entropy

MDPI

Article

Several Basic Elements of Entropic Statistics

Zhiyi Zhang

Department of Mathematics and Statistics, UNC Charlotte, Charlotte, NC 28223, USA; zzhang@charlotte.edu

Abstract: Inspired by the development in modern data science, a shift is increasingly visible in the foundation of statistical inference, away from a real space, where random variables reside, toward a nonmetrized and nonordinal alphabet, where more general random elements reside. While statistical inferences based on random variables are theoretically well supported in the rich literature of probability and statistics, inferences on alphabets, mostly by way of various entropies and their estimation, are less systematically supported in theory. Without the familiar notions of neighborhood, real or complex moments, tails, et cetera, associated with random variables, probability and statistics based on random elements on alphabets need more attention to foster a sound framework for rigorous development of entropy-based statistical exercises. In this article, several basic elements of entropic statistics are introduced and discussed, including notions of general entropies, entropic sample spaces, entropic distributions, entropic statistics, entropic multinomial distributions, entropic moments, and entropic basis, among other entropic objects. In particular, an entropic-moment-generating function is defined and it is shown to uniquely characterize the underlying distribution in entropic perspective, and, hence, all entropies. An entropic version of the Glivenko–Cantelli convergence theorem is also established.

Keywords: entropies; entropy estimation; entropic-moment-generating function; entropic statistics

1. Introduction and Summary

Let $\mathscr{X} = \{\ell_k; k \geq 1\}$ be a countable alphabet and let $\mathbf{p} = \{p_k; k \geq 1\}$ be a probability distribution on $\mathscr{X}$. Let $\mathscr{P}$ be the collection of all probability distributions on $\mathscr{X}$. Let $\mathbf{p}_\downarrow = \{p_{(k)}; k \geq 1\}$ be the nonincreasingly rearranged $\mathbf{p}$, that is, $p_{(k)} \geq p_{(k+1)}$ for every $k \geq 1$. Let $\mathscr{P}_\downarrow$ be the collection of all possible $\mathbf{p}_\downarrow$. It follows that $\mathscr{P}_\downarrow \subset \mathscr{P}$ is an aggregated version of $\mathscr{P}$ in the sense that $\mathscr{P}$ is partitioned and represented by $\mathbf{p}_\downarrow \in \mathscr{P}_\downarrow$.

Across a wide spectrum of scientific investigation, a random system is often described as a probability distribution on a countable alphabet, $\{\mathscr{X}, \mathbf{p}\}$; however many complex system properties of interest, such as those studied in information theory and statistical mechanics, are often described by functions of $\mathbf{p}_\downarrow$, for example, the Shannon entropy

$$H = -\sum_{k\geq 1} p_k \ln p_k$$

as in [1], the members of the Rényi entropy family

$$R_\alpha = \ln \sum_{k\geq 1} p_k^\alpha / (1-\alpha)$$

where $\alpha \in (0,1) \cup (1.\infty)$, as in [2], and the members of the Tsallis entropy family

$$T_\alpha = (1 - \sum_{k\geq 1} p_k^\alpha)/(\alpha - 1)$$

check for updates

Citation: Zhang, Z. Several Basic Elements of Entropic Statistics. *Entropy* **2023**, *25*, 1060. https://doi.org/10.3390/e25071060

Academic Editor: Nikolai Leonenko

Received: 14 June 2023
Revised: 11 July 2023
Accepted: 12 July 2023
Published: 13 July 2023

where $\alpha \in (-\infty, 1) \cup (1, \infty)$, as in [3]. Other similar functions come under the names of diversity indices, for example, the Gini–Simpson index

$$\zeta = 1 - \sum_{k \geq 1} p_k^2$$

as in [4], the generalized Simpson's indices

$$\zeta_{u,v} = \sum_{k \geq 1} p_k^u (1 - p_k)^v$$

where $u \geq 1$ and $v \geq 0$ are integers, as described in [5], Hill's diversity numbers

$$H_\alpha = (\sum_{k \geq 1} p_k^\alpha)^{1/(1-\alpha)}$$

where $\alpha \in (0, 1) \cup (1, \infty)$, as in [6], Emlen's index

$$D = \sum_{k \geq 1} p_k e^{-p_k},$$

as in [7], and the richness index

$$K = \sum_{k \geq 1} 1[p_k > 0],$$

where $1[\cdot]$ is the indicator function. While all the abovementioned functions each have their unique significance in their respective fields of study, they share one characteristic in common: they are all functions of $\mathbf{p}_\downarrow$.

The word *entropy* has ancient Greek roots, *en* and *tropē*, that is, inward and change respectively, in English, or internal change collectively. As such, it is a label-independent concept. For generality and conciseness of the presentation in this article, let the following definition be adopted.

Definition 1. *Let $f(\mathbf{p})$ be a function defined for every $\mathbf{p} \in \mathscr{P}$. The function $f(\mathbf{p})$ is referred to as an entropy if $f(\mathbf{p})$ depends on $\mathbf{p}$ only through $\mathbf{p}_\downarrow$, that is, $f(\mathbf{p}) = f(\mathbf{p}_\downarrow)$.*

By Definition 1, all entropies and diversity indices mentioned about are indeed entropies. In addition, $p_{(1)}$, or more generally $p_{(k)}$ for any positive integer k, is an entropy, and therefore $\mathbf{p}_\downarrow$ is an array of entropies. One important property to be noted about entropies is that $\mathbf{p}_\downarrow$ is independent of labels of the alphabet, $\{\ell_k; k \geq 1\}$. Another fact to be noted is that all entropies are uniquely determined by $\mathbf{p}_\downarrow$. For clarity of terminologies throughout this article, let it be noted that any properties of the underlying random system that are described by one or more entropies are referred to as entropic properties. Furthermore, $\mathbf{p}$ is referred to as the underlying probability distribution, or simply the distribution, of a random system, and $\mathbf{p}_\downarrow$ is referred to as the entropic distribution associated with $\mathbf{p}$. It is also to be noted that $\mathbf{p}_\downarrow = \{p_{(k)}; k \geq 1\}$ is not a probability distribution in the usual sense since it is not associated with any specific probability experiment. It is merely an array of nonincreasingly ordered positive parameters that sum up to one.

Let $\{X_1, \cdots, X_n\}$, drawn from $\mathscr{X}$ according to $\mathbf{p}$, be a random sample of size n. The sample may be summarized into $\mathbf{Y} = \{Y_k; k \geq 1\}$, where Y_k is the observed frequency of letter ℓ_k, or into $\hat{\mathbf{p}} = \{\hat{p}_k = Y_k/n; k \geq 1\}$. Let $\mathbf{Y}_\downarrow = \{Y_{(k)}; k \geq\}$ and $\hat{\mathbf{p}}_\downarrow = \{\hat{p}_{(k)} = Y_{(k)}/n; k \geq 1\}$ be the nonincreasingly rearranged $\mathbf{Y}$ and $\hat{\mathbf{p}}$, respectively, where $Y_{(k)} \geq Y_{(k+1)}$ and $\hat{p}_{(k)} \geq \hat{p}_{(k+1)}$ for every k. Under the assumption that the study interest of the underlying random system only lies with the properties described by indices that are functions of the form $f(\mathbf{p}_\downarrow)$, that is, entropies by Definition 1, there are two conceptual perspectives to the associated with statistical inference. The first is a framework of estimating $f(\mathbf{p})$

based on $\hat{\mathbf{p}}$, and the second is one of estimating $f(\mathbf{p}) = f(\mathbf{p}_\downarrow)$ based on $\hat{\mathbf{p}}_\downarrow$. For lack of better terms, let the first framework be referred to as the classical statistics and the second framework as the entropic statistics. These two frameworks are not equivalent and, in particular, the entropic framework has its special and useful implications.

The literature of statistical estimation of entropies, mostly in the specific form of the Shannon entropy, begins with the early works, as in [8–10], and expands in width and depth in works by, for example, [11–13]. Many other worthy references on entropy estimation may be found in the literature review in [14]. The general entropies of Definition 1, however, allow a discussion on the foundational elements of the statistics in entropic perspective, or entropic statistics, in a broader sense. This article focuses on three basic basic issues.

First, a notion of entropic sample space is introduced in Section 2 below. An entropic sample space is an aggregated sample space to register; not a single data point, but an ensemble of data points. It is a sample space of the entropic statistics, $\mathbf{Y}_\downarrow$ or $\hat{\mathbf{p}}_\downarrow$, and hence is label-independent. The said label-independence in turn allows an entropic sample space to accommodate statistical sampling into a population that is not necessarily prescribed, that is, the labels of alphabet $\mathscr{X}$ need not be completely specified *a priori*. This property of an entropic sample space gives new meaning to statistical learning and lends foundational support for statistical exploration into an unknown, or partially known, universe.

Second, an entropic characteristic function, $\phi(t) = \sum_{k\geq 1} p_{(k)}^t$ for $t \geq 1$, is introduced. It is obvious that $\phi(t)$ is an entropy by Definition 1 and that it always exists. It is established in Section 3 that $\phi(t)$ in an arbitrarily small neighborhood of any interior point of $[1, \infty)$ uniquely determines the $\mathbf{p}_\downarrow \in \mathscr{P}_\downarrow$ and *vice versa*. Therefore, it is immediately implied that any and all entropic properties of a random system, including statistical inferences, may be approached by way of $\phi(t)$.

Third, it is established in Section 4 that the entropic statistics converges almost surely and uniformly to the underlying entropic distribution, that is, $\hat{\mathbf{p}}_\downarrow \xrightarrow{a.s.} \mathbf{p}_\downarrow$ uniformly, for any $\mathbf{p}_\downarrow \in \mathscr{P}_\downarrow$. In light of the entropic sampling space and an entropic characterization of the associated entropic sampling distribution, the Glivenko–Cantelli-like convergence theorem provides a fundamental support in theory for exercises in entropic statistics.

The article ends with an appendix where a lengthy proof is found.

