Scientometrics (2020) 125:925–935
https://doi.org/10.1007/s11192-020-03476-8

Mean values of skewed distributions in the bibliometric assessment of research units

Ulrich Schmoch[1]

Received: 4 November 2019 / Published online: 29 April 2020

Abstract
Nearly all distributions in bibliometrics are skewed. In particular, the distribution of citations of publications by research units is skewed. In a statistical view, the calculation of mean values can imply misleading or even wrong information. However, in citation analysis, the calculation of mean values of skewed distributions are standard. Therefore, when ranking research units, it is recommended instead to replace the calculation of standard mean values by the calculation of adjusted mean values to exclude outliers with very high citations and those with very few or no citations as well. Such an adjusted mean value is oriented on the standard activity of a research unit and results in a more adequate assessment. This approach is based on the Hirsch-index concept. The calculation results in a different ranking of research units, which may be important in cases where the distribution of finances depends on bibliometric rankings. In addition, a differentiation between standard activities and excellent results is possible, thus opening two dimensions of the assessment of research units.

Keywords Skewed distribution · Mean value · Adjusted mean value · Hirsch-index · Bibliometrics · Ranking of research units

Mathematics Subject Classification 11H60

JEL Classification C46

Introduction

Mean values and skewed distributions are two major topics in the discourse on bibliometrics. Mean values are used to assess the citation score of organisations or countries or the citation behaviour in specific fields. To calculate the standardization of citations, the mean value of citations per author/organisation is divided by the mean value of the field of publication. Mean values are regularly used for the assessment of citation performance, in particular.

✉ Ulrich Schmoch
ulrich.schmoch@isi.fraunhofer.de

[1] Fraunhofer ISI, Karlsruhe, Germany

The skewed distribution is a general observation applying to the publications per author or the citations per author. A general formulation of skewed distributions in bibliometrics was provided by Lotka, and skewed distributions can be found in all types of bibliometric analysis. Sometimes, the distributions are extremely skewed. Against this background, the question may be asked whether calculating standard mean values in bibliometrics leads to meaningful results. According to Siegel (2017):

> One of the problems with skewness in data is that… many of the most common statistical methods… require at least an approximately normal distribution. When these methods are used on skewed data, the answers can at times be misleading and (in extreme cases) just plain wrong. Even when the answers are basically correct, there is often some efficiency lost; essentially, the analysis has not made the best use of all of the information in the data set.

This paper suggests an alternative approach to calculating mean values and discusses its implications for rankings of research units.

The discourse on skewness

A fundamental early publication on skewness is that by Lotka (1926). In this contribution, the author suggests a distribution of the publications per author according to the formula:

$$X^n = C/Y$$

wherein X number of publications, C, n = constants depending on the specific field (with $n \sim 2$), Y the relative frequency of authors with X publications.

This means that, e.g. 10 authors among 100 have only 1 publication, 25 authors have about 4 publications and 1 author has 100 publications. A graph of the papers written and the percentage of authors yields a quite skewed distribution with a very small number of authors having a very high number of publications and many authors with very few publications. Various studies were conducted to verify the so-called Lotka's Law, e.g. Murphy (1973), Pao (1986) or Radhakrishnan (1973). In most cases, Lotka's Law was confirmed if the examined samples were sufficiently large. Subsequent to Lotka (1926), various other suggestions were made for skewed distributions, e.g. Chen and Leimkuhler (1986) discussed Lotka's, Bradford's and Zipf's Law; Simon (1955) suggested alternative functions, and these were then modified by Mandelbrot (1959), and a general statistical description was presented by Adamic (2002). Skewed distributions were found for many areas beyond publication productivity, e.g. for linguistics (Zipf 1949) and income distribution (Pareto 1935). A good overview is provided by Newman (2005). Skewed distributions are a frequent phenomenon in science and therefore also in bibliometrics. In particular, skewness is characteristic for citation patterns (Seglen 1992).

