Sugeno integral-based confidence intervals for the theoretical h-index

Marek Gagolewski[1,2]

[1] Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland,
gagolews@ibspan.waw.pl

[2] Faculty of Mathematics and Information Science, Warsaw University of Technology,
ul. Koszykowa 75, 00-662 Warsaw, Poland

Abstract. Sugeno integral-based confidence intervals for the theoretical h-index of a fixed-length sequence of i.i.d. random variables are derived. They are compared with other estimators of such a distribution characteristic in a Pareto i.i.d. model. It turns out that in the first case we obtain much wider intervals. It seems to be due to the fact that a Sugeno integral, which may be applied on any ordinal scale, is known to ignore too much information from cardinal-scale data being aggregated.

1 Introduction

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sequence of i.i.d. random variables with a common monotone strictly increasing c.d.f. F with support $\mathbb{I} = [0, \infty)$. The theoretical h-index, cf. [11], $\mathfrak{H}_n = \mathfrak{H}_n(X) \in (0, n)$ is a solution to:

$$1 - F(\mathfrak{H}_n) = \mathfrak{H}_n / n.$$

The theoretical h-index is a sample-size dependent location characteristic of a probability distribution. For example, if X follows a Pareto/Lomax distribution with $F(x) = 1 - 1/(1 + x)$, then $\mathfrak{H}_n = (\sqrt{4n+1} - 1)/2$.

Among estimators of $\mathfrak{H}_n$ we find the generalized Hirsch [12] index:

$$\widehat{h}_n(\mathbf{X}) = \bigvee_{i=1}^{n} X_{(n-i+1)} \wedge i = \max\left\{\min\{X_{(n)}, 1\}, \dots, \min\{X_{(1)}, n\}\right\},$$

where $X_{(i)}$ denotes the ith smallest value in $\mathbf{X}$. Statistic $\widehat{h}_n$ is an OWMax [3,4] (and thus an OM3 [7]) operator corresponding to the Sugeno [14] integral of $\mathbf{X}$ with respect to the counting measure, see also [10,15]. What is important, it has already been shown (see [9] for the proof) that $\widehat{h}_n(\mathbf{X})/n$ is an asymptotically unbiased estimator of $\mathfrak{H}_n/n$.

It is well-known that the h-index, originally defined for a sample with elements in $\mathbb{N}_0$, has many fruitful applications, for example in bibliometrics [6], quality engineering [5] and information sciences [13]. However, still little is known on the stochastic properties of such a measure. In [9,11] the properties of $\widehat{h}_n$ and other Sugeno integrals in an i.i.d. setting are considered, while in e.g. [1] its behavior in a more complex model is investigated. Moreover, in [8] a statistical test for the difference of h-indices in two Pareto-distributed random samples of equal lengths is derived and it turns out that such a tool has a very weak discriminatory power.

In this contribution we are interested in constructing Sugeno integral-based confidence intervals for the theoretical h-index, which is done in the section to follow. In Sec. 3 we provide some numeric examples for the Pareto distribution family. The obtained estimates are compared with different ones. It turns out that the $\widehat{h}_n$-based intervals are very wide, which is probably due to the fact that a Sugeno integral is known to ignore too much information from data. Finally, Sec. 4 concludes the paper.

2 Derivation of Sugeno integral-based confidence intervals

Fix n. Let $\Theta = (0, n)$ be a parameter space that induces an identifiable statistical model $(\mathbb{I}, \{\Pr_\theta : \theta \in \Theta\})^n$ in which for $X \sim \Pr_\theta$ we have $\theta = \mathfrak{H}_n(X)$ for all $\theta \in \Theta$, i.e. such that the theoretical h-index of X is equal to the value of parameter θ.

