We further argue that the use of power analysis can help guide analyses when records for an entire population are not available. Specifically, this paper addresses two issues: first, the appropriateness of using inferential statistics when the entire population of records is available (Bornmann, 2013); and second, the use of power analysis and sampling when it is impractical to gather information for all institutional citation records (Bornmann & Mutz, 2013). In particular, how does a bibliometrician go about determining how large a sample needs to be in order to achieve the goals of the analysis? Conversely, when the sample size has already been determined, how large do effects need to be in order for them to be statistically significant? Answering such questions can help the bibliometrician decide how large a sample is needed; or, if the sample has already been drawn, answering these questions can help the bibliometrician form reasonable expectations as to what the analysis can accomplish.

2. Justification for using statistical inference with citation impact data

2.1. Appropriate types of data

In the following, we discuss techniques that are appropriate when a study wishes to use percentiles of citations to measure institutional citation impact. We also note that, while we focus on the analysis of percentile data, our ideas could also be applied to other types of bibliometric statistics, such as statistics based on average citations rather than percentile-based statistics.

Cross-field and cross-time-period comparisons of citation impact for institutional evaluation purposes are only possible if the impact is normalized (standardized; Schubert & Braun, 1986). For its citation impact to be normalized, a paper needs to have a reference set: all the papers published in the same publication year and subject category. Percentiles have been proposed as a robust alternative to normalization on the basis of central tendency statistics (arithmetic averages of citation counts) (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015; Wilsdon et al., 2015). Percentiles are based on an ordered set of publications in a reference set, whereby the fraction of papers at or below the citation counts of a paper in question is used as a standardized value for the relative citation impact of the focal paper. This value can be used for cross-field and cross-time-period comparisons. If the normalized citation impact for more than one paper is needed in a research evaluation study (and this is the rule in institutional evaluations), this percentile calculation is repeated (by using corresponding reference sets for each one).

Following the practice of Incites (Thomson Reuters, http://incites.thomsonreuters.com/), we use inverted percentiles in our examples, where low percentile values mean high citation impact. Hence citation impact above the median (in the field and publication year) is defined as percentiles less than 50. With inverted percentiles it can easily be seen whether a citation percentile is within the top 10 or top 1 percent most frequently cited paper range, which we think is the sort of thing most bibliometricians will be interested in (Bornmann, 2014). Of course it is a trivial matter to use non-inverted percentiles instead if the bibliometrician prefers them or if it is appropriate given the way the data being analyzed are coded.

2.2. Using bootstrapping to verify that the statistical methods employed are appropriate for percentile data

A possible statistical problem in this study is that percentiles have an approximately uniform rather than normal distribution.[1] When variables are normally distributed, cases tend to be clustered near the mean, while extreme values in either direction are less common. With percentile rankings, however, in the population there will be just as many cases in the first percentile as there are in the 50th and the 99th. *t*-tests assume that dependent variables are normally distributed, which raises the question of whether analyses based on *t*-tests (which includes the power analyses presented here) are potentially biased.

A recent analysis by Williams and Bornmann (2014) suggests that a power analysis of percentile rankings can indeed be conducted. Bootstrapping is often used as an alternative to inference based on parametric assumptions when those assumptions are in doubt (Cameron & Trivedi, 2010). Bootstrapping resamples observations (with replacement) multiple times. Standard errors, CIs and significance tests can then be estimated from the multiple resamples. Using real data for the years 2001 and 2002 from three research institutions in German-speaking countries, Williams and Bornmann (2014) made heavy use of *t*-tests and related statistics. They used bootstrapping to double-check their results, and found that "bootstrapping produced significance tests and confidence intervals that were virtually identical to those reported in our tables, giving us confidence that our procedures are valid" (p. 269). We therefore feel confident that the statistical techniques we use in this paper are appropriate and that our findings are valid.

[1] Their distribution is uniform only approximately, depending on the number of ties in the citation distribution.

Please cite this article in press as: Williams, R., & Bornmann, L. Sampling issues in bibliometric analysis. *Journal of Informetrics* (2016), http://dx.doi.org/10.1016/j.joi.2015.11.004