Albarrán and Ruiz-Castillo (2011) and Albarrán et al. (2011) examined 22 scientific fields and 219 sub-fields respectively of the Web of Science and find highly skewed distributions for citations. In about 64% of the sub-fields power laws exist and 2% of the publications account for 13.5% of all citations. Thus, a distinct skewness is confirmed.

As skewness is such an important phenomenon in bibliometrics, well known authors have discussed it. De Solla Price (1976) described this topic in terms of cumulative advantage in detail; Narin and Hamilton (1996) emphasise the relevance of highly cited publications or patents within skewed distributions, and Glänzel and Moed (2005) discuss journal

impact factors and state a statistical reliability despite skewed distributions of the citations. In detail, they state:

> In contrast to the common misbelief statistical methods can be applied to discrete 'skewed' distributions and the statistical reliability of these statistics can be used as a basis for application of journal impact measures in comparative analyses.

However, other authors see the need for a special treatment of skewed distributions, in particular, if the skewness is strong or some values are extreme (Siegel 2017; Statistics how to 2019; Von Hippel 2005). In particular, Von Hippel (2005) states:

> In a data analysis course, it is certainly possible to continue teaching the relationship between skew, median, and mean. The treatment, however, should be more qualified than it is in current textbooks…. it should be pointed out that the rule is imperfect, and that the most common exceptions occur when the variable is discrete.

Against this background, Lundberg (2007) suggests a modified version of the so-called Crown Indicator suggested by the bibliometric group in Leiden, in particular Moed et al. (1995), The original version of the Crown Indicator is defined as follows:

$$\mathrm{CI} = \mathrm{CPP}/\mathrm{FCSm}$$

wherein CPP mean citation rate of a set of papers, FCSm mean citation rate of the field where the papers belong to.

He discusses various aspects of an appropriate calculation of the crown indicator. One issue is "that the distribution of citations over publications is highly skewed", as CPP and FCSm have skewed distributions. He suggests "to make normalizations using logarithmically transformed citation rates." He concedes that due to the logarithmic transformations of citation rates, extreme values have less impact, and suggests to provide the field normalized citation score (the crown index) in addition to consider extreme cases as well. A shortcoming of this suggestion is that in the field normalized citation score, the lower citation scores dominate the mean value and the extreme cases are not well reflected. Leydesdorff and Bornmann (2011) suggest to use percentile ranks instead of mean values to cope with skewed distributions and develop an integrated impact indicator. Rousseau (2011) discusses this approach in a theoretical perspective and confirms its validity. This approach is definitively an appropriate solution to deal with skewness, however, the application in broader studies of different research units proves to be quite intricate and complex, so that it is not used in practice to a broader extent.

For instance, Opthof and Leydesdorff (2010) suggest that the "normalization can be performed using non-parametric statistics such as comparing percentile rank scores", but in the end they still use mean values.

A prominent contribution to the quantification of an individual's scientific research output was made by Hirsch (2005). The so-called Hirsch-index or h-index is calculated by counting the number of publications h for which an author has been cited by other authors at least that same number of times. This measure de facto implies that extremely high citation values are neglected as are very low ones. The implications of the h-index are illustrated in the next section.

The main advantage of the h-index is that the complex publication and citation pattern of authors is summarized to one simple index. The main disadvantages are that the citations are not field-normalized, so that the indexes of authors from different fields are not comparable, and that there are no fixed citation windows and in consequence, older

scientists achieve much better scores than younger ones (Costas and Bordons 2007; Glänzel 2006). However, the basic idea to assess publications and citations is convincing.

H-index, skewed distributions and mean values

The determination of the h-index is illustrated using the example of a research unit in the area of applying graphene in electrical engineering. According to a search in the Web of Science, this unit published 62 articles in scientific journals in 2015, with a total of 1887 citations in the period from 2015 to 2017. This leads to an average citation rate of 30.4 (mean value). The maximum citation rate is 209. 7 publications are not cited at all until the end of 2017 (cf. Figure 1).

