Journal of Informetrics 10 (2016) 1253–1257

Contents lists available at ScienceDirect

Journal of Informetrics

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

Correspondence

# Sampling issues in bibliometric analysis: Response to discussants

## 1. Introduction

The discussants of our paper raise several important issues. Some (Mutz, Claveau) agree with us on many points but argue that our sampling procedures are flawed or could be improved upon. Others (Schneider, Nane, Waltman) have serious disagreements with us, especially with our arguments concerning statistical inference with apparent (aka super) populations. We will respond to each reviewer separately.

However, we would like to begin with a new argument concerning apparent populations suggested by these exchanges: the mere fact that there is such controversy is all the more reason to apply the methods we propose. Whichever side is right, we think research will be more persuasive if it makes arguments along the lines of what we have advocated.

As our Bielby (2013) example indicated, even when detailed employment records are available, courts want to see evidence that some apparent pattern of gender discrimination could not be attributed to chance instead. Our paper showed that they are not alone. Arguments for the super-population approach date back to at least the 1940s and, rightly or wrongly, continue to be made by many people today. Similarly, we suspect many readers of bibliometric research will also be more persuaded if they see evidence that results are unlikely to be due to chance factors alone. Most researchers are used to seeing statistical inference procedures employed. Dealing with what appears to be an entire population is unusual. We think they will be more easily persuaded by procedures they are comfortable with and that many experts say are called for.

By way of analogy, a qualitative researcher (e.g. an ethnographer) doing case studies on the effects of poverty on families may feel that quantitative methods have little value. But, that same researcher may realize that her arguments will carry more weight if it is shown that quantitative results also support her case, e.g. she can only study so many families, but statistics may suggest that what she says about them is likely to be true for many others. By showing that quantitative procedures yield results that are both statistically significant and substantively important, she may be able to add credibility to her qualitative findings.

Of course, if you do not think statistical inference is leading to the correct conclusions, you can make a case against it then (or just not use it at all). But, if you do think the conclusions are correct, you can make a stronger argument by showing that, whatever approach you think is correct, the conclusions are the same. Multiple approaches all leading to the same conclusion are more persuasive than only using one approach.

## 2. Waltman

Waltman repeatedly stresses conceptual difficulties in statistical inference for citation analysis. We suspect his arguments could just as easily be applied to any study of apparent populations, e.g. arguments about randomness could be developed for almost any topic being studied. Further, his arguments about different types of randomness might be extended to any use of statistical modeling or statistical inference, even when a sample is clearly being used. Stochastic errors can be thought of as reflecting the influence of any variables that have been omitted from a model. There may be sources of error that researchers cannot even imagine, let alone build into their models. Yet, researchers do analyses anyway.

Waltman says that “when statistical inference is used in citation analysis, it is crucial to be explicit about the type of randomness that is considered” but he never explains why. Waltman does not show that failing to identify the different types of random influences seriously impacts the accuracy of statistical inferences, especially for relatively simple analyses such as *t*-tests. Again, if it did, then virtually all analyses, not just bibliometric, would be highly suspect.

DOI of the original article:http://dx.doi.org/10.1016/j.joi.2015.11.004.

http://dx.doi.org/10.1016/j.joi.2016.09.013

Most critically, even if you accept every argument Waltman makes, it seems like you should reach the opposite conclusion of what he did. Instead of settling for descriptive statistics, you should think about how to more carefully model the process. In more complex multivariate analyses, researchers should consider what variables should be included in the model. If, for some reason, those variables cannot be included, the researcher must evaluate what the implications are for omitted variable bias. Researchers encounter these issues all the time regardless of the topic of their analysis. But, the usual response is to collect better data or develop superior models, not just settle for descriptive statistics.

In conclusion, we would say that, if you accept Waltman's arguments, you are basically arguing that the perfect should be the enemy of the good. We showed many advantages and reasons for doing statistical inference with apparent populations, and Waltman countered by arguing that we may not be doing it exactly right. He makes excellent points; but there are always questions about omitted variables and uncertainty about random processes. You model them as best you can, rather than just throwing up your hands and saying that only descriptive statistics are possible.

