On using the *h*-index to analyse species biodiversity and other count data

The *h*-index, a scientometric measure of the impact of a researcher's publications, was defined by Hirsch[1] as having a value *h* if there are at least *h* citations for each of *h* of the researcher's N_p papers, while the other $(N_p - h)$ papers have at most *h* citations each. Rousseau[2] suggested that this index could be applied to data of counts more generally, so proposed using it as a biodiversity index; more specifically as a simple, robust summary statistic that combines information about species richness and abundance.

Suppose that a sample of *N* individuals is distributed among *k* categories, each category being associated with a particular phenomenon or feature, and we have a count of the number of individuals of the sample falling into each category. Then the sample has index value *h*, if *h* of the categories shows a count greater than or equal to *h*, while the remaining *k* – *h* categories all have a count less than or equal to *h*. The categories can be either nominal, for example, the *k* distinct species of the *N* moths caught in a light-trap, or ordinal, for example, the *k* distinct numbers of faults in *N* pieces of cloth tested in a textile factory.

We can restate the definition statistically by focusing on the sample of *k* counts: the index is *h* if the *h*th order statistic of this sample has a value no less than *h*, while the (*h* + 1)th order statistic has a value no greater than *h*. Such a formulation indicates the viability of the *h*-index as a summary measure for quantities like abundance (for nominal categories such as species) or prevalence (for ordered numerical categories such as number of faults). However, the purpose of this note is to highlight some problems that limit its use – particularly if one wishes to go further than simple summary measures and attempts to develop an inferential procedure for the index.

For inferences to be possible, the calculated (i.e. sample) *h*-index needs to be an estimator of a corresponding population *h*-index. Unfortunately, whereas a unique, single *h*-index exists in a finite population this is no longer the case for an infinite population, as can be easily shown from the order statistic formulation given above. Suppose that the count data have indeed arisen as a sample from some infinite population. The theory of order statistics[3] tells us that as the sample size tends to infinity, i.e. as the sample approaches the population, every order statistic has a limiting distribution rather than a single value. Hence inferences about a single population *h*-index cannot be made; we can only use the calculated sample value as a summary measure of the data at hand.

Unfortunately, there is a problem with this usage, as the *h*-index is strongly affected by sample size. To show this we have conducted simulation experiments in which 1,000 samples have been drawn from Poisson distribution, for each combination of sample sizes 5, 10, 50, 100, 1,000, 10,000 and 100,000, and Poisson parameter (lambda) values 3, 5, 10, 15, 20, 50 and 100. Table 1 shows the minimum and maximum values of the *h*-index for the 1,000 samples at each of these combinations. Clearly, sample size strongly affects the *h*-index. For example, when population biodiversity is high (large lambda), the differences in the *h*-index between the small and large samples can be great. It therefore follows that if two or more groups are to be compared by means of the *h*-index, samples from the groups must have the same size. Otherwise, the *h* measures will not be comparable.

If it is then required to decide whether two or more sample *h* values reflect genuine rather than chance differences, it is essential to assess the variability inherent in the calculation of *h* values rather than in estimating a population value. This suggests a data-based approach such as the jackknife, whose rationale holds for variability interpreted in the sense of stability.

In conclusion, it is worth emphasizing the rather curious feature of the *h*-index, that, whereas a well-defined population value exists if the population is finite, this is no longer the case if the population is infinite. We are not aware of any other index that exhibits such behaviour, but this needs further study.

Table 1. Range (minimum–maximum) of the sample *h*-index obtained for Poisson distributions having various lambda values, with different sample sizes

	Lambda for the Poisson distribution						
Sample size	3	5	10	15	20	50	100
5	1–4	2–5	3–5	4–5	5–5	5–5	5–5
10	2–5	3–7	6–10	8–10	9–10	10–10	10–10
50	4–6	6–9	10–14	15–18	19–23	40–45	50–50
100	5–7	7–10	12–15	17–20	21–25	47–52	85–92
1,000	7–9	10–12	16–18	22–24	28–30	59–62	110–114
10,000	9–11	12–14	19–21	26–28	32–34	67–69	122–124
100,000	11–12	14–16	22–23	29–31	36–37	73–75	131–132

1. Hirsch, J. E., *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 16569–16572.
2. Rousseau, R., *Curr. Sci.*, 2009, **97**, 980–981.
3. Stuart, A. and Ord, J. K., *Kendall's Advanced Theory of Statistics*, Griffin, London, 1987, vol. 1, 5th edn.

MARCIN KOZAK[1,*]
WOJTEK KRZANOWSKI[2]

[1]*Faculty of Economics,*
University of Information Technology and Management in Rzeszow,
Poland
[2]*College of Engineering,*
Mathematics and Physical Sciences,
University of Exeter, UK
**e-mail: nyggus@gmail.com*