Using the definition of Hirsch, an index value of 27 is determined, which is illustrated in Fig. 1 by a bold cross. This definition is based on ranking the publications according to their citation level. Thus, publications with more than 27 citations are not considered in greater detail and their level has no impact on the index. The same applies to publications with few or no citations. To formulate this in a statistical perspective, the h-index does not consider high or low outliers and focuses on the standard performance of a unit. This perspective can be justified qualitatively by the observation that even high-level institutions often have publications with only a few or even no citations, e.g. those which document the outcome of intermediary working steps (Schmoch et al. 2019). Due to the constant pressure to publish, however, even these results are published. In the case of extremely high citation scores, these are often not an indication of especially outstanding performance, but may be a coincidental effect of conducive

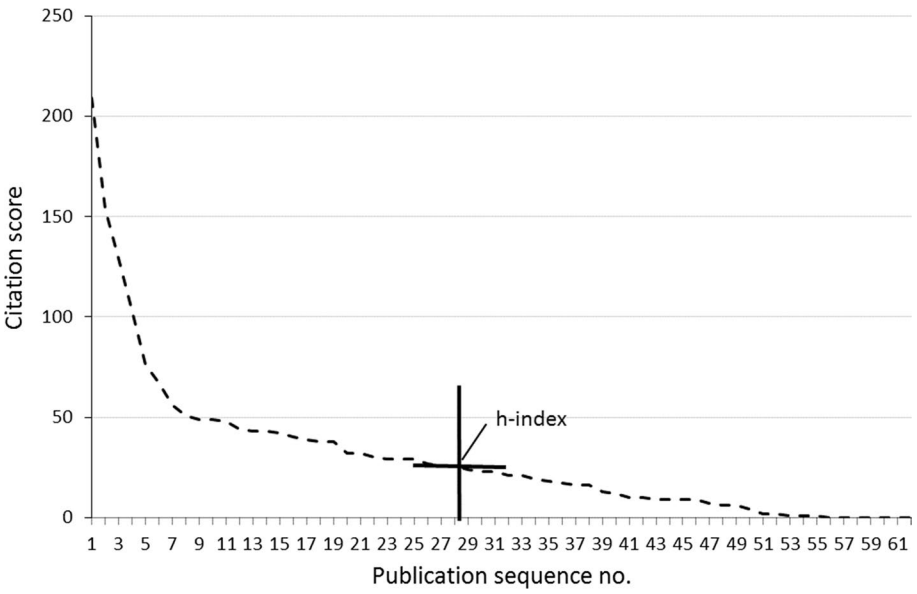

Fig. 1 Citations on the publications of a research unit in the area of graphene in electrical engineering in 2015, source: Web of Science, update 2018 (publications sorted by descending number of citations). *Source*: Web of Science, own search

circumstances. For instance, a paper may be an early contribution to a broad, long-term discourse and every subsequent publication has to cite this early one. In any case, these extreme values are not representative for the standard activity of a unit. Such a reflection forms the background to the concept of Hirsch. It is possible to transfer this reasoning into simple rules of bibliometric analysis. The topic of outliers has been quite controversial. In general, outliers are rejected and excluded from the dataset. Modern statistical theory provides an alternative to outlier rejection, in which outlying observations are retained but given less weight (Analytical Methods Committee 1989). This approach is known as robust statistics. With the logarithm, Lundberg (2007)—cited above—follows this concept.

Hirsch's concept can be simulated by a simple rule that excludes the 5% of publications with the highest citations and all publications with less than 6 citations. This rule is not based on mathematical reasoning, but on examining about 70 distributions of citations for arbitrary research units in different scientific fields. At first sight, the 5% share for publications with the highest citations seems to be high. But with the limited number of publications of research units, it is difficult to define smaller shares.

This amendment leads to the "adjusted distribution" in Fig. 2. The mean value of the adjusted distribution is 27.9, and therefore close to the Hirsch-Index; the mean value of the observed distribution is higher (30.5), but not completely different.

