

Article

On the Measurement of Randomness (Uncertainty): A More Informative Entropy

Tarald O. Kvålseth ^{1,2}

¹ Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN 55455, USA; kvals001@umn.edu; Tel.: +1-952-470-1170; Fax: +1-952-470-1169

² Department of Industrial & Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Academic Editor: Raúl Alcaraz Martínez

Received: 11 February 2016; Accepted: 15 April 2016; Published: 26 April 2016

Abstract: As a measure of randomness or uncertainty, the Boltzmann–Shannon entropy H has become one of the most widely used summary measures of a variety of attributes (characteristics) in different disciplines. This paper points out an often overlooked limitation of H : comparisons between differences in H -values are not valid. An alternative entropy H_K is introduced as a preferred member of a new family of entropies for which difference comparisons are proved to be valid by satisfying a given value-validity condition. The H_K is shown to have the appropriate properties for a randomness (uncertainty) measure, including a close linear relationship to a measurement criterion based on the Euclidean distance between probability distributions. This last point is demonstrated by means of computer generated random distributions. The results are also compared with those of another member of the entropy family. A statistical inference procedure for the entropy H_K is formulated.

Keywords: randomness; uncertainty; entropy; value validity

1. Introduction

For some probability distribution $P_n = (p_1, \dots, p_n)$, with $p_i \geq 0$ for $i = 1, \dots, n$ and $\sum_{i=1}^n p_i = 1$, the entropy $H(P_n)$, or simply H , is defined by:

$$H = - \sum_{i=1}^n p_i \log p_i \in [0, \log n] \quad (1)$$

where the logarithm is the natural (base- e) logarithm. The probabilities $p_i (i = 1, \dots, n)$ may be associated with a set of quantum states of a physical system in statistical mechanics or physics, a set of symbols or messages in a communication system, or, most generally, a set of mutually exclusive and exhaustive events. First used by Boltzmann [1] in statistical mechanics (as kH with k being the so-called Boltzmann constant) and later introduced by Shannon [2] as the basis for information theory (with base-2 logarithm and *bits* as the unit of measurement), this entropy H can appropriately be called the Boltzmann–Shannon entropy.

Although interpreted in a number of different ways, the most common and general interpretation of H is as a measure of *randomness* or *uncertainty* of a set of random events (e.g., [3] (pp. 67–97), [4], [5] (Chapter 2), [6] (pp. 12, 13, 90)). A number of alternative entropy formulations have been proposed as parameterized generalizations of H in Equation (1) (see, e.g., [7–9]), but with limited success or impact. The most notable exception is the following one-parameter family of entropies by Rényi [10]:

$$H_R = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^n p_i^\alpha, \alpha > 0, \alpha \neq 1$$

which reduces to Equation (1), with base-2 logarithm, when $\alpha \rightarrow 1$. This entropy family has, for instance, been used as a fractal dimension [11] (pp. 686–688). Another such family of entropies is that of Tsallis [12] defined as:

$$H_T = \frac{1}{\alpha - 1} \left(1 - \sum_{i=1}^n p_i^\alpha \right), \quad -\infty < \alpha < \infty$$

which includes the H in Equation (1) as the limiting case when $\alpha \rightarrow 1$.

Since its origins in statistical physics and information theory, the entropy H in Equation (1) has proved to be remarkably versatile as a summary measure of a wide variety of attributes (characteristics) within diverse fields of study, ranging from psychology (e.g., [13]) to fractal geometry [11] (pp. 678–687). However, such widespread use of H has led to misuse, improper applications, and misleading results due to the fact that, although H has a number of desirable properties [14] (Chapter 1), it does suffer from one serious limitation: comparisons between differences in H -values are not valid. The basis for this limitation is explained in the next section of this paper.

As a clarification of such comparisons in general, consider some summary measure M and probability distributions P_n, Q_m, R_t, S_u . The various types of potential comparisons can then be defined as follows:

$$\text{Size (order) comparisons : } M(P_n) > M(Q_m) \tag{2a}$$

$$\text{Difference comparisons : } M(P_n) - M(Q_m) > M(R_t) - M(S_u) \tag{2b}$$

$$\text{Proportional difference comparisons : } M(P_n) - M(Q_m) = c[M(R_t) - M(S_u)] \tag{2c}$$

where c is a constant. While, because of the properties of H in Equation (1), there is no particular reason to doubt the validity of the size comparison in Equation (2a) involving H , the difference comparisons in Equations (2b) and (2c) are not valid for H as discussed below.

In this paper, an alternative and equally simple entropy is introduced as:

$$H_K = \left(\sum_{\substack{i=1 \\ i \neq m}}^n \sqrt{p_i} \right)^2 + \sum_{\substack{i=1 \\ i \neq m}}^n p_i, \quad p_m = \max_i \{p_i\} \tag{3}$$

The term *entropy* is used for this measure of randomness or uncertainty since (a) it has many of the same properties as H in Equation (1) and (b) the *entropy* term has been used in such a variety of measurement situations for which H_K can similarly be used. As is established in this paper, H_K has the important advantage of being more informative than H in the sense that H_K meets the conditions for valid difference comparisons as in Equations (2b) and (2c). It will also be argued that the H_K is the preferred member of a family of entropies with similar properties. Statistical inference procedure for H_K will also be outlined.

2. Conditions for Valid Difference Comparisons

Consider that M is a measure of randomness (uncertainty) such that its value $M(P_n)$ for any probability distribution $P_n = (p_1, \dots, p_n)$ is bounded as:

$$M(P_n^0) \leq M(P_n) \leq M(P_n^1) \tag{4}$$

where P_n^0 and P_n^1 are the degenerate and uniform distributions

$$P_n^0 = (1, 0, \dots, 0), \quad P_n^1 = (1/n, \dots, 1/n) \tag{5}$$

and where one can set $M(P_n^0) = 0$. In order for the difference comparisons in Equations (2b) and (2c) to be permissible or valid, some condition needs to be imposed on M (see [15]). Specifically, all intermediate values $M(P_n)$ in Equation (4) have to provide numerical representations of the

extent of randomness (uncertainty) that are true or realistic with respect to some acceptable criterion. While different types of validity are used in measurement theory [16] (pp. 129–134), *value validity* will be used here for this required property of M . In order to establish specific requirements for M to have value validity, a particular probability distribution proves useful and Euclidean distances will be used as a criterion.

Therefore, consider the recently introduced *lambda distribution*:

$$P_n^\lambda = \left(1 - \lambda + \frac{\lambda}{n}, \frac{\lambda}{n}, \dots, \frac{\lambda}{n} \right), \lambda \in [0, 1] \tag{6}$$

where λ is a uniformity (evenness) parameter and with P_n^0 and P_n^1 in Equation (5) being special (extreme) cases [17]. This P_n^λ is simply the following weighted mean of P_n^0 and P_n^1 :

$$P_n^\lambda = \lambda P_n^1 + (1 - \lambda) P_n^0 \tag{7}$$

From Equations (4) and (6), it follows that, for any given P_n :

$$M(P_n) = M(P_n^\lambda) \text{ for one unique } \lambda \tag{8}$$

so that validity conditions on $M(P_n)$ can equivalently be determined in terms of $M(P_n^\lambda)$. With probability distributions considered as points (vectors) in n -dimensional space and with the Euclidean distance function d being the basis for a validity criterion, the following requirement seems most natural and obvious [17]:

$$\frac{M(P_n^1) - M(P_n^\lambda)}{M(P_n^1) - M(P_n^0)} = \frac{d(P_n^\lambda, P_n^1)}{d(P_n^0, P_n^1)} = 1 - \lambda \tag{9}$$

Since $M(P_n^0) = 0$, *i.e.*, there is no randomness when one $p_i = 1$, it follows from Equation (9) that:

$$M(P_n^\lambda) = \lambda M(P_n^1) \tag{10}$$

as a value-validity condition. This condition also follows immediately from Equation (7) as:

$$M(P_n^\lambda) = M[\lambda P_n^1 + (1 - \lambda) P_n^0] = \lambda M(P_n^1) + (1 - \lambda) M(P_n^0) \tag{11}$$

which equals Equation (10) for $M(P_n^0) = 0$.

