Testing University Rankings Statistically: Why this Perhaps is not such a Good Idea after All. Some Reflections on Statistical Power, Effect Size, Random Sampling and Imaginary Populations

Jesper W. Schneider

jws@cfa.au.dk
Danish Centre for Studies in Research & Research Policy, Department of Political Science & Government, Aarhus University, Finlandsgade 4, Aarhus, 8200 N (Denmark)

Abstract

In this paper we discuss and question the use of statistical significance tests in relation to university rankings as recently suggested. We outline the assumptions behind and interpretations of statistical significance tests and relate this to examples from the recent SCImago Institutions Ranking. By use of statistical power analyses and demonstration of effect sizes, we emphasize that importance of empirical findings lies in "differences that make a difference" and not statistical significance tests *per se*. Finally we discuss the crucial assumption of randomness and question the presumption that randomness is present in the university ranking data. We conclude that the application of statistical significance tests in relation to university rankings, as recently advocated, is problematic and can be misleading.

Introduction

In a number of recent publications (e.g., Leydsdorff & Opthof, 2010a, 2010b; Leydesdorff & *al*., 2011; Opthof & Leydesdorff, 2010) statistical significance tests have been presented as a profitable decision making tool for comparison of units of analyses in research evaluations. Most recently, two smaller publications advocate and demonstrate the use of statistical significance tests in relation to two university rankings, the Leiden Ranking 2011/2012 (Leydesdorff & Bornmann, 2011) and the SCImago Institutions Rankings 2011 (Bornmann, Moya-Anegón & Leydesdorff, *in press*), respectively. In these papers we are either told explicitly that statistical significance tests are advantageous, or this impression is implicit in the narrative. However we are not told what purpose such tests seem to serve, to what degree assumptions for such tests are met, and most importantly, *why* statistical significances tests are supposed to be advantageous. No such warrant for the application of statistical significance tests gives an impression that consensus rules in relation to their use, including interpretation of the results. But this is not the case in the social, behavioral and life sciences. Probably very few methodological issues have generated as much controversy as the use of statistical significance tests (e.g., Berkson, 1938; Rozeboom, 1960; Bakan, 1966; Carver, 1978; Meehl; 1978; 1990; Guttman, 1985; McCloskey, 1985; Oakes, 1986; Rothman, 1986; Cohen, 1990; 1994; Tukey, 1991; Gigerenzer, 1993; Goodman, 1993; 2008; McCloskey & Ziliak, 1996; Schmidt & Hunter, 1997; to name just a few critical works from different fields out of literally hundreds if not thousands).

Statistical significance tests are surrounded by myths. They are overused and are very often misunderstood and misused (see Kline, 2004 for an overview). Criticisms are numerous. Some point to the inherently logical flaws in statistical significance tests (e.g., Cohen, 1994). Others claim that such tests have no scientific relevance; in fact they may be harmful (e.g., Armstrong,

2007). Others have documented a whole catalogue of misinterpretations of statistical significance tests and especially the *p* value (e.g., Oakes, 1986). Still others have documented various different misuses, such as neglecting statistical power, indifference to randomness (probability sampling and/or random assignment), adherence to a mechanical ritual, arbitrary significance levels forcing dichotomous decision making, and implausible nil null hypotheses, to name some (e.g. Gigerenzer, 1993; Shaver, 1993).

It is not our intention in this paper to give a detailed formal discussion of statistical significance tests and their many problems. Such presentations are plenty, see for example the abovementioned references. In this paper we take a narrow point of view and argue that the use of statistical significances tests in relation to university rankings, as suggested by Leydesdorff, Bornman and Moya-Anegón (*in press*), is problematic. We discuss the assumptions behind and interpretations of statistical significance tests and relate this to examples from the SCImago Institutions Ranking. By use of statistical power analyses and demonstration of effect sizes, we emphasize that importance of empirical findings lies in "differences that make a difference" and not statistical significance tests. Finally we discuss the crucial assumption of randomness, and question the presumption that it is present in the observations in university rankings. The paper is organized as follows: The first section briefly explains the interpretation of *p* values and a couple of fallacies related to this statistic. The second section discusses an example from Leydesdorff, Bornman and Moya-Anegón (*in press*) through the lenses of statistical power. The discussion continues in section three where the example is discussed in relation to effect size. The fourth section, discusses the crucial assumption of randomness, and the final section is the conclusion.

Statistical significance tests and p values

First of all it is important to emphasize that a phrase like "[indicator values] ... can be tested statistically for significant differences" (Leydesdorff & Bornmann, 2011), by no means imply that a potential "significant difference" is important. It is not clear whether the authors imply that, but it is certainly not within the realm of statistical significance tests in themselves to be able to say anything in relation to the importance of the findings (e.g., McCloskey & Ziliak, 1996). "Significance" has a very limited meaning. The definition of *p* values, the product of statistical significance tests, is as follows:

> The probability of the observed data, plus more extreme data across all possible random samples, *if* the null hypothesis is true, *given* randomness[1] and a sample size of *n* (i.e., the sample size used in the particular study), and all assumptions of the test statistic are satisfied (e.g., Goodman, 2008, p. 136).

