There should not be any mystery: A comment on Sampling Issues in Bibliometrics

François Claveau
Université de Sherbrooke

June 30 2016

This document is the post-refereeing version of Journal of Informetrics 10-4 (2016): 1233-1240.

A research unit wants to assess whether its publications tend to be more cited than academic publications in general. This research unit could be anything from a lonesome researcher to a national research council sponsoring thousands of researchers. The unit has access to the (inverted) percentile ranks of n of its publications: each publication has an associated real number between 0 and 100, which measures the percentage of publications in its reference set that receive *at least* as many citations as its own citation count.[1] For instance, if the percentile rank of one publication is 10%, it means that 90% of publications in its reference set are less cited and 10% are cited at least as much. Now, say that the unit takes the *mean* of its n percentile ranks and reaches a value below 50%. Can the unit confidently conclude that it produces research that tends to be more cited than most publications?

The article by Richard Williams and Lutz Bornmann (in press) proposes to answer this kind of question by relying on standard statistical procedures of significance testing and power analysis. I am deeply sympathetic to this proposal. I find, however, their exposition to be sometimes clouded in mystery. I suspect that many readers will therefore be unconvinced by their proposal. In this comment, I endeavor to make a clearer case for their general strategy. It should not be mysterious why this strategy is sound. By clarifying the case, I show that some technical decisions of Williams and Bornmann are mistakes (choosing a two-tailed instead of a one-tailed test), lead to less efficient estimates (choosing the t-statistic instead of simply relying on the mean) or are not as prudent as they should be (presuming a particular standard deviation). Before making these technical points, I start with a more conceptual issue: in the next section, I dispel some confusion regarding the notion of randomness in the presentation of the authors.

About some mysterious claims on randomness

Williams and Bornmann offer an argument for using inferential statistics even when we have access to *all* the relevant data on the publications of the research unit under study (Section 3.1 in their article). An argument is needed because it is counter-intuitive to use inferential statistics when the full 'population' seems already accessible. Isn't inferential statistics about making inferences from sample to population? Why would we need its methods if all the observations are already at hand?

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.

Williams and Bornmann report this argument, but they then slip into talking about "the extent to which citations may have been influenced by random factors" (p. 9). They come up with a distinction between "random" factors and other, non-random factors. In this non-random category, we would have, primarily it seems, "the quality of the material in the papers" (p. 9). In the category of random factors, the examples

[1] see http://ipscience-help.thomsonreuters.com/incitesLiveESI/ESIGroup/fieldBaselines/PercentilesBaselines.html

given are: "how many people chose to read a particular issue of a journal or who happened to learn about a paper because somebody casually mentioned it to them" (pp. 9-10).

This distinction seems impossible to seriously uphold. How "casual" must a mention of a paper be to count as a "by chance" (p. 9) effect on citations? If most people read a given issue of a journal because they are regular readers of this journal, does the "how many people" factor become a non-random factor? And "the quality of the material" in a paper, why is this one not random? I personally cannot predict when I embark on a research project how important the output will be. Who does? The claim might instead be that a given quality translates into a number of citations in a deterministic fashion: if you can keep all the 'chancy' factors equal, an identical change in quality will be associated with a determinate change in the number of citations. This claim will sound odd to many. In contrast, I am inclined to endorse it. But I would add that, if it is plausible for the "quality" factor, it is also plausible for what the authors take to be "random" factors. For instance, if you keep all the other factors equal, a change in how many people read the paper will plausibly translate into citations.

Given how dubious the distinction between random and non-random factors is, it would be bad news if the strategy proposed by Williams and Bornmann required it. But there is no bad news: the mysterious distinction is irrelevant to the issue at stake. There is no need to be able to separate random from non-random factors to be justified in using inferential statistics even though all the realizations of the unit are given. What is needed is simply that the quantity *of interest* (here percentile ranking) is a random variable. And 'random variable' means that it is not possible in any given case to predict with certainty which value of the variable will be realized, although each value (or interval of values) has a probability of occurring. There is no need to track where the chance factors are among the factors causing the quantity.

In fact, random variables can even be the result of fully deterministic systems. Since Williams and Bornmann take the example of coin tosses as an analogy for their argument (p. 10), let me use the same case. Repeated tosses of a fair coin are properly modeled as independent realizations of a Bernouilli distribution with $p = .5$. The outcome is thus a random variable. Someone might want to say that coin tossing is affected by "chancy," "random" factors. But the statistical modeling of the phenomenon is also fully compatible with a belief that the data generating process is perfectly deterministic (Ford 1983): if only the *exact* initial conditions of the situation were known – angle of rotation, velocity, distance, etc. — and the *exact* laws of motion, the outcome could perfectly predicted. Yet, the statistical modeling of the situation remains the best option because we have limited knowledge of the initial conditions and the laws of motion, and the outcome depends so crucially on these conditions (i.e., it is a system of deterministic chaos).

