entropy

MDPI

Article

Normal Laws for Two Entropy Estimators on Infinite Alphabets

Chen Chen, Michael Grabchak, Ann Stewart, Jialin Zhang * and Zhiyi Zhang

Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA; cchen55@uncc.edu (C.C.); mgrabcha@uncc.edu (M.G.); astewa79@uncc.edu (A.S.); zzhang@uncc.edu (Z.Z.)

* Correspondence: jzhang51@uncc.edu

Received: 3 April 2018; Accepted: 10 May 2018; Published: 17 May 2018

Abstract: This paper offers sufficient conditions for the Miller–Madow estimator and the jackknife estimator of entropy to have respective asymptotic normalities on countably infinite alphabets.

Keywords: entropy; nonparametric estimator; Miller–Madow estimator; jackknife estimator; asymptotic normality

MSC: Primary 62F10; 62F12; 62G05; 62G20

1. Introduction

Let $\mathscr{X} = \{\ell_k; k \geq 1\}$ be a finite or countably infinite alphabet, let $\mathbf{p} = \{p_k; k \geq 1\}$ be a probability distribution on $\mathscr{X}$, and define $K = \sum_{k\geq 1} 1[p_k > 0]$, where $1[\cdot]$ is the indicator function, to be the effective cardinality of $\mathscr{X}$ under $\mathbf{p}$. An important quantity associated with $\mathbf{p}$ is entropy, which is defined by [1] as

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

Here and throughout, we adopt the convention that $0 \ln 0 = 0$.

Many properties of entropy and related quantities are discussed in [2]. The problem of statistical estimation of entropy has a long history (see the survey paper [3] or the recent book [4]). It is well-known that no unbiased estimators of entropy exist, and, for this reason, much energy has been focused on deriving estimators with relatively little bias (see [5] and the references therein for a discussion of some (but far from all) of these). Perhaps the most commonly used estimator is the plug-in. Its theoretical properties have been studied going back, at least, to [6], where conditions for consistency and asymptotic normality, in the case of finite alphabets, were derived. It would be almost fifty years before corresponding conditions for the countabe case would appear in the literature. Specifically, consistency, both in terms of almost sure and L^2 convergence, was verified in [7]. Later, sufficient conditions for asymptotic normality were derived in two steps in [3,8].

Despite a simple form and nice theoretical properties, the plug-in suffers from large finite sample bias, which has led to the development of modifications that aim to reduce this bias. Two of the most popular are the Miller–Madow estimator of [6] and the jackknife estimator of [9]. Theoretical properties of these have not been studied, as extensively, in the literature. In this paper, we give sufficient conditions for the asymptotic normality of these two estimators. This is important for deriving confidence intervals and hypothesis tests, and it immediately implies consistency (see e.g., [4]).

We begin by introducing some notation. We say that a distribution $\mathbf{p} = \{p_k; k \geq 1\}$ is uniform if and only if its effective cardinality $K < \infty$ and for each $k = 1, 2, \ldots$ either $p_k = 1/K$ or $p_k = 0$. We write $f \sim g$ to denote $\lim_{n\to\infty} f(n)/g(n) = 1$ and we write $f = \mathcal{O}(g(n))$ to denote $\limsup_{n\to\infty} |f(n)/g(n)| < \infty$. Furthermore, we write $\xrightarrow{L}$ to denote convergence in law and $\xrightarrow{p}$ to

Entropy **2018**, *20*, 371; doi:10.3390/e20050371 www.mdpi.com/journal/entropy

denote convergence in probability. If a and b are real number, we write $a \vee b$ to denote the maximum of a and b. When it is not specified, all limits are assumed to be taken as $n \to \infty$.

Let $X_1, \dots, X_n$ be independent and identically distributed (*iid*) random variables on $\mathscr{X}$ under $\mathbf{p}$. Let $\{Y_k; k \geq 1\}$ be the observed letter counts in the sample, i.e., $Y_k = \sum_{i=1}^n 1[X_i = \ell_k]$, and let $\hat{\mathbf{p}} = \{\hat{p}_k; k \geq 1\}$, where $\hat{p}_k = Y_k/n$, be the corresponding relative frequencies. Perhaps the most intuitive estimator of H is the plug-in, which is given by

$$\hat{H} = -\sum_{k \geq 1} \hat{p}_k \ln \hat{p}_k. \tag{2}$$

When the effective cardinality, K, is finite, [10] showed that the bias of $\hat{H}$ is

$$\mathrm{E}(\hat{H}) - H = -\frac{K-1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{k=1}^{K} \frac{1}{p_k}\right) + \mathcal{O}\left(n^{-3}\right). \tag{3}$$

One of the simplest and earliest approaches aiming to reduce the bias of $\hat{H}$ is to estimate the first order term. Specifically, let $\hat{m} = \sum_{k \geq 1} 1[Y_k > 0]$ be the number of letters observed in the sample and consider an estimator of the form,

$$\hat{H}_{MM} = \hat{H} + \frac{\hat{m}-1}{2n}. \tag{4}$$

This estimator is often attributed to [6] and is known as the Miller–Madow estimator. Note that, for finite K,

$$\mathrm{E}\left(\frac{\hat{m}-1}{2n}\right) = \frac{K-1}{2n} - \frac{\sum_k (1-p_k)^n}{2n}.$$

Since $\sum_k (1-p_k)^n \leq K(1-p_\wedge)^n$, where $p_\wedge = \min\{p_k : p_k > 0\}$, decays exponentially fast, it follows that, for finite K, the bias of $\hat{H}_{MM}$ is

$$\mathrm{E}(\hat{H}_{MM}) - H = \frac{1}{12n^2}\left(1 - \sum_{k=1}^{K} \frac{1}{p_k}\right) + \mathcal{O}\left(n^{-3}\right).$$

Among the many estimators in the literature aimed at reducing bias in entropy estimation, the Miller–Madow estimator is one of the most commonly used. Its popularity is due to its simplicity, its intuitive appeal, and, more importantly, its good performance across a wide range of different distributions including those on countably infinite alphabets. See, for instance, the simulation study in [5].