Notice, "across all possible random samples" is a long-run frequentist interpretation. The general form can be written as: p (Data$|H_0$). While the mathematical definition of the *p* value is rather simple, its meaning has shown to be very difficult to interpret correctly. Carver (1978), Kline (2004) and Goodman (2008) list many misconceptions about *p* values. For example, the incorrect interpretation that if $p = .05$, the null hypothesis has only a 5% chance of being true. As the *p* value is calculated under the assumption that the null hypothesis *is* true, it cannot simultaneously be a probability that the null hypothesis is false. Cohen, summed up this confusion by concluding that significance testing "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (Cohen,

[1] We use randomness to include both random sampling and random assignment.

1994, p. 997). Cronbach and Snow (1977) succinctly summarize this particular fantasy in terms of probability statements:

> A *p* value reached by classical methods is not a summary of the data. Nor does the *p* values attached to a result tell how strong or dependable the particular result is ... Writers and readers are all too likely to read 0.05 as *p*(H|E), 'the probability that the Hypothesis is true, given the Evidence.' As textbooks on statistics reiterate almost in vain, *p* is *p*(E|H), the probability that this Evidence would arise if the [null] hypothesis is true. Only Bayesian statistics yield statements about *p*(H|E). (p. 52).

It is commonly assumed that the *p* value is indicative of the meaningfulness or importance of a finding. The meaningfulness of a finding, however, can only be evaluated subjectively in the context of theory and/or application. The meaningfulness of findings is reflected in parameter estimates—effect sizes—and these estimates can have large or small *p* values, depending on the sample size.

Researchers often believe that the use of statistical significance tests provides an objective way to make decisions about data. However, the fact that researchers do not rely upon formal guidelines for selecting *N* allows them to support or refute the null hypothesis simply by adjusting the sample size. If the null hypothesis is false, we can refute it by sampling a large number of subjects or corroborate it by sampling too few subjects.

Regardless of how the *p* value is interpreted, hypothesis testing is illogical, because the null hypothesis of no association or no difference is almost always false in observational studies, there are no truly zero effects in nature (Lykken, 1968). The more relevant issue is not whether there is an effect, but rather how large the effect is and to what degree it makes a difference if any. Unfortunately, the *p* value alone provides us with no information about the direction or size of the effect or, given sampling variability, the range of estimated values. Depending, *inter alia*, on sample size and variability, an outcome statistic with $p < .05$ could represent an effect that is theoretically, practically, or mechanistically irrelevant. Conversely, a non-significant result does not necessarily imply that there is no worthwhile effect, as a combination of small sample size and large measurement variability may mask important effects.

Ranking examples and statistical power

It is important to emphasize that the following comments are not restricted to the specific test suggested for use with the indicator for the top 10% highly cited papers in the SCImago Institutions Ranking (Bornmann, Moya-Anegón & Leydesdorff, *in press*). The comments are general and holds for all statistical significance tests. In the two university ranking examples, the *z*-test is recommended for testing whether the observed proportion of top 10% highly cited papers for an institution differs significantly from the expected distribution, or whether the proportion for two institutions are significantly different. In the Scimago Institutions Rankings, for example, the difference between UCLA and Stanford University is tested. In the case of UCLA and Stanford University, one could ask what is the purpose of testing the difference between the two institutions; is it 1) to make a decision whether the difference is important?, 2) to test whether we can generalize the result to some unknown population in the future or outside Scopus? or 3) both of these questions? We use the UCLA and Stanford University example in the following discussion.

It is well known that *p* values are essentially a function of effect and sample sizes. In principle, a large effect combined with a small sample size can result in statistical significance and the same goes for a small effect combined with a large sample size.

In the SCImago Institutions Ranking example (Bornmann, Moya-Anegón & Leydesdorff, *in press*), the difference between UCLA and Stanford University is used as an exemplary case for the proposed statistical test for determining differences between institutions in university rankings. The indicator value tested is the "excellence indicator," which is the institution's percentage of highly cited papers among the top 10% highly cited papers within its fields of publication. When comparing two institutions, the difference in proportions is therefore tested and the *z*-test for two independent proportions may be appropriate, though the independence assumptions is most likely violated due, for example, to common publications. Leave that aside. We are told that UCLA is ranked at the 17th position in the SCImago Institutions Ranking and that the publication output is 37,994 papers (*n*) and the excellence indicator is 28.9%. Stanford University is ranked 19th with 37,885 papers and an excellence indicator of 29.1%. The observed difference in proportions is thus .2 percentage points. Using the *z*-test, the difference between UCLA and Stanford University is now tested and the conclusion goes that "the difference between these two institutions ... is not statistically significant" (the *p* value is .54 but not reported in the paper).

What is the purpose of comparing these two institutions (or any two institutions) by use of statistical significance tests? What information do we think we get? Many apply statistical significance tests as a decision making tool to determine whether an effect is "real", and if so, it is judged "significant" implying importance, where "real" refers to uncertainty and "statistical significant" means that the effect is "real" in the population. Unfortunately this common understanding is flawed and builds on several crucial assumptions as outlined in the previous section.

