Journal of Informetrics 10 (2016) 1131–1134

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Correspondence

To infer or not to infer? A comment on Williams and Bornmann

1. Introduction

Since the beginning of the twentieth century, statistics has completely revolutionized the manner in which scientific research is performed. Statistical techniques emerged based on the intrinsic random nature of the data and, equally important, on the scarce data availability. The techniques were both mathematically challenging and computationally intensive. While the computation issue for various statistical techniques was rapidly solved with the mighty help of computers, the implementation of these techniques still required a deep understanding of the procedures. Researchers needed to take into account the underlying assumptions and possible limitations. When the implementation issue for standard techniques had also been taken care of by various software programs, Statistics promise-land opened widely its gates. Researchers, practitioners, analysts, students, basically anyone in possession of a dataset could put his/her own dataset through the great, magic “Statistics machinery” and obtain the desired results.

From the statistical analysis point of view, numerous fields mainly operate on an automated recipe-based protocol. The field of scientometrics makes no difference, and a common practice of researchers is to take a dataset, perform some statistical analysis and finally, to generalize the conclusions of the analysis. See, for example, Abramo, Cicero, and D’Angelo (2015), Kozak, Bornmann, and Leydesdorff (2015) and Letchford, Preis, and Moat (2016), selected at random from the latest issues of the journals in the field. Usually there is no discussion on the available data from a statistical point of view and assumptions that need to be met often do not seem to be acknowledged. More importantly, the aim of the statistical analysis, that is, why is a particular statistical tool necessary and appropriate for the analysis seems to be frequently implicit.

Williams and Bornmann (2016; henceforth WB) adhere to this statistical recipe-based movement, even though their paper is devoted to practical matters in power analysis and sample size calculation. Power analysis and sample size calculations are steps to be taken in performing statistical inference, that could be influenced by the target of inference. I believe that acknowledging the target of inference is not only imperative, but also initiates the course of the statistical analysis. I will discuss this important issue in Section 2. Another important item I will address is the concept of apparent population (Berk, Western, & Weiss, 1995) and, more importantly, accommodating this concept into the statistical inference apparatus. While, similar to WB, I believe there are situations when this type of inference is valid and valuable, the arguments used by WB seem to be inadequate. A proper justification is in place and Section 3 is dedicated to this. While the main focus of WB regards sampling issues, my comments mainly address the more fundamental issue of statistical inference with a special focus for apparent populations. Nonetheless, WB omit important aspects to be considered in the sample size calculations. I briefly mention them in Section 4.

2. The target of statistical inference

Statistical inference is the act of drawing conclusions about a statistical population based on the data at hand. The available data is a subset of the statistical population and is usually referred to as the sample (of observations). The statistical population is rarely mentioned in practice. Despite its rare appearances, the proper definition or acknowledgement of the statistical population is vital. With this respect, the concept of target population (Smith, 1993) is commonly used to emphasize that the conclusions of the statistical inference target that specific population.

Though subtle, the concept of target population is not a trivial matter and should not be overlooked. In fact, Smith (1993) regards it as the most important statistical concept, since “the concept of target population is the best way to define the target for inference” and moreover, “specifying the target population and the selection mechanism should be the starting point for any act of statistical inference”. When the available data differs from the target population, statistical inference is appropriate and desirable. Statistical inference emerges thus from a necessity, and is not a choice the analyst makes. Therefore a “justification for using statistical inference with citation impact data”, in the title of Section 2 of the WB manuscript can be seen by readers as a misinterpretation of statistical inference.

http://dx.doi.org/10.1016/j.joi.2016.09.010

WB consider statistical inference in three situations. Three examples, based on percentile citation data are chosen to resemble the situations described in WB.