The jackknife entropy estimator is another commonly used estimator designed to reduce the bias of the plug-in. It is calculated in three steps:

1. for each $i \in \{1, 2, \dots, n\}$ construct $\hat{H}^{(i)}$, which is a plug-in estimator based on a sub-sample of size $n-1$ obtained by leaving the ith observation out;
2. obtain $\hat{H}_{(i)} = n\hat{H} - (n-1)\hat{H}^{(i)}$ for $i = 1, \cdots, n$; and then
3. compute the jackknife estimator

$$\hat{H}_{JK} = \frac{\sum_{i=1}^n \hat{H}_{(i)}}{n}. \tag{5}$$

Equivalently, (5) can be written as

$$\hat{H}_{JK} = n\hat{H} - (n-1)\frac{\sum_{i=1}^n \hat{H}^{(i)}}{n}. \tag{6}$$

The jackknife estimator of entropy was first described by [9]. From (2), it may be verified that, when $K < \infty$, the bias of $\hat{H}_{JK}$ is

$$\mathrm{E}\left(\hat{H}_{JK}\right) - H = \mathcal{O}\left(n^{-2}\right). \tag{7}$$

Both the Miller–Madow and the jackknife estimators are adjusted versions of the plug-in. When the effective cardinality is finite, i.e., $K < \infty$, the asymptotic normalities of both can be easily verified. A question of theoretical interest is whether these normalities still hold when the effective cardinality is countably infinite. In this paper, we give sufficient conditions for $\sqrt{n}(\hat{H}_{MM} - H)$ and $\sqrt{n}(\hat{H}_{JK} - H)$ to have asymptotic normalities on countably infinite alphabets and provide several illustrative examples. The rest of paper is organized as follows. Our main results for both the Miller–Madow and the jackknife estimators are given in Section 2. A small simulation study is given in Section 3. This is followed by a brief discussion in Section 4. Proofs are postponed to Section 5.

2. Main Results

We begin by recalling a sufficient condition due to [8] for the asymptotic normality of the plug-in estimator.

Condition 1. *The distribution,* $\mathbf{p} = \{p_k; k \geq 1\}$, *satisfies*

$$\sum_{k\geq 1} p_k \ln^2 p_k < \infty, \tag{8}$$

and there exists an integer-valued function $K(n)$ *such that, as* $n \to \infty$,

1. $K(n) \to \infty$,
2. $K(n)/\sqrt{n} \to 0$, *and*
3. $\sqrt{n}\sum_{k\geq K(n)} p_k \ln p_k \to 0$.

Note that, by Jensen's inequality (see e.g., [2]), (8) implies that

$$H^2 = \left(-\sum_{k\geq 1} p_k \ln p_k\right)^2 \leq \sum_{k\geq 1} p_k \ln^2 p_k < \infty,$$

where equality holds, i.e., $H^2 = \sum_{k\geq 1} p_k \ln^2 p_k$, if and only if $\mathbf{p}$ is a uniform distribution. Thus, when (8) holds, we have $H < \infty$. The following result is given in [8].

Lemma 1. *Let* $\mathbf{p} = \{p_k; k \geq 1\}$ *be a distribution, which is not uniform, and set*

$$\sigma^2 = \sum_{k\geq 1} p_k \ln^2 p_k - H^2 \text{ and } \hat{\sigma}^2 = \sum_{k\geq 1} \hat{p}_k \ln^2 \hat{p}_k - \hat{H}^2. \tag{9}$$

If $\mathbf{p}$ *satisfies Condition 1, then* $\hat{\sigma} \xrightarrow{p} \sigma$,

$$\frac{\sqrt{n}(\hat{H} - H)}{\sigma} \xrightarrow{L} N(0,1),$$

and

$$\frac{\sqrt{n}(\hat{H} - H)}{\hat{\sigma}} \xrightarrow{L} N(0,1).$$

The following is useful for checking when Condition 1 holds.

Lemma 2. *Let* $\mathbf{p} = \{p_k; k \geq 1\}$ *and* $\mathbf{p}' = \{p'_k; k \geq 1\}$ *be two distributions and assume that* $\mathbf{p}'$ *satisfies Condition 1. If there exists a* $C > 0$ *such that, for large enough* k,

$$p_k \leq C p'_k,$$

then $\mathbf{p}$ *satisfies Condition 1 as well.*

In [8], it is shown that Condition 1 holds for $\mathbf{p} = \{p_k; k \geq 1\}$ with

$$p_k = \frac{C}{k^2 \ln^2 k}, \quad k = 1, 2, \ldots,$$

where $C > 0$ is a normalizing constant. It follows from Lemma 2 that any distribution with tails lighter than this satisfies Condition 1 as well.

We are interested in finding conditions under which the result of Lemma 1 can be extended to bias adjusted modifications of $\hat{H}$. Let $\hat{H}_*$ be any bias-adjusted estimator of the form

$$\hat{H}_* = \hat{H} + \hat{B}_*, \tag{10}$$

where $\hat{B}_*$ is an estimate of the bias. Combining Lemma 1 with Slutsky's theorem immediately gives the following.

Theorem 1. *Let $\mathbf{p} = \{p_k; k \geq 1\}$ be a distribution, which is not uniform, and let σ^2 and $\hat{\sigma}^2$ be as in (9). If Condition 1 holds and $\sqrt{n}\hat{B}_* \xrightarrow{p} 0$, then $\hat{\sigma} \xrightarrow{p} \sigma$,*

$$\frac{\sqrt{n}(\hat{H}_* - H)}{\sigma} \xrightarrow{L} N(0,1),$$

and

$$\frac{\sqrt{n}(\hat{H}_* - H)}{\hat{\sigma}} \xrightarrow{L} N(0,1).$$

For the Miller–Madow estimator and the jackknife estimator, respectively, the bias correction term, $\hat{B}_*$, in (10) takes the form

$$\text{Miller–Madow:} \qquad \hat{B}_{MM} = \frac{\hat{m} - 1}{2n},$$

$$\text{Jackknife:} \qquad \hat{B}_{JK} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{H} - \hat{H}^{(i)} \right).$$

Below, we give sufficient conditions for when $\sqrt{n}\hat{B}_{MM} \xrightarrow{p} 0$ and when $\sqrt{n}\hat{B}_{JK} \xrightarrow{p} 0$.