For the case when $\lambda = 0.5$, $P_n^{0.5}$ is the midpoint of P_n^0 and P_n^1 with coordinates $(1/2 + 1/2n, 1/2n, \dots, 1/2n)$. Then:

$$M(P_n^{0.5}) = M\left(\frac{P_n^0 + P_n^1}{2}\right) = \frac{M(P_n^0) + M(P_n^1)}{2} \tag{12}$$

$$= \frac{M(P_n^1)}{2} \text{ for } M(P_n^0) = 0 \tag{13}$$

which is exactly as stated in Equations (10) and (11) with $\lambda = 0.5$. Of course, Equations (12) and (13) represent a weaker value-validity condition than Equations (10) and (11). Note also that it is not assumed *a priori* that M is a linear function of λ . This linearity is a consequence of Equations (7)–(9).

The entropy H in Equation (1) with $H(P_n^0) = 0$ and $H(P_n^1) = \log n$ does not meet these validity conditions. For example, for $n = 2$ and $\lambda = 0.5$, $H(P_2^{0.5}) = H(0.75, 0.25) = 0.56$, which far exceeds the requirement $\log 2/2 = 0.35$ in Equation (13). Similarly, $H(P_4^{0.5}) = H(0.625, 0.125, 0.125, 0.125) = 1.07 \gg \log 4/2 = 0.69$ and $H(P_{20}^{0.5}) = 2.09 \gg \log 20/2 = 1.50$. It can similarly be verified that $H(P_n^\lambda)$ for all n and $0 < \lambda < 1$, and hence $H(P_n)$ for all P_n from Equation (8), overstates the true or realistic extent of the randomness (uncertainty) that H is supposed to measure. Consequently, difference

comparisons as in Equations (2b) and (2c) based on H are invalid. An alternative measure that meets the validity conditions for such comparisons will be introduced next.

3. The New Entropy

3.1. Derivation of H_K

The logic or reasoning behind the H_K in Equation (3) as a measure of randomness or uncertainty may be outlined as follows:

- (1) As a matter of convenience and as used throughout the rest of the paper, all individual probabilities will be considered ordered such that:

$$p_1 \geq p_2 \geq \dots \geq p_n \tag{14}$$

- (2) Due to the constraint that $\sum_{i=1}^n p_i = 1$, there is no loss of generality or information by focusing on p_2, \dots, p_n .
- (3) Instead of considering $\sum_{i=2}^n f(p_i)$ or a weighted mean or sum of $f(p_2), \dots, f(p_n)$ for some function f of the individual p_i 's, one could consider the sum of the means of all pairs of the $p_i (i = 2, \dots, n)$. Since an entropy measure needs to be zero-indifferent (expansible), *i.e.*, unaffected by the addition of events with zero probabilities (e.g., [14] (Chapter 1)), a logical choice of pairwise means would be the geometric means $\sqrt{p_i p_j}$ for all $i, j = 2, \dots, n$ (since obviously $\sqrt{p_i 0} = 0$). Therefore, the measure consisting of the means $\sqrt{p_i p_j}$, including those for $i = j$, can be expressed as:

$$H_K = 2 \sum_{2 \leq i \leq j \leq n} \sum \sqrt{p_i p_j} = \sum_{i=2}^n \sum_{j=2}^n \sqrt{p_i p_j} + \sum_{i=2}^n p_i \tag{15}$$

where the multiplication factor 2 is included so that $H_K(P_n^1) = H(1/n, \dots, 1/n) = n - 1$ instead of $(n-1)/2$. With p_1 being the modal (largest) probability, this H_K in Equation (15) is twice the sum of the pairwise geometric means of all the non-modal probabilities. Furthermore, from the fact that, for a set of numbers $\{x_i\}$, $\left(\sum_{i=2}^n x_i\right)^2 = \sum_{i=2}^n \sum_{j=2}^n x_i x_j$ and then setting $x_i = \sqrt{p_i}$, it follows from the second expression in Equation (15) that:

$$H_K = \left(\sum_{i=2}^n \sqrt{p_i}\right)^2 + \sum_{i=2}^n p_i \tag{16}$$

which is the same as the formula in Equation (3).

As an alternative approach, one could begin by considering the power sum, or sum of order α , $S_\alpha = \left(\sum_i p_i^\alpha\right)^{1/\alpha}$ (e.g., [18] (pp. 138–139)). Strict Schur-concavity, which is discussed below as an important property of an entropy and one that H in Equation (1) has, requires that the parameter $\alpha < 1$ [18] (pp. 138–139). Since $p_i \geq 0 (i = 1, \dots, n)$, a further restriction is that α be positive and hence $0 < \alpha < 1$ for the power sum S_α . In order for S_α to comply with the value-validity condition in Equation (11), it is clear that S_α can only be the power sum of the non-modal probabilities so that:

$$S_\alpha(p_2, \dots, p_n) = \left(\sum_{i=2}^n p_i^\alpha\right)^{1/\alpha}, 0 < \alpha < 1 \tag{17}$$

with $S_\alpha(P_n^0) = 0$ and $S_\alpha(P_n^1) = (n - 1)^{1/\alpha}/n$ for the probability distributions in Equation (5). A reasonable upper bound would be $S_\alpha(p_n^1) = n - 1$. This requirement is met for $\alpha = 1/2$ in Equation (17) and by the addition of $\sum_{i=2}^n p_i$, resulting in:

$$S_{1/2}(p_2, \dots, p_n) + S_1(p_2, \dots, p_n) = \left(\sum_{i=2}^n \sqrt{p_i} \right)^2 + \sum_{i=2}^n p_i \tag{18}$$

which is the same as Equation (16) and for which $H_K(P_n^1) = n - 1$.

3.2. Properties of H_K

The properties of H_K in Equation (16), some of which are readily apparent from its definition, can be outlined as follows:

Property 1. H_K is a continuous function of all its individual arguments p_1, \dots, p_n .

Property 2. H_K is (permutation) symmetric in the $p_i (i = 1, \dots, n)$.

Property 3. H_K is zero-indifferent (expansible), *i.e.*,

$$H_K(p_1, \dots, p_n, 0, \dots, 0) = H_K(p_1, \dots, p_n)$$

Property 4. For any given $P_n = (p_1, \dots, p_n)$ and the P_n^0 and P_n^1 in Equation (5):

$$H_K(P_n^0) \leq H_K(P_n) \leq H_K(P_n^1); H_K(P_n^0) = 0, H_K(P_n^1) = n - 1 \tag{19}$$

Property 5. From Equation (19), $H_K(P_n^1)$ is strictly increasing in n .

Property 6. H_K is strictly Schur-concave so that, if $P_n = (p_1, \dots, p_n)$ is majorized by $Q_n = (q_1, \dots, q_n)$ (denoted by $<$):

$$P_n < Q_n \Rightarrow H_K(P_n) \geq H_K(Q_n) \tag{20}$$

with strict inequality unless P_n is simply a permutation of Q_n .