## 3. Mutz

Our paper dealt with fairly basic but important analyses of bibliometric data. Mutz agrees with our major points. Further, he offers valuable insights for points to consider when more complex analyses are being conducted. Mutz notes that publications are not independent units. For example, there are co-authorships and field dependencies. A more complex analysis should take this interdependence into account, possibly stratifying the sample to reflect different strata within the population.

Mutz also suggests additional reasons why a sample should be drawn even when a larger set of records is available. More complicated models with large data sets (e.g. for analyses of research networking) could be very difficult computationally. As Mutz says, "By reducing the sample size, sampling could facilitate more complex statistical modeling (be it Bayesian or frequentist) without significant loss of estimation accuracy."

In short we agree with most of what Mutz says. Further we think his points reinforce our response to Waltman. Just because something is complicated does not mean we should only do descriptive statistics. Instead we should try to think through the complexities and how best to handle them. At the same time, we again caution that the perfect should not be the enemy of the good. Many survey research data sets come with detailed information on how cases should be weighted and how standard errors should be adjusted to account for stratification and clustering. Most bibliometric data sets do not. The kinds of complex sampling schemes that Mutz calls for may be difficult to develop. If they are too difficult, we would rather see simpler sampling schemes used than to do nothing at all.

## 4. Claveau

Claveau indicates that he is "deeply sympathetic" to our general strategy. He further adds:

> "The general argument – a compelling one according to me – is that these observations are realizations of an underlying data generating process constitutive of the research unit. The goal is to learn properties of the data generating process. The set of observations to which we have access, although they are all the actual realizations of the process, do not constitute the set of all possible realizations. In consequence, we face the standard situation of having to infer from an accessible set of observations – what is normally called the sample – to a larger, inaccessible one – the population. Inferential statistics are thus pertinent."

Despite his sympathy to our main arguments, Claveau feels we are not as clear as we could be and that there are various technical errors in our suggested approaches. We think he makes many good points, but we do take some minor exceptions.

We are impressed that Claveau ran 1,000,000 Monte Carlo simulations and apparently analyzed around 40,000 actual documents. However, that is a lot of work! Perhaps too much work for most researchers. If somebody had the resources and skills to conduct the kind of analyses that Claveau did, customized for their own situation, they might not even need to be worried that much about how large of a sample they could afford. Further, he showed that, in most cases, the more refined figures did not make that much of a difference, with a noteworthy exception being power calculations for elite institutions. Given his insights, we think it would sometimes be useful to modify our original suggestions to estimate power for different plausible values for the standard deviation. Indeed, the values from his Table 1 could be used in the calculations.

Claveau further argues that our use of two-tailed tests is inappropriate, and correctly points out that the calculations change a bit if one-tailed hypotheses are tested instead. The specific example we gave would indeed justify the use of a one-tailed hypothesis, and in retrospect it might have been a good idea to tweak our wording a bit. However, we think caution should be used before basing calculations assuming one-tailed tests. For one thing, the difference between the observed and predicted values may be in the opposite direction of what the researcher expected. In that case the researcher will want to have enough cases and enough power to say whether the institution is significantly below average. In other cases a researcher may just wish to see whether the observed value is significantly different than the predicted value, e.g. maybe the researcher just wants to know if overall an institution is in or near the 30th percentile. If the institution is doing as well as predicted or even better, then great, but if it is not as good as predicted it would be good to know whether the difference is statistically significant or not.

Claveau also says "In fact, random variables can even be the result of fully deterministic systems." Again, we have no particular problem with this argument. However, stochastic error terms are often thought of as reflecting all the influences that are not explicitly included in a model. We gave a few examples of what they might be, but there could be an infinite number more. We think our phrasing concerning random variables and influences is fairly common, but if Claveau's phrasing makes the point clearer to people we are fine with it.

Finally we note that only part of our paper required power analysis. Unlike some of the other authors Claveau does not seem to have much concern with our arguments on apparent populations, other than saying he thinks our argument could be clearer.