2.1. Results for the Miller–Madow Estimator

Condition 2. *The distribution, $\mathbf{p} = \{p_k; k \geq 1\}$, satisfies that, for sufficiently large k,*

$$p_k \leq \frac{1}{a(k)b(k)k^3}, \tag{11}$$

where $a(k) > 0$ and $b(k) > 0$ are two sequences such that

1. *$a(k) \to \infty$ as $k \to \infty$, and, furthermore,*
   - *(a) the function $a(k)$ is eventually nondecreasing, and*
   - *(b) there exists an $\varepsilon > 0$ such that*

$$\limsup_{k \to \infty} \frac{(a(k))^{2\varepsilon}}{a\left(\frac{\sqrt{k}}{(a(k))^{\varepsilon}}\right)} < \infty; \tag{12}$$

2.

$$\sum_{k \geq 1} \frac{1}{kb(k)} < \infty. \tag{13}$$

Since this condition only requires that p_k, for sufficiently large k, is upper bounded in the appropriate way, we immediately get the following.

Lemma 3. *Let $\mathbf{p} = \{p_k; k \geq 1\}$ and $\mathbf{p}' = \{p'_k; k \geq 1\}$ be two distributions and assume that $\mathbf{p}'$ satisfies Condition 2. If there exists a $C > 0$ such that, for large enough k,*

$$p_k \leq C p'_k,$$

then $\mathbf{p}$ satisfies Condition 2 as well.

We now give our main results for the Miller–Madow Estimator.

Theorem 2. *Let $\mathbf{p} = \{p_k; k \geq 1\}$ be a distribution, which is not uniform, and let σ^2 and $\hat{\sigma}^2$ be as in (9). If Condition 2 holds, then $\hat{\sigma} \xrightarrow{p} \sigma$,*

$$\frac{\sqrt{n}(\hat{H}_{MM} - H)}{\sigma} \xrightarrow{L} N(0,1)$$

and

$$\frac{\sqrt{n}(\hat{H}_{MM} - H)}{\hat{\sigma}} \xrightarrow{L} N(0,1).$$

In the proof of the theorem, we will show that Condition 2 implies that Condition 1 holds. Condition 2 requires p_k to decay slightly faster than k^{-3} by two factors $1/a(k)$ and $1/b(k)$, where $a(k)$ and $b(k)$ satisfy (12) and (13) respectively. While (13) is clear in its implication on $b(k)$, (12) is much less so on $a(k)$. To have a better understanding of (12), we give an important situation where (12) holds. Consider the case $a(n) = \ln n$. In this case, for any $\varepsilon \in (0, 0.5)$

$$\frac{(a(n))^{2\varepsilon}}{a\left(\frac{\sqrt{n}}{(a(n))^{\varepsilon}}\right)} = \frac{(\ln n)^{2\varepsilon}}{0.5 \ln n - \varepsilon \ln \ln n} \sim \frac{(\ln n)^{2\varepsilon}}{0.5 \ln n} \longrightarrow 0.$$

We now give a more general situation, which shows just how slow $a(k)$ can be. First, we recall the iterated logarithm function. Define $\ln^{(r)}(x)$, recursively for sufficiently large $x > 0$, by $\ln^{(0)}(x) = x$ and $\ln^{(r)}(x) = \ln\left(\ln^{(r-1)} x\right)$ for $r \geq 1$. By induction, it can be shown that $\frac{\mathrm{d}}{\mathrm{d}x} \ln^{(r)}(x) = \left(\prod_{i=0}^{r-1} \ln^{(i)}(x)\right)^{-1}$ for $r \geq 1$.

Lemma 4. *The function $a(n) = \ln^{(r)}(n)$ satisfies (12) with $\varepsilon = 0.5$ for any $r \geq 2$.*

We now give three examples.

Example 1. *Let $\mathbf{p} = \{p_k; k \geq 1\}$ be such that for sufficiently large k,*

$$p_k \leq \frac{C}{k^3 (\ln k)(\ln \ln k)^{2+\varepsilon}},$$

where $\varepsilon > 0$ and $C > 0$ are fixed constants. In this case, Condition 2 holds with $a(k) = \ln \ln k$ and $b(k) = (\ln k)(\ln \ln k)^{1+\varepsilon}/C$ in (11).

We can consider a more general form, which allows for even heavier tails.

Example 2. *Let r be an integer with $r \geq 2$ and let $\mathbf{p} = \{p_k; k \geq 1\}$ be such that, for sufficiently large k,*

$$p_k \leq \frac{C}{k^3 \left(\prod_{i=1}^{r-1} \ln^{(i)} k\right) (\ln^{(r)} k)^{2+\varepsilon}}$$

where $\varepsilon > 0$ and $C > 0$ are fixed constants. In this case, Condition 2 holds with $a(k) = \ln^{(r)} k$ and $b(k) = \left(\prod_{i=1}^{r-1} \ln^{(i)} k\right) (\ln^{(r)} k)^{1+\varepsilon}/C$ in (11). The fact that $b(k)$ satisfies (13) follows by the integral test for convergence.

It follows from Lemma 3 that any distribution with tails lighter than those in this example must satisfy Condition 2. On the other hand, the tails cannot get too much heavier.

Example 3. *Let $\mathbf{p} = \{p_k; k \geq 1\}$ be such that $p_k = Ck^{-3}$, where $C > 0$ is a normalizing constant. In this case, Condition 2 does not hold. However, Condition 1 does hold.*

2.2. Results for the Jackknife Estimator

For any distribution $\mathbf{p}$, let $B_n = \mathrm{E}(\hat{H}) - H$ be the bias of the plug-in based on a sample of size n.