Property 7. H_K is concave, but not strictly concave.

Property 8. H_K meets the value-validity condition in Equation (10) with $H_K(P_n^\lambda) = \lambda(n - 1)$.

Proof of Property 6. The strict Schur-concavity of $H_K = \left(\sum_{i=2}^n \sqrt{p_i} \right)^2 + 1 - p_1$ follows immediately from the partial derivatives:

$$\frac{\partial H_K}{\partial p_1} = -1; \frac{\partial H_K}{\partial p_i} = p_i^{-1/2} \sum_{j=2}^n \sqrt{p_j}, \quad i = 2, \dots, n$$

and the fact that $\partial H_K / \partial p_i$ is strictly increasing in $i = 1, \dots, n$ (unless $p_i = p_{i+1}$) ([18] (p. 84)). The majorization $P_n < Q_n$ in Equation (20) is a more precise statement than the vague notion that the components of P_n are “more nearly equal” or “less spread out” than are those of Q_n . By definition, if $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i$ (and with the ordering in Equation (14) for P_n and Q_n):

$$P_n < Q_n \text{ if } \sum_{i=1}^j p_i \leq \sum_{i=1}^j q_i, \quad j = 1, \dots, n - 1$$

([18] (p. 8)). □

Proof of Property 7. From Equation (16) and for the probability distributions P_n and Q_n and all $\lambda \in [0, 1]$:

$$H_K[\lambda P_n + (1 - \lambda)Q_n] = \left[\sum_{i=2}^n (\lambda p_i + (1 - \lambda)q_i)^{1/2} \right]^2 + \lambda \sum_{i=2}^n p_i + (1 - \lambda) \sum_{i=1}^n q_i \tag{21}$$

From Minkowski's inequality (e.g., [19] (p. 175)):

$$\left[\sum_{i=2}^n (\lambda p_i + (1 - \lambda)q_i)^{1/2} \right]^2 \geq \left[\sum_{i=2}^n (\lambda p_i)^{1/2} \right]^2 + \left[\sum_{i=2}^n ((1 - \lambda)q_i)^{1/2} \right]^2 \tag{22}$$

so that, from Equations (21) and (22):

$$H_K[\lambda P_n + (1 - \lambda)Q_n] \geq \lambda \left[\left(\sum_{i=2}^n \sqrt{p_i} \right)^2 + \sum_{i=2}^n p_i \right] + (1 - \lambda) \left[\left(\sum_{i=2}^n \sqrt{q_i} \right)^2 + \sum_{i=2}^n q_i \right] = \lambda H_K(P_n) + (1 - \lambda)H_K(Q_n) \tag{23}$$

proving that H_K is concave. However, and importantly, H_K is not strictly concave since the inequality in Equation (23) is not strict for all P_n and Q_n such as for $P_n = P_n^1$ and $Q_n = P_n^0$ in Equation (5) when

$$H_K(P_n^\lambda) = H_K[\lambda P_n^1 + (1 - \lambda)P_n^0] = \lambda H_K(P_n^1) + (1 - \lambda)H_K(P_n^0) = \lambda(n - 1) \tag{24}$$

as required by the value-validity conditions in Equations (10) and (11). □

- *Note 1:* If a measure (function) M is strictly concave so that, instead of Equation (23), the inequality $M[\lambda P_n + (1 - \lambda)Q_n] > \lambda M(P_n) + (1 - \lambda)M(Q_n)$ is strict for all P_n, Q_n , and $\lambda \in (0, 1)$, then the condition in Equation (11) cannot be met. The H in Equation (1) is one such measure.
- *Note 2:* The extremal values $H_K(P_n^1) = H_K(1/n, \dots, 1/n) = n - 1$ for a measure of randomness or uncertainty is also a logical requirement for valid difference comparisons. As a particular case of the proportional difference comparisons in Equation (2c), and for any integer $m < n$:

$$H_K(P_{n+m}^1) - H_K(P_n^1) = H_K(P_n^1) - H_K(P_{n-m}^1) \tag{25}$$

i.e., adding an amount m to n results in the same absolute change in the value of H_K as does subtracting m from n in the equiprobable case. Or, in terms of the function f where $H_K(P_n^1) = f(n)$, Equation (25) can be expressed more conveniently as:

$$f(n + m) - f(n) = f(n) - f(n - m) \tag{26}$$

The general solution of the functional equation in Equation (26) is $f(n) = an + b$ with real constants a and b [20] (p. 82), which equals $H_K(P_n^1)$ for $a = 1$ and $b = -1$.

- *Note 3:* For the binary case of $n = 2$, $H_K(P_2^1) = H_K(0.5, 0.5) = 1$, which equals $H(0.5, 0.5)$ in Equation (1) if the base-2 logarithm is used. In fact, $H(0.5, 0.5) = 1$ is an axiom or required property, the *normalization axiom*, frequently used in information theory to justify the use of the base-2 logarithm in Equation (1) and *bits* as the unit of measurement [14] (Chapter 1). The *binary entropy* $H_K(1 - p, p) = 2p$ for $p \leq 0.5$ or:

$$H_K(1 - p, p) = 1 - |1 - 2p| \text{ for all } p \in [0, 1]$$

4. Generalization of H_K

Instead of the pairwise geometric means in Equation (15), one could consider power means, or arithmetic means of order α , and hence the following parameterized family of entropies:

$$H_{K\alpha} = \sum_{i=2}^n \sum_{j=2}^n M_\alpha(p_i, p_j) + \sum_{i=2}^n p_i; \quad M_\alpha(p_i, p_j) = \left(\frac{p_i^\alpha + p_j^\alpha}{2} \right)^{1/\alpha}, \quad -\infty < \alpha \leq 0 \quad (27)$$

of which H_K in Equations (15) and (16) is the particular member H_{K0} as $\alpha \rightarrow 0$. Since a measure of randomness (uncertainty) should be zero-indifferent (see Property 3 of H_K), it is clear from the formula in Equation (27) that α cannot be positive, i.e., $\alpha \leq 0$ where $\alpha = 0$ means the limit when $\alpha \rightarrow 0$. If $p_i = 0, p_j = 0$, or $p_i = p_j = 0$, $M_\alpha(p_i, p_j)$ is taken to be 0 for $\alpha \leq 0$ (see, e.g., [21] (Chapter 2) for the properties of such power means). One of the important properties of $M_\alpha(p_i, p_j)$ is that it is a non-decreasing function of α (for $-\infty < \alpha < \infty$) and is strictly increasing unless $p_i = p_j$. Besides this M_α , there are other types of means that could be considered (e.g., [18] (pp. 139–145), [22]).

Since $M_\alpha(p_i, p_j)$ is strictly increasing in α (if $p_i \neq p_j$), it follows from Equation (27) that, for any probability distribution $P_n = (p_1, \dots, p_n)$:

$$H_L(P_n) \leq H_{K\alpha}(P_n) \leq H_K(P_n) \text{ for all } \alpha \in (-\infty, 0) \quad (28)$$

where the lower limit $H_L(P_n) = H_{K(-\infty)}(P_n)$ is the limit of $H_{K\alpha}$ as $\alpha \rightarrow -\infty$ and $H_K(P_n)$ is defined in Equations (15) and (16) and is the limit of $H_{K\alpha}(P_n)$ as $\alpha \rightarrow 0$. The inequalities in Equation (28) are strict unless P_n equals P_n^0 or P_n^1 in Equation (5).