In short, we are pleased that Claveau agrees with our general argument. We think caution should be used before basing calculations on one-tailed hypotheses but if the theory is strong enough that may be fine. Claveau's own analyses suggest that, in many cases, using 28.87 as the standard deviation value will be fine and any harm minimal. In those cases where it is most likely to make a difference, he lays out useful guidelines for what values should be used instead.

## 5. Nane

Nane argues that "one should clearly distinguish between practical significance and statistical significance." We of course agree with that (Williams & Bornmann, 2014). Indeed, we warned in our paper that large institutions might get statistically significant results even when the substantive differences were trivial. We further noted that, in some situations, it may be easy to assess whether a result was substantively important; but in less clear cases standardized measures of effect size can be used. We would further add that any assessment of an institution's performance should certainly go well beyond just looking at the number of citations it receives. We agree that bibliometric results should also be interpreted by bibliometricians (against their experiences with other bibliometric results) and by experts in the corresponding field (against their experiences with impact differences from other bibliometric studies in the field). Thus, we think that both are necessary: statistical/practical significance tests, and assessments by peers. The danger with only using assessments by experts is that these judgements might be biased by personal preferences. Citation counts can supplement subjective analyses by providing a widely used and objective measure of performance.

With regards to superpopulations, Nane indicates that she agrees with our main arguments in some situations but disagrees in others. According to Nane, it is critical to determine what the "target of inference" and the "target population" are. She gives an example of where "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." Nane says if the interest is only in comparing the performance over that given period of time, the target population is readily available, there is no need for statistical inference, and the analyst only needs to decide whether the observed difference (in her example, 1.5%) is important from a practical point of view.

Suppose, however, that instead the goal of the analysis is evaluating the "performance in general of the two universities." She says that in that case the target population is no longer available, and that she sees "the necessity of statistical inference in this case."

We have a few responses to Nane. We are not clear when, how, or why such distinctions would be made. We suspect that researchers would often, perhaps usually, be interested in evaluating "performance in general." But even if we accept her distinctions, we think the case for the use of statistical inference still holds. As we argued before, many arguments can be made for the super-population approach (e.g. the data are treated as a "realization" of some set of social process that could have in principle produced a very large number of other realizations). We are not sure why these arguments are suddenly no longer relevant in the examples Nane gives. If anything was on the line, we imagine that any institution that was found to be lacking in some way would want proof that luck and chance were unlikely to be responsible for the findings.

Nane further says that simply showing that something was a "common" practice was not a good enough reason for using it. We think we also showed that statistical inference with apparent populations is a good practice. Things could have come out differently than they did; apparent differences from predicted means or across institutions might be due to chance factors alone. Further, as we have just now argued, whether you agree with us on super-populations or not, your argument is made stronger if you can show that statistical inference approaches support the claims you are making.

Similar to Mutz, Nane argues that factors such as clustering should be considered when drawing a sample. We suspect this is unlikely to be a major problem in relatively simple analyses. But certainly we agree that this could be a concern in more complicated work. As we said before, we do not think this negates our arguments; rather, it indicates additional concerns that should be addressed when going beyond the relatively simple sets of analyses presented here. In a Letter to the Editor, Bornmann and Mutz (2013) proposed cluster samples. Mutz pointed out in this respect (see above) that dependencies in bibliometric data are a big challenge in using sampling techniques.

In short, we think Nane's arguments often reinforce our own. She suggests technical enhancements, such as clustering, that we agree would be good ideas if possible. She agrees with us that there are at least some situations where statistical inference with apparent populations is desirable. Our main disagreements with her are over the situations where she says statistical inference is not justified. We suspect that many researchers will be interested in evaluating what she calls "performance in general" and hence her distinctions may not matter that much in practice. But, even in the other situations she describes, we think it would be a mistake to ignore the role that chance factors could have had in observed outcomes.

## 6. Schneider

As we noted in our original paper, Schneider has previously expressed strong opposition to the position we have presented (Bornmann & Leydesdorff, 2013; Schneider, 2012). We are therefore not surprised that we have not changed his mind! Further it is clear we are in very good company. Schneider apparently rejects not only our specific suggestions concerning bibliometric analyses but any analyses which make an argument for apparent/super populations. We do not expect that this response will fare much better with him. But, we will do our best to address his concerns.