Definition 1. *Let $\alpha \in [0, 1]$. A random interval $\left(\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X})\right)$ is called an $(1-\alpha)$-confidence interval for parameter θ if:*

$$(\forall \theta \in \Theta) \quad \Pr{}_\theta \left(\underline{\theta}(\mathbf{X}) \le \theta \le \overline{\theta}(\mathbf{X})\right) \ge 1 - \alpha.$$

Of course, here we are interested in constructing the smallest confidence intervals which bounds are determined solely by the observed value of $\widehat{h}_n$. Additionally, we will assume a kind of symmetry of the intervals. The lower bound, $\underline{\theta}(\mathbf{X})$, will be defined via the smallest function $d_\alpha : (0, n) \to (0, n)$ such that for all $\theta \in (0, n)$ it holds

$$\Pr{}_\theta \left(\widehat{h}_n(\mathbf{X}) \le d_\alpha(\theta)\right) \ge 1 - \alpha/2.$$

Given the observed random sample realization $\mathbf{x}$ and $h = \widehat{h}_n(\mathbf{x})$, the lower bound will be determined by calculating $d_\alpha^{-1}(h) = \sup\{\theta : d_\alpha(\theta) \le h\}$. Thanks to such a setting we will have $\Pr_\theta(d_\alpha^{-1}(\widehat{h}_n(\mathbf{X})) \le \theta) \ge 1 - \alpha/2$.

On the other hand, the upper bound shall be given by the greatest function g_α such that

$$\Pr{}_\theta \left(\widehat{h}_n(\mathbf{X}) \ge g_\alpha(\theta)\right) \ge 1 - \alpha/2,$$

which is equivalent to $\Pr_\theta \left(\widehat{h}_n(\mathbf{X}) < g_\alpha(\theta)\right) \le \alpha/2$. This will provide us with $\Pr_\theta(\theta \le g_\alpha^{-1}(\widehat{h}_n(\mathbf{X}))) \ge 1 - \alpha/2$.

By [9, Lemma 2] we have:

$$\Pr{}_\theta(\widehat{h}_n(\mathbf{X}) \le h) = \mathcal{I}(\Pr{}_\theta(X \le h); n - \lfloor h \rfloor, \lfloor h \rfloor + 1),$$

where $\mathcal{I}(p; a, b)$ denotes the incomplete beta function of p with parameters a, b. We see that the c.d.f. of $\widehat{h}_n$ can be discontinuous even for continuous c.d.f. of X. Therefore,

$$\underline{\theta}(\mathbf{x}) = d_\alpha^{-1}(h) = \sup\left\{\theta : \mathcal{I}\left(\Pr{}_\theta(X < h); n - \lfloor h \rfloor, \lfloor h \rfloor + 1\right) \geq 1 - \alpha/2\right\},$$

and

$$\overline{\theta}(\mathbf{x}) = g_\alpha^{-1}(h) = \inf\left\{\theta : \mathcal{I}\left(\Pr{}_\theta(X \leq h); n - \lfloor h \rfloor, \lfloor h \rfloor + 1\right) \leq \alpha/2\right\}.$$

Unfortunately, in most cases the confidence interval bounds can only be calculated numerically.

3 Numerical examples

For the sake of illustration let us consider the Pareto distribution family, $\mathcal{P}(k)$, with scale parameter $k > 0$. Such a distribution is sometimes used, cf. [11], in modeling empirical phenomena in the application scope of the h-index.

The cumulative distribution function of $X \sim \mathcal{P}(k)$ is defined by:

$$F(x) = 1 - \frac{1}{(x+1)^k} \quad (x \geq 0).$$

We have $\mathbb{E}\,X = 1/(k-1)$ for $k > 1$ and $\operatorname{supp} X = [0, \infty)$.

In order to guarantee that this family of distributions fits our statistical model's assumptions, we should introduce the following reparametrization. Let $\vartheta_n(k) = \mathfrak{H}_n(X)$ for $X \sim \mathcal{P}(k)$. Such a function may easily be calculated numerically with very good accuracy using some nonlinear root finding algorithm. Thus, we may consider $\mathcal{P}'(\theta) \equiv \mathcal{P}(\vartheta^{-1}(\theta))$, $\theta \in (0, n)$.