2. Things Entropic

2.1. Sample Spaces in Different Resolutions

Consider the experiment of randomly drawing a marble from urn 1, which contains marbles of $K = 3$ known colors, *red*, *white*, and *blue*. In anticipating the outcome of the experiment, one may introduce an index k, $k = 1, 2, 3$, to label the possible outcomes by $\ell_1 = red$, $\ell_2 = white$, and $\ell_3 = blue$, and denote the corresponding proportions by p_1, p_2, and p_3. In this case, the sample space is

$$\Omega_1 = \{\ell_1, \ell_2, \ell_3\}, \tag{1}$$

the event space is $\mathcal{B} = \{\emptyset, \{\ell_1\}, \{\ell_2\}, \{\ell_3\}, \{\ell_1, \ell_2\}, \{\ell_1, \ell_3\}, \{\ell_2, \ell_3\}, \{\ell_1, \ell_2, \ell_3\}\}$, and the point mass probability measure $\mu(\cdot)$ assigns p_1 to ℓ_1, p_2 to ℓ_2, and p_3 to ℓ_3. Let X denote the random outcome of the experiment. The following model of probability distribution,

X	ℓ_1	ℓ_2	ℓ_3
$\mathrm{P}(x)$	p_1	p_2	p_3

(2)

or in a different form $\mathbf{p} = \{p_1, p_2, p_3\}$ on $\mathscr{X} = \Omega_1 = \{\ell_1, \ell_2, \ell_3\}$, is well defined with three parameters, p_1, p_2, and p_3, subject to the constraints, $0 \leq p_k \leq 1$ for each k and $\sum_{k=1}^3 p_k = 1$. The result of drawing $n = 1$ marble from the urn may also be represented by a triplet of random variables $\mathbf{Y} = \{1[X = \ell_1], 1[X = \ell_2], 1[X = \ell_3]\}$. If $\mathbf{Y}$ is used to represent the outcome of the experiment, the sample space may be denoted as $\Omega_1 = \{\{1,0,0\}, \{0,1,0\}, \{0,0,1\}\}$ with corresponding probability distribution $\mathrm{P}(\mathbf{Y} = \{1,0,0\}) = p_1$, $\mathrm{P}(\mathbf{Y} = \{0,1,0\}) = p_2$ and $\mathrm{P}(\mathbf{Y} = \{0,0,1\}) = p_3$. For clarity in terminology, X is

referred to as a random element but $\mathbf{Y}$ is a set of random variables. In general, random results of an experiment that are represented by numerical values are referred to as *random variables*, and those by non-numerical symbols are *random elements*.

For a given experiment, the sample space may be chosen at different levels of resolution depending on the experimenter's interest in the study. Suppose the experimenter is to randomly draw $n = 3$ marbles from urn 1 with replacement in sequence, resulting in $\mathbf{X} = \{X_1, X_2, X_3\}$ where X_i, $i = 1, 2, 3$, is the color of the ith marble drawn in the sequence. The sample space associated with $\mathbf{X}$ may be represented by

$$\Omega_s = \left\{ \begin{array}{lll} \{\ell_1,\ell_1,\ell_1\}, & \{\ell_2,\ell_2,\ell_2\}, & \{\ell_3,\ell_3,\ell_3\}, \\ \{\ell_1,\ell_1,\ell_2\}, & \{\ell_2,\ell_1,\ell_1\}, & \{\ell_1,\ell_2,\ell_1\}, \\ \{\ell_1,\ell_1,\ell_3\}, & \{\ell_3,\ell_1,\ell_1\}, & \{\ell_1,\ell_3,\ell_1\}, \\ \{\ell_2,\ell_2,\ell_1\}, & \{\ell_1,\ell_2,\ell_2\}, & \{\ell_2,\ell_1,\ell_2\}, \\ \{\ell_2,\ell_2,\ell_3\}, & \{\ell_3,\ell_2,\ell_2\}, & \{\ell_2,\ell_3,\ell_2\}, \\ \{\ell_3,\ell_3,\ell_1\}, & \{\ell_1,\ell_3,\ell_3\}, & \{\ell_3,\ell_1,\ell_3\}, \\ \{\ell_3,\ell_3,\ell_2\}, & \{\ell_2,\ell_3,\ell_3\}, & \{\ell_3,\ell_2,\ell_3\}, \\ \{\ell_1,\ell_2,\ell_3\}, & \{\ell_1,\ell_3,\ell_2\}, & \{\ell_2,\ell_1,\ell_3\}, \\ \{\ell_2,\ell_3,\ell_1\}, & \{\ell_3,\ell_1,\ell_2\}, & \{\ell_3,\ell_2,\ell_1\} \end{array} \right\}, \tag{3}$$

where the subscript "s" stands for sequential. There are 27 distinct elements in (3). In this case, the sample space may also be expressed as $\Omega_s = \{\ell_1, \ell_2, \ell_3\}^3$. This sample space may be adopted if the order of the $n = 3$ observations is observable and is of interest.

Suppose in the above experiment the order of the observations is not observable or not of interest. Then the relevant information in $\mathbf{X} = \{X_1, X_2, X_3\}$ may be represented in the form of $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$, where Y_k, $k = 1, 2$, and 3 is the number of ℓ_ks observed in the sample. The sample space associated with $\mathbf{Y}$ is

$$\Omega_m = \left\{ \begin{array}{lllll} \{3,0,0\}, & \{0,3,0\}, & \{0,0,3\}, & \{2,1,0\}, & \{2,0,1\}, \\ \{0,2,1\}, & \{1,2,0\}, & \{0,1,2\}, & \{1,0,2\}, & \{1,1,1\} \end{array} \right\}, \tag{4}$$

where the subscript "m" stands for multinomial. There are 10 distinct elements in (4). In fact, $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$ is the usual multinomial random vector with $K = 3$ categories and category probabilities p_1, p_2, and p_3.

The two sample spaces, Ω_s and Ω_m, serve different statistical interests in various situations. Ω_s is well defined if $\mathbf{X} = \{X_1, X_2, X_3\}$ is observable. Ω_m is well defined if $\mathbf{X} = \{X_1, X_2, X_3\}$ is observable or only $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$ is observable. Noting that $\mathbf{X} = \{X_1, X_2, X_3\}$ implies $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$, a lower-resolution sample space may always be adopted if a higher-resolution sample space may, but not vice versa. For example, if the order of the draws is not observable, then only Ω_m is appropriate since $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$ is not linked uniquely to the elements of Ω_s.

Ω_m is an aggregated form of Ω_s and is hence of lower resolution; however, Ω_m may be further reduced in resolution. Let

$$\mathbf{Y}_\downarrow = \{Y_{(1)}, Y_{(2)}, Y_{(3)}\}, \tag{5}$$

where $Y_{(1)}, Y_{(2)}, Y_{(3)}$ are nonincreasingly ordered observed frequencies of the three colors. The sample space associated with $\mathbf{Y}_\downarrow$ is

$$\Omega_e = \left\{ \begin{array}{lll} \{3,0,0\}, & \{2,1,0\}, & \{1,1,1\} \end{array} \right\}, \tag{6}$$

where the subscript "e" stands for entropic. Ω_e is yet an aggregated form of Ω_m and hence of lower resolution still than that of Ω_m. Noting that Ω_e is label-independent, it is an example of entropic sample space.

It is easily verified that the probability distribution of $\mathbf{Y}_{\downarrow}$ may be expressed in terms of $\mathbf{p}_{\downarrow}$ as follows.

$\mathbf{Y}_{\downarrow}$	$\{3,0,0\}$	$\{2,1,0\}$	$\{1,1,1\}$
$\mathrm{P}(\mathbf{y}_{\downarrow})$	$\sum_{k\geq 1} p_{(k)}^3$	$3\sum_{k\geq 1} p_{(k)}^2\left(1-p_{(k)}\right)$	$6p_{(1)}p_{(2)}p_{(3)}$

(7)

Let it be noted that all the probabilities in (7) are label-independent, and therefore they are entropies by Definition 1.

In the case of sampling $n = 3$ marbles from urn 1 in sequence, a subscription to the entropic sample space, Ω_e, is by choice since both Ω_m and Ω_e are available. There are situations when the subscription to an entropic sample space may be by necessity.

Consider the experiment of randomly drawing $n = 3$ marbles in sequence from urn 2, which contains marbles of $K = 3$ unknown but distinguishable colors. In this case, the sample spaces, Ω_1 of (1) and Ω_m of (3), are not well defined due to the lack of knowledge of the color labels. However, the entropic sample space, Ω_e, is available for subscription regardless of what the colors are, known or unknown, as long as they are distinguishable.

In general, consider drawing a random sample of size n from $\mathscr{X} = \{\ell_k; k \geq 1\}$ under $\mathbf{p} = \{p_k; k \geq 1\}$ in sequence. The sequential sample space is of the form $\Omega_1 = \mathscr{X}^n$. The aggregated sample space,

$$\Omega_m = \{\{y_k; k \geq 1\} : y_k \geq 0 \text{ for every } k \geq 1 \text{ and } \textstyle\sum_{k\geq 1} y_k = n\}, \tag{8}$$

is that of the mutinomial array, $\mathbf{Y} = \{Y_k; k \geq 1\}$, with probability mass function

$$\mathrm{P}(\{y_k; k \geq 1\}) = \frac{n!}{\prod_{k\geq 1} y_k!} \prod_{k\geq 1} p_k^{y_k} \tag{9}$$

where $0 \leq y_k \leq n$ for every $k \geq 0$ and $\sum_{k\geq 1} y_k = n$. Moreover, Ω_m may be further aggregated into a sample space, Ω_e, for $\mathbf{Y}_{\downarrow} = \{Y_{(k)}; k \geq 1\}$, that is,

$$\Omega_e = \{\{y_{(k)}; k \geq 1\} : y_{(k)} \geq 0 \text{ and } y_{(k)} \geq y_{(k+1)} \text{ for every } k \geq 1, \text{ and } \textstyle\sum_{k\geq 1} y_{(k)} = n\}. \tag{10}$$

Let Ω_e of (10) be referred to as the entropic sample space. The associated probability distribution is

$$\mathrm{P}(\{y_{(k)}; k \geq 1\}) = \sum_{*} \mathrm{P}(\{y_k; k \geq 1\}) \tag{11}$$

where $\sum_*$ is summation of (9) over all $\{y_k; k \geq 1\}$s in Ω_m sharing the same given $\{y_{(k)}; k \geq 1\}$.

Given a $\mathbf{y}_{\downarrow} = \{y_{(k)}; k \geq 1\}$, (11) is an entropy. This may be seen in two steps. First, let $\hat{K} = \sum_{k\geq 1} 1[y_{(k)} \geq 1]$ be the number of distinct letters of $\mathscr{X}$ represented in a sample of size n, and let $\mathbf{z} = \{z_1, \cdots, z_{\hat{K}}\}$ be the set of $\hat{K}$ positive integer values of $\mathbf{y}_{\downarrow}$. $\hat{K}$ is a positive finite integer. Let the cardinality of $\mathscr{X}$ be denoted as $K = \sum_{k\geq 1} 1[p_k > 0]$. $K \geq 1$ may be finite or countably infinite. Consider an array $\mathbf{a}(\mathbf{y}_{\downarrow}) = \{a_k(\mathbf{y}_{\downarrow}); k \geq 1\}$ of length K whose entries are a particular allocation of the $\hat{K}$ values of z_j, $j = 1, \ldots, \hat{K}$, with the other $K - \hat{K}$ values of $\mathbf{a}(\mathbf{y}_{\downarrow})$ being zeros. Let $\mathbf{A}(\mathbf{y}_{\downarrow})$ be the complete collection of all such distinct $\mathbf{a}(\mathbf{y}_{\downarrow})$s. Then it is clear that $\mathbf{y}_{\downarrow}$ uniquely implies $\mathbf{A}(\mathbf{y}_{\downarrow})$.