Condition 3. *The distribution, $\mathbf{p} = \{p_k; k \geq 1\}$, satisfies*

$$\lim_{n\to\infty} n^{3/2} \left(B_n - B_{n-1}\right) = 0.$$

Theorem 3. *Let $\mathbf{p} = \{p_k; k \geq 1\}$ be a distribution, which is not uniform, and let σ^2 and $\hat{\sigma}^2$ be as in (9). If Conditions 1 and 3 hold, then $\hat{\sigma} \xrightarrow{p} \sigma$,*

$$\frac{\sqrt{n}(\hat{H}_{JK} - H)}{\sigma} \xrightarrow{L} N(0,1)$$

and

$$\frac{\sqrt{n}(\hat{H}_{JK} - H)}{\hat{\sigma}} \xrightarrow{L} N(0,1).$$

It is not clear to us whether Conditions 1 and 3 are equivalent or, if not, which is more stringent. For that reason, in the statement of Theorem 3, both conditions are imposed. The proof of the theorem uses the following lemma, which gives some insight into $\hat{B}_{JK}$ and Condition 3.

Lemma 5. *For any probability distribution $\mathbf{p} = \{p_k; k \geq 1\}$, we have*

$$\hat{B}_{JK} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{H} - \hat{H}^{(i)}\right) \geq 0$$

and

$$\mathrm{E}\left[\hat{B}_{JK}\right] = (n-1)\left(B_n - B_{n-1}\right) \geq 0.$$

We now give a condition, which implies Condition 3 and tends to be easier to check.

Proposition 1. *If the distribution $\mathbf{p} = \{p_k; k \geq 1\}$ is such that there exists an $\varepsilon \in (1/2, 1)$ with $\sum_{k\geq 1} p_k^{1-\varepsilon} < \infty$, then Condition 3 is satisfied.*

We now give an example where this holds.

Example 4. *Let $\mathbf{p} = \{Ck^{-(2+\delta)}; k \geq 1\}$, where $\delta > 0$ is fixed and $C > 0$ is a normalizing constant. In this case, the assumption of Proposition 1 holds and thus Condition 3 is satisfied.*

To see that the assumption of Proposition 1 holds in this case, fix $\varepsilon \in (1/2, (1+\delta)/(2+\delta))$. Note that $-(1+\delta/2) < -(2+\delta)(1-\varepsilon) < -1$, and thus

$$\sum_{k \geq 1} p_k^{1-\varepsilon} = C^{1-\varepsilon} \sum_{k \geq 1} k^{-(2+\delta)(1-\varepsilon)} < \infty.$$

3. Simulations

The main application of the asymptotic normality results given in this paper is the construction of asymptotic confidence intervals and hypothesis tests. For instance, if $\mathbf{p}$ satisfies the assumptions of Theorem 2, then an asymptotic $(1-\alpha)100\%$ confidence interval for H is given by

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

where $z_{\alpha/2}$ is a number such that $P(Z > z_{\alpha/2}) = \alpha/2$ and Z is a standard normal random variable. Similarly, if the assumptions of Theorem 3 are satisfied, then we can replace $\hat{H}_{MM}$ with $\hat{H}_{JK}$, and if the assumptions of Lemma 1 are satisfied, then we can replace $\hat{H}_{MM}$ with $\hat{H}$. In this section, we give a small-scale simulation study to evaluate the finite sample performance of these confidence intervals.

For concreteness, we focus on the geometric distribution, which corresponds to

$$p_k = p(1-p)^{k-1}, \ k = 1, 2, \ldots,$$

where $p \in (0,1)$ is a parameter. The true entropy of this distribution is given by $H = -p^{-1}(p \ln p + (1-p)\ln(1-p))$. In this case, Conditions 1, 2, and 3 all hold. For our simulations, we took $p = 0.5$. The simulations were performed as follows. We began by simulating a random sample of size n and used it to evaluated a 95% confidence interval for the given estimator. We then checked to see if the true value of H 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. This proportion should be close to 0.95 when the confidence interval works well. We repeated this for sample sizes ranging from 20 to 1000 in increments of 10. The results are given in Figure 1. We can see that the Miller–Madow and jackknife estimators consistently outperform the plug-in. It may be interesting to note that, although the proofs of Theorems 1–3 are based on showing that the bias correction term approaches zero, it does not mean that the bias correction term is not useful. On the contrary, bias correction improves the finite sample performance of the asymptotic confidence intervals.