The same holds for the data generating processes responsible for percentile ranks. Statistical modeling would remain a relevant modeling choice even if it is believed that there is nothing that happens "by chance" here. The phenomenon is best modeled statistically because there is limited knowledge of the actual factors present in each specific situation and how these factors combine to produce the result. There is no need of a mysterious quest for the "random factors." And there is room for inferential statistics: since realizations of percentile ranks follow a probability distribution, the full set of actual realizations (the sample) is not equivalent to the set of all possible realizations (the population). There is thus a gap that can be filled by inferential methods.

Why a specific t-test is fine

I now turn to discuss the technique proposed by Williams and Bornmann: a one sample t-test to decide whether the research unit in my introduction tends to produce research with more impact than most publications. It is a simple test, which is indeed appropriate here. Let me set up the problem and work through it to show the appropriateness of the solution. I will also argue that Williams and Bornmann turn out to use the wrong t-test and that, in fact, they could have done something even simpler.

Percentile ranks are attributed to each publication in a reference set by ordering these publications by their number of citations (from most cited to least cited) and mapping this ordering on the interval from 0 to 100. Given this procedure, the distribution of publications in the reference set is known: their distribution is

approximately uniform on the interval $[0, 100]$, with mean $\mu_0 = 50$ and standard deviation $\sigma_0 = 28.87$.[2] If the output of the research unit is neither more nor less impact prone than its reference set, its percentile ranks will follow the same uniform distribution. This possibility is the null hypothesis.[3]

What the researcher has access to is a sample of n percentile ranks for the research unit, $\{r_1, r_2, ..., r_n\}$. This sample can be the full set of realized ranks for the unit or only a subset of these ranks. As argued in the previous section, inferential techniques are relevant in both cases. Under the null hypothesis, once it is assumed that the elements of the sample have not been cherry-picked, it is known that the distribution of the mean of these ranks (the sampling distribution of $\bar{r}$) will, as the sample size grows, converge to a normal distribution with mean $\mu_{\bar{r}} = \mu_0$ and variance $\sigma^2_{\bar{r}} = \sigma^2_0/n$. This property is known thanks to the central limit theorem. The sampling distribution approaches normality really fast when the population is distributed uniformly, as the top graphs of Figure 1 illustrate. Panel (a) of Figure 1 plots the density of the sampling distribution with a sample size $n = 3$ against the normal distribution; the fit is already quite good.[4] Panel (b) does the same for $n = 10$; the two distributions are now virtually indistinguishable. Since Williams and Bornmann are working with samples that are far larger (the smallest sample being $n = 68$), normality of sampling can be safely assumed under the null hypothesis.

Someone could thus use the normality of the sampling distribution under the null hypothesis to compute the probability of type I error (the probability of rejecting the null hypothesis when it is true). With their t-test, Williams and Bornmann take a more roundabout method. This choice is, I must say, mysterious. The t-test is designed for observations that are assumed to follow a *normal* distribution with *unknown* variance. In the present case, the observations are, under the null hypothesis, *uniformly* distributed with *known* variance. But practically speaking – especially since t-statistics[5] are computed for us by statistical programs – there is little difference in the range of sample sizes considered by Williams and Bornmann. As Panels (c) and (d) at the

[2] For a continuous variable following a perfectly uniform distribution on the interval $[0, 100]$, the value for the mean is obviously 50. The value for the variance σ^2_0 (you take the square root for the standard deviation) can be computed based on the probability density, which is simply $f(x) = 0.01$ for $x \in [0, 100]$ and 0 otherwise:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_0^{100} (x-50)^2 0.01dx = 0.01 \int_0^{100} (x^2 - 100x + 2500)dx$$
$$= 0.01 \left(\frac{100^3}{3} - \frac{100 \times 100^2}{2} + 2500 \times 100 \right) = \frac{100^2}{3} - \frac{100^2}{2} + 2500 = 2500 - \frac{10000}{6} = \frac{5000}{6} = 833.\bar{3} \approx 28.87^2$$

As Williams and Bornmann note (fn. 1), the distribution of percentile ranks for the reference set is only *approximately* uniform because ties in citation counts result in publications sharing the same percentile rank. In particular, reference sets typically have a disproportionate amount of documents with no citation; all these documents have, according to the standard definition of percentile ranks, the same rank of 100.

