South African Statistical Journal
2018, Vol. 52, No. 2, 129–137

ON THE GINI–SIMPSON INDEX AND ITS GENERALISATION — A HISTORIC NOTE

Ferdinand Österreicher[1]
University of Salzburg, Department of Mathematics, Salzburg, Austria
e-mail: *ferdinand.oesterreicher@sbg.ac.at*

José A. P. Casquilho
Universidade Nacional Timor Lorosa'e, Dili, República Democrática de Timor-Leste

Let $P = (p_1, \ldots, p_n)$ be a probability distribution on a set $\Omega = \{\omega_1, \ldots, \omega_n\}$ with n elements, $n \in \mathbb{N}\backslash\{1\}$. Then the term $S_2(P) = 1 - \sum_{i=1}^n p_i^2$, frequently called the *Gini–Simpson index*, or, in information theory, *quadratic entropy*, is used in many different areas of research resp. applications and was, therefore, reinvented several times. In this note we give a concise history of this index and closely related measures, as well as its generalisation to all values of the parameter of the class of entropies of order $\alpha \in (0, \infty)\backslash\{1\}$ introduced by *Havrda* and *Charvát* (1967) and reinvented by *Tsallis* (1988) for the use of this index in statistical physics, for which the limiting case for $\alpha \to 1$ is *Shannon's entropy* $S_1(P) = -\sum_{i=1}^n p_i \ln p_i$. We also give a brief note on weighted versions of the Gini–Simpson index.

In addition to these central historic features our note also presents contributions on the axiomatics of entropies and on the early history of the application of the concept of entropy in thermodynamics. We also provide an entry on *Rényi*'s class of entropies, linked with *Hill*'s diversity numbers.

Key words: Gini–Simpson index, Measures of entropy.

1. Introduction

Let $\mathcal{P}_n$ be the set of all probability distributions $P = (p_1, \ldots, p_n)$ on a set $\Omega = \{\omega_1, \ldots, \omega_n\}$ with n elements, $n \in \mathbb{N}\backslash\{1\}$. Then the term

$$S_2(P) = 1 - \sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{n} p_i \cdot (1 - p_i)\,,$$

frequently called the *Gini–Simpson index*, or, in information theory, *quadratic entropy*, is used in many different areas of research resp. applications and was, therefore, reinvented several times.

In his fundamental paper (1948, p. 11), following *Ralph Vinton Lyon Hartley*'s (1888–1970) paper (1928), *Claude Shannon* (1916–2001) defined his entropy by

$$S_1(P) = -K \cdot \sum_{i=1}^{n} p_i \cdot \log_2(p_i)\,,$$

[1]Corresponding author.

MSC2010 subject classifications. 01A85, 62B10, 94A17.

K being a positive constant[2]. In fact an entropy is a *measure of uncertainty* resp. *of lack of concentration*.

Apart from the so-called property of *additivity*, the further crucial properties of Shannon's entropy $H(P) = S_1(P)$ are, as well as those of the quadratic entropy $S_2(P)$, the following:

- *Positivity*: $H(P) \geq 0$;
- *Expansibility*: An "expansion" of $P = (p_1, ..., p_n) \in \mathcal{P}_n$ to $P = (p_1, ..., p_n, 0) \in \mathcal{P}_{n+1}$ does not change $H(P)$;
- *Symmetry*: $H(P)$ is invariant under permutations of $p_1, ..., p_n$;
- *Continuity*: $H(P)$ is a continuous function of P (for fixed n);
- *Maximum property*: $H(P)$ is maximal if and only if $P = P_n = (\frac{1}{n}, ..., \frac{1}{n})$.

For a more detailed discussion of the properties of entropies see e.g. *Csiszár*'s review paper (2008, p. 263).

Provided that $H(P)$ is an entropy, which satisfies at least the latter five properties, then - most naturally - the difference

$$H(P_n) - H(P)$$

is the *corresponding measure of concentration*.