Each member of $H_{K\alpha}$ has the same types of properties as those of H_K discussed above. The strict Schur-concavity of $H_{K\alpha}$ follows from the fact that (a) $H_{K\alpha}$ is (permutation) symmetric in the p_i ($i = 1, \dots, n$) and (b) the partial derivatives, after setting $\sum_{i=2}^n p_i = 1 - p_1$ in Equation (27):

$$\frac{\partial H_{K\alpha}}{\partial p_1} = -1, \quad \frac{\partial H_{K\alpha}}{\partial p_i} = \sum_{j=2}^n \left(\frac{p_j^\alpha p_i^{-\alpha} + 1}{2} \right)^{(1-\alpha)/\alpha}, \quad i = 2, \dots, n$$

are clearly strictly increasing in $i = 1, \dots, n$ (for $p_i > p_{i+1}$) for all $\alpha \in (-\infty, 1)$. The case when $\alpha \rightarrow 0$ was proved in the preceding subsection.

As with any reasonable measure of randomness or uncertainty, each member of $H_{K\alpha}$ in Equation (27) is a compound measure consisting of two components: the dimension of the distribution or vector P_n and the uniformity (evenness) with which the elements of P_n are distributed. For any probability distribution $P_n = (p_1, \dots, p_n)$, this fact can be most simply represented by:

$$H_{K\alpha}(P_n) = H_{K\alpha}(P_n^1) H_{K\alpha}^*(P_n), \quad H_{K\alpha}^*(P_n) \in [0, 1] \quad (29)$$

where $H_{K\alpha}(P_n^1) = n - 1$ for the uniform distribution P_n^1 in Equation (5) and where $H_{K\alpha}^*(P_n) = H_{K\alpha}(P_n)/(n - 1)$ reflects the uniformity (evenness) of P_n . The $H_{K\alpha}^*$ basically controls for n . For the distribution in Equation (6), $H_{K\alpha}^*(P_n^\lambda) = \lambda$.

The limiting member of H_L in Equation (28) as $\alpha \rightarrow -\infty$ is defined by:

$$H_L = \sum_{i=2}^n \sum_{j=2}^n \min \{p_i, p_j\} + \sum_{i=2}^n p_i = \sum_{i=1}^n \sum_{j=1}^n \min \{p_i, p_j\} - 1 \quad (30a)$$

$$= 2 \sum_{i=2}^n (i - 1) p_i \quad (30b)$$

where the expression in Equation (30b) can easily be seen to follow directly from Equation (30a) (remembering again the order in Equation (14)). The second expression in Equation (30a) has been briefly mentioned by Morales *et al.* [23] and the form in Equation (30b), divided by 2, has been suggested by Patil and Taillie [24] as one potential measure of diversity.

5. Comparative Analysis

5.1. Why the Preference for H_K

From a practical point of view, what sets H_K in Equation (3) or Equation (16) apart from any other member of the family $H_{K\alpha}$ in Equation (27) is its ease of computation. Its values can easily be computed on a hand calculator for any probability distribution P_n even when the dimension n is quite large. For other members of $H_{K\alpha}$, $(n - 1)^2$ pairwise means have to be computed, which becomes practically impossible without the use of a computer program even when n is not large. The computational effort for the member H_L (when $\alpha \rightarrow -\infty$) is somewhat less than for other members. Nevertheless, the apparently simpler formula for H_L in Equation (30b) requires that all p_i be ordered as in Equation (14), which can be very tedious if done manually and nearly impossible if n is large.

The H_K is also favored over other members of $H_{K\alpha}$ when considering the agreement with some other measure based on Euclidean distance and the familiar standard deviation. Specifically, for any probability distribution $P_n = (p_1, \dots, p_n)$ and $P_n^1 = (1/n, \dots, 1/n)$, consider the following linearly decreasing function of the distance $d(P_n, P_n^1)$:

$$D(P_n) = (n - 1) \left[1 - \sqrt{\frac{n}{n - 1}} d(P_n, P_n^1) \right] = (n - 1) \left(1 - \frac{n}{\sqrt{n - 1}} s_n \right) = (n - 1) CNV \quad (31)$$

where s_n is the standard deviation of p_1, \dots, p_n (with divisor n rather than $n - 1$) and CNV is the *coefficient of nominal variation* [25,26]. It is clear from Equation (31) that, for P_n^0 and P_n^1 in Equation (5), $D(P_n^0) = 0$ and $D(P_n^1) = n - 1$. Also, for the lambda distribution in Equation (6), $D(P_n^\lambda) = \lambda D(P_n^1) = \lambda(n - 1)$, so that D satisfies the value-validity condition in Equation (10). Of course, D is not zero-indifferent (see Property 3 for H_K).

Since the Euclidean distance and the standard deviation are such universally used measures, it is to be expected that an acceptable measure of randomness (uncertainty) should not differ substantially from D in Equation (31). From numerical examples, it is seen that values of H_K in Equation (16) tend to be closer to those of D in Equation (31) than are the values of any other member of the $H_{K\alpha}$ family in Equation (27). In order to demonstrate this fact, a computer simulation was used to generate a number of random distributions using the following algorithm. For each randomly generated probability distribution $P_n = (p_1, \dots, p_n)$, n was first generated as a random integer between 3 and 20, inclusive. Then, with the ordering in Equation (14), each p_i was generated as a random number (to 5 decimal places) within the following intervals:

$$\frac{1}{n} \leq p_i \leq 1 - \frac{\sum_{j=1}^{i-1} p_j}{n - (i - 1)}, i = 2, \dots, n - 1$$

and finally

$$p_n = 1 - \sum_{i=1}^{n-1} p_i$$

For each such generated distribution, the values of H_K , H_L , and D were computed according to the formulas in Equations (16), (30b) and (31) as were their corresponding uniformity (evenness) indices from Equation (29). After excluding some (five) distributions P_n that were nearly equal to P_n^0 or P_n^1 in Equation (5), the results for 30 different distributions are summarized in Table 1.

Table 1. Values of H_K , H_L , and D in Equations (16), (30b), and (31) and of their corresponding uniformity (evenness) indices from Equation (29) based on 30 randomly generated probability distributions.