It is beyond the scope of this comment to systematically assess how closely the distribution of actual percentile ranks for a variety of reference sets tracks the uniform distribution. Yet, I have computed the means and standard deviations of five reference sets constructed from a corpus of economics documents indexed by Web of Science. Each reference set is made of documents for a single year between 1996 and 2000 inclusively. The sizes of the sets range from 7977 to 8356 documents. As expected, the actual means of percentile ranks lie slightly above the assumed value of 50, the maximal value being 52.7. The standard deviations also lie above the assumed 28.87, the maximal value being 31.2. These departures from the assumed values are arguably small. Given that there is evidence that reference sets in economics are further from the uniform distribution than reference sets in major disciplines such biology, physics and so on (Waltman and Schreiber 2013, table 1), my results about economics give some justification to the claim that the uniform distribution is a pretty close approximation to a large proportion of reference sets out there. I will thus proceed for the rest of this comment by assuming the uniform distribution at the level of the reference set.

[3] More precisely, the condition that the unit's output is neither more nor less impact prone than the reference set implies that the mean of the unit's distribution of percentile ranks will be 50. I get that this distribution also has a standard deviation of 28.87 by assuming that the density function of the distribution is monotonic. For a justification of this monotonicity assumption, see the text preceding table 1 below. Table 1 reports that 28.87 is the only value of the standard deviation compatible with the monotonicity assumption and the assumption that the mean is 50. I thank an anonymous referee for pushing me to explain why I assume, together with Williams and Bornmann, that the standard deviation under the null hypothesis is 28.87. Obviously, if someone is not willing to make the monotonicity assumption, the statistical procedure would need to be modified to account for the uncertainty relative to the standard deviation of the research unit's distribution.

[4] All the sampling distributions come from Monte-Carlo experiments (1,000,000 repetitions). The R code used can be supplied upon request.

[5] As a reminder, the t-statistic is given by:

$$t = \frac{\bar{r} - \mu_0}{s_r/\sqrt{n}}$$

where μ_0 is the mean of the distribution under the null hypothesis (here 50), $\bar{r}$ is the sample mean, n is the sample size, and s_r

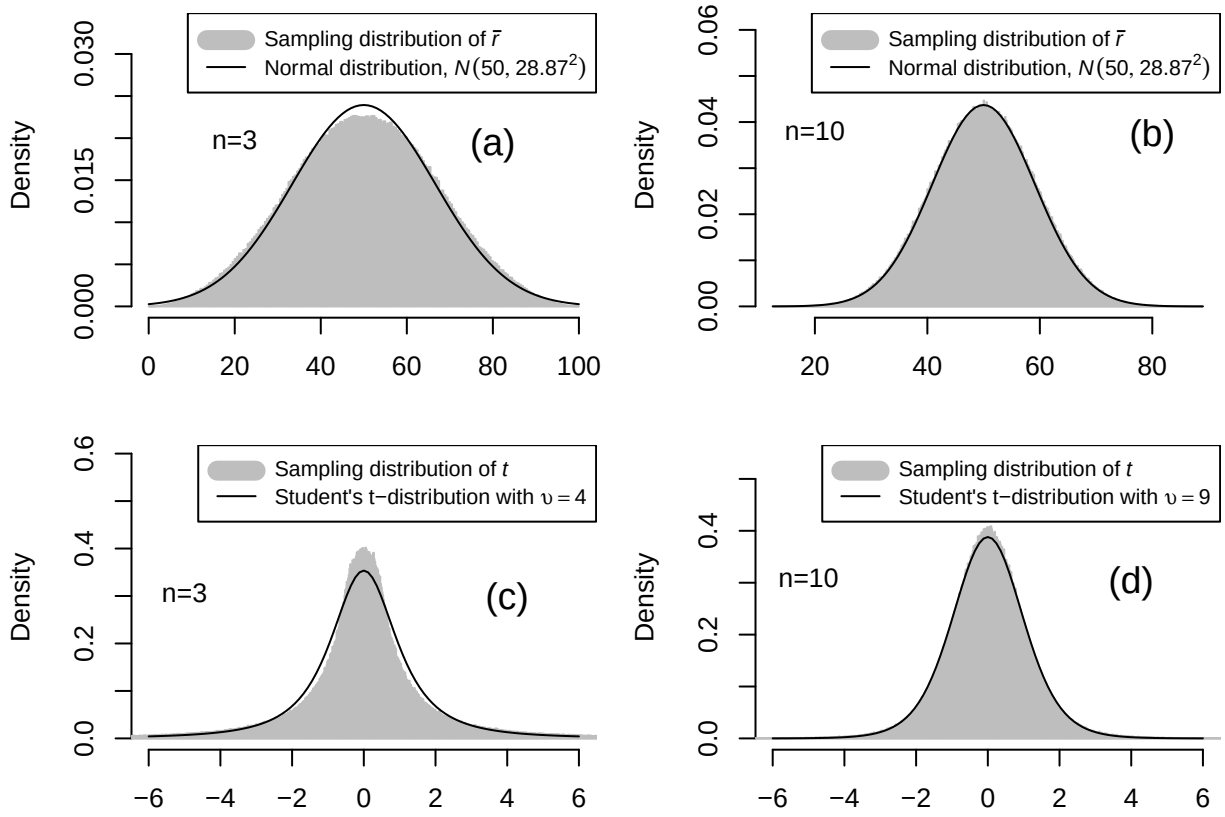

Figure 1: Comparison of sampling distributions with theoretical distributions for $n = 3$ and $n = 10$. The two panels above compare the distribution of means with the normal distribution. The two panels below compare the t-statistic with Student's t-distribution with the proper number of degrees of freedom.

bottom of Figure 1 show, the sampling distribution diverges significantly from Student's t-distribution at $n = 3$ and still somewhat at $n = 10$, but the difference vanishes at higher sample sizes.