Second, the probability in (11) may be re-expressed as

$$\mathrm{P}(\{y_{(k)}; k \geq 1\}) = \frac{n!}{\prod_{k\geq 1} y_{(k)}!} \sum_{**} \left(\prod_{k\geq 1} p_{(k)}^{a_k(\mathbf{y}_{\downarrow})} \right) \tag{12}$$

where $\sum_{**}$ is summation over all $\mathbf{a}(\mathbf{y}_{\downarrow}) \in \mathbf{A}(\mathbf{y}_{\downarrow})$, given a $\mathbf{y}_{\downarrow} = \{y_{(k)}; k \geq 1\}$. Equation (12) implies that $\mathrm{P}(\{y_{(k)}; k \geq 1\})$ is a function of $\mathbf{p}_{\downarrow}$ and hence an entropy. Let $\mathrm{P}(\{y_{(k)}; k \geq 1\})$ of (11) or (12) be referred to as the *entropic distribution* associated with the entropic sample space, Ω_e.

2.2. Entropic Objects

Let the adjective "entropic" be used to describe objects that are label-independent. Several such objects are defined or summarized below.

- A function, $f(\mathbf{p}) = f(\mathbf{p}_\downarrow)$ for all $\mathbf{p}_\downarrow \in \mathscr{P}_\downarrow$, is an entropy.
- The elements of $\mathbf{p}_\downarrow = \{p_{(k)}; k \geq 1\}$ are the entropic parameters, as compared to the elements of $\mathbf{p} = \{p_k; k \geq 1\}$, which are multinomial parameters.
- The elements of $\mathbf{Y}_\downarrow = \{Y_{(k)}; k \geq 1\}$ or equivalently of $\hat{\mathbf{p}}_\downarrow = \{\hat{p}_{(k)}; k \geq 1\}$ are entropic statistics, as compared to the elements of $\mathbf{Y} = \{Y_k; k \geq 1\}$ or equivalently $\hat{\mathbf{p}} = \{\hat{p}_k; k \geq 1\}$, which are multinomial statistics.
- Ω_e of (10) is the entropic (multinomial) sample space, as compared to Ω_m of (8), which is the multinomial sample space.
- The distribution $\mathrm{P}(\{y_{(k)}; k \geq 1\})$ of (11) or (12), is the entropic probability distribution, while $\mathrm{P}(\{y_k; k \geq 1\})$ of (9) is the multinomial probability distribution.
- Entropic statistics is the collection of statistical methodologies that help to make inference on the characteristics of a random system exclusively via entropies.

In addition, there are several useful other entropic objects. First, letting $\zeta_v = \sum_{k\geq 1} p_k(1-p_k)^v$ for all non-negative integers $v \geq 0$, $\boldsymbol{\zeta} = \{\zeta_v; v \geq 0\}$ is referred to as the *entropic basis*. The name comes from the fact that, for any well-behaved function, $h(p)$ for $p \in [0,1]$, an entropy of the form $H = \sum_{k\geq 1} p_k h(p_k)$ may be expressed as a linear combination $H = \sum_{v\geq 1} w(v)\zeta_v$. For example, the Shannon entropy, provided that it is finite, may be written as

$$H = -\sum_{k\geq 1} p_k \ln p_k = \sum_{v\geq 1} (1/v)\zeta_{v-1}.$$

The entropic basis is useful because it unfolds many entropies into simple and linearly additive forms.

Second, letting $\eta_u = \sum_{k\geq 1} p_k^u$ for all positive integers $u \geq 1$, $\boldsymbol{\eta} = \{\eta_u; u \geq 1\}$ is often referred to as the entropic moment. The elements of both $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ have good estimators. A detailed discussion may be found in [14].

Definition 2. *Let X be an random element on a countable alphabet $\mathscr{X} = \{\ell_k; k \geq 1\}$ with a corresponding probability distribution $\mathbf{p} \in \mathscr{P}$ and its associated entropic distribution $\mathbf{p}_\downarrow \in \mathscr{P}_\downarrow$. The function,*

$$\phi(t) = \sum_{k\geq 1} p_k^t, \quad \text{for } t \geq 0, \tag{13}$$

is referred to as the entropic-moment-generating function of X, of $\mathbf{p}$, or of $\mathbf{p}_\downarrow$. The two complementary parts of its domain, $[1,\infty)$ and $[0,1)$, are, respectively, referred to as the primary domain and the secondary domain of the entropic-moment-generating function.

Depending on context, $\phi(t)$ may be denoted as $\phi_X(t)$, $\phi_{\mathbf{p}}(t)$, or $\phi_{\mathbf{p}_\downarrow}(t)$ whenever appropriate. Obviously, $\phi(t)$ is uniformly bounded above by one for all $\mathbf{p} \in \mathscr{P}$ in the primary domain but is not necessarily finitely defined in the secondary domain. However, in the case of a finite alphabet, that is, $K = \sum_{k\geq 1} 1[p_k > 0] < \infty$, $\phi(t)$ is finitely defined for each and every $t \in \mathbb{R}$, in particular for $t \geq 0$. The characteristic utility of $\phi(t)$ is further explored in Section 3 below.

2.3. Examples of Entropic Statistics

Example 1. *Consider the Bernoulli experiment of tossing a coin, where $\mathrm{P}(h) = p$ and $\mathrm{P}(t) = 1-p$. The question of whether the coin is fair may be formulated in the usual classical sense, that is, whether $p = 0.5$. The question may be approached by estimating p based on a sample proportion, $\hat{p}$, if it is observable which trials lead to "h" and which lead to "t". The question may alternatively be formulated by an equivalent entropic statement, for example, whether $H = p(1-p) = 0.25$. More generally, if $K = \sum_{k\geq 1} 1[p_k]$ is finite and known, then the uniformity of $\mathbf{p}$ on $\mathscr{X}$ may be formulated entropically by, for example, $H = \sum_{k\geq 1} p_k^2 = 1/K$,*

$H = \sum_{k\geq 1} p_k(1-p_k) = (K-1)/K$, *or* $H = -\sum_{k\geq 1} p_k \ln p_k = \ln K$. *The validity of these entropic statements may then be gauged statistically.*

Example 2. *Consider a two-stage sampling scheme: a random sample of size n,* $\{X_1, \cdots, X_n\}$, *and then a single extra observation* X_{n+1} *are taken. The sample of size n may be summarized into letter frequencies,* $\mathbf{Y} = \{Y_k; k \geq 1\}$. *Let* $\pi_0 = \sum_{k\geq 1} p_k 1[Y_k = 0]$. *Clearly,* π_0 *is label-independent and therefore an entropic random variable. Given the sample of size n,* π_0 *may be thought of as the probability of that* X_{n+1} *assumes a letter in* $\mathscr{X}$ *that is not represented in the sample of size n. In some context,* π_0 *may be thought of as the probability of new discovery. Let* $N_1 = \sum_{k\geq 1} 1[Y_k = 1]$ *and* $T_n = N_1/n$. T_n *is commonly known as Turing's formula, introduced in [15], but credited largely to Alan Turing. It is to be noted that* N_1 *is label-independent and, therefore, so is* T_n. T_n *is an good estimator of* π_0 *and a discussion on many of its statistical properties may be found in [14].*

Example 3. *In developing a decision tree classifier, the data space is partitioned into an ensemble of small subspaces, in each of which a local classification rule is sought. The central spirit of every local classification may be described by a two-step scheme.*

1. *First, a random sample of size n,* $\{X_1, \cdots, X_n\}$, *is taken from* $\mathscr{X} = \{\ell_k; k \geq 1\}$, *under an unknown* $\mathbf{p} = \{p_k; k \geq 1\}$, *which is summarized into* $\mathbf{Y} = \{Y_k; k \geq 1\}$.
2. *The data-based local classification rule is as follows: the next observation,* X_{n+1}, *is predicted to be the letter which is observed most frequently in the sample of size n. For simplicity, let it be assumed that* $p_{(1)} > p_{(2)}$, *and a letter with the sample maximum frequency is unique (if not, some randomization may be employed).*

Obviously, the designated letter based on a sample is not necessarily the letter associated with the letter corresponding to the maximum of p_k*s. In such a setup, the performance of the tree classifier may be gauged by evaluating (calculating or estimating) the probability of the event that "the designated letter is the same letter of* $\mathscr{X}$ *with probability* $p_{(1)}$*", that is,*

$$\mathrm{P}\left(\arg\max_{\ell_k; k\geq 1}\{p(\ell_k); k \geq 1\} = \arg\max_{\ell_k; k\geq 1}\{\hat{p}(\ell_k); k \geq 1\}\right). \tag{14}$$

Note that the event in (14) is label-independent and hence the probability is an entropy, which may be estimated. The probability in (14) may reasonably called the confidence level of the simple classifier.

For illustration purpose, consider the special case of a binary X, with $n = 2m+1$ *for some positive integer m. For simplicity, n is chosen to be odd here so that* $Y_{(1)} > Y_{(2)}$ *always holds true. Suppose that* $p_1 = p_{(1)} > p_{(2)} = 1 - p_{(1)}$. *The event that a classifier based on the sample of n correctly identifies the letter of maximum probability may be equivalently expressed as* $Y_1 \geq m+1$. *The probability of such an event, (14), is*

$$\begin{aligned}\mathrm{P}\left(\ell_1 = \arg\max_{\ell_1,\ell_2}\{\hat{p}_1, 1-\hat{p}_1\}\right) &= \mathrm{P}(Y_1 \geq m+1) \\ &= \sum_{y\geq m+1} \frac{n!}{y!(n-y)!} p_{(1)}^y (1-p_{(1)})^{n-y},\end{aligned} \tag{15}$$

which is independent of the assumption that $p_1 > p_2 = 1 - p_1$ *and, therefore, is an entropy. More specifically, (15) is computed for several combinations of n and* $\mathbf{p}_\downarrow$ *and the resulting values are tabulated in Table 1. Table 1, and its likes, may be used in two different ways. First, given a fixed* $\mathbf{p}_\downarrow$, *it indicates how large a sample is needed to assure a reliability level of the classifier. On the other hand, at a given level of n and a particular* $\mathbf{p}_\downarrow$, *the classifier may be evaluated by the probabilities in the table. In practice,* $\mathbf{p}_\downarrow$ *is unknown but may be estimated.*

Table 1. Confidence Levels of Simple Binary Classifier.

$\mathbf{p}_\downarrow$	$n = 3$	$n = 5$	$n = 7$
$(1/2, 1/2)$	0.5000	0.5000	0.5000
$(2/3, 1/3)$	0.7407	0.7901	0.8267
$(3/4, 1/4)$	0.8438	0.8965	0.9294
$(4/5, 1/5)$	0.8960	0.9421	0.9667
$(5/6, 1/6)$	0.9259	0.9645	0.9824

3. Entropic Characterization

Entropic statistics focuses on making inference via entropies; it is therefore of interest to find a function which may characterize $\mathbf{p}_\downarrow \in \mathscr{P}_\downarrow$. Since the function $\phi(t)$ in its primary domain and $\boldsymbol{\eta} = \{\eta_u; u \geq 1\}$, where $\eta_u = \sum_{k\geq 1} p_k^u$ and u is an integer, imply each other (see Lemma 1 below), it follows immediately that (13) uniquely determines all entropies. However, the following theorem claims that the characteristic property of the entropic-moment-generating function, $\phi(t)$, remains intact in any arbitrarily neighborhood of any $t \in (1, \infty)$.