| Data Set | n | H_K | H_L | D | H_K^* | H_L^* | D^* |
|----------|-----|-------|-------|-------|---------|---------|-------|
| 1 | 16 | 2.17 | 1.28 | 3.89 | 0.14 | 0.09 | 0.26 |
| 2 | 19 | 11.14 | 10.32 | 11.22 | 0.62 | 0.57 | 0.62 |
| 3 | 3 | 0.96 | 0.82 | 0.97 | 0.48 | 0.41 | 0.49 |
| 4 | 18 | 4.28 | 3.21 | 5.05 | 0.25 | 0.19 | 0.15 |
| 5 | 15 | 1.84 | 1.40 | 2.08 | 0.13 | 0.10 | 0.15 |
| 6 | 15 | 12.65 | 11.42 | 12.57 | 0.90 | 0.82 | 0.90 |
| 7 | 13 | 10.37 | 9.93 | 10.34 | 0.86 | 0.83 | 0.86 |
| 8 | 15 | 4.13 | 3.01 | 3.76 | 0.30 | 0.22 | 0.27 |
| 9 | 7 | 5.38 | 4.83 | 5.25 | 0.90 | 0.81 | 0.88 |
| 10 | 4 | 0.78 | 0.65 | 0.86 | 0.26 | 0.22 | 0.29 |
| 11 | 12 | 1.67 | 1.23 | 3.83 | 0.15 | 0.11 | 0.35 |
| 12 | 17 | 14.71 | 14.56 | 14.71 | 0.92 | 0.91 | 0.92 |
| 13 | 14 | 4.09 | 3.88 | 4.10 | 0.31 | 0.30 | 0.32 |
| 14 | 8 | 6.04 | 5.39 | 5.95 | 0.86 | 0.77 | 0.85 |
| 15 | 17 | 10.01 | 9.03 | 10.13 | 0.63 | 0.56 | 0.63 |
| 16 | 5 | 0.57 | 0.41 | 0.74 | 0.14 | 0.10 | 0.19 |
| 17 | 10 | 4.85 | 3.82 | 5.10 | 0.54 | 0.42 | 0.57 |
| 18 | 5 | 1.00 | 0.73 | 1.27 | 0.25 | 0.18 | 0.32 |
| 19 | 5 | 2.05 | 1.16 | 2.09 | 0.51 | 0.40 | 0.52 |
| 20 | 19 | 12.00 | 11.05 | 12.06 | 0.67 | 0.61 | 0.67 |
| 21 | 19 | 13.78 | 13.64 | 13.78 | 0.77 | 0.76 | 0.77 |
| 22 | 17 | 6.93 | 6.33 | 7.40 | 0.43 | 0.40 | 0.46 |
| 23 | 12 | 8.16 | 7.71 | 8.17 | 0.74 | 0.70 | 0.74 |
| 24 | 20 | 13.56 | 11.98 | 13.83 | 0.71 | 0.63 | 0.73 |
| 25 | 17 | 3.47 | 3.07 | 3.58 | 0.22 | 0.19 | 0.22 |
| 26 | 13 | 1.15 | 0.72 | 2.29 | 0.10 | 0.06 | 0.19 |
| 27 | 14 | 10.69 | 9.66 | 10.65 | 0.82 | 0.74 | 0.82 |
| 28 | 12 | 1.68 | 1.66 | 1.68 | 0.15 | 0.15 | 0.15 |
| 29 | 20 | 9.72 | 6.88 | 11.09 | 0.51 | 0.36 | 0.58 |
| 30 | 9 | 5.23 | 4.20 | 5.36 | 0.65 | 0.53 | 0.67 |

It is apparent from the data in Table 1 that H_K agrees quite closely with D and clearly more so than does H_L . Exceptions are Data Sets 1, 11, and 26 when the H_K -values differ considerably from those of D , but still less so than do the H_L -values. If D is used to predict H_K (i.e., for the fitted model $\hat{H}_K = D$), it is found for the 30 data sets in Table 1 that the coefficient of determination, when properly computed as $R^2 = 1 - \sum (H_K - D)^2 / \sum (H_K - \bar{H}_K)^2$ [27], becomes $R^2 = 0.98$ (i.e., 98% of the variation of H_K is explained by the fitted model $\hat{H}_K = D$) as compared to $R^2 = 0.91$ in the case of H_L . Also, the root mean square (RMS) of the differences between the values of H_K and D is found to be 0.64 as compared to 1.33 for H_L and D . Similarly, when comparing the values of the indices H_K^* , H_L^* , and D^* , the H_K^* -values are considerably closer to the D^* -values than are the H_L^* -values, with $RMS(H_K^*, D^*) = 0.05$ and $RMS(H_L^*, D^*) = 0.10$.

No other member of $H_{K\alpha}$ in Equation (27) is generally in as close agreement with D as is H_K , but more so than H_L . This can be explained by the fact that (a) whenever there is a notable difference between the values of H_K and D , those of H_K tend to be less than those of D as seen from Table 1; and (b) $H_{K\alpha}(P_n)$ is a strictly increasing function of α for any given P_n other than P_n^0 and P_n^1 in Equation (5).

5.2. Comparative Weights on p_1, \dots, p_n

The difference between values of H_K and H_L as demonstrated in Table 1, or between any of the members of the $H_{K\alpha}$ family, is due to the fact that $H_{K\alpha}$ places different weights or emphases on the p_i ($i = 1, \dots, n$) depending upon α . When considering each pairwise mean $M_\alpha(p_i, p_j)$ in Equation (27), p_i and p_j are weighted equally only when $\alpha = 1$. Then, since (a) $M_\alpha(p_i, p_j)$ is strictly increasing in

α (for $p_i \neq p_j$) and (b) $H_{K\alpha}$ is zero-indifferent (Property 3 of H_K) only for $\alpha \leq 0$, the $H_{K0} = H_K$ in Equations (15) and (16) is the zero-indifferent member of $H_{K\alpha}$ that is always closest in value to H_{K1} and whose pairwise means $M_0(p_i, p_j)$ are always closest to $M_1(p_i, p_j)$ for all i and j .

Besides the weights placed on each component of all pairs (p_i, p_j) , the weights given to each individual p_i ($i = 2, \dots, n$) can also be examined by expressing the $H_{K\alpha}$ in Equation (27) as the following weighted sum:

$$H_{K\alpha} = \sum_{i=2}^n \left[\sum_{j=2}^n \left(\frac{p_i^\alpha + p_j^\alpha}{2} \right)^{1/\alpha} \left(\frac{1}{p_i} \right) + 1 \right] p_i = \sum_{i=2}^n w_{\alpha i} p_i \tag{32}$$

which shows that the weights $w_{\alpha i}$ ($i = 2, \dots, n$) are increasing in both α and i . In the case of H_K , with $\alpha \rightarrow 0$ in Equation (32), $w_{0i} = \sum_{j=2}^n \sqrt{p_j/p_i} + 1$ for $i = 2, \dots, n$ whereas, for H_L when $\alpha \rightarrow -\infty$, $w_{-\infty i} = \sum_{j=2}^n \min\{p_i, p_j\} p_i^{-1} + 1$ for $i = 2, \dots, n$. These weights for H_K are basically a compromise between the weights for H_L and those for H_{K1} . Note also that these weights for H_L from Equation (32) can differ substantially from those in Equation (30b) as can the weights for H_{K1} from Equation (32) when compared with the weights from the expression $H_{K1} = n \sum_{i=2}^n p_i$.

When comparing H_K and H_L , small p_i 's are given more weight by H_K than by H_L and the addition of low probability components to a set of events has more effect on H_K than on H_L . However, when weighting the pros and cons of such relative sensitivity to small p_i 's, it is important to keep in mind the relationship in Equation (29) and not jump to conclusion. For example, when going from $P_4 = (0.40, 0.30, 0.20, 0.10)$ to $Q_6 = (0.40, 0.30, 0.20, 0.05, 0.04, 0.01)$, H_K increases from $H_K(P_4) = 2.32$ to $H_K(Q_6) = 2.91$, a 25% increase, while $H_L(P_4) = 2.00$ and $H_L(Q_6) = 2.12$, a 6% increase. However, from Equation (29), the dimensional component of both H_K and H_L increased by 67% (from $n - 1 = 3$ to $n - 1 = 5$) whereas the uniformity (evenness) components decreased by 25% in the case of H_K (from $2.32/3$ to $2.91/5$) and 37% for H_L (from $2.00/3$ to $2.12/5$). In this regard the 25% increase in randomness (uncertainty) as measured by H_K does not appear unreasonable.