For each sample $\{r_1, ..., r_n\}$, statistical programs will compute for you the t-statistic. This value can then be compared with the t-distribution with $n - 1$ degrees of freedom. Another mysterious thing happens at this point in Williams and Bornmann's article. The context is clearly one where the research unit wants a procedure to decide whether to reject the null hypothesis in favor of the alternative hypothesis that its publications tend to have *more* impact than the reference set. They write:

> So, for example, suppose that an institution believes that it is above average in terms of how often its publications get cited. If the papers of the institution really are above average, how much above average does it need to be, and how large does the sample need to be, in order to detect statistically significant differences from the average score in the reference sets (percentile=50)? A power analysis can be used to address such questions[.] (pp. 11-12)

The table that Williams and Bornmann present as the result of their power analysis (Table 1 in their article) uses a two-tailed t-test instead of a one-tailed test (they do not state that they use a two-tailed test, but this is what you find out if you try to replicate their results). This choice implies that what they are really testing against the null hypothesis is not that the research unit is "above average," but that it tends to produce either more or *less* cited publications than the reference set. They do not use the right version of the t-test.

Correcting this mistake is crucial because it leads to estimated quantities significantly different from the true quantities. For instance, if the sample size (n) equals 200 and the t-statistic equals -1.7, the p-value relevant for the decision to reject or not the null hypothesis will be 0.09 for a two-tailed test and only 0.045 for a one-tailed test. If the significance level used is $\alpha = .05$, the null hypothesis will be rejected under the appropriate one-tailed test, but not under the test used by Williams and Bornmann. Fortunately, this mistake is extremely easy to correct in R (R Core Team 2016) or any other statistical software.[6]

So far, the (corrected) procedure sketched is simple and sound. Given a sample of publications from the research unit, a t-statistic can be calculated. Assuming that the null hypothesis is true, the approximate distribution of the t-statistic when the sample size is big enough is Student's t-distribution with $n - 1$ degrees of freedom. It is thus possible to compute the probability of having a value of t as low as or lower than the one for the sample. If this probability is lower than what we are willing to risk as a type I error, the null hypothesis is rejected in favor of the alternative hypothesis that the research unit has a tendency to produce higher impact research than its reference set.

Why power analysis is on slightly more shaky ground

Williams and Bornmann add only one layer of complexity when they bring power analysis in the picture. This layer is not as straightforwardly sound as the simple procedure of the previous section. As I will show, the issue is that power analysis relies on guessing an unknown quantity – the standard deviation under the alternative hypothesis. Fortunately, for the cases considered by the authors, the plausible values of this quantity lie in a relatively small interval. Errors in guesses thus translate into relatively minor errors in the conclusion of the power analysis.

Power analysis uses the fact that the following quantities are jointly determined (Cohen 1988):

- The sample size n;

is the standard deviation of the sample given by:

$$s_r = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2}$$

[6] In R for instance, the command will be `t.test(r,mu=50, alternative="less")` instead of simply `t.test(r,mu=50)`, where `r` in both cases is the sample of percentile ranks.

- The probability α of committing a type I error;
- The probability β of committing a type II error – i.e., not rejecting the null hypothesis although the alternative hypothesis is true. The *power* is $1 - \beta$, the probability of rejecting the null given the truth of the alternative hypothesis.
- The effect size:

$$d = \frac{\mu_1 - \mu_0}{\sigma_u}$$

where μ_1 is the mean percentile rank for the unit under the alternative hypothesis that this unit tends to produce research with more impact than its reference set; $\mu_0 = 50$ is the mean under the null hypothesis; σ_u is the standard deviation of percentile ranks for the unit.