All in all, although it is possible to calculate the mean values of skewed distributions mathematically, there are good reasons to assess the citation performance of research units based on their standard activity and to exclude extreme values at both ends of the spectrum.

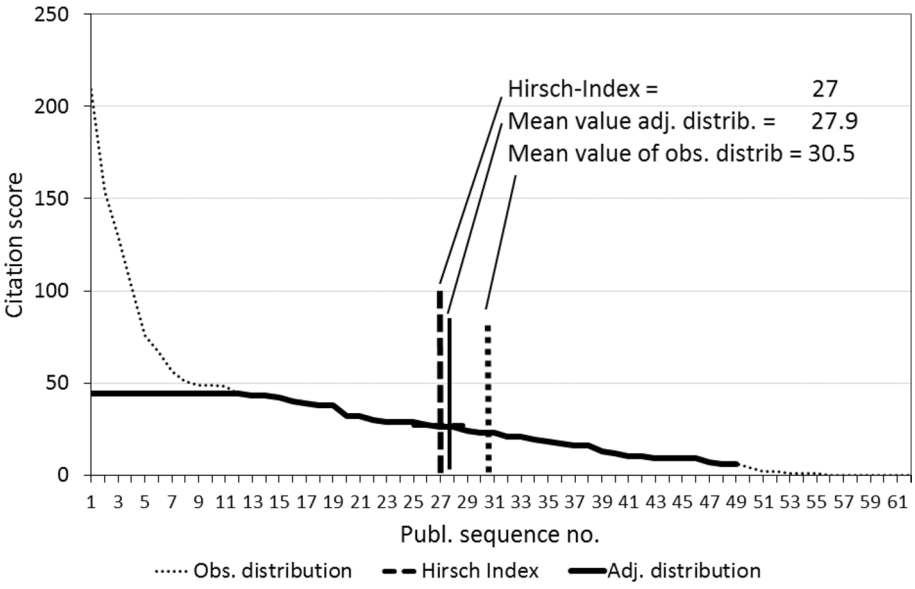

Fig. 2 Different types of mean values for citations on the publications of a research unit in the area of graphene in electrical engineering in 2015, source: Web of Science, update 2018 (publications sorted by descending number of citations). *Source*: WoS, own search

Rankings of research units based on adjusted distributions

There are good reasons to calculate the mean value of skewed distributions of citations based on adjusted rather than full range, observed distributions. However, the level of the resulting mean values is quite similar to the standard mean values—27.9 instead of 30.4—in the example shown above. Therefore, the question has to be raised whether this difference is so important that a separate calculation brings new insights and, in particular, whether the rankings of research units change. The latter issue is very important, because the distribution of research funding is based on such bibliometric rankings in many countries, e.g. in the Research Excellence Framework (REF) in the United Kingdom. Of course, it is problematic to produce strict rankings on the basis of citations, as citations can reflect other issues than scientific performance and also depend on accidental circumstances. But despite these uncertainties, bibliometric rankings are often used in practice in the context of funding in any countries.

In order to check the implications of adjusted mean values for the ranking of research units, we analysed the citation activity of ten research units in the subject category "Biotechnology & Applied Microbiology" in the Web of Science. There were about 500 citations of publications from the year 2015. In Table 1, these research units are ranked according to the mean value of their citations. Calculating the adjusted mean values leads to different values and a different ranking, too. The resulting new ranking is quite similar but, for example, Unit 3 advances to second place, Unit 2 drops to fourth and Unit 9 advances to seventh. This change in ranking is due to the fact that the distributions for all research units are skewed, but within the samples, the degree of skewness differs by unit. To conclude, calculating the adjusted mean values implies different rankings.

The example of Table 1 illustrates that the distributions of citations for research units are skewed, but that the shape of skewness differs. In particular, the level of very high citations is quite erratic due to the relatively small number of publications per research unit. In this regard, the situation of research units differs from that for scientific fields covering much higher number of publications.

