Scientometrics (2018) 116:645–653
https://doi.org/10.1007/s11192-018-2724-8

The repeat rate: from Hirschman to Stirling

Ronald Rousseau[1,2]

Received: 13 March 2018 / Published online: 18 April 2018

Abstract In this short note we recall the history and definition of the repeat rate, also known as the Hirschman–Herfindahl index or as the Simpson index, and show that its generalization to a measure that includes disparity between items, known as the Rao-Stirling index, or a monotone transformation of it, is an acceptable diversity measure which, however, does not meet the 'monotonicity of balance' requirement.

Keywords Repeat rate · Hirschman index · Simpson index · Herfindahl index · Rao-Stirling index · Measuring diversity · Interdisciplinarity

Introduction

Two events led to this article. The first was that some time ago we came upon an article in which Hirschman (1964) pointed out that the so-called Herfindahl index (sometimes even referred to as Gini index) was actually first proposed by him. The second one was the recent claim by Leydesdorff et al. (2018) that the Rao-Stirling diversity index or its monotone transformations do not take balance into account. This article brings these two observations together in one story.

How it started (really?)

In 1945 Albert Hirschman published his book *National Power and the Structure of Foreign Trade* (Hirschman 1945). In this book he introduced an index for concentration, denoted as C, and described as: *the square root of the sum of the square of the elements in a series*

✉ Ronald Rousseau
ronald.rousseau@uantwerpen.be; ronald.rousseau@kuleuven.be

[1] Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium

[2] Facultair Onderzoekscentrum ECOOM, KU Leuven, Naamsestraat 61, 3000 Louvain, Belgium

when these elements are expressed as percentages of their sums. Concretely, and expressed in mathematical terms, this becomes: *if* a_1, a_2, …, a_n *are the elements of an ungrouped statistical series and if we have* $\sum_{i=1}^{n} a_i = A$, *then the index is*

$$C = \sqrt{\sum_{i=1}^{n} \left(\frac{a_i}{A} \cdot 100\right)^2} = \frac{100}{A} \sqrt{\sum_{i=1}^{n} a_i^2} \tag{1}$$

Denoting $\frac{a_i}{A}$ as p_i, Hirschman's index C can be rewritten as:

$$C = 100 \sqrt{\sum_{i=1}^{n} p_i^2} \tag{2}$$

We note that it would have been more precise if Hirschman had written fractions instead of percentages (0.04 or 4/100 or 1/25 are—equal—fractions corresponding to a percentage of 4).

It turned out that, among economists, this index, with or without the (superfluous) factor 100, was soon attributed to Gini, although Gini did not propose this index or to Herfindahl, although Herfindahl (1950) was just a re-inventor. This led Hirschman to write a "complaint" stating that his index was named either after Gini, who did not invent it at all, or after Herfindahl who reinvented it (Hirschman 1964).

Meanwhile, in 1949, Simpson (1949) had proposed

$$\lambda = \sum_{i=1}^{n} p_i^2 \tag{3}$$

as a measure of concentration. Clearly $C = 100\sqrt{\lambda}$ or $\lambda = \left(\frac{C}{100}\right)^2$. Simpson (1949) added a version for the case the index is sampled from a finite population, including its variance. Good (1982) later pointed out that, almost certainly, Simpson learned about 'his' index from Turing (the originator of the Turing machine, a universal theoretical computer) with whom he collaborated during World War II, among other things in connection with ENIGMA, decoding the German secret code. As this work was classified Simpson could not tell the origin of the index λ. Indeed, Good writes that Turing, who referred to it as the repeat rate, used the Simpson index already in 1941. Good (1982) moreover suggests that the origin of the repeat rate may even be traced to the nineteenth century and was certainly used by Polish cryptanalysts before Turing. He concludes that it is unjust to associate λ with any one person. True as this is, it is hard to go against common usage.

It is easy to see that the repeat rate is nothing but the probability that two elements taken at random and with replacement, from the dataset of interest represent the same type (in abstract terms: belong to the same cell).

From now on we will mainly use λ (not C) and will refer to λ as the repeat rate.