5.3. Inconsistent Orderings

Although all members of the family $H_{K\alpha}$ in Equation (27) have the same types of properties, including the value-validity property in Equation (10), this does not necessarily imply that different members will always produce the same results for the comparisons in Equation (2). Such lack of consistency is inevitable whenever measures are used to summarize data sets into a single number. However, as stated by Patil and Taillie [24] (p. 551), "Inconsistent measures . . . are a familiar problem and should not be a cause for undue pessimism", pointing out the fact that, for instance, the arithmetic mean and the median are not consistent measures of average (central tendency) and the standard deviation and the mean absolute deviation are inconsistent measures of variability (spread). One type of consistent results for all members of $H_{K\alpha}$ is the size (order) comparison $H_{K\alpha}(P_n) > H_{K\alpha}(Q_n)$ in Equation (2a) whenever P_n is majorized by Q_n and P_n is not a permutation of Q_n . This is the result of Equation (20) and the fact, as proved above, that $H_{K\alpha}$ is strictly Schur-concave for all $\alpha \in (-\infty, 1)$.

It is only when two measures M_1 and M_2 have a perfect linear relationship that (a) the comparison results from Equation (2) will always be consistent and (b) the compliance by M_1 with the value-validity conditions in Equations (10) and (11) also implies compliance by M_2 . In the case of H_K and H_L , and from the simulation results in Table 1, Pearson's correlation coefficient between H_K and H_L is found to be $r = 0.993$, indicating a near perfect linear relationship between H_K and H_L . However, since the linearity is not truly perfect or exact, H_K and H_L will not always give the same results for the comparisons in Equation (2) as is evident from some of the data in Table 1.

6. Discussion

The value-validity condition in Equation (10) as a necessary requirement for valid difference comparisons as in Equation (2) is based on Euclidean distances. Such distances are also being used as a basis for the preference of H_K over other potential members of the family of entropies in Equation (27). This distance metric is the standard one in engineering and science. The use of any other “distance” measures, such as *directed divergencies* discussed below, would seem to require particular justification in the context of value-validity assessment.

As a simple numerical example illustrating the reasoning behind the value-validity arguments in Equations (6)–(13) and the use of Euclidean distances, consider the following probability distributions based on P_n^λ in Equation (6):

$$\begin{aligned}
 P_5^0 &= \{p_i^{(0)}\} = (1, 0, 0, 0, 0) \\
 P_5^{0.5} &= \{p_i^{(0.5)}\} = (0.6, 0.1, 0.1, 0.1, 0.1) \\
 P_5^1 &= \{p_i^{(1)}\} = (0.2, 0.2, 0.2, 0.2, 0.2)
 \end{aligned}$$

The Euclidean distances $d(P_5^{0.5}, P_5^0) = d(P_5^{0.5}, P_5^1)$ and $|p_i^{(0.5)} - p_i^{(0)}| = |p_i^{(0.5)} - p_i^{(1)}|$ for $i = 1, \dots, 5$. A measure of uncertainty (randomness) M that takes on reasonable numerical values within the general bounds $M(P_n^0)$ and $M(P_n^1)$ should in this example satisfy the equality $|M(P_5^{0.5}) - M(P_5^0)| = |M(P_5^{0.5}) - M(P_5^1)|$ so that, with $M(P_5^0) = 0$, $M(P_5^{0.5}) = M(P_5^1)/2$. That is, since $P_5^{0.5}$ is the same distance from P_5^0 as it is from P_5^1 and each element of $P_5^{0.5}$ is the same distance from the corresponding element of P_n^0 as it is from that of P_5^1 , M would reflect this fact by taking on the value $M(P_5^{0.5}) = M(P_5^1)/2$. The H_K in Equation (3) or Equation (16) meets this requirement with $H_K(P_5^{0.5}) = H_K(P_5^1)/2 = (5 - 1)/2 = 2$. However, in the case of H in Equation (1), $H(P_5^{0.5}) = 1.23 \gg H(P_5^1)/2 = 0.80$, a substantial overstatement of the extent of the uncertainty or randomness.

A similar comparison between H_K and H for P_5^λ with $\lambda = 0.25$ and $\lambda = 0.75$ is given in Table 2 together with the results from some other probability distributions. The results are also given in terms of the normalized measures $H_K^*(P_n) = H_K(P_n)/(n - 1)$ and $H^*(P_n) = H(P_n)/\log n$ as well as $D^*(P_n) = D(P_n)/(n - 1)$ for D in Equation (31). As seen from Table 2, while $H_K^*(P_5^\lambda) = D^*(P_5^\lambda) = \lambda$, $H^*(P_5^\lambda) \gg \lambda$ for both λ -values. For all distributions in Table 2, the values of H_K^* are quite comparable to those of D^* , but those of H^* are all considerably greater.

Table 2. Comparative results for H in Equation (1) and H_K in Equation (16) and their normalized forms as well as D^* from Equation (31) for various probability distributions.

| $P_n^{(i)}$ | H_K | H_K^* | H | H^* | D^* |
|---|-------|---------|------|-------|-------|
| $P_5^{(1)} = P_5^{0.75} = (0.40, 0.15, 0.15, 0.15, 0.15)$ | 3.00 | 0.75 | 1.50 | 0.93 | 0.75 |
| $P_5^{(2)} = P_5^{0.25} = (0.80, 0.05, 0.05, 0.05, 0.05)$ | 1.00 | 0.25 | 0.78 | 0.48 | 0.25 |
| $P_{10}^{(3)} = (0.26, 0.15, 0.15, 0.10, \dots, 0.03)$ | 6.87 | 0.76 | 2.09 | 0.91 | 0.77 |
| $P_{14}^{(4)} = (0.27, 0.25, 0.15, 0.07, \dots, 0.01)$ | 8.22 | 0.63 | 2.10 | 0.80 | 0.67 |
| $P_4^{(5)} = (0.40, 0.30, 0.20, 0.10)$ | 2.32 | 0.77 | 1.28 | 0.92 | 0.74 |
| $P_5^{(6)} = (0.50, 0.20, 0.10, 0.10, 0.10)$ | 2.45 | 0.61 | 1.36 | 0.84 | 0.61 |
| $P_5^{(7)} = (0.70, 0.10, 0.10, 0.05, 0.05)$ | 1.47 | 0.37 | 1.01 | 0.63 | 0.37 |
| $P_5^{(8)} = (0.30, 0.20, 0.20, 0.20, 0.10)$ | 3.45 | 0.86 | 1.56 | 0.97 | 0.84 |

The distributions $P_{10}^{(3)} - P_5^{(8)}$ are included in Table 2 to exemplify the types of contradictory results that may be obtained when making the difference comparisons in Equation (2) based on H_K versus H . The distributions $P_{10}^{(3)}$ and $P_{14}^{(4)}$ in Table 2 are real data for the market shares (proportions) of the carbonated soft drinks industry in the U.S. and the world-wide market shares of cell phones,

respectively (obtained by Googling “market shares” by industries). Some of the smaller market shares are not given in Table 2 because of space limitation, but were included in the computations. The H , which has been used as a measure of market concentration or rather of its converse, deconcentration (e.g., [28]), would indicate that these two industries have nearly the same market deconcentration. By contrast, when considered in terms of H_K for which such comparison is valid because of the value-validity property of H_K , the results in Table 2 show that the cell-phone industry is about 20% more deconcentrated than the soft-drink industry. Similarly, for the fictitious distributions $P_4^{(5)}-P_5^{(8)}$ in Table 2, the type of difference comparison in Equation (2b) shows that $H_K(P_5^{(8)}) - H_K(P_4^{(5)}) = 1.13 > H_K(P_5^{(6)}) - H_K(P_5^{(7)}) = 0.98$ whereas the result would have been the reverse had H been used for this comparison, with $H(P_5^{(8)}) - H(P_4^{(5)}) = 0.28 < H(P_5^{(6)}) - H(P_5^{(7)}) = 0.35$.