For Shannon's entropy the corresponding measure of concentration is

$$\kappa_1(P) = \log_2(n) - S_1(P) = I(P \parallel P_n)\,,$$

where the latter is the special case of the *Kullback–Leibler divergence*, or, in short, the *I-divergence* of P and P_n, introduced by the US-American cryptanalysts and mathematicians *Solomon Kullback* (1907–1994) and *Richard A. Leibler* (1914–2003) in their paper (1951).

Remark 1 The $\log_2$ and the ln, respectively, are typical for the use in information theory and physics.

Early history of entropy in thermodynamics

Thermodynamics was started by the interest in heat and the motion of molecules within a box filled with gas. Important corresponding work was done by the following physicists: the German *Rudolf Clausius* (1822–1888), the Scotsman *James Clerk Maxwell* (1831–1879), the US-American *Josiah Willard Gibbs* (1839-1903), the Austrian *Ludwig Boltzmann* (1844–1906) and the German *Max Planck* (1858-1947). *Einstein*'s paper (1905) on *Brownian molecular movement* perhaps marks the end of the early history of research in thermodynamics.

Remark 2 In 1865, *Clausius* gave the first mathematical version of the concept of entropy, and also gave it its name. He chose the word because the meaning (from Greek *ἐν* [en] "in" and *τροπή* [tropē] "transformation") is "content transformative" or "transformation content" (German: "Verwandlungsinhalt").

[2]In our note we typically choose $K = 1$.

Following *Boltzmann*'s paper (1877) let $P_n = (\frac{1}{n}, ..., \frac{1}{n})$ be the uniform distribution on a set $\Omega_n = \{\varepsilon_1, ..., \varepsilon_1\}$, $(m_1, ..., m_n) \in \mathbb{N}_0^n : \sum_{i=1}^n m_i = m$, let M_{m,P_n} denote the multinomial distribution with parameters m and P_n, and finally, let the vector $(X_1, ..., X_n)$ be distributed according to M_{m,P_n}. Then the probability of the event $\{(X_1, ..., X_n) = (m_1, ..., m_n)\}$ equals

$$W := W(X_1 = m_1, \cdots, X_n = m_n) = \frac{m!}{m_1! \cdot \ldots \cdot m_n!} \cdot \left(\frac{1}{n}\right)^m .$$

In addition, let $P = (p_1, ..., p_n)$ be the probability distribution given by $p_i = m_i/m$, $i \in \{1, ..., n\}$. Then by virtue of *Stirling*'s formula

$$n! \sim n^n \cdot e^{-n} \cdot \sqrt{2\pi n}$$

we get, in a first approximation,

$$\ln\left(\frac{m!}{m_1! \cdot \ldots \cdot m_n!}\right) \cong m \cdot \sum_{i=1}^n p_i \cdot \ln(p_i) .$$

Consequently, the logarithm of the inverse W^{-1} of the probability W is approximately

$$\ln W^{-1} \cong m \cdot (\ln(n) - H(P)) = m \cdot I(P \| P_n) . \quad (1)$$

From (1) one may — of course, apart from the exponent -1 — easily identify the relationship given by *Max Planck* in his paper (1901, formula (3), p. 556)

$$\boxed{S = k \cdot \ln W + const}$$

for the connection between the entropy S and the probability W, where k is *Boltzmann's constant*[3]. *Planck*'s famous formula adorns the epitaph of *Boltzmann*'s honorary grave on Vienna's Central Cemetery (see Figure 1).

Remark 3 Let E, S and T denote the energy, the entropy and the absolute temperature (measured in degrees *Kelvin*). Then the following relationship applies:

$$\frac{dE}{dS} = k \cdot T ,$$

which was discovered by *Max Planck* in a first version resp. the US-American mathematician and physicist of Hungarian origin *John von Neumann* (1903–1957) in its final form.