Variations on the repeat rate

Several other concentration or diversity indices are simple functions of the repeat rate. The most interesting one is the coefficient of variation, denoted as V. V is defined as the standard deviation, denoted as σ, divided by the average $\mu = \frac{A}{n}$. In terms of the sequence $(a_j)_{j=1,\ldots,n}$ we have:

$$V = \frac{\sqrt{\left(\frac{1}{n}\sum_{j=1}^{n} a_j^2\right) - \mu^2}}{\mu} \tag{4}$$

V can also be written as:

$$V = \frac{\sqrt{\sum_{j=1}^{n}\left(a_j - \mu\right)^2}}{\mu \cdot \sqrt{n}} = \frac{\sqrt{n \cdot \sum_{j=1}^{n}\left(a_j - \mu\right)^2}}{\sum_{j=1}^{n} a_j} = \frac{\sqrt{n \cdot \sum_{j=1}^{n}\left(a_j - \mu\right)^2}}{A} \tag{5}$$

We will next show that

$$V^2 = n \cdot \lambda - 1 \tag{6}$$

Using formula (4) we have $V^2 = \frac{\left(\frac{1}{n}\sum_{j=1}^{n} a_j^2\right) - \mu^2}{\mu^2} = \frac{\left(\frac{n^2}{n}\sum_{j=1}^{n} a_j^2\right)}{A^2} - 1 = n\sum_{j=1}^{n}\left(\frac{a_j}{A}\right)^2 - 1 = n \cdot \lambda - 1$.

Another variant is the so-called Yule coefficient

$$Y = \frac{V^2}{n} = \lambda - \frac{1}{n} \tag{7}$$

which in this form was proposed by Herdan (1966). Ray and Singer (1973) proposed the CON (for concentration) index:

$$\mathrm{CON} = \sqrt{\frac{\left(\sum_{j=1}^{n} p_j^2\right) - \frac{1}{n}}{1 - \frac{1}{n}}} \tag{8}$$

which can be rewritten as $\sqrt{\frac{n\lambda - 1}{n-1}} = \frac{V}{\sqrt{n-1}}$.

All these variants are proposed as measures of concentration. When they are bounded by 1, such as λ itself—but not V—one may propose $1 - \lambda$ as a diversity index. This diversity index is sometimes rewritten in other ways (Zhou et al. 2012; Leydesdorff 2015), which we show here for completeness and possible further reference.

$$1 - \lambda = 1 - \sum_{j=1}^{n} p_j^2 = \left(\sum_{i=1}^{n} p_i\right)\left(\sum_{j=1}^{n} p_j\right) - \sum_{j=1}^{n} p_j^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} p_i p_j \tag{9}$$

or

$$1 - \lambda = 2\left(\sum_{i=1}^{n}\left(\sum_{j=i+1}^{n} p_i p_j\right)\right) \tag{10}$$

When one is not interested in measures bounded by zero and one, and there are good reasons for this, see (Jost, 2006, 2007, 2009) and further on in this article, then one may use $1/\lambda$ as a measure of diversity. In (Nijssen et al., 1998) we have shown that the Lorenz curve represents balance or evenness, without taking the number of species (variety) into account. The corresponding evenness Simpson-type diversity measure is $1/(n \cdot \lambda)$.

Arguments used by Hirschman when introducing *C*

Hirschman (1945) wrote:

> *One of the well-known conditions of perfect competition is that no individual seller should command an important share of the total market supply; this condition implies the presence of both relative equality of distribution and of large numbers. The notion of concentration is thus seen to be more complex than the concept of income concentration. Therefore, the methods which have been devised to measure the concentration of income are inadequate for the measurement of the concentration phenomenon with which we are here concerned. An extreme case is this: if we would try to read off from a Lorenz graph the degree of concentration of an industry in which two firms divided between themselves the total output, we would have to conclude that, because the Lorenz curve would coincide with the equidistribution line, there is no concentration.*
>
> *Concentration of control or of power over a corporation over the market in one industry, or over foreign trade, is not only a direct function of the relative inequality of distribution but also a reciprocal function of the number of stockholders of producing firms in the industry, and of importing and exporting countries.*

Hirschman's argument is that the notion of concentration he needs to quantify, must combine balance as shown by a Lorenz curve and the reciprocal of variety (the number of companies in his case). He then continues introducing the C index and proves that this index satisfies the two requirements he needs. For this, he proves relation (6) in the form:

$$V^2 = n\left(\frac{C}{100}\right)^2 - 1 \quad \text{or} \quad C = 100\sqrt{\frac{V^2+1}{n}} \tag{11}$$

He argues that C (and we add, hence λ) increases with the coefficient of variation V, an accepted measure of relative dispersion, and decreases with n. The corresponding diversity measures $1 - \lambda$ and $1/\lambda$ increase with n, as expected for a diversity measure.

