entropy

MDPI

Article

Asymptotic Normality for Plug-In Estimators of Generalized Shannon's Entropy

Jialin Zhang * and Jingyi Shi

Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS 39762, USA; jshi@math.msstate.edu
* Correspondence: jzhang@math.msstate.edu

Abstract: Shannon's entropy is one of the building blocks of information theory and an essential aspect of Machine Learning (ML) methods (e.g., Random Forests). Yet, it is only finitely defined for distributions with fast decaying tails on a countable alphabet. The unboundedness of Shannon's entropy over the general class of all distributions on an alphabet prevents its potential utility from being fully realized. To fill the void in the foundation of information theory, Zhang (2020) proposed generalized Shannon's entropy, which is finitely defined everywhere. The plug-in estimator, adopted in almost all entropy-based ML method packages, is one of the most popular approaches to estimating Shannon's entropy. The asymptotic distribution for Shannon's entropy's plug-in estimator was well studied in the existing literature. This paper studies the asymptotic properties for the plug-in estimator of generalized Shannon's entropy on countable alphabets. The developed asymptotic properties require no assumptions on the original distribution. The proposed asymptotic properties allow for interval estimation and statistical tests with generalized Shannon's entropy.

Keywords: Shannon's entropy; generalized Shannon's entropy; plug-in estimation; asymptotic normality

Citation: Zhang, J.; Shi, J. Asymptotic Normality for Plug-In Estimators of Generalized Shannon's Entropy. *Entropy* **2022**, *24*, 683. https://doi.org/10.3390/e24050683

Academic Editor: Yong Deng

Received: 14 April 2022
Accepted: 11 May 2022
Published: 12 May 2022

1. Introduction

1.1. Introduction and Related Work

Shannon's entropy, introduced by [1], is one of the building blocks of Information Theory and a key aspect of Machine Learning (ML) methods (e.g., Random Forests). It is one of the most popular quantities on countable alphabet (An countable alphabet is a space that could be either finite, or countably infinite; the elements in an alphabet can be either ordinal (e.g., numbers) or non-ordinal (e.g., letters)), particularly on non-ordinal space with categorical data. For example, in [2], all reviewed feature selection methods on non-ordinal space boiled down to a function of Shannon's entropy. In addition, Shannon's entropy is one of the most important foundations for all tree-based ML algorithms, sometimes substitutable with the Gini impurity index [3–5].

As one of the essential information-theoretical quantities, Shannon's entropy and its estimation are widely studied in the past decades [6–12]. In particular, [9] proved that an unbiased estimator of Shannon's entropy does not exist. Current state-of-art Shannon's entropy point estimator was provided in [10] with the fastest bias decaying rate (exponentially-decaying).

Nevertheless, Shannon's entropy is only finitely defined for distributions with fast decaying tails [13].

It is never known if the real distribution yields a finite Shannon's entropy in practice. Furthermore, all existing results on Shannon's entropy require it to be finitely defined, which results in a usage restriction when adopting the entropy-based methods. This is, in fact, a void in the foundation of all Shannon's entropy-related results.

Example 1 (Unbounded Shannon's Entropy). *Let a distribution $\mathcal{P}$ be $\mathcal{P}_k = c/(k \ln^2 k)$ for $k \geq 2$, where c is the constant that makes $\mathcal{P}$ a valid probability distribution. Such c uniquely exists because $\sum_{k=2}^{\infty}[1/(k \ln^2 k)]$ converges. Then Shannon's entropy of $\mathcal{P}$, $H(\mathcal{P})$ is unbounded because*

$$\begin{aligned}
H(\mathcal{P}) &= -\sum_{k=2}^{\infty}[\mathcal{P}_k \ln \mathcal{P}_k] \\
&= -\sum_{k=2}^{\infty}\left[\frac{c}{k \ln^2 k} \ln \frac{c}{k \ln^2 k}\right] \\
&= -\sum_{k=2}^{\infty}\left[\frac{c}{k \ln^2 k} \ln c\right] + \sum_{k=2}^{\infty}\left[\frac{c}{k \ln^2 k} \ln k\right] + \sum_{k=2}^{\infty}\left[\frac{c}{k \ln^2 k} \ln\left(\ln^2 k\right)\right] \\
&= -\sum_{k=2}^{\infty}\left[\frac{c}{k \ln^2 k} \ln c\right] + \sum_{k=2}^{\infty}\left[\frac{c}{k \ln k}\right] + \sum_{k=2}^{\infty}\left[\frac{2c \ln \ln k}{k \ln^2 k}\right] \\
&= A\ Finite\ Value + \infty + A\ Finite\ Value = \infty.
\end{aligned}$$