1. A bibliometrician is interested in the publication performance of Universities A and B. All publications of universities A and B in a given period are recorded, say between 2010 and 2013. For the 2000 publications of University A, PP(top 10%) is 13.5%, that is 13.5% of University A's publications are among the top 10% most cited publications of their field and publication year. For the 150 publications of university B, PP(top 10%) is 15%.
2. A bibliometrician is interested in the publication performance of University C. Collecting the university's entire scientific output is not feasible and hence a sample needs to be drawn, say from the publication output in 2010–2013. The analyst decides to collect 2000 publications, for which PP(top 10%) is 10.05%.
3. A bibliometrician is interested in the publication performance of University D. A sample of 200 publications is already available, say from the publication output in 2010–2013, with PP(top 10%) of 8.5%.

Analysing and interpreting the results depends on the goal of the analysis and the data selection mechanism, as Smith (1993) underlined. The data selection mechanism is an essential aspect of statistical inference, and unfortunately WB make no reference to the matter. Brief comments on the matter are mentioned in Section 4.

In Example 1, the entire scientific output of Universities A and B in 2010–2013 is recorded. As discussed beforehand, the first question should be: What is the purpose of the statistical analysis? What is the target population? If the analyst is interested in the publication performance of the two universities in 2010–2013, then the target population is readily available. There is no need for statistical inference, and there is no need to compute significance tests or confidence intervals! Based on PP(top 10%), University A is outperforming University B. The analyst needs to decide whether the difference of 1.5% is significant, not from a statistical point of view but only from a practical point of view. Investigating whether the difference in the publication performance of 1.5% is statistically significant would incorrectly assume that the data for the two universities are random samples and not statistical populations.

Unfortunately, WB do not account for this instance, that is when the target population is available. WB mention that "even when all records are available, a power analysis can be useful [. . .]". I disagree, a power analysis is useless when the target population is available since power analysis assumes the existing data to be a sample and not the target population itself.

When the goal of the analysis is the performance in general of the two universities, the target population is no longer available to the bibliometrician. What is then the target population? What is the data at hand, from a statistical point of view, and, more importantly, how do we interpret the results? WB gather their arguments from several sources and adapt them to citation analysis. While I agree with most of the arguments and I see the necessity of statistical inference in this case, I believe several important observations need to be made. The observations and the discussion of the results are included in Section 3.

In Examples 2 and 3, the available data are samples. Despite the unambiguity in the available data, from a statistical point of view, it should be emphasized that the statistical analysis is not straightforward. Suppose that the bibliometrician is interested in the publication performance over the time period 2010–2013. The target population is thus the set of all publications of the university in the given time frame and is finite. Finite target populations require a different type of statistical analysis and emerging issues in WB sampling size calculations are addressed in Section 4. If the goal of analysis is the general publication performance of the two universities, even if the entire publication record had been available, the target population would still not have been available. In Examples 2 and 3, the collected data are, in fact, sub-samples of the target population.

A final remark on the purpose of the analysis is in place here. It needs to be emphasized that obtaining statistically significant results should definitely not be an aim in itself. The main motto should be "Put science before statistics" (Lenth, 2007). Sample-size calculations can be maneuvered to lead to statistically significant differences of 0.05, that is a PP(top 10%) of 10.05% would be statistically significantly higher than the 'world average' of 10%. How significant is the difference of 0.05% in practice? Most bibliometricians would probably not regard this difference as significant from a practical or substantive point of view. Therefore, one should clearly distinguish between practical significance and statistical significance. One might wonder though why is the statistical significance necessary once a practical significance of 5%, for example has been achieved. The answer is convoluted in the sample-population paradigm, since the statistical significance helps to distinguish if the difference of 5% (which has a practical significance) could be due to chance variations.

3. Statistical inference for apparent populations

Coming back to Example 1, suppose the analyst aims to inspect the publication performance of University A in general. The target of inference is the general publication performance of University A. What constitutes then the target population? The target population can be defined as the set of publication and citation records of University A that could result from "repeating the history", under the same initial conditions, such as number of researchers, fte's, etc. A more compact and rigorous manner is to regard the entire publication record of University A as one outcome from the set of all possible outcomes of a stochastic process. We therefore employ an underlying academic, social and economic stochastic process associated with the publication and citation output of University A. The set of all possible realizations of this stochastic process constitutes a superpopulation (Berk et al., 1995). The data at hand do not follow the standard definition of a sample. To emphasize the difference, the term apparent population (Berk et al., 1995) is used in turn and the target population is referred to as a superpopulation.