Let us scrutinize the UCLA and Stanford University example, where we failed to reject the null hypothesis of no difference between the two proportions. At face value the result seems obvious, the difference between the institutions' proportions are indeed small at .02 percentage points. But do we need a statistical significance tests to point that out? Three issues are important here, the meaning of *p* values, statistical power and uncertainty. The meaning of *p* values was discussed in the previous section and we reiterate that a *p* value is a conditional probability of the data given that H_0 is true. Failure to reject H_0, however, does not mean that H_0 is true, only that we have failed to reject it given the applied α and β levels, as well as effect and samples sizes. Failure to reject immediately raises the question about the statistical power in a given study. The power of a statistical significance test equals the probability (1 - β) of detecting a particular effect when the null hypothesis (H_0) is false. Power is the complement of the probability of a type II error.

In the present case the statistical power is .09[2] for a two-tailed test with a 5% significance level, which means that we are able to detect the difference of .02 between UCLA and Stanford University in nine out of hundred equal tests. In other words, given the very small "effect size" the current set up has a very poor chance of detecting a false null hypotheses, i.e. that there *is* a

[2] Achieved power is calculated in a post hoc power analysis for z tests of independent proportions using *G*Power 3*: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

"real" difference in the population. With almost two times 38,000 papers, the sample sizes can be considered very large indeed. The problem, of course, is the small effect size.

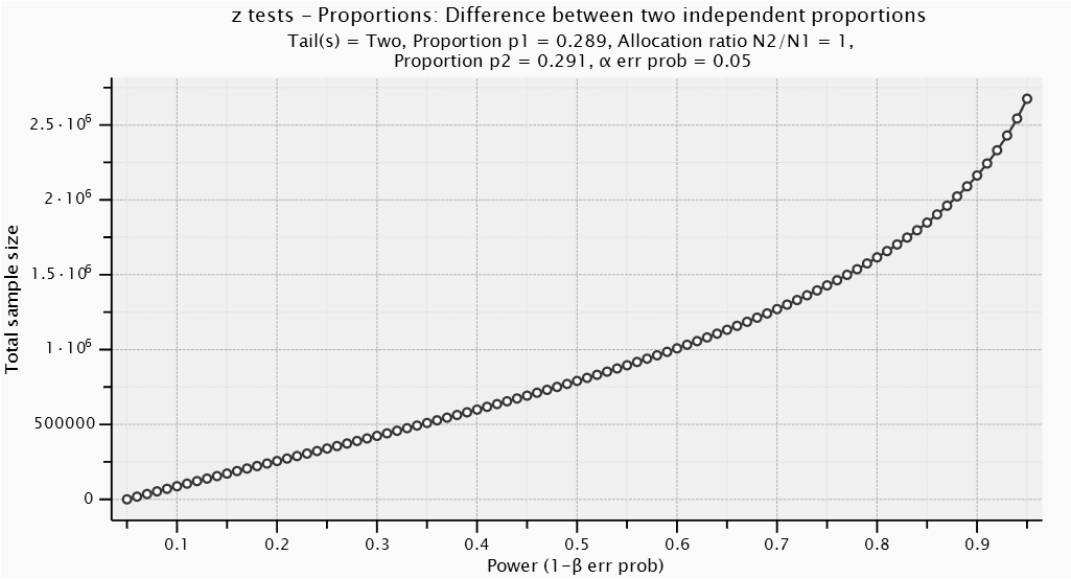

Figure 1. Statistical power as a function of sample size.

The probability of detecting an effect increases with effect size and with sample size. In the statistical literature .80 is often suggested as a reasonable degree of power. In the present case, the sample size needed to detect a difference of size .02 with a power of .80, given two independent samples of almost equal size, is approximately 808,000 papers for each institution, or a total sample size of approximately 1,616,000. papers For a power of .50, the same as tossing a coin, 395,400 papers in each sample is required. Figure 1 above show the level of power as a function of sample size.

Alternatively, the required effect size in a two-tailed test with .80 statistical power, is an "excellence indicator" of 30% for UCLA, when Stanford University's effect size remains 29.1%. That is an increase of .9 percentage point. However, it is of course nonsense to speak of "raising" the effect size. What we can speak of is variability in statistics and raising the sample size. As we will discuss below, this is also meaningless in the university ranking examples, as samples sizes in their thirty-thousands have very low variability, if any, such "samples" may well be conceived of as apparent populations. How can we randomly enlarge the sample size for an institution for a given period in the current setup? We do not think it is possible, rendering statistical significance tests meaningless. More on this in the next section.

Now we return to the question of whether H_0 is true or false. Given the conditions of statistical significance tests, H_0 must be true in the population. Usually this means that no difference in proportions is assumed in the population. Notice that this is a nil null hypothesis meaning no difference to the infinitesimal decimal (Cohen, 1994). Given the relation between sample and effect size, failing to reject H_0 should immediately prompt a question of whether the study is underpowered. It is obvious that a very small effect size requires a very large sample size to be statistical significant and significant it will be at some point because, as we argued above, most null hypotheses are in fact not true (e.g., Lykken, 1968; Meehl, 1990). A small difference always

exist also between UCLA and Stanford University. We just do not have sufficient power to detect it. If the null hypothesis is wrong, statistical significance tests become meaningless as the conditional probability breaks down. There are no Type I errors only Type II.