The effort to generalize Shannon's entropy has been long and extensive in the existing literature. As summarized in [14], the main perspective in the generalization in the existing literature is based on axiomatic characterization of Shannon's entropy [15,16]. For example, Refs. [17,18] are efforts with respect to the functional form, $H = \sum_{k \geq 1} h(p_k)$, under certain desirable axioms, $h(p) = -p \log p$ is uniquely determined up to a multiplicative constant; if the strong additivity axiom is relaxed to be one of the weaker versions, say α-additivity or composability, then $h(p)$ may be of other forms, which give rise to Rényi's entropy [19], and the Tsallis entropy [20]. However, all such generalization effort does not seem to lead to an information measure on a joint alphabet that would possess all the desirable properties of mutual information, which is supported by an argument via the Kullback–Leibler divergence [21]. Interested readers may refer to [14] for details.

To further address the deficiency of Shannon's entropy [14] proposed generalized Shannon's entropy (GSE) and showed that GSE enjoys all properties of a finite Shannon's entropy. In addition, GSE is finitely defined on all distributions. Due to the advantages of GSE and the deficiency of Shannon's entropy, the use of Shannon's entropy should eventually be transited to GSE.

1.2. Summary and Contribution

To aid the transition, the estimation of GSE needs to be studied. In practice, the plug-in estimator is one of the most popular estimation approaches. For plug-in estimation of GSE, asymptotic properties are required for statistical tests and confidence intervals. This article studies the asymptotic properties for plug-in estimators of GSE.

As a summary of the article's results, Theorem 1 and Corollary 1 provide asymptotic normality properties for the plug-in estimators of GSE for all orders (An explanation of the order is given in Definition 2) on countably infinite alphabet. Corollary 2 provides the asymptotic normality properties for the plug-in estimators of GSE for all orders on finite alphabet, except the underlying distribution being uniform (Under a uniform distribution, the estimation of GSE is reduced to an estimation of population size. Interested readers can read [22]). The presented asymptotic normality properties immediately allow interval estimation and hypothesis testing with plug-in estimators of GSE. The numerical results in Section 3 show that the developed asymptotic properties converge fast, especially when the order is 2.

The presented properties and performance of GSE plug-in estimators suggest that GSE's use is full of promising potential. One may be concerned the construction of CDOTC (Defined in Definition 1) would bring additional estimation challenges to the already-difficult estimation of Shannon's entropy. Yet, the convergence speed for GSE plug-in estimators is fast. To further unlock the potentials of GSE, additional estimation methods of GSE and asymptotic properties of functions of GSE (e.g., Generalized Mutual Information,

also originated in [14]) shall be visited. This article's results and proofs' approaches provide a solid direction toward that end.

The rest of this paper is organized as follows. Section 2 formally states the problem and gives our main results. In Section 3, we provide a small-scale simulation study. In Section 4, we discuss the potential of GSE. Proofs are postponed to Section 5.

2. Main Results

Let Z be a random element on a countable alphabet $\mathscr{Z} = \{z_k; k \geq 1\}$ with an associated distribution $\mathbf{p} = \{p_k; k \geq 1\}$. Let the cardinality or support on $\mathscr{Z}$ be denoted $K = \sum_{k\geq 1} 1[p_k > 0]$, where $1[\cdot]$ is the indicator function. K is possibly finite or infinite. Let $\mathscr{P}$ denote the family of all distributions on $\mathscr{Z}$. Shannon's entropy, H, is defined as

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

To state our main result, we need to state Definitions 1 and 2 given by [14], and Definition 3.