Theorem 1. *Let* $\mathbf{p} = \{p_k; k \geq 1\}$ *and* $\mathbf{q} = \{q_k; k \geq 1\}$ *be two probability distributions on a same countable alphabet,* $\mathscr{X} = \{\ell_k; k \geq 1\}$. *Let* $\mathbf{p}_\downarrow = \{p_{(k)}; k \geq 1\}$ *and* $\mathbf{q}_\downarrow = \{q_{(k)}; k \geq 1\}$ *be the respective corresponding entropic distributions of* $\mathbf{p}$ *and* $\mathbf{q}$. *Then* $\mathbf{p}_\downarrow = \mathbf{q}_\downarrow$ *if and only if* $\phi_\mathbf{p}(t) = \phi_\mathbf{q}(t)$ *for all* $t \in (a, b)$ *where* (a, b) *is an arbitrary interval such that* $1 \leq a < b < \infty$.

Lemma 1. *Let* $\mathbf{p} = \{p_k; k \geq 1\}$ *and* $\mathbf{q} = \{q_k; k \geq 1\}$ *be two probability distributions in* $\mathscr{P}$ *with two corresponding associated entropic distributions* $\mathbf{p}_\downarrow$ *and* $\mathbf{q}_\downarrow$ *in* $\mathscr{P}_\downarrow$. *Then* $\mathbf{p}_\downarrow = \mathbf{q}_\downarrow$ *if and only if* $\sum_{k\geq 1} p_k^n = \sum_{k\geq 1} q_k^n$ *for all positive integers* $n \geq 1$.

A proof of Lemma 1 may be found on pages 50 and 51 in [14]. To prove Theorem 1, it suffices to show that $\phi(t)$ in an arbitrarily small neighborhood of any interior point of $[1, \infty)$ determines the function globally.

Proof of Theorem 1. If $\mathbf{p}_\downarrow = \mathbf{q}_\downarrow$, then it immediately follows that $\phi_\mathbf{p}(t) = \phi_\mathbf{q}(t)$ for all $t \in [1, \infty)$ and, therefore, for $t \in (a, b)$ specifically. To prove the theorem, it suffices to show the converse.

Consider the series

$$f(z) = \sum_{k=1}^{\infty} p_k^z$$

where $z \in \mathbb{C}$ is a complex variable. Denote the real and the imaginary parts of a complex value z by $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$, respectively.

Let $\mathbb{D} = \{z : \mathrm{Re}(z) > 1\}$ be the subset of $\mathbb{C}$ such that the real part of z is greater than 1. For every $z \in \mathbb{D}$, since $p_k^{\mathrm{Re}(z-1)} \leq 1$ and $\left| p_k^{i\,\mathrm{Im}(z)} \right| = 1$, for every k, where $|z|$ is the modulus of z, it follows that

$$f(z) = \sum_{k=1}^{\infty} p_k p_k^{z-1} = \sum_{k=1}^{\infty} p_k p_k^{\mathrm{Re}(z-1)} p_k^{i\,\mathrm{Im}(z)}$$

and

$$|f(z)| \leq \sum_{k=1}^{\infty} p_k. \tag{16}$$

Letting $\alpha_k = \ln(1/p_k)$,

$$f(z) = \sum_{k=1}^{\infty} e^{-\alpha_k z}, \tag{17}$$

and the functions $e^{-\alpha_k z}$, $k \geq 1$, are analytic on $\mathbb{C}$.

Since the series in (17), for $z \in \mathbb{D}$, is dominated by the convergent series $\sum_{k\geq 1} p_k$ as in (16), by the Weierstrass uniform convergence theorem, $f(z)$ is analytic on $\mathbb{D}$. By a similar argument, $g(z) = \sum_{k=1}^{\infty} q_k^z$ is also analytic on $\mathbb{D}$.

Assuming that $\phi_{\mathbf{p}}(t) = \phi_{\mathbf{q}}(t)$ for $t \in (a, b)$ where $1 \leq a < b < \infty$, there exists a convergent sequence, $\{z_n; n \geq 1\}$ in (a, b) such that $\lim_{n\to\infty} z_n = z_0 \in (a, b)$. Noting that $(a, b) \subset \mathbb{D}$ and $f(z_n) = \phi_{\mathbf{p}}(z_n) = \phi_{\mathbf{q}}(z_n) = g(z_n)$ for $n \geq 0$, by the identity theorem for analytic functions, $f(z) = g(z)$ for all $z \in \mathbb{D}$. It follows that $\phi_{\mathbf{p}}(t) = f(t) = g(t) = \phi_{\mathbf{q}}(t)$ for all $t \in [1, \infty)$, specifically, $\sum_{k\geq 1} p_k^n = \sum_{k\geq 1} q_k^n$ for all $n \geq 1$. Finally, by Lemma 1, $\mathbf{p}_{\downarrow} = \mathbf{q}_{\downarrow}$. □

Theorem 1 immediately implies that a subfamily of the Rényi entropy R_α with $\alpha \in (a, b) \subset (1, \infty)$, a subfamily of the Tsallis entropy T_α with $\alpha \in (a, b) \subset (1, \infty)$, and a subfamily of Hill's diversity numbers H_α with $\alpha \in (a, b) \subset (1, \infty)$, respectively, characterizes $\mathbf{p}_{\downarrow}$ and, hence, characterizes all entropies.

The characterization of $\mathbf{p}_{\downarrow}$ in Theorem 1 may be equivalently stated only on any infinitely countable subset of (a, b).

Corollary 1. *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be two probability distributions on a same countable alphabet,* $\mathscr{X}$*. Let* $\mathbf{p}_{\downarrow}$ *and* $\mathbf{q}_{\downarrow}$ *be the corresponding entropic distributions of* $\mathbf{p}$ *and* $\mathbf{q}$*, respectively. Then* $\mathbf{p}_{\downarrow} = \mathbf{q}_{\downarrow}$ *if and only if* $\phi_{\mathbf{p}}(t) = \phi_{\mathbf{q}}(t)$ *on any infinite sequence of distinct values,* $\{t_n; n \geq 1\}$*, such that* $\lim_{n\to\infty} t_n = c \in (1, \infty)$*.*

Proof. Both $\phi_{\mathbf{p}}(t)$ and $\phi_{\mathbf{q}}(t)$ are analytic at $t = c$, and therefore $h(t) = \phi_{\mathbf{p}}(t) - \phi_{\mathbf{q}}(t)$ is analytic at $t = c$. Let it be first shown, by induction, that all derivatives of $h(t)$ at $t = c$ are zero, that is, $h^{(m)}(c) = 0$ for $m \geq 0$. Note first that $h(c) = h^{(0)} = 0$ by the fact that both $\phi_{\mathbf{p}}(t)$ and $\phi_{\mathbf{q}}(t)$ are continuous and $\lim_{n\to\infty} \phi_{\mathbf{p}}(t_n) = \phi_{\mathbf{p}}(c) = \phi_{\mathbf{q}}(c) = \lim_{n\to\infty} \phi_{\mathbf{q}}(t_n)$. Suppose that $h^{(0)}(c) = h^{(1)}(c) = h^{(2)}(c) = \cdots = h^{(m)}(c) = 0$ but $h^{(m+1)}(c) \neq 0$. Then there exists an interval $(c - \varepsilon, c + \varepsilon)$ such that $h(t) \neq 0$ for $t \in (c - \varepsilon, c + \varepsilon)$. However, there is at least one $t_n \in (c - \varepsilon, c + \varepsilon)$ such that $h(t_n) = 0$ by assumption. This is a contradiction and therefore $h^{(m)}(c) = 0$ for all $m \geq 1$. □

Corollary 2. *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be two probability distributions on a same countable alphabet,* $\mathscr{X}$*. Let* $\mathbf{p}_{\downarrow}$ *and* $\mathbf{q}_{\downarrow}$ *be the corresponding entropic distributions of* $\mathbf{p}$ *and* $\mathbf{q}$*, respectively. Then* $\mathbf{p}_{\downarrow} = \mathbf{q}_{\downarrow}$ *if and only if* $\phi_{\mathbf{p}}(t) = \phi_{\mathbf{q}}(t)$ *on any infinite sequence of distinct values,* $\{t_n; n \geq 1\} \in (a, b)$ *where* $1 \leq a < b < \infty$*.*

Proof. Noting that the infinitely many t_ns are in an bounded interval, there exists an infinite subset of $\{t_n; n \geq 1\}$ that converges to a constant $c \in [a, b]$. The corollary follows Corollary 1. □

Consider a pair of random elements, (X, Y), on a countable joint alphabet, $\mathscr{X} \times \mathscr{Y} = \{(l_i, m_j); i \geq 1, j \geq 1\}$, with a corresponding joint probability distribution, $\mathbf{p}_{X,Y} = \{p_{i,j}; i \geq 1, j \geq 1\}$. Let $\mathbf{p}_X = \{p_{i,\cdot}; i \geq 1\}$ and $\mathbf{p}_Y = \{p_{\cdot,j}; j \geq 1\}$, where $p_{i,\cdot} = \sum_{j\geq 1} p_{i,j}$ and $p_{\cdot,j} = \sum_{i\geq 1} p_{i,j}$, be the two marginal probability distributions of X and Y, respectively.

Corollary 3. *X and Y are independent if and only if*

$$\phi_{X,Y}(t) = \phi_X(t) \times \phi_Y(t) \tag{18}$$

for all $t \in (a, b)$, where a and b are two arbitrary real numbers such that $1 \leq a < b < \infty$.

Proof. If X and Y are independent, then (18) follows immediately. Conversely, suppose that (18) holds. Consider another pair of independent random elements, (U, V), on the same countable joint alphabet $\mathscr{X} \times \mathscr{Y}$ and with identical marginal distributions to those of

(X, Y), that is, $\mathbf{p}_X$ and $\mathbf{p}_Y$. It then follows, by (18) and Theorem 2, that $\mathbf{p}_{U,V} = \mathbf{p}_{X,Y}$, which in turn implies that X and Y are independent. □

Corollary 3 provides a characterization of independence on a general countable joint alphabet, and its utility may be explored further.

4. A Basic Convergence Theorem

From an entropic perspective, the convergence of $\hat{\mathbf{p}}_\downarrow$ to $\mathbf{p}_\downarrow$, to be distinguished from that of $\hat{\mathbf{p}}$ to $\mathbf{p}$, is of fundamental interest.

For clarity of presentation in this section, let it be noted that, whenever necessary, the subindex n may be added to $\mathbf{Y}$, Y_k, $\hat{\mathbf{p}}$, $\hat{p}_k$, $\hat{\mathbf{p}}_\downarrow$, and $\hat{p}_{(k)}$ to highlight the dynamic nature of these previously defined quantities as n changes, that is, $\mathbf{Y} = \mathbf{Y}_n$, $Y_k = Y_{k,n}$, $\hat{\mathbf{p}} = \hat{\mathbf{p}}_n$, $\hat{p}_k = \hat{p}_{k,n}$, $\hat{\mathbf{p}}_{\downarrow,n} = \hat{\mathbf{p}}_\downarrow$ and $\hat{p}_{(k)} = \hat{p}_{(k),n}$, respectively.

The main result established in this section is the uniform almost-sure convergence of $\hat{\mathbf{p}}_\downarrow$ to $\mathbf{p}_\downarrow$, which is made more precise in Theorem 2 below.