WB argue that “the use of statistical inference and significance testing is both common and desirable”. The “common and desirable/helpful” argument is repeated several times throughout the manuscript. Even though WB mention that “chance factors could have increased the number of citations a paper received or else decreased them”, I believe it would have been preferable to see more why, de facto, and not how, statistical inference for apparent population is desirable and/or helpful. Additionally, scientific research should definitely not be based on common practice. More importantly, the fact that statistical inference for apparent populations is “common” should not constitute an argument for encouraging researchers to use statistical inference in this setting.

The superpopulation concept emerges because the purpose of analysis demands it and not from the need to accommodate the statistical inference machinery in every possible situation. As mentioned earlier, the purpose of the analysis and the target population are mandatory before statistical inference for apparent populations becomes desirable or helpful.

The general publication performance of a university can be inferred via the following questions:

Q1. Can the results in 2010–2013 be generalized (used as estimates) for the general publication performance of the two universities?

Q2. Can we infer that both universities have a general publication performance above world average and that, in terms of a specific bibliometric criterion, University B outperforms University A?

Similar to WB, I agree that the above two questions can be answered satisfactorily by making use of statistical inference. Under the superpopulation framework, statistical inference methods are valid once the sample satisfies certain assumptions. Due to the assumed selection mechanism, the available data is assumed to be a random sample. Whether the sample is representative for the target population or underlying process is for the analyst to decide. An analyst might decide that the publication record for a university before certain major events, such as creating new departments or merging with other institutions is not representative for measuring the general publication performance of that university. Even though not mentioned by WB, under the superpopulation setting, Example 1 is similar to Example 3, that is a sample is available to the analyst and the sample size is pre-determined. The difference is that in Example 1 the target population is infinite, under the superpopulation setting, whereas in Example 3 the target population could be finite, if the interest is in the performance in 2010–2013. When both target populations are infinite, it is noteworthy that in Example 1 the data is a sample, whereas in Example 3, the data is a sub-sample.

We are therefore interested in the general performance of a given university, via the percentages of its publications among the top 10% of their field, that is PP(top 10%). We use the sample PP(top 10%), that is PP(top 10%) computed for a given sample as an estimator of the true, unknown PP(top 10%). For clarity, we denote the sample PP(top 10%) by PP(top 10%)^S. From Example 1, PP(top 10%)$^S{}_A = 13.5\%$ is therefore an estimate for PP(top 10%)$_A$. Similarly, PP(top 10%)$^S{}_B = 15\%$ is an estimate for PP(top 10%)$_B$. Is the analyst uncertain about these estimates? Of course she/he is!

To account for the uncertainty inherited by these estimates, confidence intervals can be computed, either via the central limit theorem, when the asymptotic distribution is known, or via bootstrapping. Sometimes specific questions such as those in Q2 are of interest and then statistical significance tests can be employed. As previously mentioned, the analyst should decide if a difference of 1.5% in the PP(top 10%) estimates is considered to be of practical significance and, consequently, if it is worthwhile testing its statistical significance. For example, an analyst might decide that a difference of 1.5% in PP(top 10%) is not meaningful to be tested, that is, it is not worthwhile investigating whether the difference of 1.5% is due to chance or not. The conclusion, based on the sample at hand would be that there is no significant practical difference between the publication performance of the two universities.

4. Other sampling issues and power analysis

WB focus on sample size computations under different sampling conditions, that are depicted in Examples 1–3 in Section 2. Even though size effects and a proper power analysis are important matters, other essential issues need to be addressed in the sample or effect size calculations. Though this is not the main focus of this manuscript, I believe these issues at least need to be acknowledged.