Figures 1 and 2 depict the 95%-confidence intervals bounds for $n = 10$ and 25, respectively. Note that the bounds are not continuous functions of $\widehat{h}_n$: they have jumps in points from the set $\{1, \ldots, n-1\}$. For example, for $n = 10$ and observed value of $\widehat{h}_n = 5$, we obtain an interval $(3.341, 7.779)$. On the other hand, for $\widehat{h}_n = 5^-$ we get $(2.840, 7.021)$.

We should also keep in mind that even though the obtained intervals are the smallest possible (at a confidence level of 95%), in fact the true probability of covering a theoretical h-index may sometimes be greater that 95%. This phenomenon, depicted in Figures 3 and 4, is of course consistent with the provided definition of a confidence interval. A similar behavior is observed e.g. for the Neyman-Clopper-Pearson (beta distribution-based, see [2]) confidence intervals for the probability of success in a Bernoulli experiment, cf. [16].

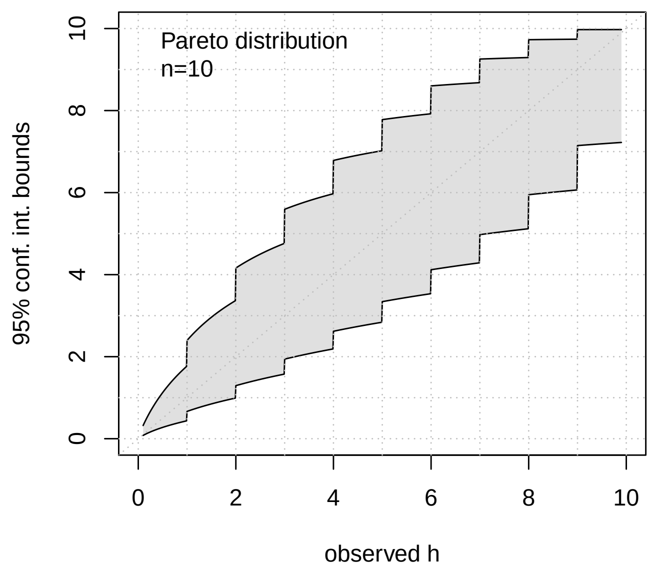

Fig. 1. Bounds for the Sugeno integral-based 95%-confidence intervals for the theoretical h-index; Pareto distribution family; $n = 10$.

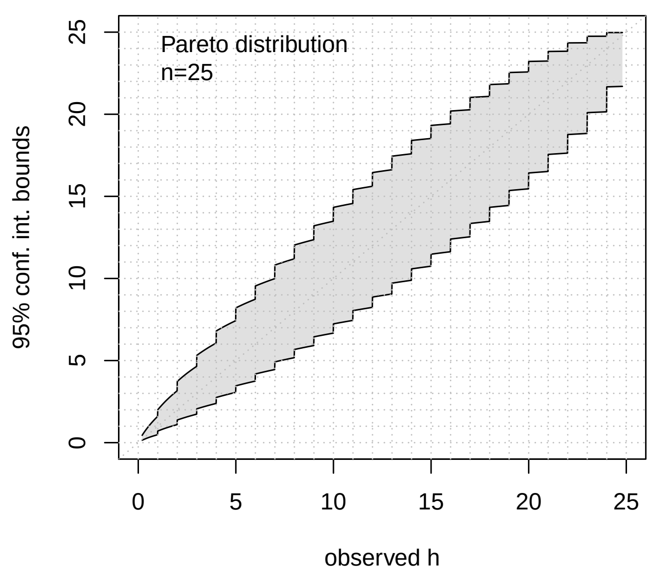

Fig. 2. Bounds for the Sugeno integral-based 95%-confidence intervals for the theoretical h-index; Pareto distribution family; $n = 25$.