So the difference between UCLA and Stanford University is not statistically significant. We know it will be at some point but only with a spectacularly large sample size, which are in fact imaginary. Can we then imply that the difference between UCLA and Stanford University is unimportant. Well, yes and no. No because we cannot infer the importance of a result from the statistical significance test. And yes, because the difference is unimportant, the effect size is miniscule, but that information was actually evident prior to testing.

If we instead compare Stanford University with Cambridge University, then the difference turns out to be statistically significant. Cambridge University is ranked 27th in the SCImago Institutions Ranking, with 32,900 papers and an "excellence indicator" of 26.7%. Obviously power is 1, however, if we considered a power calculation with the given effect size we would see that a total sample size of approximately 8,600 (split in two) is needed for a power of .80 or 8 out of 10 times detecting the difference. Is the difference of 2.4 percentage points between Stanford University and Cambridge University a "difference that makes a difference"? Judging which differences that make a difference or the importance of results should be based on effect sizes. The next section will discuss the importance of effect sizes.

Ranking examples and effect sizes

Effect sizes come either as unstandardized statistics, such as means or regression coefficients, or as standardized statistics, such as correlation coefficients and Cohen's *d* (Vacha-Hasse & Thompson, 2004). Effect sizes assess the magnitude or strength of the findings, critical information that cannot be obtained solely by focusing on a particular *p* value such as .05. There is no straightforward relationship between a *p* value and the magnitude of effect. A small *p* value can relate to a low, medium, or high effect. Moreover, there is no straightforward relationship between the magnitude of an effect and its importance. Depending on the circumstances, an effect of lower magnitude on one outcome can be more important than an effect of higher magnitude on another outcome.

A number of different standardized effect sizes exist. Cohen's *h* statistic is the most useful measure of effect sizes when data is proportions, such as the *z*-statistic for independent proportions (Cohen, 1988). The *h* statistic employs an arcsine transformation of the percentages to correct for the fact that the parameters of a distribution of percentages cannot be known, which makes the determination of a standard deviation impossible. According to Cohen (1988), a rough rule of thumb would start with the assumption that $h \approx .20$ is a "small" effect size of negligible practical importance, corresponding to a correlation coefficient of .10; that $h \approx .50$ is a "medium" effect size of moderate practical importance, equivalent to an *r* of .25; and that $h \approx .80$ is a "large" effect size of crucial practical importance, implying an *r* of .37-.39. However, he also stresses that judgments about the relative importance of effect sizes must be determined by experience within each area of investigation. In principle, we are skeptical of this categorization as it encourages "mechanical thinking" similar to statistical significance tests. Practical importance of an effect depends entirely on its relative costs and benefits. A "one size fit all" yardstick is problematic as Cohen himself argued. However, for the principled arguments in this paper the categorization may do.

Consider Figure 2 below. We have calculated differences in "excellence indicator" values between Stanford University and all other institutions on the SCImago Institutions Ranking and subsequently calculated their Cohen's *h* statistics. Figure 2 shows that approximately 39% of the observed differences in "excellence" indicator values between Stanford University and one of the other 3023 institutions, can be considered as "small effects" according to Cohen's (1988) rule of thumb.

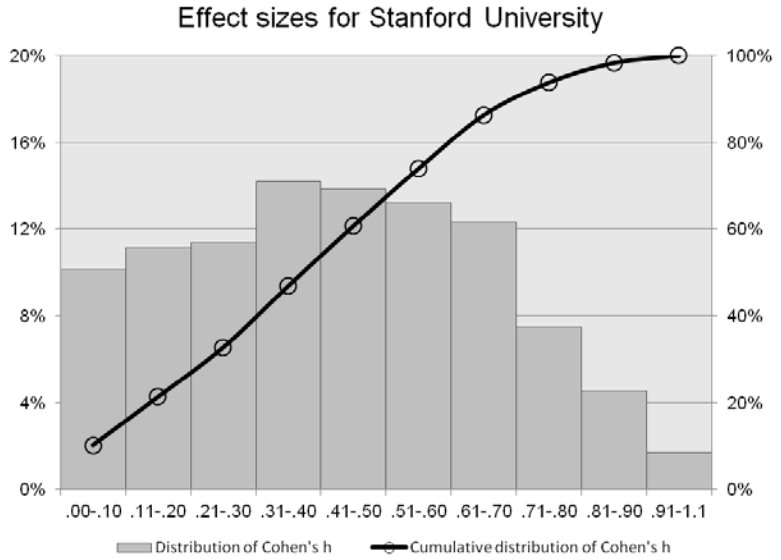

Figure 2. Cohen's *h*, effect size, for Stanford University compared to other institutions on the SCImago Institutions Rankings.

Further, 33% of the observed differences can be considered of "medium size," and 6% of "large size." Finally, approximately 21% of all comparisons between Stanford University and other institutions can be considered of such "trivial effect," according to Cohen's rule of thumb, that they are most likely irrelevant.