In this note we give a concise history of the *Gini–Simpson index* and closely related measures, as well as its generalisation to all values of the parameter of the class of entropies of order $\alpha \in [0, \infty)\backslash\{1\}$ introduced by *Havrda* and *Charvát* (1967) and reinvented by *Tsallis* (1988) for its use in statistical physics. Finally, we provide a brief account concerning weighted versions of the Gini–Simpson index and its generalisations. In addition, an entry on *Rényi*'s class of entropies is presented, combined with *Hill*'s diversity numbers $N_\alpha(P)$.

[3] In fact *Planck*'s term W equals *Boltzmann*'s and our term W^{-1}, which roughly is the number of *a priori* equal probable microscopic states. Note that the German word for probability is *Wahrscheinlichkeit*.

Figure 1. Photo of *Boltzmann's* epitaph.

2. Gini–Simpson index and related quantities

(A) The eminent Italian statistician *Corrado Gini* (1884–1965) introduced the index $S_2(P)$ in equations (141) and (143) of his Section "*Gli indici di mutatilità per serie sconnesse*", i.e. *index of mutability for disconnected (qualitative) variables*, in his book *Variabilità e Mutabilità* (1912). He applied the term $(1 - \frac{1}{n})^{-1} \cdot S_2(P)$, which he named "*differenza media*" ("*mean difference*"), using relative frequencies of qualitative data, exemplified with the colour of eyes and hair of Italian soldiers from the different Italian provinces.

(B) *Solomon Kullback* introduced and used the quantity

$$\kappa(P) = \sum_{i=1}^{n} p_i^2 ,$$

which he named *probability of monographic coincidence*, in his technical paper (1938, pp. 81–84) on cryptography and cryptanalysis. This quantity was used by the US Signal Intelligence Service also for breaking into codes — typically so-called *Vigenère ciphers* — used by some of the 'bootleggers' smuggling alcohol from Canada into the United States during the Prohibition era (1920–1933). For further corresponding information on this subject, see e.g. *Österreicher* (2008, Section 2.4).

(C) The US-American economist of German origin, *Albert O. Hirschman* (1915–2012), introduced and used the measure

$$\kappa^{1/2}(P) = \sqrt{\sum_{i=1}^{n} p_i^2}$$

in his book (1945, Chapter VI and pp. 157–162) on political economy.

(D) *Edward H. Simpson* (1922–), an Englishman, who as a young man was also a cryptanalyst[4],

[4] He was working in the British cryptanalytic group with *Alan Turing* (1912–1954) at *Blatchley Park* during 1942–45.

introduced the original term $\kappa(P)$ in his seminal paper (1949) as a measure of concentration in terms of population constants.

(E) The US-American economist *Orris C. Herfindahl* (1918–1972) reinvented the quantity $\kappa(P)$ in his Ph.D. thesis (1950, p. 19).

(F) *Jack P. Gibbs* and *Walter T. Martin* reinvented and used the quantity $S_2(P)$ in their paper (1962, p. 670) as an index of diversity (or lack of concentration) in sociology.

(G) *Igor Vajda* (1942–2010), a colleague and friend of one of us (*F.Ö.*), independently introduced this quantity in his paper (1968) in the context of information theory and used it as a bound for the probability of error for testing multiple hypotheses. In his paper (1969, pp. 515–516) he coined the name *quadratic entropy* for the term $S_2(P)$ and used it also for pattern recognition. *C. R. Rao* (1984, p. 76) named the Gini–Simpson index *second order entropy*. This term, however, should not be confused with Rényi's *entropy of order* $\alpha = 2$ (1961, formula (1.21), p. 549).

(H) The US-American sociologist of Austrian origin, *Peter M. Blau* (1918–2002), reintroduced the quantity $S_2(P)$, which he called *measure of heterogeneity*, in his influential book (1977, p. 9) on sociology. This is the reason that it is frequently called *Blau index* in this field.