If it can be assumed that the sampling distribution of the t-statistic follows Student's t-distribution with $n - 1$ degrees of freedom (an acceptable assumption for large enough samples as shown above), the value of one of the four quantities (n, α, $1 - \beta$, d) can be determined by giving values to the three others. For instance, someone could fix the sample size to $n = 200$, set the desired level of statistical significance to $\alpha = 0.05$, hypothesize that the research unit has a mean percentile rank 5 point lower than the reference set ($\mu_1 - \mu_0 = -5$) and, finally, postulate that the distribution producing the percentile ranks of the research unit has a standard deviation equal to the reference set ($\sigma_u = 28.87$). Given all this information, the probability that the sample to be drawn will lead to the rejection of the null hypothesis can be computed – i.e., the power of the test. The reason why power can be inferred from the previous information is illustrated in Figure 2. The sample size gives the degrees of freedom of Student's t-distribution, the effect size gives the displacement of the density from the null hypothesis (black line) to the alternative hypothesis (gray line), the level of statistical significance gives the critical value of the t-statistic. All of this information taken together gives the proportion under the alternative t-distribution that lies in the rejection region ($1 - \beta = .787$ in this case).

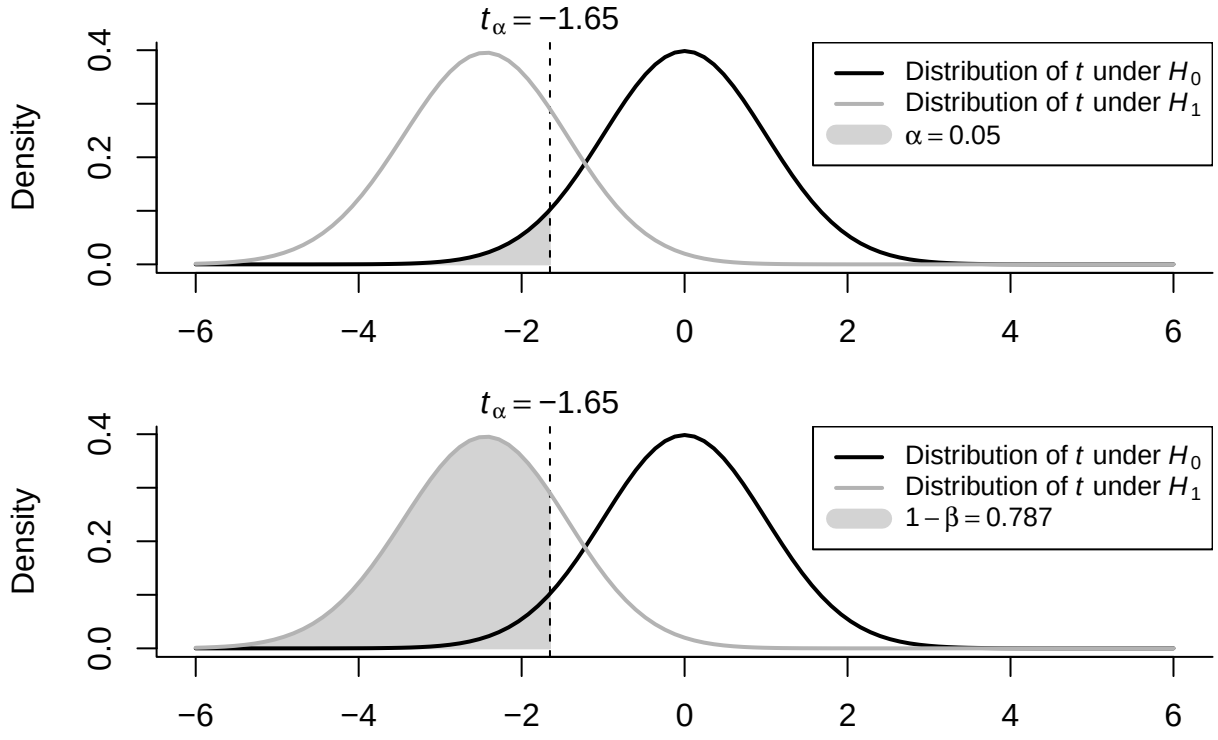

Figure 2: Density of the t-statistic showing the relationship between level of significance and power when sample size and effect size are given. Here, $n = 200$ and $d = (45 - 50)/28.87$.

Power analysis is quite simple and widely used. But it must be noticed that an assumption with little justification has been included in the process of computing the missing quantity: the standard deviation of

the distribution from which the sample of percentile ranks is drawn is made to be equal to the standard deviation of the reference set ($\sigma_u = \sigma_0 = 28.87$). In their paper, Williams and Bornmann write that "[i]t is common to assume that the sample standard deviation will be the same, although the bibliometrician could choose some other value if there were reason to believe otherwise." (p. 12) They say nothing about this mysterious "reason to believe otherwise." To dispel this mystery, I want to address two questions. First, Williams and Bornmann's burden of proof should be reversed: Are there reasons to believe that the "common" assumption is close to the truth? Second, do plausible departures from this assumption have important implications for the conclusion of the power analysis?