Instead of using the Euclidean metric to formulate the value-validity conditions in Section 2, one could perhaps consider other potential “distance” measures such as *divergencies*, also referred to as “statistical distances”. The best known such measure of the divergence of the distribution $P_n = (p_1, \dots, p_n)$ from the distribution $Q_n = (q_1, \dots, q_n)$ is the Kullback-Leibler divergence [29] defined as:

$$KLD(P_n : Q_n) = \sum_{i=1}^n p_i \log \left(\frac{p_i}{q_i} \right)$$

This measure is directional or asymmetric in P_n and Q_n . A symmetric measure is the so-called Jensen–Shannon divergence (*JSD*) (e.g., [30–33]), which can be expressed in terms of the Kullback–Leibler divergence (*KLD*) as:

$$JSD(P_n, Q_n) = \frac{1}{2}KLD(P_n : M_n) + \frac{1}{2}KLD(Q_n : M_n)$$

where $M_n = (P_n + Q_n)/2$. Neither *KLD* nor *JSD* are metrics, but \sqrt{JSD} is [34].

Consider now the family of distributions P_n^λ in Equation (6) and the extreme members of P_n^0 and P_n^1 in Equation (5). For the case of $n = 5$, for example, it is found that $KLD(P_5^0 : P_5^{0.5}) = 0.51$ ($KLD(P_5^{0.5} : P_5^0)$ is undefined) and $KLD(P_5^1 : P_5^{0.5}) = 0.33$. In the case of *JSD*, $M_5(P_5^{0.5}, P_5^0) = (0.8, 0.05, \dots, 0.05)$ so that $JSD(P_5^{0.5}, P_5^0) = 0.16$. Similarly, $M_5(P_5^{0.5}, P_5^1) = (0.40, 0.15, \dots, 0.15)$ and $JSD(P_5^{0.5}, P_5^1) = 0.09$. These results differ greatly from those based on Euclidean distances for which $d(P_5^{0.5}, P_5^0) = d(P_5^{0.5}, P_5^1)$. The fact that $d(P_n^{0.5}, P_n^0) = d(P_n^{0.5}, P_n^1)$ for all n , which is also reflected by the normalized $H_K^*(P_n^{0.5}) = 0.5$, corresponds to the fact that each component of $P_n^{0.5}$ is of equal distance from the corresponding components of P_n^0 and P_n^1 . However, no such correspondence exists for the divergence measures *KLD* and *JSD*.

The derivation of H_K in Equation (3) or Equation (16) is based on the exclusion of the modal probability p_1 . Of course, p_1 can enter the expression for H_K since $\sum_{i=2}^n p_i = 1 - p_1$. One may wonder what the result would be if a different p_i were to be excluded. If the smallest p_i is excluded, the measure would not be zero-indifferent (expansible). If any p_i other than p_1 is excluded, then the measure would not be strictly Schur-concave as can be verified from the proof of Property 6 of H_K . This property is essential for any measure of uncertainty (randomness). In fact, the exclusion of p_1 makes H_K unique in this regard.

It should also be emphasized that even though the entropy H in Equation (1) lacks the value-validity property, it has many of the same properties as H_K and has undoubtedly numerous useful and appropriate applications as demonstrated in the extensive published literature. The problems with H arise when it is used uncritically and indiscriminately in fields far from its origin: a statistical concept of communication theory. Both Shannon [35] and Wiener [36] cautioned against such uncritical applications. It is when H or its normalized form H^* is used as a summary measure (statistic) of various attributes (characteristics) and when its values are interpreted and compared that its lack of the value-validity property can lead to incorrect and misleading results and conclusions.

This is the motivation for introducing the new entropy H_K as a measure that overcomes the lack of value validity by H .

7. Statistical Inference about H_K

Consider now the case when each $p_i = n_i/N$ is the multinomial sample estimate of the unknown population probability π_i for $i = 1, \dots, n$ and based on sample size $N = \sum_{i=1}^n n_i$. It may then be of interest to investigate the potential statistical bias of $H_K(P_n)$ and to construct confidence intervals for the population measure $H_K(\Pi_n)$ for the population probability distribution $\Pi_n = (\pi_1, \dots, \pi_n)$.

7.1. Bias

Using bold letters to distinguish random variables from their sample values and expanding $H_K(P_n)$ in Equation (16) into a Taylor series about $p_i = \pi_i$ for $i = 2, \dots, n$, the following result is obtained:

$$\begin{aligned}
 H_K(P_n) = & H_K(\Pi_n) + \sum_{i=2}^n \left(\frac{1}{\sqrt{\pi_i}} \sum_{i=2}^n \sqrt{\pi_i} + 1 \right) (p_i - \pi_i) \\
 & + \frac{1}{4} \sum_{i=2}^n \left(-\pi_i^{-3/2} \sum_{i=2}^n \sqrt{\pi_i} + \pi_i^{-1} \right) (p_i - \pi_i)^2 \\
 & + \frac{1}{4} \sum_{i=2}^n \sum_{\substack{j=2 \\ i \neq j}}^n \left(\frac{1}{\sqrt{\pi_i \pi_j}} \right) (p_i - \pi_i) (p_j - \pi_j) + \dots
 \end{aligned} \tag{33}$$

Taking the expected value of each side of Equation (33) and using the well-known expectations $E(p_i - \pi_i) = 0$, $E(p_i - \pi_i)^2 = \pi_i(1 - \pi_i)/N$, and $E[(p_i - \pi_i)(p_j - \pi_j)] = -\pi_i \pi_j / N$ for $i, j = 2, \dots, n$, it is found that:

$$E[H_K(P_n)] = H_K(\Pi_n) + \frac{1}{4N} \left[- \left(\sum_{i=2}^n \sqrt{\pi_i} \right) \sum_{i=2}^n \frac{1}{\sqrt{\pi_i}} + n - 1 \right] + O(N^{-3/2}) \tag{34}$$

Equation (34) shows that the estimator $H_K(P_n)$, while asymptotically unbiased, does have a small bias for finite sample size N . However, unless N is small, this bias can effectively be ignored for all practical purposes.

7.2. Confidence Interval Construction

Under multinomial sampling and based on the *delta method* of large sample theory (e.g., [37] (Chapter 14)), the following convergence to the normal distribution holds:

$$\sqrt{N}[H_K(P_n) - H_K(\Pi_n)] \xrightarrow{d} Normal(0, \sigma^2) \tag{35}$$

where, in terms of partial derivatives:

$$\sigma^2 = \sum_{i=1}^n \pi_i \left(\frac{\partial H_K(\Pi_n)}{\partial \pi_i} \right)^2 - \left[\sum_{i=1}^n \pi_i \left(\frac{\partial H_K(\Pi_n)}{\partial \pi_i} \right) \right]^2 \tag{36}$$

That is, for large N , $H_K(P_n)$ is approximately normally distributed with mean $H_K(\Pi_n)$ and variance $\text{Var}[H_K(P_n)] = \sigma^2/N$ or standard error $SE = \sigma/\sqrt{N}$. The limiting normal distribution in Equation (35) also holds when, as is necessary in practice, the σ^2 is replaced with the estimated variance $\hat{\sigma}^2$ by substituting the sample estimates (proportions) p_i for the population probabilities π_i .

for $i = 1, \dots, n$, resulting in the estimated standard error $\hat{SE} = \hat{\sigma}/\sqrt{N}$. It then readily follows from Equations (16) and (36) that:

$$\hat{SE} = N^{-1/2} \left\{ n \left[H_K(P_n) - \sum_{i=2}^n p_i \right] + H_K(P_n) [1 - H_K(P_n)] \right\}^{1/2} \quad (37)$$

As a simple numerical example, consider the sample distribution $P_4 = (0.4, 0.3, 0.25, 0.05)$ based on sample size $N = 100$. From Equation (16), $H_K(P_4) = 2.22$ and, from Equation (37), $\hat{SE} = 0.19$. Therefore, because of Equation (35), an approximate 95% confidence interval for the population measure $H_K(\Pi_n)$ becomes $2.22 \pm 1.96(0.19)$, or $[1.85, 2.59]$. Statistical hypotheses such as $H_0 : H_K(\Pi_n) = H_K(\Pi_m)$ versus $H_1 : H_K(\Pi_n) > H_K(\Pi_m)$ can also be tested based on Equations (35) and (37).