3. Generalised entropies

(I) The first generalisation of Shannon's entropy was introduced and investigated by the eminent Hungarian mathematician *Alfréd Rényi* (1921–1971) in his seminal paper (1961) defined by

$$R_\alpha(P) = \frac{1}{1-\alpha}\log_2\Big(\sum_{i=1}^{n} p_i^\alpha\Big), \quad \alpha \in (0,\infty)\backslash\{1\},$$

the so-called *Rényi's class of entropies of order* α, Shannon's entropy being the limiting case for $\alpha \to 1$. As a matter of fact, *M. O. Hill* used in his *paper* (1973, p. 428) an exponential form of *Rényi*'s entropies in order to derive his class of diversity numbers, namely by $N_\alpha(P) = (\sum_{i=1}^n p_i^\alpha)^{1/(1-\alpha)} \in [1,n]$, $\alpha \in (0,\infty)\backslash\{1\}$, the particular case for $\alpha = 2$ being $N_2(P) = \kappa(P)^{-1}$.

The corresponding measures of concentration are therefore

$$\kappa_\alpha^R(P) = \log_2(n) - \frac{1}{1-\alpha}\log_2\Big(\sum_{i=1}^{n} p_i^\alpha\Big) = I_\alpha(P \mid P_n),$$

where the latter is Rényi's *information of order* α *obtained if* P_n *is replaced by* P (see Rényi, 1961, formula (3.3), p. 554).

(J) *Jan Havrda* and *František Charvát* (1967) generalised the measure $S_2(P)$ to the class

$$\check{S}_\alpha(P) = \check{c}(\alpha)\cdot\Big(1 - \sum_{i=1}^{n} p_i^\alpha\Big) = \check{c}(\alpha)\cdot\sum_{i=1}^{n} p_i\cdot(1 - p_i^{\alpha-1}),$$

including all parameters $\alpha \in (0,\infty)\backslash\{1\}$, with the standardisation

$$\check{c}(\alpha) = \frac{1}{1-2^{1-\alpha}},$$

for use in information theory; the limiting case

$$\lim_{\alpha\to 1} \check{S}_\alpha(P) = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i)$$

being *Shannon's* entropy. It is remarkable that *Havrda* and *Charvát* derived their class of information measures (entropies) in an axiomatic way. Cf. also *Csiszár's* review paper (2008, Section 2.4).

(K) *Constantino Tsallis* (1988) reinvented the generalisation

$$S_\alpha(P) = c(\alpha)\cdot\Big(1-\sum_{i=1}^{n} p_i^\alpha\Big) = c(\alpha)\cdot\sum_{i=1}^{n} p_i\cdot(1-p_i^{\alpha-1})$$

with the standardisation

$$c(\alpha) = \frac{1}{\alpha-1}, \quad \alpha\in(0,\infty)\backslash\{1\},$$

especially for its use in thermodynamics. Note that owing to $c(2) = 1$ the class $S_\alpha(P)$ is an extension of $S_2(P)$ in the literal sense (note that, however $\check{c}(2) = 2$) and, of course, *Shannon's entropy*[5] is the limiting case for $\alpha \to 1$ (cf. also *De Wet and Österreicher*, 2017, Section 2).

The corresponding measures of concentration are

$$\kappa_\alpha^T(P) = \frac{\sum_{i=1}^n p_i^\alpha - n^{1-\alpha}}{\alpha-1} .$$

(L) For the special case $\alpha = 2$, i.e. the quadratic entropy or the Gini–Simpson index $S_2(P)$ the corresponding measure of concentration is

$$\kappa_2^T(P) = \sum_{i=1}^{n} p_i^2 - \frac{1}{n} = \sum_{i=1}^{n}\Big(p_i-\frac{1}{n}\Big)^2 .$$

Časlav Brukner and *Anton Zeilinger* introduced this quantity as an appropriate measure of information for the specific use in quantum measurement in their paper (1999, p. 2).