Table 1 Ranking of research units in Bbiotechnology & Applied Microbiology" with about 500 citations to publications of 2015, source: Web of Science, update 2018. *Source*: WoS, own search

Research unit	# Publications	# Citations	Mean value	Adjusted mean value	Rank adj. mean
Unit 1	10	506	50.60	63.00	1
Unit 2	22	517	23.50	25.60	**4**
Unit 3	25	531	21.24	30.00	**2**
Unit 4	24	507	21.13	26.16	**3**
Unit 5	28	512	18.29	21.74	5
Unit 6	32	532	16.63	18.61	6
Unit 7	41	516	12.59	17.89	8
Unit 8	42	514	12.24	16.66	9
Unit 9	42	511	12.17	18.00	**7**
Unit 10	49	515	10.51	15.35	10

Bold numbers indicate a change of rank of the unit for adusted mean values compared to standard mean values

To illustrate the change in ranking, Fig. 3 shows the citation distributions for Unit 2 and Unit 3. In the standard calculation of mean values, Unit 2 is ranked higher due to two publications with very high citations. In the adjusted calculation, Unit 3 achieves a higher position, as it has some publications with higher citation scores than Unit 2 in the area of standard activities. Here, one has to decide whether a research institution is best characterised by a few very high citations or a larger number of citations in its standard activities.

To further illustrate the effect of adjusted means, another example is shown in Fig. 4. Here, various research units in physics were compared, in this case, units with about 300 publications in 2013. From this sample, two units were selected. To highlight the distribution of the citations in the standard area, Fig. 4 is cut at 30 publications. In the observed distribution, Unit A has 313 publications, Unit B 306 publications. Due to the high number of publications with no citations, the standard mean values of both units are very low and almost equal. Unit B has a slight lead due to two high citation scores of 124 and 122. In the adjusted calculation, Unit A is ranked above Unit B due to the higher citation scores in its standard activities. This example shows that the adjustment can bring about a fairer comparison of research units, and that a few highly cited publications should not outweigh relevant standard activities. Furthermore, the standard mean values for research units with long tails of uncited publications are dominated by these tails and imply significant distortions. In any case, the important diminution of the mean value due to the long tail of publications with no citations is doubtful, because some research units publish every intermediate result as a response to the pressure to publish frequently.

The adjusted mean value reflects the standard activities of a research unit, the excellent results by the mean values of the upper 5% of the publications with the highest citations. The latter indicator implies different rankings of research units, e.g., in the example of Table 1, the research units on the ranks six and nine exchange their positions. In any