Definition 1 (Conditional Distribution of Total Collision (CDOTC)). *Given $\mathscr{Z} = \{z_k; k \geq 1\}$ and $\mathbf{p} = \{p_k\}$, consider the experiment of drawing an identically and independently distributed (iid) sample of size m ($m \geq 2$). Let C_m denote the event that all observations of the sample take on a same letter in $\mathscr{Z}$, and let C_m be referred to as the event of total collision. The conditional probability, given C_m, that the total collision occurs at letter z_k is*

$$p_{m,k} = \frac{p_k^m}{\sum_{i\geq 1} p_i^m},$$

where $m \geq 2$. $\mathbf{p}_m = \{p_{m,k}\}$ is defined as the m-th order CDOTC.

Definition 2 (Generalized Shannon's Entropy (GSE)). *Given $\mathscr{Z} = \{z_k; k \geq 1\}$, $\mathbf{p} = \{p_k\}$, and $\mathbf{p}_m = \{p_{m,k}\}$, generalized Shannon's entropy (GSE) is defined as*

$$H_m(Z) = -\sum_{k\geq 1} p_{m,k} \ln p_{m,k},$$

where $p_{m,k}$ is defined in Definition 1, and $m = 2, 3, \ldots$ is the order of GSE. GSE with order m is referred to as the m-th order GSE.

It is clear that $\mathbf{p}_m$ is a probability distribution induced from $\mathbf{p} = \{p_k\}$. Furthermore, for each m, $m \geq 2$, $\mathbf{p}$ and $\mathbf{p}_m$ uniquely determined each other (Lemma 1 in [14]). To help understand Definitions 1 and 2, Examples 2 and 3 are provided as follows.

Example 2 (The 2nd order CDOTC). *Given $\mathscr{Z} = \{z_k; k \geq 1\}$ and $\mathbf{p} = \{p_k\} = \{6k^{-2}/\pi^2; k = 1, 2, 3, \ldots\}$, the 2nd order CDOTC is then defined as*

$$\mathbf{p}_2 = \{p_{2,k}\},$$

where

$$p_{2,k} = \frac{p_k^2}{\sum_{i\geq 1} p_i^2} = \frac{36k^{-4}/\pi^4}{\sum_{i\geq 1}[36i^{-4}/\pi^4]} = \frac{k^{-4}}{\sum_{i\geq 1} i^{-4}}$$

for $k = 1, 2, 3, \ldots$.

Example 3 (The 2nd order GSE). *Given $\mathscr{Z} = \{z_k; k \geq 1\}$, $\mathbf{p} = \{p_k\} = \{6k^{-2}/\pi^2; k = 1, 2, 3, \ldots\}$, and $\mathbf{p}_2 = \{p_{2,k}\} = \{\frac{k^{-4}}{\sum_{i\geq 1} i^{-4}}; k = 1, 2, \ldots\}$, the 2nd order GSE, $H_2(Z)$, is then defined as*

$$H_2(Z) = -\sum_{k\geq 1} p_{2,k} \ln p_{2,k},$$

where $p_{2,k}$ is given in Example 2.

The definition of the plug-in estimator of GSE is stated in Definition 3.

Definition 3 (Plug-in estimator of GSE). *Let $Z_1, Z_2, \ldots, Z_n$ be independent and identically distributed (iid) random variables taking values in $\mathscr{Z} = \{z_k; k \geq 1\}$ with distribution $\mathbf{p}$. For each $k = 1, 2, \ldots$, let $Y_k = \sum_{i=1}^{n} 1[Z_i = z_k]$ be the sample count of observations in category z_k, and let $\hat{p}_k = Y_k/n$ be the sample proportion. The plug-in estimator for the m-th order GSE, $\hat{H}_m(Z)$, is defined as*

$$\begin{aligned}\hat{H}_m(Z) &= -\sum_{k\geq 1}[\hat{p}_{m,k} \ln \hat{p}_{m,k}] \\ &= -\sum_{k\geq 1}\left[\frac{\hat{p}_k^m}{\sum_{i\geq 1}\hat{p}_i^m} \ln \frac{\hat{p}_k^m}{\sum_{i\geq 1}\hat{p}_i^m}\right].\end{aligned}$$

Our main results are stated in Theorem 1, Corollary 1 and 2.