8. Conclusions

Since the ubiquitous Boltzmann–Shannon entropy H is only valid for making size (order or “larger than”) comparisons, the entropy H_K is being introduced as an alternative measure of randomness (uncertainty) that is more informative than H in the sense that H_K can also be used for making valid difference comparisons as in Equations (2b) and (2c). The H_K , which is a particular member of the family of entropies $H_{K\alpha}$ and is basically a compromise between the members H_{K1} and $H_L = H_{K(-\infty)}$, has the types of desirable properties one would reasonably expect of a randomness (uncertainty) measure. One of the differences between H_K and H_L is that small probabilities have a greater influence on H_K than on H_L . The addition of some small probability events causes a larger increase in H_K than in H_L , but causes a smaller decrease in the uniformity (evenness) index H_K^* than in H_L^* as defined in Equation (29).

Besides being computationally most simple, which is certainly a practical advantage, H_K is also that member of $H_{K\alpha}$ that appears to be most nearly linearly (and decreasingly) related to the Euclidean distance between the points $P_n = (p_1, \dots, p_n)$ and $P_n^1 = (1/n, \dots, 1/n)$ or to the standard deviation s_n of p_1, \dots, p_n . The s_n is the usual measure of variability (spread) for a set of data, although it is not resistant against “outliers” (extreme and suspect data points). However, “outliers” are not a concern when dealing with probabilities p_i ($i = 1, \dots, n$). Therefore, s_n cannot justifiably be criticized for being excessively influenced by large or small p_i 's, with the same argument extending to H_K .

Acknowledgments: The author would like to thank the reviewers for constructive and helpful comments.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Von Boltzmann, L. Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. In *Sitzungsberichte der Kaiserliche Akademie der Wissenschaften, II Abteil*; (Vol. 66, Pt. 2); K.-K. Hof- und Staatsdruckerei in Commission bei C. Gerold's Sohn: Wien, Austria, 1872; pp. 275–370. (In German)
2. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [[CrossRef](#)]
3. Klir, G.J. *Uncertainty and Information: Foundations of a Generalized Information Theory*; Wiley: Hoboken, NJ, USA, 2006.
4. Ruelle, D. *Chance and Chaos*; Princeton University Press: Princeton, NJ, USA, 1991.
5. Han, T.S.; Kobayashi, K. *Mathematics of Information and Coding*; American Mathematical Society: Providence, RI, USA, 2002.
6. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, USA, 1949.
7. Arndt, C. *Information Measures: Information and Its Description in Science and Engineering*; Springer: Berlin/Heidelberg, Germany, 2004.

8. Kapur, J.N. *Measures of Information and Their Applications*; Wiley: New Delhi, India, 1994.
9. Kvålseth, T.O. Entropy. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; Part 5; pp. 436–439.
10. Rényi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, 20 June–30 July 1961; University of California Press: Berkeley, CA, USA, 1961; Volume 1, pp. 547–561.
11. Peitgen, H.-O.; Jürgens, H.; Saupe, D. *Chaos and Fractals: New Frontiers of Science*, 2nd ed.; Springer: New York, NY, USA, 2004.
12. Tsallis, C. Possible generalization of Boltzmann–Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. [[CrossRef](#)]
13. Norwich, K.H. *Information, Sensation, and Perception*; Academic Press: San Diego, CA, USA, 1993.
14. Aczél, J.; Daróczy, Z. *On Measures of Information and Their Characterizations*; Academic Press: New York, NY, USA, 1975.
15. Kvålseth, T.O. Entropy evaluation based on value validity. *Entropy* **2014**, *16*, 4855–4873. [[CrossRef](#)]
16. Hand, D.J. *Measurement Theory and Practice*; Wiley: London, UK, 2004.
17. Kvålseth, T.O. The Lambda distribution and its applications to categorical summary measures. *Adv. Appl. Stat.* **2011**, *24*, 83–106.
18. Marshall, A.W.; Olkin, I.; Arnold, B.C. *Inequalities: Theory of Majorization and Its Applications*, 2nd ed.; Springer: New York, NY, USA, 2011.
19. Bullen, P.S. *A Dictionary of Inequalities*; Addison Wesley Longman: Essex, UK, 1998.
20. Aczél, J. *Lectures on Functional Equations and Their Applications*; Academic Press: New York, NY, USA, 1966.
21. Hardy, G.H.; Littlewood, J.E.; Pólya, G. *Inequalities*; Cambridge University Press: Cambridge, UK, 1934.
22. Ebanks, B. Looking for a few good means. *Am. Math. Mon.* **2012**, *119*, 658–669.
23. Morales, D.; Pardo, L.; Vajda, I. Uncertainty of discrete stochastic systems: General theory and statistical inference. *IEEE Trans. Syst. Man Cyber Part A* **1996**, *26*, 681–697. [[CrossRef](#)]
24. Patil, G.P.; Taillie, C. Diversity as a concept and its measurement. *J. Am. Stat. Assoc.* **1982**, *77*, 548–567. [[CrossRef](#)]
25. Kvålseth, T.O. Coefficients of variation for nominal and ordinal categorical data. *Percept. Mot. Skills* **1995**, *80*, 843–847. [[CrossRef](#)]
26. Kvålseth, T.O. Variation for categorical variables. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; Part 22; pp. 1642–1645.
27. Kvålseth, T.O. Cautionary note about R^2 . *Am. Stat.* **1985**, *39*, 279–285.
28. Nawrocki, D.; Carter, W. Industry competitiveness using Herfindahl and entropy concentration indices with firm market capitalization data. *Appl. Econ.* **2010**, *42*, 2855–2863. [[CrossRef](#)]
29. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
30. Lin, J. Divergence measures based on Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [[CrossRef](#)]
31. Wong, A.K.C.; You, M. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1985**, PAMI-7, 599–609. [[CrossRef](#)]
32. Sagar, R.P.; Laguna, H.G.; Guevara, N.L. Electron pair density information measures in atomic systems. *Int. J. Quantum Chem.* **2011**, *111*, 3497–3504. [[CrossRef](#)]
33. Antolin, J.; Angulo, J.C.; López-Rosa, S. Fisher and Jensen–Shannon divergences: Quantitative comparisons among distributions. Application to position and momentum atomic densities. *J. Chem. Phys.* **2009**, *130*, 074110. [[CrossRef](#)] [[PubMed](#)]
34. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [[CrossRef](#)]
35. Shannon, C.E. The bandwagon. *IRE Trans. Inf. Theory* **1956**, *2*, 3. [[CrossRef](#)]
36. Weiner, N. What is information theory? *IRE Trans. Inf. Theory* **1956**, *2*, 48. [[CrossRef](#)]
37. Bishop, Y.M.M.; Fienberg, S.E.; Holland, P.W. *Discrete Multivariate Analysis: Theory and Practice*; MIT Press: Cambridge, MA, USA, 1975.