So far, so good, but …

There are two remarks we would like to make now. The first is about the argument used by Hirschman to prove that his index is a correct index for measuring concentration. This argument is based on the idea that the coefficient of variation is a correct measure of balance (he writes relative dispersion, but we do not know what he exactly means by this expression; we have the impression that he uses it in the general sense of 'relative variability'). In that way his index would combine balance and variety. The second is the case that cells are allowed to be empty. Consider the case (1, 0) and the case (1, 0, 0, 0, 0, 0, 0, 0, 0, 0). Clearly the second case represents a much more concentrated situation than the first one (one among 10 has everything, versus one among two has everything). Yet, for both the repeat rate is equal to one. The coefficients of variation for these two cases are 1 and 3 respectively. It is easy to show that for the case of one monopolist and $n - 1$ empty cells the coefficient of variation is $\sqrt{n-1}$.

In (Nijssen et al. 1998) we claimed that the Lorenz curve represents evenness or balance. The main argument is that replicated data arrays have the same Lorenz curves. By replicated data arrays we mean arrays of the form (4, 2, 1, 1), (4, 4, 2, 2, 1, 1, 1, 1) or (4, 4, 4, 2, 2, 2, 1, 1, 1, 1, 1, 1). In that article we showed that V is indeed a measure of evenness

or balance. So, this observation proves that the repeat rate combines in an acceptable way variety and balance, and, by the way, the coefficient of variation and the Gini-index (which clearly respects the balance requirement, but not variety) do not.

As to the second remark: the repeat rate does not satisfy the requirement that adding one empty cell must strictly increase the concentration value. Indeed, adding one (or more) empty cells does not change the repeat rate. We conclude that the repeat rate is only an acceptable measure if there are no empty cells. This implies that when data result from observations then there is never a problem with the repeat rate as one never sees a non-observed item. If, however one wants to find out if e.g., Olympic Gold Medals are distributed over participating countries more or less unequally since World War II, then the repeat rate is not good as there are always countries without gold medals and, moreover, the number of participating countries changed over the years. We note further that for a fixed number of cells (fixed variety) being an acceptable measure of concentration or diversity only depends on the balance requirement.

Further developments related to variety and balance

In this section we briefly mention two developments, referring the interested reader to the original publications for detailed arguments. The first refers to the requirement that indicators are to be preferred if statements such as "the diversity has decreased by 25%" are correctly reflected by them. The second relates to the sensitivity of an indicator. Some indicators may be more sensitive to rare species (referring to an ecological context) or to poor households (in a socio-economic context) while others may be more sensitive to the 'average' case. How can this be taken into account?

As to the first requirement, we refer to Jost (2006, 2007, 2009) who convincingly argued that a useful measure of diversity must lead to a value which is 50% higher for e.g., (4, 4, 4) than for (4, 4). Yet, using $1 - \lambda$ as diversity measure gives a value 2/3 in the first case and ½ in the second. He refers to diversity measures which satisfy his requirements as 'true' diversity measures. Clearly $1/\lambda$ is a true diversity measure: it yields 3 for the first case and 2 for the second. We note though that $1/\lambda$ is not bounded anymore.

For the second argument we refer to Hill (1973) who proposed a series of diversity measures, actually of "true diversities":

$$\left(\sum_{i=1}^{N} p_i^q\right)^{1/(1-q)} \tag{12}$$

where q is a parameter with values ranging from 0 to infinity (the cases $q = 1$ and $q = \infty$ are obtained as limits). For $q = 2$, this is: $\left(\sum_{i=1}^{N} p_i^2\right)^{-1} = \frac{1}{\sum_{i=1}^{N} p_i^2} = \frac{1}{\lambda}$. Calculating the Hill measures for a given situation leads to a diversity profile, a set of values, one for each parameter value q, reflecting the sensitivity of the indicator. Examples of diversity profiles can be found e.g. in Leinster and Cobbold (2012) who provide profiles for coral diversity, butterfly diversity and flounder diet diversity.