Theorem 1. *Let $\mathbf{p} = \{p_k\}$ be a probability distribution supported by a countably infinite alphabet, without any further conditions,*

$$\sqrt{n}\big(\hat{H}_m(Z) - H_m(Z)\big) \xrightarrow{d} N(0, \sigma_m^2),$$

where

$$\sigma_m^2 = \sum_{k=1}^{\infty}\left[\frac{m^2}{p_k}\left(p_{m,k} \ln p_{m,k} + p_{m,k} H_m(Z)\right)\right]^2.$$

Corollary 1. *Let $\mathbf{p} = \{p_k\}$ be a probability distribution supported by a countably infinite alphabet, without any further conditions,*

$$\sqrt{n}\left(\frac{\hat{H}_m(Z) - H_m(Z)}{\hat{\sigma}_m}\right) \xrightarrow{d} N(0, 1),$$

where

$$\hat{\sigma}_m^2 = \sum_{k=1}^{\infty}\left[\frac{m^2}{\hat{p}_k}\left(\hat{p}_{m,k} \ln \hat{p}_{m,k} + \hat{p}_{m,k} \hat{H}_m(Z)\right)\right]^2. \tag{1}$$

Corollary 2. *Let $\mathbf{p} = \{p_k; k = 1, 2, \ldots, K\}$ be a non-uniform probability distribution on a finite alphabet, without any further conditions,*

$$\sqrt{n}\left(\frac{\hat{H}_m(Z) - H_m(Z)}{\hat{\sigma}_m}\right) \xrightarrow{d} N(0, 1),$$

where

$$\hat{\sigma}_m^2 = \sum_{k=1}^{K}\left[\frac{m^2}{\hat{p}_k}\left(\hat{p}_{m,k} \ln \hat{p}_{m,k} + \hat{p}_{m,k} \hat{H}_m(Z)\right)\right]^2.$$

Corollary 2 is a special case of Theorem 1. All proofs are provided in Section 5.

3. Simulations

One of the main applications of our results is the ability to construct confidence intervals, and hence testing hypothesis. Specifically, Corollary 1 implies that an asymptotic $(1-\alpha)100\%$ confidence interval for H_m is given by

$$\left(\hat{H}_m - z_{\alpha/2}\frac{\hat{\sigma}_m}{\sqrt{n}}, \hat{H}_m + z_{\alpha/2}\frac{\hat{\sigma}_m}{\sqrt{n}}\right), \tag{2}$$

where $\hat{\sigma}_m$ is given by (1) and $z_{\alpha/2}$ is a number such that $P(Z > z_{\alpha/2}) = \alpha/2$ and $Z \sim N(0,1)$. In this section, we give a small scale simulation study to check the finite sample performance of this confidence interval.

We consider Zeta distribution

$$P(x = k) = \frac{1}{\zeta(s)}k^{-s}, \quad k \in \{1, 2, \dots\}$$

with $s = 1.5$ and 2.5, where $\zeta(s)$ is the Riemann zeta function given by

$$\zeta(s) = \sum_{n=1}^{\infty}\frac{1}{n^s}.$$

The simulations were performed as follows. For the given distribution, we obtained a random sample of size n and used it to evaluate a 95% confidence interval for a given index using (2). We then checked to see if the true value of the H_m was in the interval or not. This was repeated 5000 times, and the proportion of times when the true value was in the interval was calculated. When the asymptotics works well, this proportion should be close to 0.95. We repeated this for sample sizes ranging from 10 to 1000 in increments of 10. The results for $s = 1.5$, order $m = 2$ and $m = 3$ are given in Figures 1 and 2; the results for $s = 2.5$, order $m = 2$ and $m = 3$ are given in Figures 3 and 4.