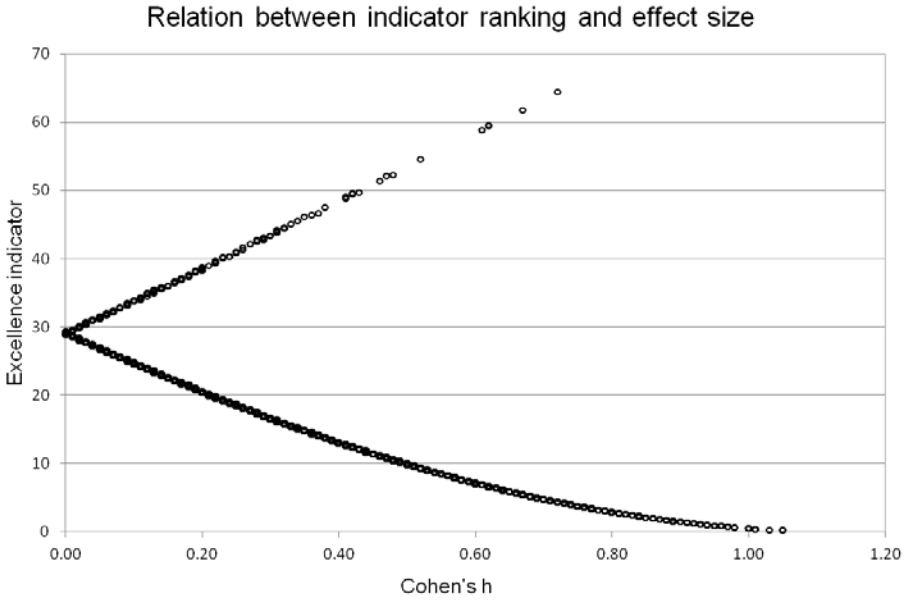

Figure 3. Relation between "excellence" indicators values (rankings) and Cohen's *h*; the case of Stanford University.

If we look a Figure 3 above, we can see that we need to go beyond the ranking interval of 20-38% for the "excellence indicator" values on the *y* axis in order to obtain at least "small" effect sizes for Stanford University when compared with one of the other institutions (.20 on the *x* axis). In other words, Stanford University is ranked 219 according to their "excellence indicator" of 29.1%. In total 678 institutions immediately above or below Stanford University's ranking—ranking positions 61 to 738—have "trivial" differences in indicator values when compared to Stanford University, according to Cohen's qualitative categories. In order to obtain "medium" effect sizes, the indicator for the institution compared to Stanford University would have be on or above 54.6% or on or below 9.8% for the "excellence indicator" values on the *y* axis.

The example given by Bornmann, Leydesdorff and Moya-Anegón failed to detect a statistical significant difference in "excellence indicators" between UCLA (28.9%) and Stanford University (29.1%). Cohen's *h* for this effect is actually .004, which is extremely small and as we demonstrated above a gigantic sample size is needed for detection of this difference. The difference is most certainly of no real importance. We knew, or could have known that before we commenced with the statistical significance tests. But most importantly, the statistical significance test, i.e., failure to reject H_0, cannot say anything about the probability of the hypothesis of no importance, given the data. It is the other way around, the probability of the data given that the null hypothesis is true, which we claim is most likely not true, as there is an extremely small effect in the population. If we accept this claim, then, we are dealing with a Type II error instead. Leaving that aside, we also tested the difference between Stanford University, 29.1%, and Cambridge University, 26.7%. This time the difference was statistically different, but the effect size is only .07. Still a miniscule effect size, nowhere near the "small" effect delineation defined by Cohen. In other words, statistical significance in this case simply means that with have a sufficiently large sample to detect the actual difference. The importance of the result cannot be inferred from the fact that the result is statistically significant.

Stanford University's profile, as shown in Figure 3, is most likely common for most institutions on the SCImago ranking. The important point is that one should consider what an important "effect size" might be before one sets out to detect "significant differences." An a priori power analysis can compute the required sample size for a given effect size, with a certain power level such as .80. If statistical significance tests are to be used at all, such a priori calculations are necessary. Most importantly they inform about the study's power to detect differences, but they will probably also reveal for the investigator that many differences between institutions in the rankings are so small that statistical significance tests become completely irrelevant. The tests simply cannot detect such small effect sizes given the actual sample available, if indeed the samples can be considered random samples. Further, a failure to reject H_0, which would be the consequence, can easily be misinterpreted as no effect.

Forcing investigators to think about effect sizes is very important. Normally, we apply statistical significance tests in research. In research, before testing commences, we should judge potential effect sizes based upon theory, former research, and the potential relevance and importance of findings. Judgment should not be dichotomous, i.e. whether something is significant or not. Magnitudes and importance is a matter of degree. In research evaluation matters are different. We do not have any theory to guide our interpretation of effect sizes. We can construct and calculate indicators, which is a research activity; but we fail to explain how and when indicator values become important. What constitutes an important difference in indicators between two

institutions on a university ranking? University rankings and league tables combined with statistical significance tests are simple but unsatisfactory solutions. We simply cannot say what is an important difference between university *x* and university *y* measured with the "excellence indicator"? We do not know. It is a serious mistake to think that statistical significance test can do the work for us. Interpretation of such results can easily be misleading. For the sake of argument, we have used Cohen's qualitative categories for interpreting the potential importance of findings, but we stress that these categories have no theoretical or empirical warrant in relation to university rankings.