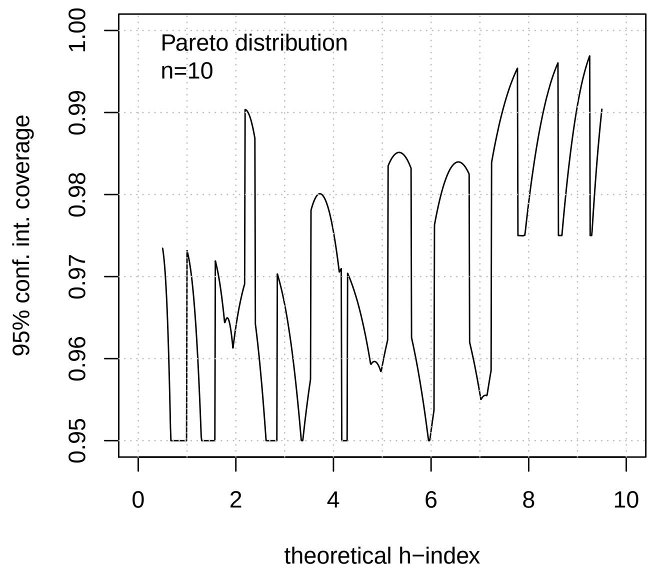

Fig. 3. Actual coverage of the true $\mathfrak{H}_n$ by Sugeno integral-based 95%-confidence intervals; Pareto distribution family; $n = 10$.

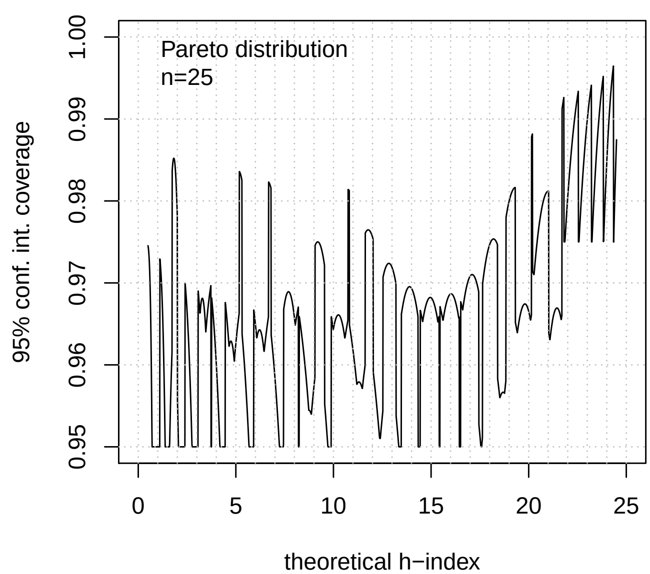

Fig. 4. Actual coverage of the true $\mathfrak{H}_n$ by Sugeno integral-based 95%-confidence intervals; Pareto distribution family; $n = 25$.

Comparison to other estimates. It might easily be shown that for $(X_1, \dots, X_n)$ i.i.d. $\mathcal{P}(k)$ the statistic

$$\widehat{k}_n^*(\mathbf{X}) = (n-1)/\sum_{i=1}^n \log(1+X_i)$$

is an unbiased and consistent estimator of k. What is more, $\sum_{i=1}^n \log(1+X_i) \sim \Gamma(n,k)$.

We may thus try using $\widehat{h}_n^* = \vartheta_n(\widehat{k}_n^*)$ as an estimator of $\mathfrak{H}_n$. Numerical results indicate that $\widehat{h}_n^*/n$ may only be asymptotically unbiased estimator of $\mathfrak{H}_n/n$. By the above-mentioned fact, if $(X_1, \dots, X_n)$ i.i.d. $\mathcal{P}'(\vartheta(k))$, then

$$\Pr{}_{\vartheta(k)}(\widehat{h}_n^*(\mathbf{X}) \le h) = 1 - G_{n,k}\left(\frac{n-1}{\vartheta^{-1}(h)}\right),$$

where $G_{n,k}$ is the c.d.f. of the gamma distribution $\Gamma(n,k)$. This time, such an estimator has a continuous distribution.