Consider the experiment of repeatedly and independently drawing a letter from $\mathscr{X}$ under $\mathbf{p}$, resulting in a sequence of randomly selected letters, $\omega = \{x_1, x_2, \cdots\}$. Let the collection of all possible such sequences or paths be denoted Ω. A sample of size n is a partial sequence of the first n randomly selected letters in an ω, $\{x_1, \cdots, x_n\}$.

Let $\mathbf{p}_\downarrow = \{p_{(k)}; k \geq 1\}$ and $\hat{\mathbf{p}}_\downarrow = \{\hat{p}_{(k)}; k \geq 1\}$ be defined as above. It is to be specifically noted that the rearrangement of the observed relative frequencies, $\hat{\mathbf{p}}_\downarrow$, is performed independently based on the observed values of $\hat{p}_k$ for all $k \geq 1$, with no regard to the arrangement of the probabilities, $\mathbf{p}_\downarrow = \{p_{(k)}; k \geq 1\}$. Consequently, the letter of which the relative frequency $\hat{p}_{(k)}$ is observed is not necessarily the same letter with which the probability $p_{(k)}$ is associated. This is, in fact, the essence of entropic perspective.

Theorem 2. *For any $\mathbf{p} \in \mathscr{P}$, let $\mathbf{p}_\downarrow$, $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}_\downarrow$ be defined as above. Then*

$$\max_{k\geq 1} \left| \hat{p}_{(k)} - p_{(k)} \right| \xrightarrow{a.s.} 0. \tag{19}$$

A proof of Theorem 2 requires Lemmas 2 and 3 below.

Lemma 2. *For any $\mathbf{p} \in \mathscr{P}$, let $\hat{\mathbf{p}}$ be as defined above. Then*

$$\max_{k\geq 1} |\hat{p}_k - p_k| \xrightarrow{a.s.} 0. \tag{20}$$

Proof. For each k, by the strong law of large numbers, $\hat{p}_k \xrightarrow{a.s.} p_k$ or equivalently $|\hat{p}_k - p_k| \xrightarrow{a.s} 0$. Let the collection of paths $\omega = \{x_1, x_2, \cdots\}$ in Ω that satisfies $\lim_{n\to\infty} |\hat{p}_k - p_k| = 0$ be denoted as $\Omega_k \subseteq \Omega$. It follows that $\mathrm{P}(\Omega_k) = 1$, that the complement of Ω_k, Ω_k', is of probability zero, that $\cup_{k\geq 1}\Omega_k'$ is of probability zero, and that, letting $\Omega^* = \cap_{k\geq 1}\Omega_k$, $\mathrm{P}(\Omega^*) = 1 - \mathrm{P}(\cup_{k\geq 1}\Omega_k') = 1$.

For each and every path $\omega \in \Omega^*$ and every k, $\lim_{n\to\infty} |\hat{p}_k - p_k| = 0$. Note the fact that $|\hat{p}_k - p_k| \leq \hat{p}_k + p_k$ and, therefore , $\sum_{k\geq 1} |\hat{p}_k - p_k| \leq \sum_{k\geq 1}(\hat{p}_k + p_k) = 2$, by the bounded convergence theorem,

$$\lim_{n\to\infty} \sum_{k\geq 1} |\hat{p}_k - p_k| = \sum_{k\geq 1} \lim_{n\to\infty} |\hat{p}_k - p_k| = 0, \tag{21}$$

that is, $\sum_{k\geq 1} |\hat{p}_k - p_k| \xrightarrow{a.s.} 0$. By (21), the lemma follows from the fact that

$$\max_{k\geq 1} |\hat{p}_k - p_k| \leq \sum_{k\geq 1} |\hat{p}_k - p_k| \xrightarrow{a.s.} 0.$$

□

Lemma 2 may be viewed as a version of the Glivenko–Cantelli theorem on countable alphabets with respect to observed data from a classical multinomial experiment. The uniformity of the convergence in (20) is of essential importance in the proof of Theorem 2, which is given below by way of Lemma 3.

Lemma 3. *For each* $k \geq 1$,

$$\hat{p}_{(k)} - p_{(k)} \xrightarrow{a.s.} 0. \tag{22}$$

A proof of Lemma 3 is given in Appendix A. Let it be noted that Ω is the sample space of a perpetual multinomial *iid* sampling scheme on $\mathscr{X}$ under a probability distribution $\mathbf{p} \in \mathscr{P}$. Each path in Ω may be represented by $\{\hat{\mathbf{p}}_n; n \geq 1\}$ where $\hat{\mathbf{p}}_n = \{\hat{p}_{k,n}; k \geq 1\}$. For each such path $\{\hat{\mathbf{p}}_n; n \geq 1\} \in \Omega$, there exists a corresponding path $\{\hat{\mathbf{p}}_{\downarrow,n}; n \geq 1\}$, which is the rearranged $\{\hat{\mathbf{p}}_n; n \geq 1\}$ over all k for every n. Let the total collection of all rearranged paths of Ω be denoted as $\Omega_\downarrow$, and let the collection of all rearranged paths of Ω^* be denoted as $\Omega^*_\downarrow$. It follows that $\mathrm{P}(\Omega^*_\downarrow) = \mathrm{P}(\Omega^*) = 1$. Lemma 3 states that, in each path of $\Omega^*_\downarrow$, the k th component of $\hat{\mathbf{p}}_{\downarrow,n}$ converges to the k th component of $\mathbf{p}_\downarrow$, namely, $p_{(k)}$, for each k.

Proof of Theorem 2. For any $\omega \in \Omega^*_\downarrow$, note that $\sum_{k\geq 1} |\hat{p}_{(k)} - p_{(k)}| \leq 2$, by the bounded convergence theorem and Lemma 3, $\lim_{n\to\infty} \max_{k\geq 1} |\hat{p}_{(k)} - p_{(k)}| \leq \lim_{n\to\infty} \sum_{k\geq 1} |\hat{p}_{(k)} - p_{(k)}| = \sum_{k\geq 1} \lim_{n\to\infty} |\hat{p}_{(k)} - p_{(k)}| = 0$. The theorem follows the fact that $\mathrm{P}(\Omega^*_\downarrow) = 1$. □

Theorem 2 may be viewed as a version of the Glivenko–Cantelli theorem on countable alphabets with respect to observed data from an entropic multinomial experiment. Theorem 2 immediately implies almost sure convergence for estimators of several key quantities in classification procedures.

Example 4. $\hat{p}_{(1)} \xrightarrow{a.s.} p_{(1)}$.

Example 5. *Suppose that* $p_{(1)} > p_{(2)}$, *that is, there exists a unique letter in* $\mathscr{X}$, *denoted* ℓ_0, *associated with probability* $p_{(1)}$. *Then the probability of a correct classification, that is,* $\ell_0 = \arg\max_{\mathscr{X}} \{\hat{p}_k; k \geq 1\}$, *converges almost surely to one. This is so because, for any path in* $\Omega^*_\downarrow$ *and any* $\varepsilon < (p_{(1)} - p_{(2)})/2$, *there exists an* N *such that, for any* $n > N$, $|\hat{p}_{(1)} - p_{(1)}| < \varepsilon$ *and* $|\hat{p}_{(1)} - p_{(k)}| > \varepsilon$ *for all* $k \geq 2$.

The results of Examples 4 and 5 lend fundamental support for classification algorithms based on maximum observed frequency, used widely in exercises of modern data science, for example, decision trees, as mentioned in Example 3.

Many entropies of interest across a wide spectrum of studies are of the additive form, $H(\mathbf{p}_\downarrow) = \sum_{k\geq 1} g(p_{(k)})h(p_{(k)})$, where $g(p) \geq 0$ and $h(p) \geq 0$ are functions of $p \in [0,1]$. The almost-sure convergence of Theorem 2 may be passed on to the plug-in estimators of some such entropies by way of a rather trivial statement in the proposition below.

Proposition 1. *Let* $H(\mathbf{p}_\downarrow) = \sum_{k\geq 1} g(p_{(k)})h(p_{(k)})$ *where* $g(p) \geq 0$ *and* $0 \leq h(p) \leq M$ *for some* $M > 0$ *are continuous functions of* $p \in I = [0,1]$. *Suppose that* $\mathbf{p} \in \mathscr{P}$ *such that*

1. $\sum_{k\geq 1} g(p_{(k)}) < \infty$, *and*
2. $\sum_{k\geq 1} g(\hat{p}_{(k),n}) \xrightarrow{a.s.} \sum_{k\geq 1} g(p_{(k)})$.

Then $H(\hat{\mathbf{p}}_\downarrow) \xrightarrow{a.s.} H(\mathbf{p}_\downarrow)$.

Proof. Noting that $H(\mathbf{p}_\downarrow) = \sum_{k\geq 1} g(p_{(k)})h(p_{(k)}) \leq M \sum_{k\geq 1} g(p_{(k)}) < \infty$, it follows by Conditions 1 and 2 that

$$|H(\hat{\mathbf{p}}_\downarrow) - H(\mathbf{p}_\downarrow)| \leq M \sum_{k\geq 1} g(\hat{p}_{(k)}) + M \sum_{k\geq 1} g(p_{(k)}) < \infty. \tag{23}$$

Let $\Omega_{\downarrow}^{**} \subseteq \Omega_{\downarrow}$ be the total collection of paths such that Condition 2 holds. For each path, $\{\hat{\mathbf{p}}_{n,\downarrow}; n \geq 1\} \in \Omega_{\downarrow}^{**}$, by (23), the proposition follows by the bounded convergence theorem and the fact that $\mathrm{P}(\Omega_{\downarrow}^{**}) = 1$. □

Example 6. *Let $H(\mathbf{p}_{\downarrow}) = \sum_{k\geq 1} p_{(k)}^{s}(1-p_{(k)})^{t}$ where $s \geq 1$ and $t \geq 0$ are two real constants. In the setup of Proposition 1, $h(p) = (1-p)^{t} \leq 1$ on $I = [0,1]$, and $g(p) = p^{s}$ satisfying $\sum_{k\geq 1} p_{(k)}^{s} \leq 1$ for all $\mathbf{p}_{\downarrow} \in \mathscr{P}_{\downarrow}$ without qualification, and therefore also for $\hat{\mathbf{p}}_{\downarrow} \in \mathscr{P}_{\downarrow}$, that is, $\sum_{k\geq 1} \hat{p}_{(k)}^{s} \leq 1$, which implies, by the bounded convergence theorem, $\sum_{k\geq 1} \hat{p}_{(k)}^{s} \to \sum_{k\geq 1} p_{(k)}^{s}$ along each and every path in Ω^{*}. By Proposition 1, $\sum_{k\geq 1} \hat{p}_{k}^{s}(1-\hat{p}_{(k)})^{t} \overset{a.s.}{\to} \sum_{k\geq 1} p_{k}^{s}(1-p_{(k)})^{t}$. More specifically, when s and t take integers $u \geq 1$ and $v \geq 0$, the plug-in estimator of the generalized Simpson's diversity index $H(\mathbf{p}_{\downarrow}) = \sum_{k\geq 1} p_{(k)}^{u}(1-p_{(k)})^{v}$ (see [5,16]) converges almost surely.*

Example 6 implies that the plug-in estimator of $H(\mathbf{p}_{\downarrow}) = \sum_{k\geq 1} p_{(k)}^{s}$ where $s \geq 1$ converges almost surely, which in turn implies that the plug-in estimators of members of the Rényi entropy family and the Tsallis entropy family converge almost surely for all $\mathbf{p}_{\downarrow} \in \mathscr{P}_{\downarrow}$ without qualification when $\alpha \geq 1$. However, it is not known whether the plug-in estimators of the members of the families with $\alpha \in (0,1)$ converge almost surely when $\mathbf{p}_{\downarrow} \in \mathscr{P}_{\downarrow}$ without other qualification (also, see [17]).