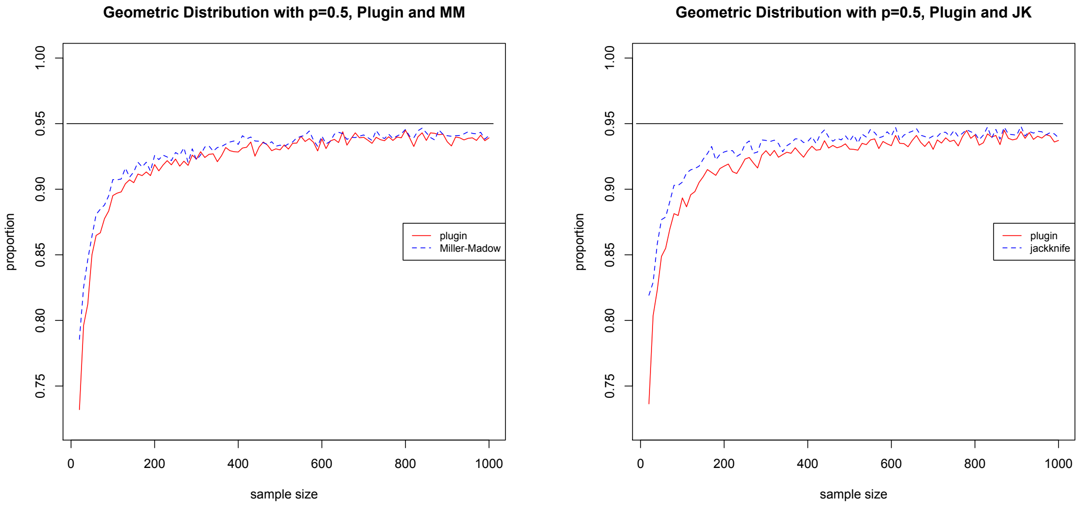

Figure 1. *Cont.*

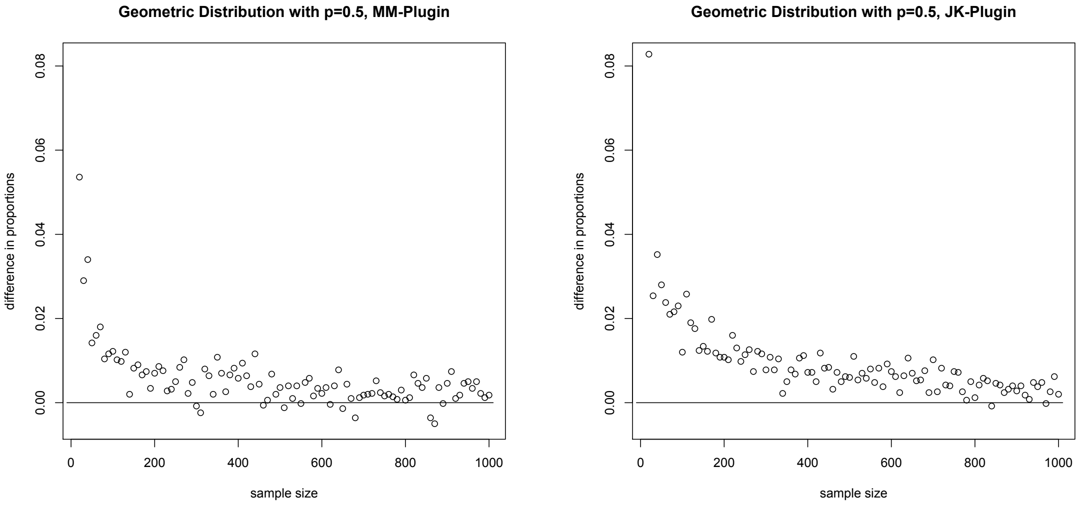

Figure 1. Effectiveness of the 95% confidence intervals as a function of sample size. The plot on the top left gives the proportions for the Miller–Madow and the plug-in estimators, while the one on the top right gives the proportions for the jackknife and the plug-in estimators. The horizontal line is at 0.95. The closer the proportion is to this line, the better the performance. The plot on the bottom left gives the proportion for Miller–Madow minus the proportion for the plug-in, while the one of the bottom right gives the proportion for the jackknife minus the proportion for the plug-in. The larger the value, the greater the improvement due to bias correction. Here, the horizontal line is at 0.

4. Discussion

In this paper, we gave sufficient conditions for the asymptotic normality of the Miller–Madow and the Jackknife estimators of entropy. While our focus is on the case of countably infinite alphabets, our results are formulated and proved in the case where the effective cardinality K may be finite or countably infinite. As such, they hold in the case of finite alphabets as well. In fact, for finite alphabets, Conditions 1–3 always hold and we have asymptotic normality so long as the underlying distribution is not uniform. The difficulty with the uniform distribution is that it is the unique distribution for which σ^2, as given by (9), is zero (see the discussion just below Condition 1). When the distribution is uniform, the asymptotic distribution is chi-squared with $(K-1)$ degrees of freedom (see [6]).

In general, we do not know if our conditions are necessary. However, they cover most distributions of interest. The only distributions, which they preclude, are ones with extremely heavy tails. However, in complete generality, Conditions 1–3 may look complicated, and they are easily checked in many situations. For instance, Condition 2 always holds when, for large enough k, $p_k \le Ck^{-3-\delta}$ for some $C, \delta > 0$, i.e., when

$$\sum_{k=1}^{\infty} k^2 p_k < \infty.$$

If the alphabet $\mathscr{X} = \mathbb{N}$ is the set of natural numbers, then this is equivalent to the distribution $\mathbf{p}$ having a finite variance. Similarly, Conditions 1 and 3 both holding is the case when, for large enough k, $p_k \le Ck^{-2-\delta}$ for some $C, \delta > 0$, i.e., when

$$\sum_{k=1}^{\infty} k p_k < \infty.$$

If the alphabet $\mathscr{X} = \mathbb{N}$ is the set of natural numbers, then this is equivalent to the distribution $\mathbf{p}$ having a finite mean.

5. Proofs

Proof of Lemma 2. Without loss of generality, assume that $C > 1$ and thus that $\ln C > 0$. Let $f(x) = x \ln x$ for $x \in (0,1)$. It is readily checked that f is negative and decreasing for $x \in (0, e^{-1})$. Since $Cp'_k \to 0$ as $k \to \infty$, it follows that $Cp'_k < e^{-1}$ for large enough k. Now, let $K(n)$ be the sequence that works for $\mathbf{p}'$ in Condition 1. For large enough n,

$$\begin{aligned} 0 &\geq \sqrt{n} \sum_{k \geq K(n)} p_k \ln p_k \geq C\sqrt{n} \sum_{k \geq K(n)} p'_k \ln(Cp'_k) \\ &\geq C \ln(C)\sqrt{n} \sum_{k \geq K(n)} p'_k + C\sqrt{n} \sum_{k \geq K(n)} p'_k \ln(p'_k) \\ &\geq C\left(\ln(C) + 1\right)\sqrt{n} \sum_{k \geq K(n)} p'_k \ln(p'_k) \longrightarrow 0. \end{aligned}$$

Similarly, the function $g(x) = x \ln^2 x$ for $x \in (0,1)$ is positive and increasing for $x \in (0, e^{-2})$. Thus, there is an integer $M > 0$ such that if $k \geq M$, then $Cp'_k < e^{-2}$ and

$$\begin{aligned} 0 \leq \sum_{k \geq 1} p_k \ln^2 p_k &\leq \sum_{k=1}^{M-1} p_k \ln^2 p_k + C \sum_{k=M}^{\infty} p'_k \ln^2(Cp'_k) \\ &= \sum_{k=1}^{M-1} p_k \ln^2 p_k + C \sum_{k=M}^{\infty} p'_k \ln^2(p'_k) \\ &\quad + C \ln^2(C) \sum_{k=M}^{\infty} p'_k + 2C \ln(C) \sum_{k=M}^{\infty} p'_k \ln(p'_k) < \infty, \end{aligned}$$

as required. □

To prove Theorem 2, the following Lemma is needed.