A $\widehat{h}_n^*$-based $(1-\alpha)$-confidence interval may be derived in a manner similar (but much simpler due to continuity) to the previously considered one. It is a random interval $(\underline{\theta}^*(\mathbf{X}), \overline{\theta}^*(\mathbf{X}))$ such that $\underline{\theta}^*(\mathbf{X}) = d_\alpha^{-1*}(h)$ and $\overline{\theta}^*(\mathbf{X}) = g_\alpha^{-1*}(h)$ for which it holds

$$\begin{aligned}
&\Pr{}_{d_\alpha^{-1*}(h)}(\widehat{h}_n^*(\mathbf{X}) \le h) = \alpha/2, \\
&\Pr{}_{g_\alpha^{-1*}(h)}(\widehat{h}_n^*(\mathbf{X}) \le h) = 1 - \alpha/2.
\end{aligned}$$

Again, these equations may be solved numerically with a nonlinear root finder. This time we obtain a confidence interval which is exactly at a confidence level of $1-\alpha$ for each θ.

Figure 5 depicts $\widehat{h}_n^*$-based 95%-confidence interval bounds for $n = 25$. We observe that they are of smaller length than those presented in Figure 2. Moreover, interval lengths for different sample sizes are given in Figure 6. We note that $\widehat{h}_n^*$ are better quality estimates than the Sugeno integral-based ones.

4 Conclusions

In this paper we derived Sugeno integral-based confidence intervals for the theoretical h-index, which is a location-type characteristic of a probability distribution. Large widths of the Sugeno integral-based intervals for a sample from the Pareto distribution family may possibly be due to the fact that this aggregation method is known not to utilize "full information" in input data. For example, for $n = 6$, $\widehat{h}_n(\mathbf{x}) = 3$ is obtained for $\mathbf{x} = (3,3,3,0,0,0)$ as well as for $\mathbf{x} = (\infty,\infty,\infty,3,3,3)$.

Taking into account the close relationship between confidence intervals and statistical hypothesis tests, the presented results are consistent with conclusions of [8]: the nature of Sugeno integral allows its application on any ordinal scale, but the prize we are paying for its robustness is the lack of good performance for cardinal scales.

Acknowledgments

The author would like to thank the anonymous reviewers for comments and suggestions.

References

1. Burrell, Q.L.: Hirsch's h-index: A stochastic model. Journal of Informetrics 1, 16–25 (2007)
2. Clopper, C., Pearson, E.: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26, 404–413 (1934)
3. Dubois, D., Prade, H.: Semantics of quotient operators in fuzzy relational databases. Fuzzy Sets and Systems 78(1), 89–93 (1996)
4. Dubois, D., Prade, H., Testemale, C.: Weighted fuzzy pattern matching. Fuzzy Sets and Systems 28, 313–331 (1988)
5. Franceschini, F., Maisano, D.A.: The Hirsch index in manufacturing and quality engineering. Quality and Reliability Engineering International 25, 987–995 (2009)
6. Franceschini, F., Maisano, D.A.: Structured evaluation of the scientific output of academic research groups by recent h-based indicators. Journal of Informetrics 5, 64–74 (2011)
7. Gagolewski, M.: On the relationship between symmetric maxitive, minitive, and modular aggregation operators. Information Sciences 221, 170–180 (2013)
8. Gagolewski, M.: Statistical hypothesis test for the difference between Hirsch indices of two Pareto-distributed random samples. In: Kruse, R., et al. (eds.) Synergies of Soft Computing and Statistics for Intelligent Data Analysis, vol. 190, pp. 359–367. Springer-Verlag (2013)
9. Gagolewski, M., Grzegorzewski, P.: S-statistics and their basic properties. In: Borgelt, C., et al. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis, pp. 281–288. Springer-Verlag (2010)
10. Gagolewski, M., Mesiar, R.: Monotone measures and universal integrals in a uniform framework for the scientific impact assessment problem. Information Sciences 263, 166–174 (2014)
11. Glänzel, W.: On some new bibliometric applications of statistics related to the h-index. Scientometrics 77(1), 187–196 (2008)
12. Hirsch, J.E.: An index to quantify individual's scientific research output. Proceedings of the National Academy of Sciences 102(46), 16569–16572 (2005)
13. Hovden, R.: Bibliometrics for internet media: Applying the h-index to YouTube. Journal of the American Society for Information Science and Technology 64(11), 2326–2331 (2013)
14. Sugeno, M.: Theory of fuzzy integrals and its applications. Ph.D. thesis, Tokyo Institute of Technology (1974)
15. Torra, V., Narukawa, Y.: The h-index and the number of citations: Two fuzzy integrals. IEEE Transactions on Fuzzy Systems 16(3), 795–797 (2008)
16. Zieliński, R.: Confidence intervals for proportions (Przedziały ufności dla frakcji, In Polish). Matematyka Stosowana 10, 51–68 (2009)