What can be said with certainty is that the true standard deviation of the research unit's distribution lies somewhere in the interval: $\sigma_u \in \left[0, \sqrt{\mu_1(100-\mu_1)}\right]$. The lower bound of $\sigma_u = 0$ corresponds to the implausible situation where a unit is such that it always produces publications situated at the same percentile rank. The first graph of Figure 3 is the density of a distribution approaching this total concentration. The upper bound $\sigma_u = \sqrt{\mu_1(100-\mu_1)}$ corresponds to another implausible situation where all the percentile ranks lie at the two extremes of the interval $[0, 100]$;[7] a case approaching this situation is illustrated by the middle graph of Figure 3. Somewhere in between these two extremes lies a distribution with a mean different from the mean of the reference set ($\mu_1 \neq \mu_0 = 50$), but with an identical standard deviation ($\sigma_u = \sigma_0 = 28.87$). The last graph of Figure 3 illustrates this situation.

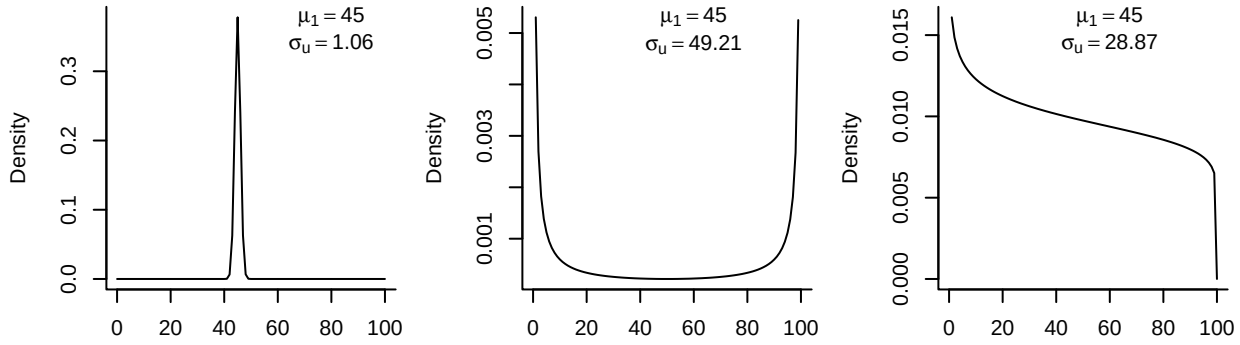

Figure 3: Densities of possible distributions for the percentile ranks r_i of the research unit. All distributions have the same mean, but different standard deviations. The underlying distributions are beta distributions with different values of the two parameters.

The range of *possible* values of σ_u is thus quite large; for the values of μ_1 considered by Williams and Bornmann's (between 40 and 47.5), the interval of possible values is close to $[0, 50]$. Although all these values are possible, they are not all *plausible*. It would indeed be extremely surprising that a research unit's distribution of percentile ranks resembles the first two cases in Figure 3. I submit that plausible distributions when $\mu_1 < \mu_0$ must meet the following condition: the density of percentile ranks must be decreasing over the full interval $[0, 100]$. This condition is met by the last distribution of Figure 3, but not by the first two. When it is not met, at least one of two counterintuitive results occurs. First, the probability of realizing a percentile might diminish when approaching higher-impact percentiles. This behavior is exemplified by the first graph of Figure 3: on average, the publications of the unit have more impact than the reference set, but the unit cannot produce publications in the highest-impact percentiles, say $[0, 10]$. Second, the probability of realizing a percentile might rise when moving toward lower-impact percentiles. The second graph of Figure 3 exemplifies this possibility: the unit, although it produces on average higher impact publications than the reference set, also produces a disproportionate amount of extremely low-impact publications, say in the region $[90, 100]$.

[7] In this case, $\sigma_u = \sqrt{\mu_1(100-\mu_1)}$ because, in order to have the hypothesized μ_1, the probability of having a realization at the lower bound of the interval $[0, 100]$ must be $p_{\text{lb}} = (100-\mu_1)/100$ and the probability at the upper bound must be $p_{\text{ub}} = \mu_1/100$. The variance of the distribution is thus:

$$\sigma_u^2 = p_{\text{lb}}(0-\mu_1)^2 + p_{\text{ub}}(100-\mu_1)^2 = \frac{(100-\mu_1)\mu_1^2}{100} + \frac{\mu_1(100-\mu_1)^2}{100} = \mu_1(100-\mu_1)$$

I don't want to be read as maintaining that distributions with the two peculiar behaviors described above will never be realized, but I maintain that they are far less likely than distributions that meet the condition of decreasing density.[8] In consequence, I grant to Williams and Bornmann that the *likely* value of σ_u is not anywhere on the large interval $\left[0, \sqrt{\mu_1(100-\mu_1)}\right]$, but on a smaller interval determined by the condition of decreasing density. If I make the relatively weak assumption that the distribution of $r_i/100$ is a beta distribution,[9] this smaller interval can be determined exactly:[10]

$$\sigma_u \in \left[\sqrt{\frac{(100-\mu_1)\mu_1^2}{100+\mu_1}}, \sqrt{\frac{\mu_1(100-\mu_1)^2}{200-\mu_1}}\right]$$