Lemma 6. *If Condition 2 holds, then there exists a $K_1 > 0$ such that for all $k \geq K_1$*

$$p_k \leq \frac{1}{a(k)b(k)k^3} \leq 1 - \left(1 - \frac{2}{kb(k)}\right)^{\frac{1}{a(k)k^2}}. \tag{14}$$

Proof. Observing that $e^{-x} \geq 1 - x$ holds for all real x and that $\lim_{x \to 0}(1 - e^{-x})/x = 1$, we have $e^{-2/(kb(k))} \geq 1 - 2/(kb(k))$, and hence

$$1 - \left(1 - \frac{2}{kb(k)}\right)^{\frac{1}{a(k)k^2}} \geq 1 - e^{-\frac{2}{a(k)b(k)k^3}} \sim \frac{2}{a(k)b(k)k^3}.$$

This implies that there is a $K_1 > 0$ such that for all $k \geq K_1$ (11) holds and

$$\frac{\frac{2}{a(k)b(k)k^3}}{1 - \left(1 - \frac{2}{kb(k)}\right)^{\frac{1}{a(k)k^2}}} \leq 2.$$

It follows that, for such k,

$$p_k \leq \frac{1}{a(k)b(k)k^3} = .5 \frac{\frac{2}{a(k)b(k)k^3}}{1 - \left(1 - \frac{2}{kb(k)}\right)^{\frac{1}{a(k)k^2}}} \left[1 - \left(1 - \frac{2}{kb(k)}\right)^{\frac{1}{a(k)k^2}}\right] \leq 1 - \left(1 - \frac{2}{kb(k)}\right)^{\frac{1}{a(k)k^2}}$$

as required. □

Proof of Theorem 2. By Theorem 1, it suffices to show that Condition 2 implies that both Condition 1 and $\sqrt{n}\hat{B}_{MM} \xrightarrow{p} 0$ hold. The fact that Condition 2 implies Condition 1 follows by Example 3, Lemmas 2 and 3. We now show that $\sqrt{n}\hat{B}_{MM} \xrightarrow{p} 0$.

Fix $\varepsilon_0 \in (0, \varepsilon)$. From (12) and the facts that $a(k)$ is positive, eventually nondecreasing, and approaches infinity, it follows that

$$\begin{aligned}\limsup_{k\to\infty} \frac{(a(k))^{2\varepsilon_0}}{a\left(\frac{\sqrt{k}}{(a(k))^{\varepsilon_0}}\right)} &= \limsup_{k\to\infty} (a(k))^{-2(\varepsilon-\varepsilon_0)} \frac{(a(k))^{2\varepsilon}}{a\left(\frac{\sqrt{k}}{(a(k))^{\varepsilon_0}}\right)} \\ &\leq \limsup_{k\to\infty} (a(k))^{-2(\varepsilon-\varepsilon_0)} \frac{(a(k))^{2\varepsilon}}{a\left(\frac{\sqrt{k}}{(a(k))^{\varepsilon}}\right)} = 0. \qquad (15)\end{aligned}$$

Let K_2 be a positive integer such that, for all $n \geq K_2$, $a(n)$ is nondecreasing, and let $r_n = \left(\sqrt{n}/(a(n))^{\varepsilon_0}\right) \vee K_3$, where $K_3 = K_1 \vee K_2$ and K_1 is as in Lemma 6. It follows that

$$\begin{aligned}\mathrm{E}\left(\sqrt{n}\hat{B}_{MM}\right) &= \sqrt{n}\,\mathrm{E}\left(\frac{\hat{m}-1}{2n}\right) \leq \frac{1}{\sqrt{n}}\mathrm{E}(\hat{m}) = \frac{1}{\sqrt{n}}\sum_{k\geq 1}\left[1-(1-p_k)^n\right] \\ &\leq \frac{1}{\sqrt{n}}\sum_{k\leq r_n} 1 + \frac{1}{\sqrt{n}}\sum_{k>r_n}\left[1-\left(1-\frac{2}{kb(k)}\right)^{\frac{n}{a(k)k^2}}\right] =: S_1 + S_2.\end{aligned}$$

We have

$$\begin{aligned}S_1 &\leq \frac{r_n}{\sqrt{n}} = \left((a(n))^{-\varepsilon_0}\right) \vee \frac{K_3}{\sqrt{n}} \to 0; \text{ and} \\ S_2 &\leq \frac{1}{\sqrt{n}}\sum_{k>r_n}\left[1-\left(1-\frac{2}{kb(k)}\right)^{\frac{n}{a(r_n)r_n^2}}\right] \leq \frac{1}{\sqrt{n}}\sum_{k>r_n}\left[1-\left(1-\frac{2}{kb(k)}\right)^{\frac{(a(n))^{2\varepsilon_0}}{a\left(\frac{\sqrt{n}}{(a(n))^{\varepsilon_0}}\right)}}\right].\end{aligned}$$