Remark 4 The one-to-one relationship between $S_\alpha(P)$ and $R_\alpha(P)$ is

$$R_\alpha(P) = \frac{1}{1-\alpha}\cdot\log_2(1+(1-\alpha)\cdot S_\alpha(P)) .$$

4. Weighted versions

(M) For this subject matter, let the assumptions and terminology of (J) be valid and let, in addition, $W = (w_1,\ldots,w_n) \in (0,\infty)^n$ be a vector with positive weights. Then *Shama, Mitter* and *Mohan* (1978, p. 334) called the entities

$$H_\alpha(W,P) = \check{c}(\alpha)\cdot\sum_{i=1}^{n} w_i\cdot p_i\cdot(1-p_i^{\alpha-1}), \quad \alpha\in(0,\infty)\backslash\{1\},$$

[5] with ln instead of $\log_2$

and

$$H_1(W,P) = \lim_{\alpha\to 1} H_\alpha(W,P) = -\sum_{i=1}^{n} w_i \cdot p_i \cdot \log_2(p_i)\,,$$

generalised useful information of degree α, identical to what *Aggarwal* and *Picard* (1978, p. 175) named *entropy of degree* α *with utility of an experiment*; in every case referring to the previous work of *Emptoz* (1976) on this subject. From the general expression we get, owing to $\check{c}(2) = 2$,

$$H_2(W,P) = 2 \cdot S_2^W(P) \quad \text{with} \quad S_2^W(P) = \sum_{i=1}^{n} w_i \cdot p_i \cdot (1 - p_i),$$

which will be referred to as the *weighted Gini–Simpson index*.

(N) Explicit versions of the weighted Gini–Simpson index $S_2^W(P)$ do not seem to have appeared before the end of the last century. *Casquilho* (1999, pp. 91–124 and 2016) formulated $S_2^W(P)$ as a sum of variances of interdependent Bernoulli variables and studied the range, including optimal solutions, written within the scope of ecological and economic applications, while *Sen* (1999) addressed $S_2^W(P)$ within the realm of utility-oriented indices concerning the economics of poverty. Also, *Guiaşu* and *Guiaşu* (2003) independently reinvented and studied $S_2^W(P)$ under the scope of conditional and weighted measures of ecological diversity. A more detailed historic study on the weighted Gini–Simpson indices is still to be done.

5. Conclusion

Owing to the simple analytic structure of the Gini–Simpson index and of the related quantities, it is not surprising that these quantities were often reinvented in various disciplines. Their variety is nevertheless amazing.

Major findings of our historic study of these entities are the following:

1. As *Corrado Gini* has created many indices in his book (1912) and because a considerable amount of research has been devoted to these indices, our identification of the quantity $S_2(P)$ in equations (141) and (143) of Section "*Gli indici di mutabilità per serie sconesse*" came as a nice confirmation.

2. In the relevant literature the term *index of coincidenc*e is often identified with the quantity $\kappa(P)$ and both are attributed to the prominent US-American cryptanalyst of Russian origin *William F. Friedman* (1891–1969) and to his fundamental treatise (1922) on cryptopgraphy. A detailed study of the latter showed, however, that *Friedman*'s index of coincidence is completely different from $\kappa(P)$ and that *Friedman* does not seem to be, with a reasonable amount of certainty, the originator of the latter.

3. It has rather turned out that *Friedman*'s young colleague *Solomon Kullback* invented $\kappa(P)$ and presented it in his technical paper (1938).

Acknowledgement. The authors would like to thank Prof. *Tertius de Wet* for his valuable suggestions, which substantially improved, among other aspects, both the structure of the paper and the design of the introduction.

References

Aggarwal, N. L. and Picard, C.-F. (1978). Functional equations and information measures with preference. *Kybernetica*, **14**, 174–181.

Blau, P. M. (1977). *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. The Free Press, New York.