If we tend to forget or disregard the basic assumptions of statistical significance tests, i.e., randomness, power analyses may reveal the problem for us. The case of university rankings is striking. In what way can UCLA's 37,994 papers or Stanford University's 37,885 papers be considered a random sample? What population are they a random sample of? How are we to enlarge these samples in order to get more power to be able to detect the genuine but minuscule difference of .02 percentage points? We cannot, because these "samples" in our opinion are either apparent populations or convenience samples. We will discuss this in the following section.

Randomness

Statistical significance tests concern sampling error and we sample in order to make statistical inferences, either descriptive inferences from sample to population or causal claims (Greenland, 1990). Statistical inference relies on probability theory. In order for probability theory and statistical tests to work *randomness* is required. This is a mathematical necessity as standard errors and *p* values are estimated in distributions that assume random sampling from well-defined populations (Berk & Freedman, 2003). Information on how data is generated becomes critical when we go beyond description. In other words, when we make statistical inferences we assume that data are generated by a stochastic mechanism and/or that data are assigned to treatments randomly. The empirical world has a structure that typically negates the possibility of random selection unless random sampling is imposed. Ideally, random sampling ensures that sample units are selected independently and with a known nonzero chance of being selected (Shaver, 1993). As a consequence, random samples should come from well-defined finite populations, not "imaginary" or "super-populations" (Berk & Freedman, 2003). With random sampling an important empirical matter is resolved. Without random sampling, we must legitimate that the nature produced the equivalent of a random sample or constructed the data in a manner that can be accurately represented by a convenient and well-understood model. Alternatively, data may constitute a convenience sample or an apparent population (or a census from a population) (Berk, Western & Weiss, 1995).

Very few observational studies using inferential statistics in the social sciences clarify how data are generated, what chance mechanism is assumed if any, or define the population to which results are generalized, whether explicitly or implicitly. Presumably, most observational studies, also in our field, are based on convenience and not probability samples (Kline, 2004). Albeit many social scientists do it, it is nevertheless a category mistake to make statistical inferences based upon samples of convenience. With convenience samples, bias is to be expected and independence becomes problematic (Copas & Li, 1997). When independence is lacking conventional estimation procedures will likely provide incorrect standard errors and *p* values can be grossly misleading. Beck and Freedman (2003) suggest that standard errors and *p* values will

be too small, and that many research results are held to be statistically significant when they are the mere product of chance variation. Indeed, there really is no point in addressing sampling error when there is no random mechanism to ensure that the probability and mathematical theory behind the calibration is working consistently.

In what sense does the observations used for statistical significance tests in the university rankings constitute a probability sample? And how is the population defined? It is most likely that the 37,885 papers from Stanford University are selected because they constitute all eligible publications of certain publication types in the database for that institution in the given period, or rather all publications that can be identified. Data therefore constitutes all the available observations from the "apparent" population of publications affiliated with Stanford University or in the general case, all other institutions. If so, frequentist inference based on long-run interpretation of some repeatable data mechanism is not appropriate. There is no uncertainty due to variation in repeated sampling from the population.

A counter argument could be that "the data are just one of many possible data sets that could have been generated if the publication history of the institution were to be replayed many times over." But this does not clarify what sampling mechanism selected the history we happened to observe. No one knows, or can know. It is simply not relevant for the problem at hand to think of observations as draws from a random process when further realizations are impossible in practice and lack meaning even as abstract propositions. Adhering to a frequentist conception of probability in the face of non-repeatable data and in a non-stochastic setting seems dubious.

Neither can the set of publications identified for the SCImago Institutions Ranking in the specific citation database be considered a random draw from the finite population of all papers affiliated with the institution, including those external to the database. It is unlikely that the data generation mechanism can be stochastic when governed by indexing policies in one database. Most likely, the data set constitutes a convenience sample of specific publication types coincidentally indexed in the specific database. Convenience samples are often treated as if they were a random realization from some large, poorly-defined population. This unsupported assumption is sometimes called the "super-population model" (Cochran, 1953). While some authors argue that "super-populations" are justifiable for statistical significance test (e.g., Bollen, 1995), we do not find such arguments convincing for frequentist statistics of non-experimental data. Super-populations are defined in a circular way as the population from which the data would have come if the data were a random sample (Berk & Freedman, 2003). Super-populations are imaginary with no empirical existence, as a consequence, they do not generate real statistics and inferences to them do not directly answer any empirical questions. What draw from an "imaginary super-population" does the real-world sample we have in hand represent? We simply cannot know. Inferences to imaginary populations are also imaginary (Berk & Freedman, 2003).[3]

[3] Notice, there is an important difference between imaginary populations that plausibly could exist and those that could not. An imaginary population is produced by some real and well-defined stochastic process. The conditioning circumstances and stochastic processes are clearly articulated often in mathematical terms. In the natural sciences such imaginary populations are common. This is not the case in the social sciences, yet super-populations are very often assumed, but seldom justified.