By (15), it follows that, for large enough n,

$$S_2 \leq \frac{1}{\sqrt{n}}\sum_{k>r_n}\left[1-\left(1-\frac{2}{kb(k)}\right)\right] = \frac{1}{\sqrt{n}}\sum_{k>r_n}\frac{2}{kb(k)} \leq \left(\sum_{k\geq 1}\frac{1}{kb(k)}\right)\frac{2}{\sqrt{n}} \to 0.$$

From here, Markov's inequality implies that $\sqrt{n}\hat{B}_{MM} \xrightarrow{p} 0$. □

Proof of Lemma 4. First note that $\ln^{(r-1)}\left(0.5\ln n - 0.5\ln^{(v)} n\right) \sim \ln^{(r-1)}(0.5\ln n)$ for any $v \geq 2$ and $r \geq 1$. This can be shown by induction on r. Specifically, the result is immediate for $r = 1$. If the result is true for $r = m$, then for $r = m + 1$

$$\begin{aligned}\ln^{(m)}\left(0.5\ln n - 0.5\ln^{(v)} n\right) &= \ln\left(\ln^{(m-1)}\left(0.5\ln n - 0.5\ln^{(v)} n\right)\right) \\ &= \ln\left(\frac{\ln^{(m-1)}\left(0.5\ln n - 0.5\ln^{(v)} n\right)}{\ln^{(m-1)}(0.5\ln n)}\right) + \ln^{(m)}(0.5\ln n) \\ &\sim \ln^{(m)}(0.5\ln n).\end{aligned}$$

It follows that for $r \geq 2$

$$\lim_{n\to\infty}\frac{\ln^{(r)} n}{\ln^{(r)}\left(\sqrt{n/(\ln^{(r)} n)}\right)} = \lim_{n\to\infty}\frac{\ln^{(r)} n}{\ln^{(r-1)}\left(0.5\ln n - 0.5\ln^{(r+1)} n\right)}$$

$$= \lim_{n\to\infty} \frac{\ln^{(r-1)}(\ln n)}{\ln^{(r-1)}(0.5\ln n)} = 1,$$

where the final equality follows by the fact that $\ln^{(r-1)}(x)$ is a slowly varying function. Recall that a positive-valued function ℓ is called slowly varying if for any $t > 0$

$$\lim_{x\to\infty} \frac{\ell(xt)}{\ell(x)} = 1$$

(see [11] for a standard reference). To see that $\ln^{(r-1)}(x)$ is slowly varying, note that ln is slowly varying and that compositions of slowly varying functions are slowly varying by Proposition 1.3.6 in [11]. □

Proof of Lemma 5. Observing the convention that $0\ln 0 = 0$,

$$\begin{aligned}
\sum_{i=1}^{n} \hat{H}^{(i)} &= \sum_{k\geq 1}\sum_{i:X_i=\ell_k} \hat{H}^{(i)} \\
&= \sum_{k\geq 1} Y_k \left(-\frac{Y_k-1}{n-1}\ln\frac{Y_k-1}{n-1} - \sum_{j:j\geq 1, j\neq k} \frac{Y_j}{n-1}\ln\frac{Y_j}{n-1} \right) \\
&= \sum_{k\geq 1} Y_k \left[-\frac{Y_k-1}{n-1}\left(\ln\frac{Y_k-1}{Y_k} + \ln\frac{Y_k}{n-1}\right) - \sum_{j:j\geq 1, j\neq k} \frac{Y_j}{n-1}\ln\frac{Y_j}{n-1} \right] \\
&= \sum_{k\geq 1} Y_k \left[-\frac{Y_k-1}{n-1}\ln\frac{Y_k-1}{Y_k} + \frac{1}{n-1}\ln\frac{Y_k}{n-1} - \sum_{j\geq 1} \frac{Y_j}{n-1}\ln\frac{Y_j}{n-1} \right] \\
&= -\frac{1}{n-1}\sum_{k\geq 1} Y_k(Y_k-1)\ln\frac{Y_k-1}{Y_k} + \sum_{k\geq 1}\frac{Y_k}{n-1}\ln\frac{Y_k}{n-1} - \sum_{k\geq 1} Y_k \sum_{j\geq 1}\frac{Y_j}{n-1}\ln\frac{Y_j}{n-1} \\
&= -\frac{1}{n-1}\sum_{k\geq 1} Y_k(Y_k-1)\ln\frac{Y_k-1}{Y_k} - (n-1)\sum_{k\geq 1}\frac{Y_k}{n-1}\ln\frac{Y_k}{n-1} \\
&= -\frac{1}{n-1}\sum_{k\geq 1} Y_k(Y_k-1)\ln\frac{Y_k-1}{Y_k} - \sum_{k\geq 1} Y_k\left(\ln\frac{Y_k}{n} + \ln\frac{n}{n-1}\right) \\
&= -\frac{1}{n-1}\sum_{k\geq 1} Y_k(Y_k-1)\ln\frac{Y_k-1}{Y_k} - n\ln\frac{n}{n-1} + n\hat{H}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{n}\left(\hat{H} - \hat{H}^{(i)}\right) &= \frac{1}{n-1}\sum_{k\geq 1} Y_k(Y_k-1)\ln\frac{Y_k-1}{Y_k} + n\ln\frac{n}{n-1} \\
&= \frac{1}{n-1}\sum_{k\geq 1} Y_k\left[(Y_k-1)\ln\frac{Y_k-1}{Y_k} + (n-1)\ln\frac{n}{n-1}\right] \\
&= \frac{1}{n-1}\sum_{k\geq 1} Y_k\left[(n-1)\ln\frac{n}{n-1} - (Y_k-1)\ln\frac{Y_k}{Y_k-1}\right].
\end{aligned}$$

It suffices to show that for any $y \in \{1, 2, \cdots, n\}$

$$(y-1)\ln y - (y-1)\ln(y-1) \leq (n-1)\ln n - (n-1)\ln(n-1). \qquad (16)$$

Towards that end, first note that the inequality of (16) holds for $y = 1$. Now, let

$$f(y) = (y-1)\ln y - (y-1)\ln(y-1)$$

and, therefore, letting $s = 1 - 1/y$,

$$f'(y) = \ln \frac{y}{y-1} - \frac{1}{y} = \left(1 - \frac{1}{y} - 1\right) - \ln\left(1 - \frac{1}{y}\right) = (s-1) - \ln s.$$

Since $s - 1 \geq \ln s$ for all $s > 0$ (see e.g., 4.1.36 in [12]), $f(y) \geq 0$ for all y, $1 < y \leq n$, which implies (16).