Boltzmann, L. (1877). Über die Beziehung zwischen dem zweiten Hauptsatz der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Sitzungsberichte der mathematisch-naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften Wien, II Abtg.*, 373–435.

Brukner, Č. and Zeilinger, A. (1999). Operationally invariant information in quantum measurements. *Physical Review Letters*, **83**, 1–4.

Casquilho, J. A. P. (1999). *Ecomosaico: índices para o diagnóstico de proporções de composição (Ecomosaic: indices for diagnosis of compositional proportions)*. Ph.D. thesis, Universidade Técnica de Lisboa.

Casquilho, J. A. P. (2016). A methodology to determine the maximum value of weighted Gini–Simpson index. *SpringerPlus*, **5**, 1–10. doi:10.1186/s40064-016-2754-8.

Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy*, **10**, 261–273.

De Wet, T. and Österreicher, F. (2017). A note on extended Arimoto's entropies. *South African Statistical Journal*, **51**, 285–294.

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, **322**, 549–560.

Empotz, H. (1976). Information de type β intégrant un concept d'utilité. *C. R. Acad. Sci. Paris*, **282A**, 911–914.

Friedman, W. F. (1922). The index of coincidence and its applications in cryptography. *Riverbank Publication No 22, Riverbank Labs*. Reprinted by Aegean Park Press, 1987.

Gibbs, J. P. and Martin, W. T. (1962). Urbanization, technology, and the division of labor: International patterns. *American Sociological Review*, **27**, 667–677.

Gini, C. (1912). Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche. *In Studi Economici-giuridici della Regia Facoltà di Giurisprudenza della R. Università di Cagliari*, anno III, parte 2. Tipografia di Paolo Cuppini, Bologna.

Guiaşu, R. C. and Guiaşu, S. (2003). Conditional and weighted measures of ecological diversity. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **11**, 283–300.

Hartley, R. V. (1928). Transmission of information. *Bell System Technical Journal*, **7**, 535–563.

Havrda, J. and Charvát, F. (1967). Quantification method of classification processes. Concept of structural α-entropy. *Kybernetika*, **3**, 30–35.

Herfindahl, O. C. (1950). *Concentration in the U.S. steel industry*. Ph.D. thesis, Columbia University, New York.

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

Hirschmann, A. O. (1945). *National Power and the Structure of Foreign Trade*. University of

California, Berkeley, California.

KULLBACK, S. (1938). Statistical methods in cryptanalysis – revised edition. Technical report, Signal Intelligence Service, War Department, Washington D.C.

KULLBACK, S. AND LEIBLER, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

ÖSTERREICHER, F. (2008). Informationstheorie. Lecture notes, University of Salzburg.

PLANCK, M. (1901). Ueber das Gesetz der Energieverteilung im Normalspectrum. *Annalen der Physik*, **4**, 553–563.

RAO, C. R. (1984). Convexity properties of entropy functions and analysis of diversity. *Lecture Notes – Monograph Series*, **5**, 68–77.

RÉNYI, A. (1961). On measures of entropy and information. *In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley.

SEN, P. K. (1999). Utility-oriented Simpson-type indexes and inequality measures. *Calcutta Statistical Association Bulletin*, **49**, 1–22.

SHANNON, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423 and 623–656.

SHARMA, B. D., MITTER, J., AND MOHAN, M. (1978). On measures of "useful" information. *Information and Control*, **39**, 323–336.

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

TSALLIS, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–487.

VAJDA, I. (1968). Bounds of the minimal error probability on checking a finite or countable number of hypotheses (in Russian). *Problemy Peredachi Informatsii*, **4**, 9–19.

VAJDA, I. (1969). A contribution to the informational analysis of pattern. *In* WANTABE, S. (Editor) *Methodologies of Pattern Recognition*. Academic Press, New York, 509–519.

Manuscript received 2018-02-15, revised 2018-03-05, accepted 2018-03-08.