As can be seen in Table 1, the interval of *plausible* values of σ_u is much smaller than the interval of *possible* values. For the values of the mean under the alternative hypothesis that Williams and Bornmann consider (values between 47.5 and 40), assuming as they do that $\sigma_u = 28.87$ can in the worst of the plausible cases cause an overestimation of the true standard deviation by 10 % or an underestimation of 4 %. Note, however, that for values of μ_1 that are much lower than 50 (i.e., the last rows of Table 1), the assumed value of $\sigma_u = 28.87$ does not even lie inside the plausible interval.

Table 1: Interval of plausible values of the standard deviation given a hypothesized mean.

mean	lower bound	upper bound
50.0	28.87	28.87
47.5	28.34	29.30
45.0	27.71	29.63
42.5	27.00	29.87
40.0	26.19	30.00
30.0	22.01	29.41
20.0	16.33	26.67
10.0	9.05	20.65

In sum, my answer to the first question – whether there are reasons to believe that Williams and Bornmann's assumption $\sigma_u = 28.87$ is close to the truth – is that this assumption is most probably quite accurate for the mean values that they consider, but that it drops in credibility when 'elite' research units are considered (the ones having mean percentile ranks far below 50).

Now to the second question: Do plausible departures from the assumption $\sigma_u = 28.87$ have important implications for the conclusion of the power analysis? Table 2 offers an answer to this question. It reports different values of power for the significance level $\alpha = .05$, a hypothesized mean (first column) and a sample size (second column). The 'assumed' power comes from assuming that $\sigma_u = 28.87$ and that the sampling distribution of the t-statistic follows a (noncentral) Student's t-distribution with $n-1$ degrees of freedom.[11]

[8] A systematic empirical analysis of how frequently percentile distributions of research units do exhibit these peculiar behaviors would be relevant, but it lies beyond the scope of this comment.

[9] Since the beta distribution – a two parameter distribution for a continuous variable with support on the interval $[0, 1]$ – is a highly flexible distribution, I suspect that the results to follow are true more generally. Unfortunately, I cannot supply a more general analysis at the moment.

[10] This result comes from the fact that the beta distribution is decreasing if and only if its two parameters p and q are such that $p \leq 1$ and $q \geq 1$. The two parameters are related to the mean and standard deviation of percentile ranks in the following way:

$$\mu_r = \frac{100p}{p+q} \qquad \sigma_r = \sqrt{\frac{10000pq}{(p+q)^2(p+q+1)}}$$

From the mean, one parameter can be expressed in terms of the other: $p = q\mu_r/(100-\mu_r)$ and $q = (100-\mu_r)p/\mu_r$. The bounds of σ_r under the condition of decreasing density can then be found by solving for $p = 1$ in one case and $q = 1$ in the other.

[11] In R, the code to retrieve this power is:

The two other columns reporting powers do not make these two assumptions. First, they use the standard deviations in Table 1 instead of 28.87. Second, they are arrived at through Monte-Carlo experiments (100,000 repetitions) using samples drawn from the relevant beta distribution. The reported powers are thus the proportion of runs in which the null hypothesis is rejected.

Table 2: Assumed and real powers of the t-test with a 0.05 significance level. The assumed power uses the assumption of a standard deviation equal to 28.87, the other two powers use the standard deviations in Table 1. Sample size n is chosen to have assumed power be approximately 0.8.

mean	n	assumed	lower-bound std. dev.	upper-bound std. dev.
47.5	826	0.800	0.811	0.791
45.0	208	0.801	0.827	0.779
42.5	93	0.800	0.840	0.773
40.0	53	0.801	0.853	0.764
30.0	15	0.817	0.933	0.768
20.0	8	0.840	0.989	0.806
10.0	5	0.813	1.000	0.833

Table 2 shows that the assumed power is quite close to all plausible powers for values of mean percentiles lower but still close to 50. For $\mu_1 \geq 40$, assumed power, at worst, overestimates true power by around 5 % or underestimates it by around 6.5 %. But again, elite research units are different: if the mean percentile rank is far below 50, true power can be underestimated by a large margin.[12]

Fail-safe tables with one-tailed t-tests

In their article, Williams and Bornmann present three tables based on their specification of power analysis. In the first two tables, they estimate the minimum sample size required to detect with sufficient power a given difference in means. In the last table, they estimate the minimum difference in means that can be detected with a specific power given a sample size. I have indicated above two main reasons why the estimated values they report can be biased. The first reason is that, most simply and most importantly, they do not use the appropriate t-test: they use a two-tailed, one-sample t-test while they should be using a one-tailed, one-sample t-test. This choice gives an upward bias to their estimated sample size and their estimated difference in means. The second reason is that the true standard deviation of the research unit's distribution, σ_u, might be different from the one they postulate (28.87). If σ_u turns out to be higher than presupposed, the estimated sample size and the estimated minimum difference in means will suffer from a downward bias. The two reasons thus produce biases in opposite directions; they partly, but not perfectly cancel each other.