One could of course treat data as an apparent population. In this non-stochastic setting statistical inference is unnecessary because all the available information is collected. Nonetheless, we often still produce standard errors and significance tests for such settings, but their contextual meaning is obscure. There is no sampling error, means, percentages and variances are population parameters. Notice, population parameters can still be inaccurate due to measurement error, an issue seldom discussed in relation to citation indicators. Leaving measurement error aside for a moment, what we are left with is the indicator, the actual parameter, what used to be the estimated statistic. Informed judgment, based on effect sizes, not mechanical decision making, is now needed to compare and evaluate the indicator value of the institutions.

Notwithstanding the basic violation of assumptions, we think it is questionable to put so must trust in significance tests with sharp margins of failure when our data and measurements most likely at best are imprecise. In practice sampling is complicated and because even well-designed probability samples are usually implemented imperfectly, the usefulness of statistical inference will usually be a matter of degree. Nevertheless, this is rarely reflected upon. In practice sampling assumptions are most often left unconsidered. We believe that the reason why the assumption of randomness is often ignored is the widespread and indiscriminate misuse of statistical significance tests which may have created a cognitive illusion where assumptions behind such tests have "elapsed" from our minds and their results are thought to be something they are not, namely decision statements about the importance of the findings. One can always make inferences but statistical inferences come with restrictive assumptions and frequentist inference is not applicable in non-stochastic settings. The incontrovertible fact is that in nonprobability sampling, it is not possible to estimate sampling errors. Therefore, validity of inferences to a population cannot be ascertained. If one's sample is very large, then "chance" becomes irrelevant. Language should be driven entirely by the magnitude of the expected or observed effect. Consequently, statistical significance tests applied to university rankings as suggested by Bornmann, Leydesdorff, Moya-Anegón (*in press*) is also meaningless from the point of view of randomness, as observations cannot be considered as random draws of some large known population.

Conclusions

Despite their numerous applications, statistical significance tests are controversial. The information they produce is quite limited, conditional on some crucial assumptions, and they need to be interpreted within a frequentist statistical framework. Most often information from such tests are used for dichotomous decision making. Too often "significant" results are taken to be important results. This is not necessarily the case and certainly not a question statistical significance tests can give an algorithmic answer to. Statistical significance tests only say something about the probability of data conditional on the null hypothesis being true and that the assumption of randomness is fulfilled. Notice, it is a long run interpretation, entailing repeated random sampling. The latter is crucial. Without a random sample from a known population, the probability calculus breaks down, and external validity becomes problematic. The sole purpose of statistical significance tests is addressing sampling error. This is not possible with convenience samples. Consequently, statistical significance tests, if used at all, should be used properly within a domain that satisfies the most crucial assumptions.

For matters of importance, emphasis should be placed on effect sizes, former research, theory and the research design. Uncertainty—sampling error—is better addressed by confidence intervals and/or re-sampling techniques if randomness is present, but the most efficient way to deal with uncertainty is to replicate studies.

In this paper we have questioned the use of statistical significance tests in relation to testing differences between institutions on university rankings as recently suggested in two small papers by Bornmann, Leydedorff and Moya-Anegón. Leaving the general controversies concerning such tests aside, in this paper we have addressed a number of specific issues related to their application in relation to university rankings. The issues we have discussed are deeply interrelated. We claim that statistical significance tests are most often applied mechanically in order to bring about a dichotomous decision whether a difference or association is statistically significant. The latter is most often taken to mean an important difference. Whereas the size of the difference, if it is statistically significant is seldom reflected upon.

Looking at university rankings and the application of statistical significance tests in that contexts, we have pointed out some important issues that needs to be considered. We argue that an institution's number of papers for a given period, which go into the calculation of indicators and is represented in the rankings, are *not* a random sample from some known population. They either constitute an apparent population for the institution in the given period or simply a convenience sample. Assuming that the number of papers constitute a random sample is problematic because we need to know a random sample from what? We cannot know whether the stochastic mechanisms work if we do not have a clue as to what we draw from. It is a general problem for our field. Making statistical inferences from a specific citation database to a population outside this database is illusionary. Such a population is imaginary, empirically non-existing, and statistical inferences to it is also imaginary. Within a citation databases we can draw random samples from defined populations or we can work with the whole population. In the latter case, statistical significances tests become irrelevant as there is no sampling error.

The question about the importance of differences, for example, between institutions in a university ranking should be based on informed human judgment, where the size of the difference is interpreted in relation to former research, consequences, relevance and theory. The problem in relation to theory and research evaluation is that the field is predominantly a-theoretic. We cannot find good evidence in theories for why a difference between two indicators and thus two research institutions should be important and others not. Bringing in statistical significance tests as the mechanical judge is extremely problematic. Statistical significance tests cannot address such issues and in the current example of university rankings they become meaningless as they are applied to populations.

If uncertainty is the issue, then an approach like the “stability interval” in the Leiden Ranking is laudable. Notice, it is not a confidence interval in the frequentist definition. Parameters are known as we are dealing with populations. What is tested by resampling techniques is the stability of the indicators in relation to the underlying set of documents, i.e. documents in the extreme ends of the distribution.