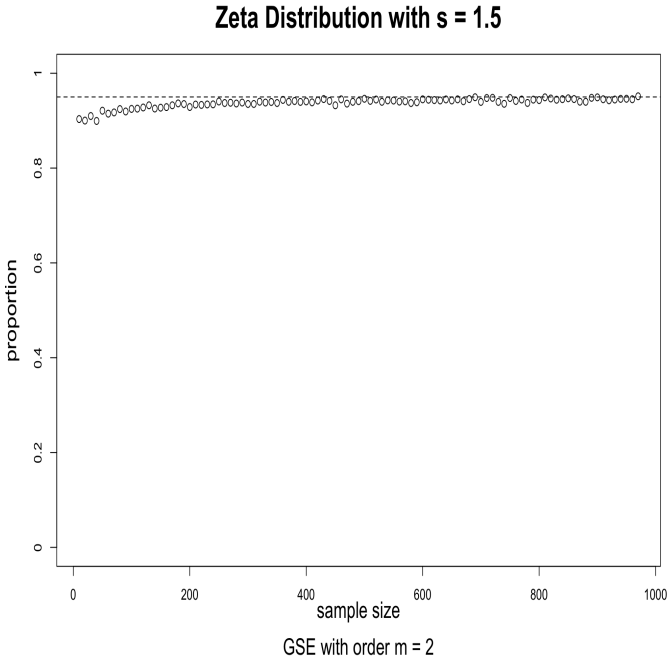

Figure 1. Effectiveness of the 95% confidence intervals as a function of sample size. Simulations from Zeta distribution with $s = 1.5$ and GSE with order $m = 2$. The horizontal dashed line is at 0.95.

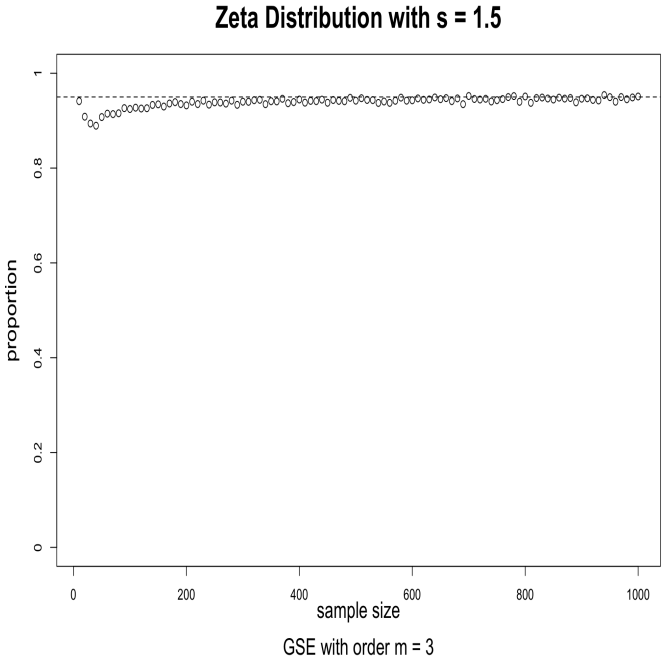

Figure 2. Effectiveness of the 95% confidence intervals as a function of sample size. Simulations from Zeta distribution with $s = 1.5$ and GSE with order $m = 3$. The horizontal dashed line is at 0.95.

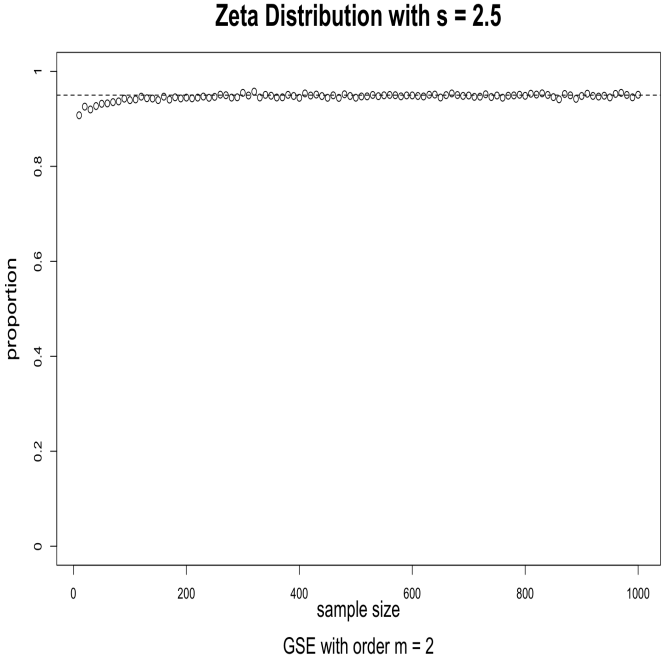

Figure 3. Effectiveness of the 95% confidence intervals as a function of sample size. Simulations from Zeta distribution with $s = 2.5$ and GSE with order $m = 2$. The horizontal dashed line is at 0.95.