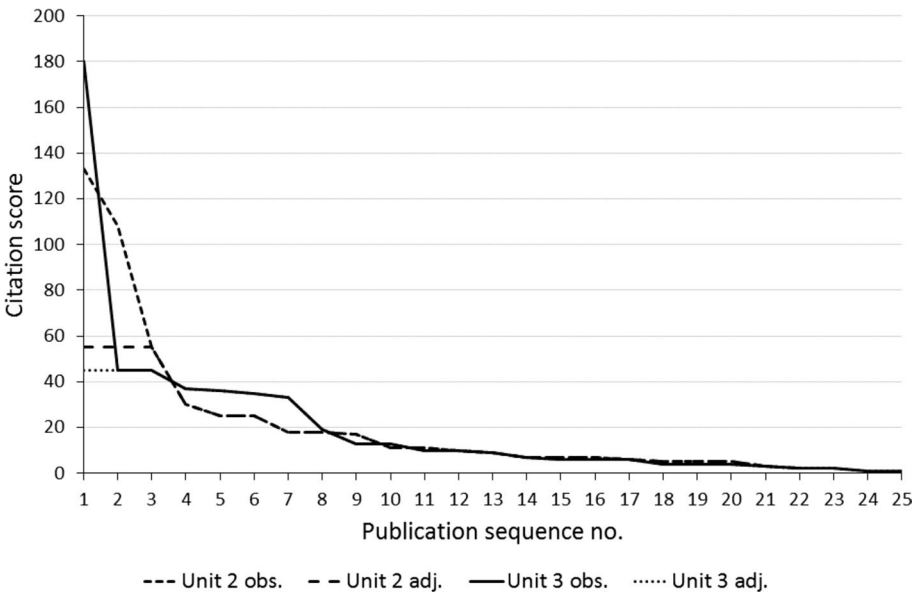

Fig. 3 Distribution of citations of two selected research units in "Biotechnology & Applied Microbiology, source: Web of Science, update 2018 (publications sorted by descending number of citations). *Source*: WoS, own search

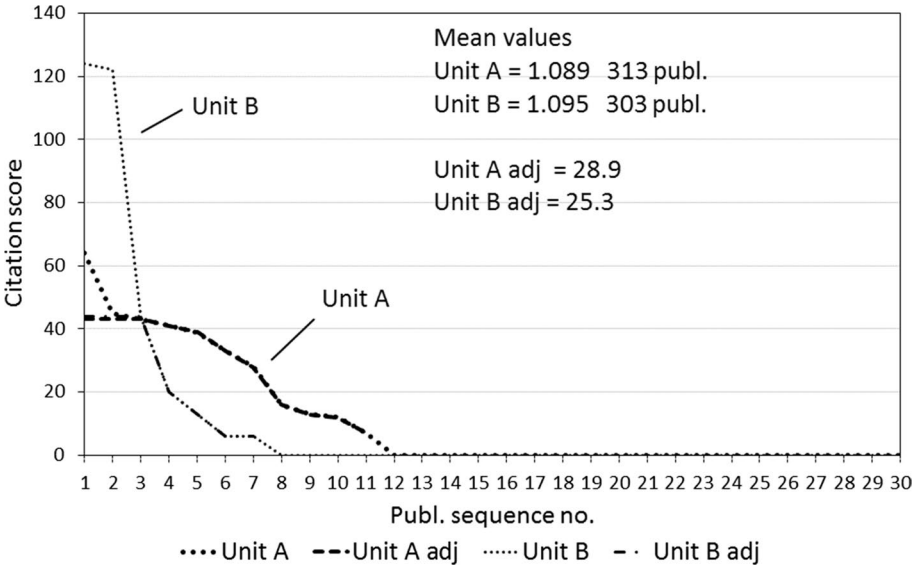

Fig. 4 Distribution of citations of two selected research units in "Physics", source: Web of Science, update 2018 (publications sorted by descending number of citations). *Source*: WoS, own search

case, these two types of indicators are more distinct than those suggested by Lundberg and can be used in a combined way. For a more in-depth analysis of excellence, it is interesting to compile the high citation values of several subsequent years for verifying the regularity of excellent publications.

Mean values of large skewed distributions

A final question is why an adjusted mean is suggested for research units instead of a Hirsch-index. The reason is that a normalisation by the field averages is necessary to compare units in different fields of activity. This implies that a normalisation of the total field by the Hirsch-index has to be calculated. However, the Hirsch-index was conceived for smaller samples of analysis. For instance, for the field of "Biotechnology & Applied Microbiology", about 2900 publications appeared in 2015 (Fig. 5). For this large sample, a Hirsch-index of 96 is calculated, which is far beyond the mean value of 9.4, and not useful in this context. In contrast, the adjusted mean value at 14.5 is distinctly above the standard mean value, but is much more rational than the h-index, because about 50% of all publications have less than 6 citations. These publications are not included in the adjusted mean value. In this case, excluding very high citations is less important for the level of the adjusted mean, although the first publication with high citations received 2335 citations in 3 years, the second 628, the third 559 and the fourth already 414. Thus, the first publication is not at all characteristic for standard activities in this area and its exclusion in the adjusted calculation is justified.