Example 7. *The plug-in estimator of the Shannon entropy, $H(\mathbf{p}_{\downarrow}) = -\sum_{k\geq 1} p_{(k)} \ln p_{(k)}$, converges almost surely when $\mathbf{p}_{\downarrow}$ is such that $K = \sum_{k\geq 1} 1_{[p_{(k)}>0]} < \infty$. In this case, even though $-\ln p$ is not bounded above on $I = [0,1]$, $h(p) = -p^{\alpha} \ln p \leq 1/(\alpha e)$ is for any $\alpha \in (0,1)$. Writing $H(\mathbf{p}_{\downarrow}) = \sum_{k\geq 1} g(p_{(k)})h(p_{(k)})$ where $g(p) = p^{1-\alpha}$ and $h(p) = -p^{\alpha} \ln p$, it suffices to show $\sum_{k\geq 1} \hat{p}_{(k)}^{1-\alpha}$ converges almost surely. However, this is the case since, by Theorem 2, for every path $\{\hat{\mathbf{p}}_{n,\downarrow}; n \geq 1\} \in \Omega_{\downarrow}^{*}$, $\sum_{k\geq 1} \hat{p}_{(k)}^{1-\alpha} \to \sum_{k\geq 1} p_{(k)}^{1-\alpha}$, due to the fact that $K < \infty$ and $\mathrm{P}(\Omega_{\downarrow}^{*}) = 1$.*

It is not known whether the plug-in estimator of the Shannon entropy converges almost surely when $\mathbf{p}_{\downarrow} \in \mathscr{P}_{\downarrow}$ without further qualification.

The Shannon entropy has utilities across a wide spectrum of scientific investigations (see [18]). However, it is not finitely defined for all distributions in $\mathscr{P}$. A family of the generalized Shannon entropies, for any $\mathbf{p}_{\downarrow} \in \mathscr{P}_{\downarrow}$, is proposed as follows:

$$H_{m}(\mathbf{p}_{\downarrow}) = -\sum_{k\geq 1} \left(\frac{p_{(k)}^{m}}{\sum_{j\geq 1} p_{(j)}^{m}} \right) \ln \left(\frac{p_{(k)}^{m}}{\sum_{j\geq 1} p_{(j)}^{m}} \right) \tag{24}$$

in [19], where $m \geq 1$ is an integer. The Shannon entropy is a special family member corresponding to $m = 1$. It may be verified that each member of the family, except the Shannon entropy, is finitely defined for all $\mathbf{p} \in \mathscr{P}$ and offers all important utilities that the Shannon entropy offers, including the fact that the mutual information derived based on each member with $m \geq 2$ is zero if and only if the two underlying random elements are independent.

Example 8. *The plug-in estimator of (24) converges almost surely for any $\mathbf{p}_{\downarrow} \in \mathscr{P}_{\downarrow}$ whenever $m \geq 2$. To see this, let it be first noted that the plug-in estimator of $-\sum_{k\geq 1} p_{k}^{m} \ln p_{k}$ converges almost surely. This fact follows from Proposition 1 with $g(p) = p$, $h(p) = -p^{m-1} \ln p$ which is uniformly bounded above on $I = [0,1]$. The claimed almost-sure convergence then follows the fact that, in the re-expression of (24) below,*

$$H_{m}(\mathbf{p}_{\downarrow}) = \left(\frac{1}{\sum_{j\geq 1} p_{(j)}^{m}} \right) \left[\left(-m \sum_{k\geq 1} p_{(k)}^{m} \ln p_{(k)} \right) + \left(\sum_{k\geq 1} p_{(k)}^{m} \right) \left(\ln \sum_{j\geq 1} p_{(j)}^{m} \right) \right], \tag{25}$$

and the fact that the plug-in estimator of each of the four series converges almost surely.

5. Conclusions and Discussion

This article introduces a perspective termed entropic statistics. One of the motivations of the perspective is to accommodate probability experiments on sample spaces which may include outcomes that are known to exist (and therefore are prescribed) and those whose existence is not known (and therefore not prescribable). Such a framework allows statistical exploration into a general population with possibly infinitely many previously unobserved and unknown outcomes, or new discoveries. The key concept to foster such a framework is the label-independence, that is, all parameters and statistics do not depend of the labels of an alphabet as long as they are distinguishable. Consequently, in this article an array of label-independent objects are defined and termed entropic objects. In particular, a general entropy, entropic parameters, entropic statistics, entropic sample spaces, entropic probability distributions, and an entropic-moment-generating function are defined.

Based on the defined entropic objects, two basic theorems are established. Theorem 1 provides a characterization of the entropic probability distribution on the alphabet via the entropic-moment-generating function, and Theorem 2 establishes the almost-sure convergence of the entropic statistics to the entropic parameters and, hence, provides a foundational support to the entropic framework.

On the other hand, this article merely provides a few basic results in entropic statistics. On a broader spectrum, many other issues may be fruitfully considered on at least three fronts, namely, fundamental, probabilistic, and statistical. To begin with, the fundamental question of what constitutes entropy may be explored in many directions. One of the most cited sets of axioms is that discussed by Khinchin [20], under which the Shannon entropy is proved to be unique. However under slightly less restrictive axioms, many other entropies exist and enjoy almost all the desirable utilities of the Shannon entropy; for example, see [19]. The existing literature on generalization of entropy is extensive in physics and information theory; for example, see [21,22]. The collective effort to better understand what entropy is and how it may help to describe an underlying random system is ongoing. Further research in understanding generalized entropies and their implications could greatly enrich the framework of entropic statistics.

Entropy in general is often thought of as summary of a profile state, however measured numerically, of inner energy or chaos within a random system. As such, it is independent of any labeling systems, regardless of whether the state is observable or not. A key conceptual shift introduced in this article is from statistical inference on $\mathbf{p}$ (or a function of $\mathbf{p}$) based on the multinomial frequencies $\mathbf{Y}$ to that on $\mathbf{p}_{\downarrow}$ (or a function of $\mathbf{p}_{\downarrow}$) based on the entropic frequencies $\mathbf{Y}_{\downarrow}$. Such a framework shift, by necessity or by choice, triggers a long array of basic probability and statistics questions, under different degrees of model restriction, ranging from parametric forms of $p_k = p(k, \boldsymbol{\theta})$ for some parameter $\boldsymbol{\theta}$ to the nonparametric form, $\{p_{(k)}; k \geq 1\}$. It may be interesting to note that even for the nonparametric form, there are several qualitatively different forms, that of a known $K = \sum_{k\geq 1} 1_{[p_{(k)}>0]} < \infty$, that of an unknown $K = \sum_{k\geq 1} 1_{[p_{(k)}>0]} < \infty$, and that of $K = \sum_{k\geq 1} 1_{[p_{(k)}>0]} = \infty$. Each of these model classes could imply a very different stochastic behavior of $\mathbf{Y}_{\downarrow}$ as the sample size n increases. Even long before the notion of information entropy was coined by Shannon in [1], the behavior of $\mathbf{Y}_{\downarrow}$ had been discussed in the literature by, for example, Auerbach [23] and Zipf [24]. More recently, several articles [25,26] discussed domains of attraction in the total collection of all distributions on a countable alphabet by a tail index, $\tau_n = n \sum_{k\geq 1} p_{(k)}(1 - p_{(k)})^n$. Each domain characterizes the decay rate of the tail of the underlying entropic distribution and, in turn, dictates the rates of convergence of various statistical estimators of various entropies. Further advances on that front would enhance the understanding of probabilistic behavior of the entropic statistics and, in turn, the estimated entropies of interest.

In terms of statistical estimation, a large proportion of the existing literature mainly focuses on the Shannon entropy and variations of the plug-in estimators under various conditions, most of which are described and referenced in [14]. There are also non-plug-in estimators of different types, for example, the Bayes estimators [27–29], the hi-

erarchical Bayes estimators [30], the James–Stein estimators [31], the coverage-adjusted estimators [32–34], and an unbiased estimator based on sequential data proposed by Montgomery-Smith and Schürmann. In general, the asymptotic distributions of the plug-in estimators and their variants seem to have been studied and described to some extent; for example, see [12,35–38]. However, it is fair to say that many, if not most, of the proposed estimators of various types have not yet been assigned asymptotic distributions. Any advances in that direction could much benefit applications of these estimators.

In short, the landscape of entropic statistics is quite porous in comparison to that of richly supported classical statistics. Many basic and important questions are yet to be answered, from the axiomatic foundation, to the definitions of basic elements, to the theoretical supporting architecture, and to the relevance in applications. However, the same said porosity also offers opportunities for interesting contemplation.

Funding: This research received no external funding.

Data Availability Statement: This research involved no data.

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

Appendix A

Proof of Lemma 3. For clarity, the proof of (22) is given, respectively, in five progressively more general cases: (1) $p_1 = 1$; (2) $K = \sum_{k\geq 1} 1_{[p_k>0]} < \infty$ and all positive p_k's are distinct; (3) $K < \infty$; (4) K is infinite and all positive p_k's are distinct; (5) K is infinite.

For notation simplicity in all cases, let it be assumed without loss of generality that $\mathbf{p} = \{p_k; k \geq 1\}$ is nonincreasingly arranged to begin with, that is, $p_k \geq p_{k+1}$ for every k. With this assumption, the only rearranged object is $\hat{\mathbf{p}}_\downarrow = \{\hat{p}_{(k)}; k \geq 1\}$ with $\hat{p}_{(k)} \geq \hat{p}_{(k+1)}$ for every k.

In Case 1, the statement of (22) is trivial.

In Case 2, let $p_0 = 1$ and $p_{K+1} = 0$. It follows that

$$1 = p_0 > p_1 > p_2 > \cdots > p_{k-1} > p_k > p_{k+1} > \cdots > p_K > p_{K+1} = 0.$$

For each sequence $\omega \in \Omega^*$ as defined in the proof of Lemma 2, the uniformity of (20) implies that for any $\varepsilon > 0$, there exists an N such that for all $n > N$, $-\varepsilon < \hat{p}_k - p_k < \varepsilon$ for all $k \geq 1$. Specifically, let

$$\varepsilon_0 = \min\{(p_k - p_{k+1})/2; 1 \leq k \leq K\} > 0. \tag{A1}$$

There exists an N such that for all $n > N$, $\max\{|\hat{p}_k - p_k|; 1 \leq k \leq K\} < \varepsilon_0$, which has the following two implications.