The results suggest that convergence is fast, particularly when the order is $m = 2$. We conjecture that this may be caused by the fact that, when m is larger, the probabilities in the corresponding CDOTC are smaller and hence require a larger sample size for convergence. For the same reason, the results with $s = 1.5$ converge faster than that of $s = 2.5$, because $s = 2.5$ yields a thinner tail distribution which requires a larger sample size for convergence. Although GSE with order $m \geq 3$ may shed some light on specific information, GSE with order $m = 2$ is enough to well exist with asymptotic properties for any valid underlying probability distribution $\mathbf{p}$.

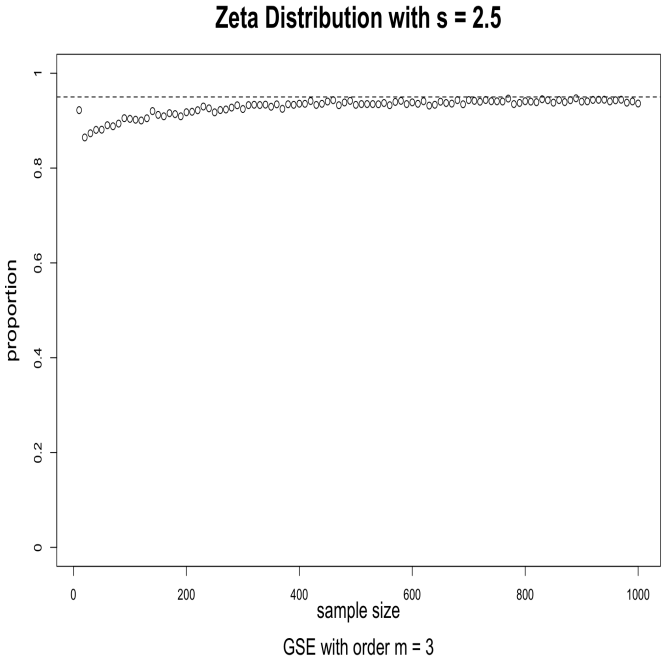

Figure 4. Effectiveness of the 95% confidence intervals as a function of sample size. Simulations from Zeta distribution with $s = 2.5$ and GSE with order $m = 3$. The horizontal dashed line is at 0.95.

4. Discussion

The proposed asymptotic properties in Corollary 1 and 2 make it possible for interval estimation and statistical tests. Based on the simulation results, the convergence is quite fast, particularly under order $m = 2$. Note that a GSE with order $m = 2$ already enjoys all asymptotic properties without any assumption on original distribution $\mathbf{p}$.

We recommend using GSE with order $m = 2$ in place of Shannon's entropy in all entropy-based methods when applicable. By replacing Shannon's entropy with GSE, one still enjoys all the benefits of Shannon's entropy with a fast convergence speed. Moreover, using GSE is risk-free compared to Shannon's entropy because Shannon's entropy (1) does not exist on some thick-tailed distributions and (2) requires thinner tail distribution for some asymptotic properties [11]. Additional research is required to aid the transition. The proposed asymptotic results allow interval estimation and statistical tests on the modified entropy-based methods that replaced Shannon's entropy with GSE. Future research should aim to provide additional estimation methods of GSE and statistical properties of functions of GSE, such as GMI. The proposed asymptotic properties in this article directly provide asymptotic normality for the plug-in estimator of GMI when the real underlying GMI is not 0. The asymptotic behavior for the plug-in estimator of GMI when the real underlying GMI is 0 remains an open question, which we will address in future work.

5. Proofs

The proofs require several lemmas. The first lemma is state below.

Lemma 1 ([11,23]). *Assume that* $\sum_{k=1}^{\infty} p_k |\log p_k|^2 < \infty$ *and that there is a deterministic sequence* $K(n)$ *with* $K(n) \to \infty$ *such that* $\lim_{n\to\infty} K(n)/\sqrt{n} \to 0$ *and*

$$\lim_{n\to\infty} \sqrt{n} \sum_{k=K(n)}^{\infty} p_k \log p_k = 0.$$

In this case

$$\sqrt{n}\left(\hat{H}_n - H\right) \xrightarrow{d} N\left(0, \sigma^2\right),$$

where

$$\sigma^2 = \sum_{k=1}^{\infty} p_k (\log p_k)^2 - \left(\sum_{k=1}^{\infty} p_k \log p_k \right)^2.$$

Furthermore, if $\sigma > 0$,

$$\sqrt{n}\left(\frac{\hat{H}_n - H}{\hat{\sigma}_n} \right) \xrightarrow{d} N(0,1)$$

where

$$\hat{\sigma}_n^2 = \sum_{k=1}^{\infty} \hat{p}_k (\log \hat{p}_k)^2 - \left(\sum_{k=1}^{\infty} \hat{p}_k \log \hat{p}_k \right)^2.$$

Different proofs of Lemma 1 are provided in [11,23].