In Table 3, sample size estimations are reported. These estimations are 'fail-safe' in the sense that they use the biggest plausible standard deviation (see Table 1) and are thus unlikely to underestimate the required sample. Although these estimations play safe, the sample sizes are still lower than the ones reported by

`power.t.test(n=n,delta = abs(mu1-50)/28.87 ,alternative="one.sided",type="one.sample")`
where `n` is the sample size and `mu1` is the mean under the alternative hypothesis.

[12] It might be worth noting that, for small differences between μ_0 and μ_1, the difference in power is almost exclusively accounted for by the fact that the standard deviation used is off the mark. For bigger differences, the assumption that t follows Student's t-distribution becomes a factor of error. This behavior can be explained by the fact that samples drop to small values when $\mu_1 - \mu_0$ grows large; and I have shown above that the distributional assumption is far less accurate with extremely small samples (see Figure 1).

Finally, I also note that Williams and Bornmann's reliance on t-statistics and t-distributions increase error in comparison to the direct use of the sample mean with the normal distribution. For instance, their procedure can, with $\mu_1 = 40$ and $n = 53$, underestimated power by .055, while a procedure based on sample mean and the normal distribution would, at worst, underestimate it by .024. R code for these results can be provided upon request.

Williams and Bornmann (compare the last two columns of Table 3). This comes from the fact that they are based on one-tailed t-tests while Williams and Bornmann overestimate required sample sizes because they use two-tailed tests.

Table 3: Estimation by power analysis of required sample sizes to have sufficient power to detect a hypothesized difference in mean. The last column gives sample sizes for one-tailed t-tests based on the worst-case plausible scenario of Table 1. The second-to-last column gives, for comparison purposes, the sample sizes proposed by Williams and Bornmann (combining Tables 1 and 2 from their article).

Alpha	Power	Mean	W&B n	Fail-safe n
0.05	0.8	47.5	1049	851
0.05	0.8	45.0	264	219
0.05	0.8	42.5	119	100
0.05	0.8	40.0	68	58
0.01	0.9	47.5	1988	1791
0.01	0.9	45.0	500	460
0.01	0.9	42.5	224	210
0.01	0.9	40.0	128	120

Table 4 reports estimates of how small the unit's mean μ_1 must be to achieve the chosen power $1 - \beta$ given α and n. Again, the 'fail-safe' μ_1 is based on the least favorable, but still plausible value of σ_u. Since the two biases described above play against each other, the difference of these 'fail-safe' μ_1 with Williams and Bornmann's estimates is quite small (compare the third-to-last and the last columns of Table 4).

Table 4: Estimation by power analysis of the maximum mean that can be detected with sufficient power given a significance level and a fixed sample size. The last column gives this target mean for one-tailed t-tests based on the worst-case plausible scenario of Table 1. The third-to-last column gives, for comparison purposes, the target mean estimated by Williams and Bornmann (Table 3 from their article).

Alpha	Power	n	W&B mean	Delta	Fail-safe mean
0.01	0.8	200	42.96	-0.2437	42.72
0.05	0.8	200	44.25	-0.1991	44.08
0.10	0.8	200	44.91	-0.1764	44.77

The general results of these corrected tables are thus not far from what Williams and Bornmann report, but that can be attributed to a form of luck on their side: they have introduced two biases that correct each other to a great extent.

Conclusion

Inferential statistics, including widespread methods such as hypothesis testing and power analysis, are well worth using in bibliometrics. But we must be careful in justifying and deploying them. Improper justifications can lead to confusion while faulty assumptions can significantly bias results. The article by Williams and Bornmann is sometimes confused in its justification (e.g., with respect to the issue of randomness) and it makes

specific choices that are clearly faulty or, at a minimum, not particularly prudent. What I have attempted here is to make a more compelling case for the general strategy proposed by Williams and Bornmann. It should not be a mystery why their general proposal is sound.

References

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2 edition. Hillsdale, N.J: Routledge.

Ford, Joseph. 1983. "How Random Is a Coin Toss?" *Physics Today* 36 (April): 40–47. doi:10.1063/1.2915570.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Waltman, Ludo, and Michael Schreiber. 2013. "On the Calculation of Percentile-Based Bibliometric Indicators." *Journal of the American Society for Information Science and Technology* 64 (2): 372–79. doi:10.1002/asi.22775.

Williams, Richard, and Lutz Bornmann. in press. "Sampling Issues in Bibliometric Analysis." *Journal of Informetrics*.