Refinements

After having reviewed a wide range of documents on diversity measurement Stirling (2007) concluded that diversity consists of three basic concepts: variety, balance and disparity, each of them being a necessary but insufficient property of diversity as a whole. Species in ecology, or scientific fields in the Science of Science, are not independent entities, but they are shaped by patterns of common developments or ancestry leading to proximities (or the opposite: disparities) between units (Stirling 2007). This brought him to propose the following measure D, since then often referred to as the Rao-Stirling measure, also known as quadratic entropy:

$$D = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} d_{ij} p_i p_j \tag{13}$$

where $(d_{ij})_{i,j}$ is a disparity matrix. We note that if one sets $d_{jj} = 0$ for every j (a natural requirement for a disparity function), then the restriction $i \neq j$ is not necessary. This measure was already proposed by Nei and Li (1979) [the formula for nucleotide diversity with number [22] in their article is exactly our formula (13)]. Independently, Rao (1982) had proposed a formula structured in a similar way, be it in a more general context. In his article d represents a non-negative symmetric function (leading to values d_{ij}), and the measure D is not defined as a sum, but as an integral over some probability space.

More recently all these ideas were brought together, including Jost's suggestion to use so-called true diversity measures, and fully elaborated by Leinster and Cobbold (2012). This led to the formulae

$${}^{q}D^{S} = \left(\sum_{i=1}^{N} p_i \left(\sum_{j=1}^{N} s_{ij} p_j \right)^{q-1} \right)^{\frac{1}{1-q}} \tag{14}$$

where q has the same meaning as in formula (12); $(s_{ij})_{i,j}$ ($0 \leq s_{ij} \leq 1$; and for all j, $s_{jj} = 1$) denotes a similarity matrix; setting $d_{ij} = 1 - s_{ij}$ leads to a disparity matrix. Using this matrix and taking $q = 2$ leads to a monotone transformation of the Rao-Stirling measure D:

$${}^{2}D^{S} = \frac{1}{\sum\limits_{i,j=1}^{N} \left(1 - d_{ij}\right) p_i p_j} = \frac{1}{\sum\limits_{i,j=1}^{N} p_i p_j - D} = \frac{1}{1 - D} \tag{15}$$

The use of formula (15) in an informetric framework was advocated in (Mugabushaka et al. 2015; Zhang et al. 2016). Detailed analyses and generalizations of the Rao-Stirling measure can be found in (Ricotta and Szeidl 2006; Pavoine 2012; Smouse et al. 2017) among others. Discussions of these further developments would lead us too far and are not the purpose of this short note.

Discussion

Following Rafols and Meyer (2010) we claim that interdisciplinarity or knowledge integration has two main aspects: diversity and coherence. Referring to Wagner et al. (2011), Leydesdorff et al. (2018) and Rousseau et al. (2018) we remark that it is still an open

question if diversity of references is the best approach to measure the diversity aspect of interdisciplinarity. In relation to the coherence aspect we note that (Rafols et al. 2012) use as a measure of coherence the ratio of the observed average distance of cross-citations and the expected average distance, where the 'expected' part is calculated using the Rao-Stirling measure (leading to another use of this measure).

The Rao-Stirling measure does not meet the 'monotonicity of balance' requirement

In Leydesdorff et al. (2018, p. 573) the authors wrote that *these diversity measures* [referring to quadratic entropy or its true diversity variant] *do not include "balance" as the third element distinguished in the definition of interdisciplinarity* by *Rafols and Meyer* (2010). By this statement the authors meant (Leydesdorff, personal communication) that the Rao-Stirling measure is a "dual concept" (Junge 1994; Stirling 2007) which includes properties in a hidden way. In this case they refer to the fact that the Rao-Stirling measure, similarly to the repeat rate, hides the balance aspect.

Yet, on reflection, we found out that the balance aspect is not hidden, but simply is not present. Concretely, the Rao-Stirling measure does not meet the 'monotonicity of balance' property (Stirling 2007) which states that for given variety and disparity (or similarity) the measure increases monotonically with balance. In order to prove this point one counterexample suffices.

Let $D = \begin{pmatrix} 0 & 0.2 & 0.9 \\ 0.2 & 0 & 0.1 \\ 0.9 & 0.1 & 0 \end{pmatrix}$ be a disparity matrix and let $P = \begin{pmatrix} 5/9 \\ 3/9 \\ 1/9 \end{pmatrix}$ be a normalized array. Calculating the Rao-Stirling formula (13) for this example leads to the value 0.1926. If we now consider the array $Q = \begin{pmatrix} 6/9 \\ 2/9 \\ 1/9 \end{pmatrix}$ then we obtain a Rao-Stirling value of 0.1975, which is larger than the previous value. Clearly, the evenness or balance of P is larger than the evenness of Q (consider their Lorenz curves). So the array with the larger evenness or balance (P) has the smaller diversity value as measured with the Rao-Stirling measure (with the same variety and similarity). This shows that the Rao-Stirling measure does not meet the "monotonicity of balance" requirement. The same statement holds for its "true diversity" variant (Zhang et al. 2016).