1. $\cap_{k=1}^{K}(p_k - \varepsilon_0, p_k + \varepsilon_0) = \emptyset$, that is, $p_k \pm \varepsilon_0$ are disjoint for all $k = 1, \ldots, K$.
2. For every k, $k = 1, \ldots, K$, $\hat{p}_k \in p_k \pm \varepsilon_0$ and it is the only observed relative frequency in $p_k \pm \varepsilon_0$.

Combining the above two implications, it follows that $\hat{p}_k = \hat{p}_{(k)}$, that is, $|\hat{p}_{(k)} - p_k| \to 0$, for every k, $k = 1, \ldots, K$. Since Ω^* is of probability one, (22) is established.

In Case 3, it is allowed that several consecutive probabilities in $\mathbf{p} = \{p_k; 1 \leq k \leq K\}$, where $K = \sum_{k\geq 1} 1[p_k > 0] < \infty$, are identical. It follows that

$$1 = p_0 \geq p_1 \geq p_2 \geq \cdots \geq p_{k-1} \geq p_k \geq p_{k+1} \geq \cdots \geq p_K > p_{K+1} = 0.$$

Noting that $\mathbf{p} = \{p_k; k \geq 1\}$ is a finite sequence of runs of identical values, collecting the first value in each run and retaining its index value, a subset of $\{p_k; k \geq 1\}$ is obtained, namely, $\{p_{k_i}; i = 1, \ldots, I\}$, where I is the number of distinct values in $\mathbf{p}$. Let r_i be the multiplicity of p_{k_i} in $\mathbf{p}$, $i = 1, \ldots, I$. It follows that

$$1 = p_0 = p_{k_0} > p_{k_1} > p_{k_2} > \cdots > p_{k_{K'}} > p_{K+1} = 0.$$

For each sequence $\omega \in \Omega^*$ as defined in the proof of Lemma 2, the uniformity of (20) implies that for any $\varepsilon > 0$, there exists an N such that for all $n > N$, $-\varepsilon < \hat{p}_k - p_k < \varepsilon$ for all k, $k = 1, \ldots, K$. Specifically, let

$$\varepsilon_1 = \min\{(p_{k_i} - p_{k_{i+1}})/2; i = 0, \ldots, I\} > 0 \tag{A2}$$

where $p_{k_0} = p_0 = 1$. There exists an N_1 such that for all $n > N_1$, $\max\{|\hat{p}_k - p_k|; k \geq 1\} < \varepsilon_1$, which has the following implications.

1. $\cap_{i=1}^{I}(p_{k_i} - \varepsilon_1, p_{k_i} + \varepsilon_1) = \emptyset$, that is, $p_{k_i} \pm \varepsilon_1$ are disjoint for all i, $i = 1, \ldots, I$.
2. For every given k, and therefore an implied i, there are exactly r_i relative frequencies among $\{\hat{p}_k; 1 \leq k \leq K\}$ found in $p_{k_i} \pm \varepsilon_1$.

It then follows that, for each given k,

$$\min\{\hat{p}_{(k_i+j)}; j = 0, \ldots, r_i - 1\} \leq \hat{p}_{(k)} \leq \max\{\hat{p}_{(k_i+j)}; j = 0, \ldots, r_i - 1\},$$

and hence $\hat{p}_{(k)} \to p_{(k)} = p_k$. Finally, (22) follows the fact that $\mathrm{P}(\Omega^*) = 1$.

In Case 4, $p_k > 0$ for all $k \geq 1$ and all probabilities are distinct. Letting $p_0 = 1$ and $p_\infty = 0$,

$$1 = p_0 > p_1 > p_2 > \cdots > p_{k-1} > p_k > p_{k+1} > \cdots > p_\infty = 0.$$

For every fixed k' such that $p_{k'} \in (0, 1)$, let $m \geq 1$ be an integer such that

$$1 - \sum_{k=1}^{m} p_k < p_{k'+1}, \text{ and} \tag{A3}$$

$$m \geq k' + 1. \tag{A4}$$

Such an m exists for any given $\mathbf{p}$ with an infinite K and a fixed $k' \geq 1$.

For each sequence $\omega \in \Omega^*$, as defined in the proof of Lemma 2, the uniformity of (20) implies that for any $\varepsilon > 0$, there exists an N such that for all $n > N$, $-\varepsilon < \hat{p}_k - p_k < \varepsilon$ for all k, $k \geq 1$. Specifically, let

$$\varepsilon_2 = \min\{(p_k - p_{k+1})/2; k = 0, \ldots, m\} > 0.$$

There exists an N_2 such that for all $n > N_2$, $\max\{|\hat{p}_k - p_k|; k \geq 1\} < \varepsilon_2$, which implies the following.

1. The first m probabilities of $\mathbf{p}$, $p_1, \cdots, p_m$, are covered, respectively, by m disjoint intervals, $p_k \pm \varepsilon_2$, $k = 1, \ldots, m$.
2. The relative frequencies corresponding to $\{p_1, \cdots, p_m\}$, namely, $\{\hat{p}_1, \cdots, \hat{p}_m\}$, are also covered, respectively, by the same disjoint intervals, $p_k \pm \varepsilon_2$, $k = 1, \ldots, m$.

On the other hand, noting the strict inequality in (A3) and the fact that k' is a fixed integer, there exists a sufficiently small ε_3 such that

$$1 - \sum_{k=1}^{m} p_k + m\varepsilon_3 < p_{k'+1} \tag{A5}$$

or equivalently

$$1 - \sum_{k=1}^{m} (p_k - \varepsilon_3) < p_{k'+1}. \tag{A6}$$

Let $\varepsilon_4 = \min\{\varepsilon_2, \varepsilon_3\}$. By Lemma 2, there exists an N_4 such that for all $n > N_4$,

$$p_k - \varepsilon_4 < \hat{p}_k < p_k + \varepsilon_4, \tag{A7}$$

for all k, $k = 1, \cdots, m$, and that the updated (A5) and (A6) hold, namely,

$$1 - \sum_{k=1}^{m} p_k + m\varepsilon_4 < p_{k'+1}$$

or, equivalently,

$$1 - \sum_{k=1}^{m} (p_k - \varepsilon_4) < p_{k'+1}. \tag{A8}$$

That is, in each of the disjoint intervals of (A7), there is at least one relative frequency. In particular, $\hat{p}_k$ is covered in $(p_k - \varepsilon_4, p_k + \varepsilon_4)$ for each k, $k = 1, \ldots, k' < m$, by (A4).

Next it is necessary to show that there may not be more than one relative frequency in $(p_k - \varepsilon_4, p_k + \varepsilon_4)$ for each k, $k = 1, \ldots, k'$. Toward that end, consider the total mass of 100% distributed among $\hat{p}_k$, $k \geq 1$, given n. From interval $(p_1 - \varepsilon_4, p_1 + \varepsilon_4)$ to interval $(p_m - \varepsilon_4, p_m + \varepsilon_4)$, the total collective mass covered is at least $\sum_{k=1}^{m} \hat{p}_k$; however, by (A7) and (A8),

$$\sum_{k=1}^{m} \hat{p}_k = \sum_{k=1}^{m} (\hat{p}_k + \varepsilon_4) - m\varepsilon_4 > \sum_{k=1}^{m} p_k - m\varepsilon_4 = \sum_{k=1}^{m} (p_k - \varepsilon_4) > 1 - p_{k'+1}$$

and the remainder of the mass is

$$1 - \sum_{k=1}^{m} \hat{p}_k < p_{k'+1} < p_{k'+1} + \varepsilon_4. \tag{A9}$$

Regardless of the mass, $1 - \sum_{k=1}^{m} \hat{p}_k$, on the left side of (A9) is allocated to one or more than one letter, other than those in $\{\ell_1, \cdots, \ell_m\}$, the corresponding $\hat{p}_k$, $k \geq m+1$, could not possibly be sufficiently large to exceed $p_{k'+1} + \varepsilon_4$, nor, therefore, $p_{k'} - \varepsilon_4$. That implies that, along the path of that selected $\omega \in \Omega^*$, for any $n > N_4$, $\hat{p}_k$ and $\hat{p}_k$ alone is covered in $(p_k - \varepsilon_4, p_k + \varepsilon_4)$ for k, $k = 1, \ldots, k'$. This immediately implies that $\hat{p}_k = \hat{p}_{(k)}$ for all k, $k = 1, \ldots, k'$, and in particular $\hat{p}_{k'} = \hat{p}_{(k')}$. $\hat{p}_{(k')} \to p_{k'}$ since $\hat{p}_{k'} \to p_{k'}$. Finally (22) follows the fact that $\mathrm{P}(\Omega^*) = 1$.

In Case 5, $p_k > 0$ for all $k \geq 1$ but the probabilities in $\mathbf{p} = \{p_k; k \geq 1\}$ are allowed to have multiplicities. Letting $p_0 = 1$ and $p_\infty = 0$,

$$1 = p_0 > p_1 \geq p_2 \geq \cdots p_k \geq \cdots > p_\infty = 0. \tag{A10}$$

$\mathbf{p} = \{p_k; k \geq 1\}$ has a special pattern: its maximum value runs for r_1 times; then its second largest value runs for r_2 times, and so on and so forth. In general, its i th largest value runs for r_i times followed by a run of its $i-1$ st largest value. Collect the first value in each run and record its index, k_i, $i \geq 1$, resulting in a strictly decreasing subsequence, $\{p_{k_i}; i \geq 1\}$. Letting $k_0 = 0$ and $k_\infty = \infty$,

$$1 = p_0 = p_{k_0} > p_{k_1} > p_{k_2} > \cdots p_{k_i} > \cdots > p_{k_\infty} = p_\infty = 0.$$

Consequently, $\mathbf{p} = \{p_k; k \geq 1\}$ may be viewed as a sequence containing p_{k_i} for $i \geq 1$ with $r_i - 1$ p_{k_i}s between p_{k_i} and $p_{k_{i+1}}$.

Given a value of k, say k', there is an i' such that $p_{k'} = p_{k_{i'}}$ and k' must be one of the values from the list $\{k_{i'}, k_{i'} + 1, \cdots, k_{i'} + r_{i'} - 1\}$, noting $p_{k_{i'}+r_{i'}} = p_{k_{i'}+1} < p_{k'}$. Let m be such that

$$1 - \sum_{i=1}^{m} r_i p_{k_i} < p_{k_{i'}+1}, \text{ and} \tag{A11}$$

$$\sum_{i=1}^{m} r_i \geq k_{i'} + 1. \tag{A12}$$

Such an m exists for any given $\mathbf{p}$ and a fixed $k' \geq 1$, which fixes an i'.

For each sequence $\omega \in \Omega^*$ as defined in the proof of Lemma 2, the uniformity of (20) implies that for any $\varepsilon > 0$, there exists an N such that for all $n > N$, $-\varepsilon < \hat{p}_k - p_k < \varepsilon$ for all k, $k \geq 1$. Specifically let

$$\varepsilon_5 = \min\{(p_{k_i} - p_{k_{i+1}})/2; i = 0, \ldots, m\} > 0.$$

There exists an N_5 such that for all $n > N_5$, $\max\{|\hat{p}_k - p_k|; k \geq 1\} < \varepsilon_5$, which has the following two implications.