Schneider says that our approach is senseless since "further realizations are impossible in practice." But, we would argue that further realizations are occurring all the time; every day more and more citations get made. As we argued earlier, any year could be viewed as "a one-time sample of the unfolding of the universe."

We would also contend that Schneider seems to be arguing for a deterministic world; what actually did happen had to happen. So, if a swimmer wins an Olympic gold medal by 1/100th of a second, does that mean that victory was guaranteed? Of course not. Luck or random factors may well have made a difference on that particular day. Similarly, if an institution underperformed or over performed with regards to citation counts in a particular year, luck or chance could have been factors. In subsequent years the institution may be less or more lucky. For example, if an institution fared better in its citation counts between one year and the next, do we really want to declare that it got "better" as opposed to just "luckier?" When evaluating observed outcomes, we think it is a mistake to simply act as though things could not have come out any other way. They could have, and in the future they may come out differently even if nothing has fundamentally changed.

A good example for the randomness in citation processes is the linking of cited references to the corresponding cited publications (Marx & Bornmann, in press). Citations of database documents are lost more or less often. There are cited references variants in the databases which lead to the problem of citation linking or matching citations in bibliometrics (Olensky, Schmidt, & van Eck, 2015). Inaccurate cited references result in missed matches in the Web of Science and Scopus databases, which lead to reduced citation counts for papers. Depending on the definition of the errors, the scope of the data sets investigated, and the field of the cited references, missed matches in the Web of Science of between 5.6% (Olensky, 2015) and 12% (Hildebrandt & Larsen, 2008) are reported.

We think we have shown many arguments and benefits to using the super-population approach. Conversely, we don't think Schneider has shown any clear harms to using it. Indeed, Schneider himself has used statistical significance tests with bibliometric data (see, e.g., page 9 in Bloch, Schneider, & Sinkjær, 2016). Schneider does warn that there can be ethical problems when, for example, claims are made about false precision in estimates. But, if anything, we would argue that the problems of false precision are even greater when it is implied that the results are what they are and the possible role of chance factors should not be considered.

Researchers will have to decide for themselves which side of the issues they come down on. We think the case for our approach is very strong. But even those who disagree should consider doing as Schneider has done and use inferential statistics when they can reinforce the arguments they feel are already valid.

## 7. Conclusions

We find ourselves in substantial agreement with the discussants on many points. Indeed, many of the comments strike us as being enhancements or extensions of the techniques we propose. We discuss relatively simple t-tests. The results of those t-tests, however, can help to motivate far more complex analyses, e.g. if there are no differences between institutions you may not want to explore things any further; but if one institution is doing significantly better you may want to find out why. What are the variables that cause institutions to differ? We offered a few illustrative examples of variables that might be important but certainly greater theoretical and empirical depth would be desirable for a more in-depth analysis. For example, Bornmann, Mutz, Marx, Schier and Daniel (2011) use multilevel modeling techniques to examine the determinants of publication decisions by editors, while Mutz, Wolbring, & Daniel (2016) conduct a propensity score assessment of VIP (Very Important Paper) designation on citation impact. Rather than just refer to random or stochastic elements in citation processes, researchers can try to explicitly model the variables they think are important (see Tahamtan, Safipour Afshar, & Ahamdzadeh, 2016, for a review of several factors that may affect citations). Rather than draw a simple random sample or even a complete set of records, clustering and stratification could and often should be employed in the sampling process.

We are not sure how many minds will be changed by this debate over the handling of apparent populations. We have cited many sources who agree with us, but others remain unpersuaded. In our defense, we would primarily stress that what did happen is not what had to happen; when evaluating performance the potential role of chance and random factors should be considered. We reject the idea that further realizations are impossible; further realizations occur every day as more and more citations get made and more and more articles and books get published. We agree that proper modeling and sample selection can get complicated and that improvements on what we suggest are possible; but we think that this is an argument for even more sophisticated analyses, rather than falling back on simpler ones. We reiterate that the perfect should not be the enemy of the good; even in situations where our proposals may be imperfect, we think they will often be good or at least good enough, and almost certainly better than only settling for descriptive statistics.