The spirit for proof of Theorem 1 is to regard CDOTC as an original distribution and utilize the result from Lemma 1. Toward that end, several lemmas are needed and stated below.

Lemma 2 (Equivalent conditions in Lemma 1)**.** *For any valid distribution* $\mathbf{p}$*, let the corresponding CDOTC with order* m *be denoted as* $\mathbf{p}_m$*, then*

$$\sum_{k=1}^{\infty} p_{m,k} \left| \log p_{m,k} \right|^2 < \infty$$

and that there is a deterministic sequence $K(n)$ *with* $K(n) \to \infty$ *such that* $\lim_{n\to\infty} K(n)/\sqrt{n} \to 0$ *and*

$$\lim_{n\to\infty} \sqrt{n} \sum_{k=K(n)}^{\infty} p_{m,k} \log p_{m,k} = 0.$$

Lemma 3 (σ_m^2 in Theorem 1)**.** *In Theorem 1,*

$$\sigma_m^2 = \sum_{k=1}^{\infty} \left[\frac{m^2}{p_k} \left(p_{m,k} \ln p_{m,k} + p_{m,k} H_m(Z) \right) \right]^2.$$

Lemma 4 ($\hat{\sigma}_m^2$ in Corollary 1)**.** *In Corollary 1,*

$$\hat{\sigma}_m^2 = \sum_{k=1}^{\infty} \left[\frac{m^2}{\hat{p}_k} \left(\hat{p}_{m,k} \ln \hat{p}_{m,k} + \hat{p}_{m,k} \hat{H}_m(Z) \right) \right]^2.$$

Proof of Lemma 2. Note that for any $\mathbf{p}$ to be a valid distribution, the tail of $\mathbf{p}$ must be thicker than $1/(k \ln k)$ because $\sum_{k\geq 2} 1/(k \ln k)$ diverges. Hence $\mathbf{p}_m$ is thicker than $1/(k^2 \ln^2 k)$ for any $m \geq 2$ by definition. It is shown in Example 3 of [11] that such tail satisfies the mentioned conditions. □

Proof of Lemma 3. Because of Lemma 2, σ^2 can be obtained under finite K and then let $K \to \infty$. For a finite K, it can be verified that for $i = 1, \ldots, K-1$,

$$\frac{\partial H_m}{\partial p_i} = (\ln p_{m,K} - \ln p_{m,i}) \frac{m p_{m,i}}{p_i} - m \left(\frac{p_{m,i}}{p_i} - \frac{p_{m,K}}{p_K} \right) (H_m + \ln p_{m,K}).$$

Let

$$v = (p_1, \cdots, p_{K-1})^\tau,$$
$$\hat{v} = (\hat{p}_1, \cdots, \hat{p}_{K-1})^\tau.$$

We have $\sqrt{n}(\hat{v} - v) \xrightarrow{L} MVN(0, \Sigma(v))$, where $\Sigma(v)$ is the $(K-1) \times (K-1)$ covariance matrix given by

$$\Sigma(v) = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{K-1} \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ -p_{K-1}p_1 & -p_{K-1}p_2 & \cdots & p_{K-1}(1-p_{K-1}) \end{pmatrix}$$

According to the first-order Delta method,

$$\sigma_K^2 = \nabla H_m^T \Sigma \nabla H_m = \sum_{k=1}^{K} \left[\frac{m^2}{p_k} (p_{m,k} \ln p_{m,k} + p_{m,k} H_m(Z)) \right]^2.$$

Given Lemma 2, let $K \to \infty$,

$$\sigma^2 = \sum_{k=1}^{\infty} \left[\frac{m^2}{p_k} (p_{m,k} \ln p_{m,k} + p_{m,k} H_m(Z)) \right]^2.$$

□

Proof of Lemma 4. Lemma 4 is because of $\hat{\sigma}_m^2 \xrightarrow{p} \sigma_m^2$. □