1. The first $\sum_{i=1}^{m} r_i$ probabilities of $\mathbf{p}$, $p_1, \cdots, p_{\sum_{i=1}^{m} r_i}$, are covered, respectively, by $k_{i'}$ disjoint intervals, $p_{k_i} \pm \varepsilon_5$, $i = 1, \ldots, i'$.
2. The relative frequencies corresponding to $\{p_1, \cdots, p_{\sum_{i=1}^{m} r_i}\}$, namely, $\{\hat{p}_1, \cdots, \hat{p}_{\sum_{i=1}^{m} r_i}\}$, are also covered, respectively, by the same disjoint intervals, $p_{k_i} \pm \varepsilon_5$, $i = 1, \ldots, i'$.

On the other hand, noting the strict inequality in (A11) and the fact that k' is a fixed integer, there exists a sufficiently small ε_6 such that

$$1 - \sum_{i=1}^{m} r_i p_{k_i} + \varepsilon_6 \sum_{i=1}^{m} r_i < p_{k_{i'}+1} \tag{A13}$$

or, equivalently,

$$1 - \sum_{i=1}^{m} r_i (p_{k_i} - \varepsilon_6) < p_{k_{i'}+1}. \tag{A14}$$

Let $\varepsilon_7 = \min\{\varepsilon_5, \varepsilon_6\}$. By Lemma 2, there exists an N_7 such that for all $n > N_7$, all relative frequencies sharing the same p_{k_i}, namely, $\hat{p}_{k_i}, \hat{p}_{k_i+1}, \cdots, \hat{p}_{k_i+r_i-1}$, are found in

$$(p_{k_i} - \varepsilon_7, p_{k_i} + \varepsilon_7) \tag{A15}$$

for all i, $i = 1, \cdots, m$, and the updated (A13) and (A14) are

$$1 - \sum_{i=1}^{m} r_i p_{k_i} + \varepsilon_7 \sum_{i=1}^{m} r_i < p_{k_{i'}+1}$$

or, equivalently,

$$1 - \sum_{i=1}^{m} r_i (p_{k_i} - \varepsilon_7) < p_{k_{i'}+1}. \tag{A16}$$

That is, in each of the disjoint intervals of (A15), there are at least r_i relative frequencies. In particular, the r_i relative frequencies, $\{\hat{p}_{k_i}, \hat{p}_{k_i+1}, \cdots, \hat{p}_{k_i+r_i-1}\}$, are covered in $(p_{k_i} - \varepsilon_7, p_{k_i} + \varepsilon_7)$ for each i, $i = 1, \ldots, i' \leq m$, by (A12).

Next it necessary is to show that there may not be more than r_i relative frequencies in $(p_{k_i} - \varepsilon_7, p_{k_i} + \varepsilon_7)$ for each i, $i = 1, \ldots, i'$. Toward that end, consider the total mass of 100% distributed among $\hat{p}_k$, $k \geq 1$, given n. From interval $(p_1 - \varepsilon_7, p_1 + \varepsilon_7)$ to interval $(p_{\sum_{i=1}^{m} r_i} - \varepsilon_7, p_{\sum_{i=1}^{m} r_i} + \varepsilon_7)$, the total collective mass covered is at least $\sum_{i=1}^{m} r_i \hat{p}_{k_i}$; however, by (A15) and (A16),

$$\sum_{i=1}^{m} r_i \hat{p}_{k_i} > \sum_{i=1}^{m} r_i (p_{k_i} - \varepsilon_7) > 1 - p_{k_{i'}+1}$$

and the remainder of the mass is

$$1 - \sum_{k=1}^{m} \hat{p}_k < p_{k_{i'}+1} < p_{k_{i'}+1} + \varepsilon_7. \tag{A17}$$

Regardless of if the mass on the left side of (A17) is allocated to one or more than one letter, other than those in $\{\ell_1, \cdots, \ell_{\sum_{i=1}^{m} r_i}\}$, the corresponding $\hat{p}_k$, $k \geq \sum_{i=1}^{m} r_i + 1$, could

not possibly be sufficiently large to exceed $p_{k_{i'}+1} + \varepsilon_7$, nor, therefore, $p_{k'} - \varepsilon_7$. That implies that, along the path of that selected $\omega \in \Omega^*$, for any $n > N_7$, $\{\hat{p}_{k_i}, \hat{p}_{k_i+1}, \cdots, \hat{p}_{k_i+r_i-1}\}$ and only $\{\hat{p}_{k_i}, \hat{p}_{k_i+1}, \cdots, \hat{p}_{k_i+r_i-1}\}$ are covered in $(p_{k_i} - \varepsilon_7, p_{k_i} + \varepsilon_7)$ for i, $i = 1, \ldots, i'$. This immediately implies that

1. $\{\hat{p}_{k_{i'}}, \hat{p}_{k_{i'}+1}, \cdots, \hat{p}_{k_{i'}+r_i-1}\} = \{\hat{p}_{(k_{i'})}, \hat{p}_{(k_{i'}+1)}, \cdots, \hat{p}_{(k_{i'}+r_i-1)}\}$ but is not necessarily equal component-wise;
2. $|\hat{p}_{(k_{i'}+j)} - p_{k'}| < \varepsilon_7$ for all $j = 0, 1, \ldots, r_{i'} - 1$;
3. In particular, $|\hat{p}_{(k')} - p_{k'}| < \varepsilon_7$.

Finally (22) follows the fact that $\mathrm{P}(\Omega^*) = 1$. □

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]
2. Rényi, A. On measures of information and entropy. In Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability, Berkeley, CA, USA, 20–30 June 1961; pp. 547–561.
3. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. [CrossRef]
4. Simpson, E.H. Measurement of diversity. *Nature* **1949**, *163*, 688. [CrossRef]
5. Zhang, Z.; Zhou, J. Re-parameterization of multinomial distribution and diversity indices. *J. Stat. Plan. Inference* **2010**, *140*, 1731–1738. [CrossRef]
6. Hill, M.O. Diversity and evenness: A unifying notation and its consequences. *Ecology* **1973**, *54*, 427–432. [CrossRef]
7. Emlen, J.M. *Ecology: An Evolutionary Approach*; Addison-Wesley: Reading, MA, USA, 1973.
8. Miller, G.A.; Madow, W.G. *On the Maximum Likelihood Estimate of the Shannon-Weaver Measure of Information*; Air Force Cambridge Research Center Technical Report AFCRC-TR-54-75; Operational Applications Laboratory, Air Force, Cambridge Research Center, Air Research and Development Command: New York, NY, USA, 1954.
9. Miller, G.A. Note on the bias of information estimates. *Inf. Theory Psychol. Probl. Methods* **1955**, *11-B*, 95–100.
10. Harris, B. *The Statistical Estimation of Entropy in the Non-Parametric Case*; Wisconsin University—Madison Mathematics Research Center: Madison, WI, USA, 1975.
11. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms* **2001**, *19*, 163–193. [CrossRef]
12. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
13. Silva, J.F. Shannon entropy estimation in ∞-alphabets from convergence results: Studying plug-in estimators. *Entropy* **2018**, *20*, 397. [CrossRef]
14. Zhang, Z. *Statistical Implications of Turing's Formula*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017.
15. Good, I.J. The population frequencies of species and estimation of population parameters. *Biometrika* **1953**, *40*, 237–264. [CrossRef]
16. Grabchak, M.; Marcon, G.; Lang, G.; Zhang, Z. The generalized Simpson's entropy is a measure of biodiversity. *PLoS ONE* **2017**, *12*, e0173305. [CrossRef] [PubMed]
17. Contreras-Reyes, J.E. Mutual information matrix based on Rényi entropy and application. *Nonlinear Dyn.* **2022**, *110*, 623–633. [CrossRef]
18. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley & Son, Inc.: New York, NY, USA, 2006.
19. Zhang, Z. Generalized Mutual Information. *Stats* **2020**, *3*, 158–165. [CrossRef]
20. Khinchin, A.I. *Mathematical Foundations of Information Theory*; Dover Publications: New York, NY, USA, 1957.
21. Amigó, J.M.; Balogh, S.G.; Hernández, S. A Brief Review of Generalized Entropies. *Entropy* **2018**, *20*, 813. [CrossRef] [PubMed]
22. Ilić, V.M.; Korbel, J.; Gupta, G.; Scarfone, A.M. An overview of generalized entropic forms. *Europhys. Lett.* **2021**, *133*, 50005. [CrossRef]
23. Auerbach, F. Das Gesetz der Bevölkerungskonzentration. *Petermann's Geogr. Mitteilungen* **1913**, *59*, 74–76.
24. Zipf, G.K. *Selected Studies of the Principle of Relative Frequency in Language*; Harvard University Press: Cambridge, MA, USA; London, UK, 1932.
25. Zhang, Z. Domains of attraction on countable alphabets. *Bernoulli* **2018**, *24*, 873–894. [CrossRef]
26. Molchanov, S.; Zhang, Z.; Zheng, L. Entropic Moments and Domains of Attraction on Countable Alphabets. *Math. Meth. Stat.* **2018**, *27*, 60–70. [CrossRef]
27. Krichevsky, R.E.; Trofimov, V.K. The Performance of Universal Encoding. *IEEE Trans. Inf. Theory* **1981**, *27*, 199–207. [CrossRef]
28. Holste, D.; Große, I.; Herzel, H. Bayes' estimators of generalized entropies. *J. Phys. A Math. Gen.* **1998**, *31*, 2551–2566. [CrossRef]
29. Schurmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos* **1996**, *6*, 414–427. [CrossRef] [PubMed]
30. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems*; Dietterich, T.G., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2002; Volume 14.
31. Hausser, J.; Strimmer, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.* **2009**, *10*, 1469–1484.
32. Chao, A.; Shen, T.-J. Non-parametric estimation of Shannon's Index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **2003**, *10*, 429–443. [CrossRef]

33. Vu, V.Q.; Yu, B.; Kass, R.E. Coverage-adjusted entropy estimation. *Stat. Med.* **2007**, *26*, 4039–4060. [CrossRef] [PubMed]
34. Zhang, Z. Entropy estimation in Turing's perspective. *Neural Comput.* **2012**, *24*, 1368–1389. [CrossRef]
35. Zhang, Z.; Zhang, X. A normal law for the plug-in estimator of entropy. *IEEE Trans. Inf. Theory* **2012**, *58*, 2745–2747. [CrossRef]
36. Zhang, Z. Asymptotic normality of an entropy estimator with exponentially decaying bias. *IEEE Trans. Inf. Theory* **2013**, *59*, 504–508. [CrossRef]
37. Chen, C.; Grabchak, M.; Stewart, A.; Zhang, J.; Zhang, Z. Normal Laws for Two Entropy Estimators on Infinite Alphabets. *Entropy* **2018**, *20*, 371. [CrossRef]
38. Grabchak, M.; Zhang, Z. Asymptotic Normality for Plug-in Estimators of Diversity Indices on Countable Alphabet. *J. Nonparametric Stat.* **2018**, *30*, 774–795. [CrossRef]