The example of the distribution of citations in the field of biotechnology illustrates that the skewness in scientific fields is very strong, as examined by Albarrán et al. (2011) for 219 sub-fields. The skewness for research units is less strong (cf. Tables 1 to 4) and the

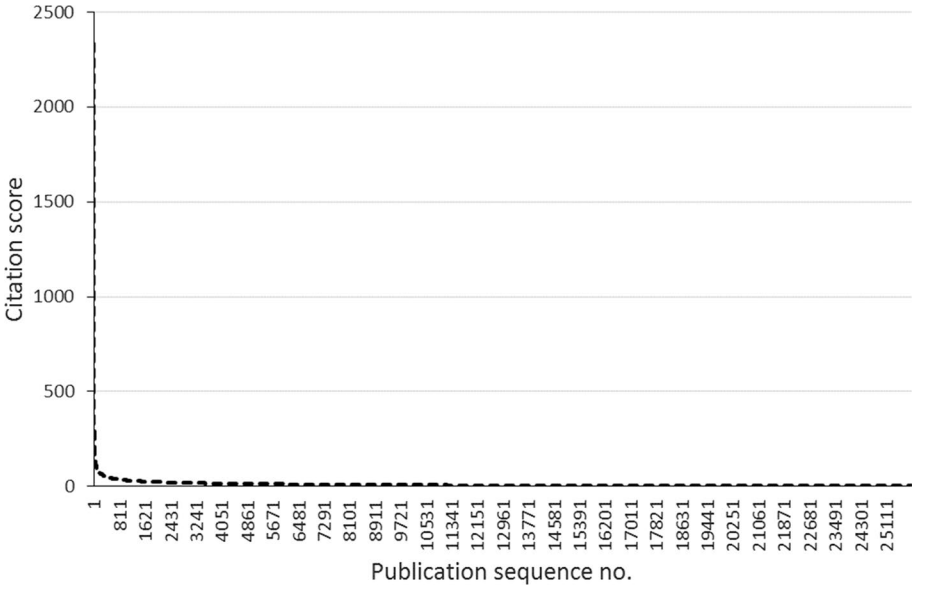

Fig. 5 Distribution of citations of all publications of 2015 in the subject category "Biotechnology & Applied Microbiology", source: Web of Science, update 2018 (publications sorted by descending number of citations). *Sourc*e: WoS, own search

highest citation scores are generally less extreme. That is why the sector for the highest citations was fixed at a level of 5%.

Conclusions

Nearly all distributions in bibliometrics are skewed. In particular, the distributions of citations of publications by research units are skewed, often highly skewed. To rank research units, it is recommended to replace the calculation of standard mean values by adjusted mean values, which exclude outliers with very high citations as well as those with very low or no citations. Such an adjusted mean value is oriented on the standard activity of a research unit and leads to a more adequate assessment. This approach is based on the concept of the Hirsch-index. This calculation often results in a different ranking of research units and is important in cases where the distribution of finances to research units depends on bibliometric rankings.

The decision to use adjusted mean values is not based on objective mathematical criteria. Instead, it is a rather subjective reflection of what activities are important for the assessment of research units. Some funders are primarily interested in excellent results that are distinctly above the average activities in a field. In this case, they should base their assessment on the upper 5% of publications with very high citations. In the present practice of calculating mean values of skewed distributions, extreme outliers are mixed with standard activities in a non-transparent way. I, on the other hand, am in favour of using standard activities for assessment, because, in many cases, extremely high citations are simply outliers.

A more comprehensive analysis of research units can be achieved by analysing the standard activities by adjusted mean values and excellent results by the mean values of the upper 5% of the citations. These two indicators are quite distinct and reflect different aspects of the activities of research units. In this perspective, the suggested indicators allow for a more differentiated analysis that conventional mean values. This approach can also be used for assessing the work of authors.

By the introduction of two dimensions of assessment, it gets obvious that the final outcome of an assessment is a question of an appropriate interpretation and cannot be objectively found by correct statistics.