For those who remain unpersuaded, we close with the argument this response began with: How can you best persuade those who disagree with you? You presumably have your own ideas on how populations of records should be discussed. Is it possible that your own arguments will be more persuasive if coupled with inferential statistics? Many people will expect to

see, indeed may even demand, that inferential statistics be included in an analysis. Given that many experts agree with them, it may be hard to convince them they are wrong. But, it may also be unnecessary. If you can show that multiple methods (even methods: that you consider flawed) all lead to the same conclusions, your case will be enhanced.

## References


Bielby, W., (2013). Managerial discretion, employment discrimination, and title VII class actions: Is there life after dukes v. Wal-Mart? Talk given at the University of Notre Dame, November 1, 2013.

Bloch, C., Schneider, J. W., & Sinkjær, T. (2016). Size, accumulation and performance for research grants: Examining the role of size for centres of excellence. *PLoS One*, *11*(2), e0147726. http://dx.doi.org/10.1371/journal.pone.0147726

Bornmann, L., & Leydesdorff, L. (2013). Statistical tests and research assessments: A comment on Schneider (2012). *Journal of the American Society for Information Science and Technology*, *64*(6), 1306–1308. http://dx.doi.org/10.1002/asi.22860

Bornmann, L., & Mutz, R. (2013). The advantage of the use of samples in evaluative bibliometric studies. *Journal of Informetrics*, *7*(1), 89–90. http://dx.doi.org/10.1016/j.joi.2012.08.002

Bornmann L., Mutz R., Marx W., Schier H., & Daniel H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high-profile journal select manuscripts that are highly cited after publication? Journal of the Royal Statistical Society—Series A (Statistics in Society), 174(4), 857–879. 10.1111/j.1467-985X.2011.00689.x.

Hildebrandt, A.L., & Larsen, B., (2008). Reference and citation errors: A study of three law journals. Presented at the 13th Nordic Workshop on Bibliometrics and Research Policy. 11–12 September 2008, Tampere (Finland).

Marx, W., Bornmann, L., (in press) Change of perspective: Bibliometrics from the point of view of cited references. A literature overview on approaches to the evaluation of cited references in bibliometrics. *Scientometrics*.

Mutz, R., Wolbring, T., & Daniel, H.-D. (2016). The effect of the very important paper (VIP) designation in Angewandte Chemie International Edition on citation impact: A propensity score matching analysis. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23701 [n/a-n/a]

Olensky, M., Schmidt, M., & van Eck, N. J. (2015). Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of science. *Journal of the Association for Information Science and Technology*, 21–22.

Olensky, M. (2015). Data accuracy in bibliometric data sources and its impact on citation matching. Doctoral dissertation. Humboldt-Universität zu Berlin (Germany). Retrieved May 19, 2016, from http://edoc.hu-berlin.de/dissertationen/olensky-marlies-2014-12-17/PDF/olensky.pdf.

Schneider, J. (2012). Testing university rankings statistically: Why this is not such a good idea after all. Some reflections on statistical power, effect sizes, random sampling and imaginary populations. In E. Archambault, Y. Gingras, & V. Lariviere (Eds.), *The 17th international conference on science and technology indicators* (pp. 719–732).

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, *107*(3) http://dx.doi.org/10.1007/s11192-016-1889-2

Williams, R., & Bornmann, L. (2014). The substantive and practical significance of citation impact differences between institutions: Guidelines for the analysis of percentiles using effect sizes and confidence intervals. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact:. methods and practice* (pp. S259–S281). Heidelberg, Germany: Springer.



Richard Williams*
*Department of Sociology, 810 Flanner Hall, University of Notre Dame, Notre Dame, IN 46556, USA*

Lutz Bornmann
*Division for Science and Innovation Studies Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539, Munich, Germany*

* Corresponding author.
*E-mail address:* Richard.A.Williams.5@nd.edu (R. Williams)

Available online 15 October 2016