Conclusion

This short note provided a short history of the repeat rate, also known as the Hirschman–Herfindahl or as the Simpson index. This led us to the Rao-Stirling index and an example that it does not meet the 'monotonicity of balance' property. Yet, following Leinster and Cobbold (2012), we do not consider balance as an essential factor for diversity.

We conclude that, if using diversity of references is at least one (among others) acceptable approach to study knowledge integration, then we see no reason why the Rao-Stirling true diversity should not be used. Yet, one must realize that this measure does not meet the "monotonicity of balance" requirement. Of course, this does not exclude that other indicators may exist, with, perhaps, better—extra—properties.

Acknowledgement We thank Loet Leydesdorff for helpful observations about an earlier version of this note.

References

Good, I. J. (1982). Comment [on Patil & Taillie, 1982]. *Journal of the American Statistical Association, 77*(379), 561–563.

Herdan, G. (1966). *The Advanced theory of language as choice and chance*. Berlin: Springer-Verlag.

Herfindahl, O. C. (1950). *Concentration in the U.S. steel industry*. Doctoral dissertation, Columbia University.

Hill, M. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology, 54*(2), 427–432.

Hirschman, A. O. (1945). *National power and the structure of foreign trade*. Berkeley: University of California Press.

Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review, 54*(5), 761–762.

Jost, L. (2006). Entropy and diversity. *Oikos, 113*(2), 363–375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology, 88*(10), 2427–2439.

Jost, L. (2009). Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). *Ecological Economics, 68*(4), 925–928.

Junge, K. (1994). Diversity of ideas about diversity measurement. *Scandinavian Journal of Psychology*, *35*(1), 16–26.

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: The importance of species similarity. *Ecology, 93*(3), 477–489.

Leydesdorff, L. (2015). Can technology life-cycles be indicated by diversity in patent classifications? The crucial role of variety. *Scientometrics, 105*(3), 1441–1451.

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2018). Betweenness and diversity in journal citation networks as measures of interdisciplinarity—A tribute to Eugene Garfield. *Scientometrics, 114*(2), 567–592.

Mugabushaka, A.-M., Kyriakou, A., & Papazoglou, T. (2015). Bibliometric indicators of interdisciplinarity exploring new class of diversity measures. In A. A. Salah, Y. Tonta, A. A. A. Salah, C. Sugimoto, & U. Al (Eds.), *Proceedings of ISSI 2015* (pp. 397–402). Istanbul: Boğaziçi University Printhouse.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Science USA, 76*(10), 5269–5273.

Nijssen, D., Rousseau, R., & Van Hecke, P. (1998). The Lorenz curve: A graphical representation of evenness. *Coenoses, 13*(1), 33–38.

Patil, G. P., & Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association, 77*(379), 548–561.

Pavoine, S. (2012). Clarifying and developing analyses of biodiversity: Towards a generalisation of current approaches. *Methods in Ecology and Evolution, 3*(3), 509–518.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management. *Research Policy, 41*(7), 1262–1282.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology, 21*(1), 24–43.

Ray, J. L., & Singer, J. D. (1973). Measuring the concentration of power in the international system. *Sociological Methods and Research, 1*(4), 403–437.

Ricotta, C., & Szeidl, L. (2006). Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology, 70*(3), 237–243.

Rousseau, R., Hu, X. J., & Zhang, L. (2018). Knowledge integration: Its meaning and measurement. In: W. Glänzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. (**To appear**).

Simpson, E. H. (1949). Measurement of diversity. *Nature, 163*(4148), 688.

Smouse, P. E., Banks, S. C., & Peakall, R. (2017). Converting quadratic entropy to diversity: Both animals and alleles are diverse, but some are more diverse than others. *PLoS ONE, 12*(10), e0185499.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society, Interface, 4*(15), 707–719.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics, 5*(1), 14–26.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator for interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology, 67*(5), 1257–1265.

Zhou, Q. J., Rousseau, R., Yang, L. Y., Yue, T., & Yang, G. L. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics, 93*(3), 787–812.