Acknowledgements This article is an extended version of a paper presented at the ISSI conference 2019 in Rome. The author thanks the audience for various helpful comments. I thank two anonymous reviewers for their very helpful comments and suggestions. Certain data included in this paper are derived from the Science Citation Index Expanded (SCIE), the Social Science Citation Index (SSCI), the Arts and Humanities Citation Index (AHCI), and the Index to Social Sciences & Humanities Proceedings (ISSHP) prepared by Clarivate Analytics, Philadelphia, Pennsylvania, USA.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Adamic, L. A. (2002). Zipf, power-laws, and Pareto—A ranking tutorial, Resource document. Information Dynamics Lab. http://www.labs.hp.com/research/idl/papers/ranking/ranking.html. Retrieved 22 October 2019.

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics, 88*(2), 385–397.

Albarrán, P., & Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology, 62*(1), 40–49.

Analytical Methods Committee. (1989). Robust statistics—How not to reject outliers, part 1. *Basic Concepts, Analyst, 114,* 1693–1697.

Chen, Y. S., & Leimkuhler, F. F. (1986). A relationship between Lotka's Law, Bradford's Law, and Zipf's Law. *Journal of the Association for Information Science and Technology, 37*(5), 307–314.

Costas, R., & Bordons, M. (2007). The H-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics, 1,* 193–203.

De Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *The Journal of the Association for Information Science and Technology, 27*(5), 292–306.

Glänzel, W. (2006). On the opportunities and limitations of the H-index. *Science focus, 1*(1), 10–11.

Glänzel, W., & Moed, H. (2005). Journal impact measures in bibliometric research. *Scientometrics, 53*(2), 171–193.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS, 102*(46), 16569–16572.

Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators compared with impact factors: A alternative research design with policy impact. *Journal of the American Society for Information Science and Technology, 62*(11), 2133–2146.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16,* 317–323.

Lundberg, J. (2007). Lifting the crown—Citation *z*-score. *Journal of Informetrics, 1*(2), 245–254.
Mandelbrot, B. (1959). A note on a class of skew distribution functions: Analysis and critique of a paper by H. A. Simon. *Information and Control, 2,* 90–99.
Moed, H. F., De Bruin, R. E., & Van Leuven, T. N. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.
Murphy, L. J. (1973). Lotka's law in the humanities? *Journal of the Association for Information Science and Technology, 24*(6), 461–462.
Narin, F., & Hamilton, K. S. (1996). Bibliometric performance measures. *Scientometrics, 36*(3), 293–310.
Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323–351.
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.
Pao, M. L. (1986). An empirical examination of Lotka's Law. *Journal of the Association for Information Science and Technology, 37*(1), 26–33.
Pareto, V. (1935). *The mind and society*. New York: Harcourt, Brace and Company.
Radhakrishnan, T. (1973). Lotka's Law and computer science literature. *Journal of the Association for Information Science and Technology, 24*(6), 461–462.
Rousseau R. (2011). Percentile rank scores are congruous indicators of relative performance, or aren't they? https://arxiv.org/abs/1108.1860.
Schmoch, U., Beckert, B., & Schaper-Rinkel, P. (2019). Impact assessment of a support programme of science-based emerging technologies. *Scientometrics, 118*(3), 1141–1161.
Seglen, O. (1992). The skewness of science. *Journal of the Association for Information Science and Technology, 43*(9), 628–638.
Siegel, A. F. (2017). Practical business statistics. Resource document. Academic Press. https://www.sciencedirect.com/topics/mathematics/skewed-distributions. Retrieved 23 October 2019.
Simon, H. A. (1955). A note on a class of skew distribution functions. *Biometrika, 42*(3/4), 5–440.
Statistics how to. (2019). Skewed distribution: Definition, example. Resource document. Statistics how to. https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/. Retrieved 25 October 2019.
Von Hippel, P. T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education*. https://doi.org/10.1080/10691898.2005.11910556.
Zipf, G. H. (1949). *Human behaviour and the principle of least effort*. Reading, Massachusetts: Addison-Wesley Press.