Random or non-random sampling, with or without replacement, cluster, stratified are some of the considerations that need to be made when designing a sampling framework or when analysing a sample that has already been drawn. WB briefly mention, in the introduction, a two-stage sampling design, in which a random cluster is sampled at the first stage and all the bibliometric data are collected in the second stage (Bornmann & Mutz, 2013). For standard statistical inferential procedures, the sample needs to be random, with independent and identically distributed observations. An equally important matter is the representativeness of the sample. Needless to say that if the sample is not representative for the target population, then statistical inference is futile.

The size of the target population can in some circumstances plays an important role. Take for example the results in Table 1 in WB. For the $\mu_A = 47.5$ and the small effect of -0.086 a sample of $N = 1049$ needs to be taken. What if the size of the target population is 1000? This question addresses a more general issue that needs to be accounted for, that is statistical inference for finite populations. When the target population is finite, certain corrections need to be applied in computing standard errors or confidence intervals based on asymptotic distributions or bootstrapping, and in performing statistical significance tests. In the literature, attention has been given to the finite population setting, starting with the sampling from a finite population (Hájek & Dupač, 1981) and design and inference in finite population sampling (Heyadat & Sinha, 1991), etc. Ignoring the finiteness of the target population can lead to large errors when performing statistical inference, as shown, for example, by Kozak (2008) in biological applications. Nonetheless, not all functions in statistical software packages have finite population corrections. For example, the function "svy bootstrap" in Stata has no option for such a correction.

5. Conclusions

Sampling techniques have received tremendous attention in various fields. This shows that along with key ingredients addressed by WB, the proper sampling framework remains a challenging aspect of study design and should definitely receive more attention within bibliometric analysis.

Nonetheless, sampling design is deeply embedded in the statistical inference framework. The target of statistical inference should trigger the course of any study design and, implicitly the sampling framework. While I adhere to the rationale behind statistical inference for apparent populations employed by WB, I think more emphasis should be placed on motivation. This comment is a call for researchers to dedicate more time in their research and space in their publications to motivating their statistical decisions.

Finally, about to the Hamletian dilemma in the title, WB seem to answer: yes, anytime and anyhow. Even when all available data are at hand, it is "both common and desirable". My answer folds the main message of this comment: there should be no dilemma, the necessity of statistical inference should emerge from the target of statistical analysis.

Acknowledgements

I would like to thank the members of the Advanced Bibliometric Methods working group at the Centre for Science and Technology Studies of Leiden University for the helpful discussions and especially Ludo Waltman for his input that significantly improved this comment.

References

Abramo, G., Cicero, T., & D'Angelo, C. A. (2015). Should the research performance of scientists be distinguished by gender? *Journal of Informetrics*, *9*, 25–28.
Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent populations (with discussions). *Sociological Methodology*, *25*, 481–485.
Bornmann, L., & Mutz, R. (2013). The advantage of the use of samples in evaluative bibliometric studies. *Journal of Informetrics*, *7*(1), 89–90.
Hájek, J., & Dupač, V. (1981). *Sampling from a finite population*. New York: M. Dekker.
Heyadat, A., & Sinha, B. K. (1991). *Design and inference in finite population sampling*. New York: Wiley.
Kozak, M., Bornmann, L., & Leydesdorff, L. (2015). How have the Eastern European countries of the former Warsaw Pact developed since 1990? A bibliometric study. *Scientometrics*, *102*, 1101–1117.
Kozak, M. (2008). Finite and infinite populations in biological statistics: Should we distinguish them? *The Journal of American Science*, *4*(1), 59–62.
Lenth, R. V. (2007). Statistical power calculations. *Journal of Animal Science*, *85*, E24–E29.
Letchford, A., Preis, T., & Moat, H. S. (2016). The advantage of simple paper abstracts. *Journal of Informetrics*, *10*, 1–8.
Smith, T. M. F. (1993). Populations and selection: Limitations of statistics. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *156*(2), 144–166.
Williams, R., & Bornmann, L. (2016). Sampling issues in bibliometric analysis. *Journal of Informetrics*, *10*, 1225–1232.

Gabriela F. Nane
Delft University of Technology, Netherlands
E-mail address: G.F.Nane@tudelft.nl
Available online 17 October 2016