References

Armstrong, J.S. (2007). Significance Tests Harm Progress in Forecasting. *International Journal of Forecasting*, 23 (2), 321-327.

Bakan, D. (1966). The Test of Significance in Psychological Research. *Psychological Bulletin*, 66 (6), 423-437.

Berk, R.A., Wester, B. & Weiss, R.E. (1995). Statistical Inference for Apparent Populations. *Sociological Methodology*, 25, p. 421-458.

Berk, R.A. & Freedman, D.A. (2003). Statistical Assumptions as Empirical Commitments. In T. G. Blomberg & S. Cohen (Eds.), *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger* (pp. 235-254). New York: Aldine.

Berkson, J. (1938). Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test. *Journal of the American Statistical Association*, 33 (203), 526-536.

Bollen, K.A. (1995). Apparent and Nonapparent Significance Tests. *Sociological Methodology*, 25, p. 459-268.

Bornmann, L., de Moya-Anegón, F. & Leydesdorff, L. (in press). The New Excellence Indicator in the World Report of the SCImago Institutions Rankings 2011. *Journal of Informetrics*, preprint available at http://arxiv.org/abs/1110.2305.

Carver, R. P. (1978). The Case Against Statistical Significance Testing. *Harvard Educational Review*, 48 (3), 378-399.

Cochran, W.G. (1953). *Sampling Techniques*. New York: Wiley.

Cohen, J. (1990). Things I Have Learned (so far). *American Psychologist*, 45 (12), 1304-1312.

Cohen, J. (1994). The Earth is Round ($p<.05$). *American Psychologist*, 49 (12), 997-1003.

Copas, J.B. & Li, H.G. (1997). Inference for Non-Random Samples (with discussion). *Journal Royal Statistical Society*, B, 59 (1), 55-95.

Cronbach, L.J. & Snow, R.E. (1977). *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York: Irvington.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in Statistical Reasoning. In G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological* Issues (pp. 311-339). Hillsdale: Erlbaum.

Goodman, S.N. (1993). *P* Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate. *American Journal of Epidemiology*, 137 (5), 485-496.

Goodman, S.N. (2008). A Dirty Dozen: Twelve *p*-Value Misconceptions. *Seminars in Hematology*, 45 (3), 135-140.

Guttman, L. (1985). The Illogic of Statistical Inference for Cumulative Science. *Applied Stochastic Models and Data Analysis*, 1 (1), 3-10.

Greenland, S. (1990). Randomization, Statistics, and Causal Inference. *Epidemiology*, 1 (6), 421-429.

Kline, R.B. (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research.* Washington: American Psychological Association.

Leydesdorff, L. & Opthof, T. (2010a). Normalization at the Field Level: Fractional Counting of Citations. *Journal of Informetrics*, 4 (4), 644-646.

Leydesdorff, L. & Opthof, T. (2010b). Normalization, CWTS Indicators, and the Leiden Rankings: Differences in Citation Behavior at the Level of Fields. arXiv:1003.3977v3. http://arxiv.org/ftp/arxiv/papers/1003/1003.3977.pdf

Leydesdorff, L. & Bornmann, L. (2011). Testing Differences Statistically with the Leiden Ranking. http://arxiv.org/abs/1112.4037

Leydesdorff, L., Bornmann, L., Mutz, R. & Opthof, T. (2011). Turning the Tables on Citation Analysis one More Time: Principles for Comparing Sets of Documents. *Journal of the American Society for Information Science and Technology*, 62 (7), 1370-1381.

Lykken, D.T. (1968). Statistical Significance in Psychological Research. *Psychological Bulletin*, vol. 70 (3, part 1), 151-159.

McCloskey, D.N. (1985). The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests. *American Economic Review*, 75 (2), 201-205.
McCloskey, D.N. & Ziliak, S.T. (1996). The Standard Error of Regression. *Journal of Economic Literature*, 34 (3), 97-114.
Meehl, P.E. (1978). Theoretical Risks and Tabular Asterisk: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Counseling and Clinical Psychology*, 46 (4), 806-834.
Meehl, P.E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, 1 (2), 108-141.
Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.
Opthof, T. & Leydesdorff, L. (2010). Caveats for the Journal and Field Normalizations in the CWTS ("Leiden") Evaluations of Research Performance. *Journal of Informetrics*, 4 (3), 423-430.
Rothman, K.J. (1986). Significance Questing. *Annals of Internal Medicine*, 105 (3), 445-447.
Rozeboom, W.W. (1960). The Fallacy of the Null Hypothesis Significance Test. *Psychological Bulletin*, 57 (5), 416-428.
Schmidt, F.L. & Hunter, J.E. (1997). Eight Common but False Objections to the Discontinuation of Significance Testing in the Analysis of Research Data (pp. 37-64). In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there Were no Significance Tests?* Hillsdale: Erlbaum.
Shaver, J.P. (1993). What Statistical Significance Testing Is, and What it Is Not. *Journal of Experimental Education*, 61 (4), 293-316.
Tukey, J.W. (1991). The Philosophy of Multiple Comparisons. *Statistical Science*, 6 (1), 100-116.
Vacha-Haase, T. & Thompson, B. (2004). How to Estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology*, 51 (4), 473-481.