Proof of Theorem 1 and Corollary 1. With Lemmas 1–4, and Slutsky's theorem, Theorem 1 and Corollary 1 are proved. □

Proof of Corollary 2. Corollary 2 is a directly result of Theorem 1, except under uniform distribution when $\nabla H_m = 0$ for all $m \geq 2$. □

Author Contributions: Conceptualization, J.Z. and J.S.; methodology, J.Z. and J.S.; simulation, J.Z. and J.S.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We thank Zhiyi Zhang at the University of North Carolina at Charlotte for their profound discussions on the subject of Entropic Statistics.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CDOTC	Conditional Distribution of Total Collision
GMI	Generalized Mutual Information
GSE	Generalized Shannon's Entropy
ML	Machine Learning

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
2. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–45. [CrossRef]
3. Banerjee, M.; Reynolds, E.; Andersson, H.B.; Nallamothu, B.K. Tree-based analysis: a practical approach to create clinical decision-making tools. *Circ. Cardiovasc. Qual. Outcomes* **2019**, *12*, e004879. [CrossRef] [PubMed]
4. Mienye, I.D.; Sun, Y.; Wang, Z. Prediction performance of improved decision tree-based algorithms: A review. *Procedia Manuf.* **2019**, *35*, 698–703. [CrossRef]
5. Hssina, B.; Merbouha, A.; Ezzikouri, H.; Erritali, M. A comparative study of decision tree ID3 and C4. 5. *Int. J. Adv. Comput. Sci. Appl.* **2014**, *4*, 13–19.

6. Miller, G.A.; Madow, W.G. *On the Maximum Likelihood Estimate of the Shannon-Weiner Measure of Information*; Operational Applications Laboratory, Air Force Cambridge Research Center: Bedford, MA, USA, 1954.
7. Harris, B. *The Statistical Estimation of Entropy in the Non-Parametric Case*; Technical Report; Wisconsin Univ-Madison Mathematics Research Center: Madison, WI, USA, 1975.
8. Esty, W.W. A normal limit law for a nonparametric estimator of the coverage of a random sample. *Ann. Stat.* **1983**, *11*, 905–912. [CrossRef]
9. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
10. Zhang, Z. Entropy estimation in Turing's perspective. *Neural Comput.* **2012**, *24*, 1368–1389. [CrossRef] [PubMed]
11. Zhang, Z.; Zhang, X. A normal law for the plug-in estimator of entropy. *IEEE Trans. Inf. Theory* **2012**, *58*, 2745–2747. [CrossRef]
12. Zhang, Z. Asymptotic normality of an entropy estimator with exponentially decaying bias. *IEEE Trans. Inf. Theory* **2013**, *59*, 504–508. [CrossRef]
13. Baccetti, V.; Visser, M. Infinite shannon entropy. *J. Stat. Mech. Theory Exp.* **2013**, *2013*, P04010. [CrossRef]
14. Zhang, Z. Generalized Mutual Information. *Stats* **2020**, *3*, 158–165. [CrossRef]
15. Amigó, J.M.; Balogh, S.G.; Hernández, S. A brief review of generalized entropies. *Entropy* **2018**, *20*, 813. [CrossRef] [PubMed]
16. Csiszár, I. Axiomatic characterizations of information measures. *Entropy* **2008**, *10*, 261–273. [CrossRef]
17. Khinchin, A.Y. *Mathematical Foundations of Information Theory*; Courier Corporation: Chelmsford, MA, USA, 2013.
18. Chakrabarti, C.; Chakrabarty, I. Shannon entropy: axiomatic characterization and application. *Int. J. Math. Math. Sci.* **2005**, *2005*, 2847–2854. [CrossRef]
19. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkley, CA, USA, 1 January 1961; University of California Press: Los Angeles, CA, USA, 1961; Volume 4, pp. 547–562.
20. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. [CrossRef]
21. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
22. Zhang, Z.; Chen, C.; Zhang, J. Estimation of population size in entropic perspective. *Commun.-Stat.-Theory Methods* **2020**, *49*, 307–324. [CrossRef]
23. Grabchak, M.; Zhang, Z. Asymptotic normality for plug-in estimators of diversity indices on countable alphabets. *J. Nonparametr. Stat.* **2018**, *30*, 774–795. [CrossRef]