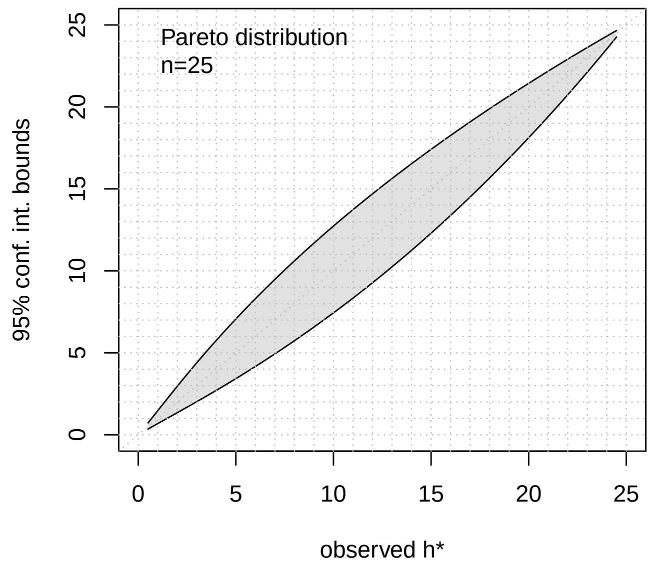

Fig. 5. Bounds for the $\widehat{h}_n^*$-based 95%-confidence intervals for the theoretical h-index; Pareto distribution family; $n = 25$.

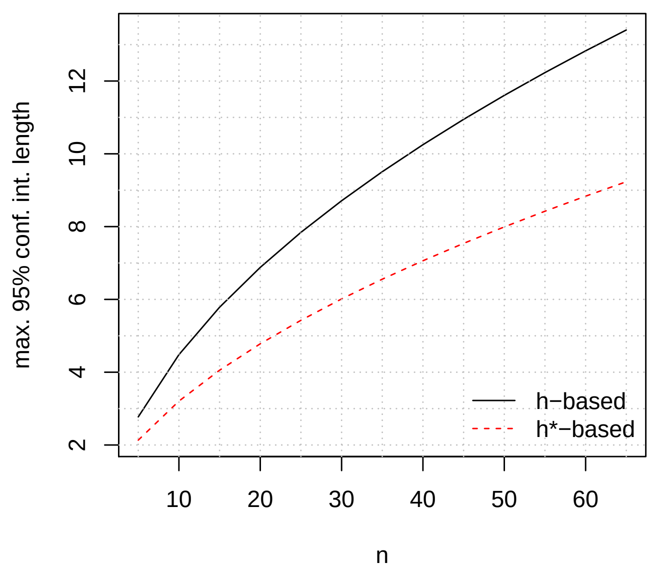

Fig. 6. Maximal widths of Sugeno integral- and $\widetilde{h}_n^*$-based 95-% confidence intervals for the theoretical h-index as a function of sample size n; Pareto distribution family.