For the second part, we use the first part to get

$$0 \leq \mathrm{E}\left[\sum_{i=1}^{n}\left(\hat{H} - \hat{H}^{(i)}\right)\right] = n\,\mathrm{E}[\hat{H} - H] - \sum_{i=1}^{n}\mathrm{E}[\hat{H}^{(i)} - H] = n(B_n - B_{n-1}),$$

where the last equality follows from the facts that for each i, $\hat{H}^{(i)}$ is a plug-in estimator of H based on a sample of size $(n-1)$ and that $\mathrm{E}\left[\hat{H}^{(i)}\right]$ does not depend on i due to symmetry. From here, the result follows. □

Proof of Theorem 3. By Theorem 1, it suffices to show $\sqrt{n}\hat{B}_{JK} \xrightarrow{p} 0$. Note that, by Lemma 5,

$$0 \leq \mathrm{E}\left[\sqrt{n}\hat{B}_{JK}\right] = \sqrt{n}(n-1)(B_n - B_{n-1}) \sim n^{3/2}(B_n - B_{n-1}) \to 0,$$

where the convergence follows by Condition 2. From here, the result follows by Markov's inequality. □

To prove Proposition 1, we need several lemmas, which may be of independent interest.

Lemma 7. *Let S_n and S_{n-1} be binomial random variables with parameters (n, p) and $(n-1, p)$, respectively. If $n \geq 2$ and $p \in (0,1)$, then*

$$\mathrm{E}(S_n \ln S_n) = \mathrm{E}\left(np\ln(S_{n-1}+1)\right). \tag{17}$$

The proof is given on page 178 in [7].

Lemma 8. *Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with parameter $p \in (0,1)$. For $m = 1, \ldots, n$ let $S_m = \sum_{i=1}^{m} X_i$, $\hat{p}_m = S_m/m$, $\hat{h}_m = -\hat{p}_m \ln \hat{p}_m$, and $\Delta_m = \mathrm{E}[\hat{h}_m - \hat{h}_{m-1}]$. Then,*

$$\Delta_n = \mathrm{E}[\hat{h}_n - \hat{h}_{n-1}] \leq \frac{p(2-p)}{(n-1)[(n-2)p+2]} \leq \frac{2p}{(n-1)[(n-2)p+2]}. \tag{18}$$

Proof. Applying Lemma 7 to Δ_n gives

$$\begin{aligned}\Delta_n &= p\ln\left(\frac{n}{n-1}\right) + p\,\mathrm{E}\left[\ln\left(\frac{S_{n-2}+1}{S_{n-1}+1}\right)\right] \\ &= p\ln\left(\frac{n}{n-1}\right) + p\,\mathrm{E}\left[\ln\left(\frac{S_{n-2}+1}{S_{n-2}+X_{n-1}+1}\right)\right].\end{aligned}$$

Conditioning on X_{n-1} gives

$$\Delta_n = p\ln\left(\frac{n}{n-1}\right) + p^2\,\mathrm{E}\left[\ln\left(\frac{S_{n-2}+1}{S_{n-2}+2}\right)\right].$$

Noting that $f(x) = \ln(x/(x+1))$ is a concave function for $x > 0$, by Jensen's inequality,

$$\Delta_n \leq p\ln\left(\frac{n}{n-1}\right) + p^2\ln\left(\frac{(n-2)p+1}{(n-2)p+2}\right).$$

Applying the following inequalities (both of which follow from 4.1.36 in [12]) to the terms of the above expression,

$$\ln\left(\frac{x}{x-1}\right) < \frac{1}{x-1} \text{ for } x > 1 \qquad \text{and} \qquad \ln\left(\frac{x}{x+1}\right) < -\frac{1}{x+1} \text{ for } x > 0,$$

it follows that

$$\Delta_n \leq \frac{p}{n-1} - \frac{p^2}{(n-2)p+2} = \frac{p(2-p)}{(n-1)[(n-2)p+2]},$$

which completes the proof. □

For fixed $\varepsilon > 0$, rewriting the upper bound of (18) gives

$$\frac{2p}{(n-1)[(n-2)p+2]} = \frac{2p^{1-\varepsilon}}{n-1}\left\{\frac{p^{\varepsilon}}{(n-2)p+2}\right\} =: \frac{2p^{1-\varepsilon}}{n-1}\{g(n,p,\varepsilon)\}. \tag{19}$$

Lemma 9. *For any $\varepsilon \in (0,1)$ and $n \geq 3$, there exists a $p_0 \in (0,1)$ such that $g(n,p,\varepsilon)$ defined in (19) is maximized at p_0 and*

$$0 \leq g(n,p_0,\varepsilon) = \mathcal{O}(n^{-\varepsilon}). \tag{20}$$

Proof. Taking the derivative of $\ln g(n,p,\varepsilon)$ with respect to p gives

$$\frac{\partial}{\partial p}\ln g(n,p,\varepsilon) = \frac{\varepsilon}{p} - \frac{n-2}{(n-2)p+2}.$$

It is readily checked that this equals zero only at

$$p_0 = \frac{2\varepsilon}{(1-\varepsilon)(n-2)}$$

and is positive for $0 < p < p_0$ and negative for $p_0 < p < 1$. Thus, p_0 is the global maximum. For a fixed ε, we have

$$g(n,p_0,\varepsilon) = \frac{(2\varepsilon)^{\varepsilon}}{(1-\varepsilon)^{\varepsilon}(n-2)^{\varepsilon}[\frac{2\varepsilon}{(1-\varepsilon)}+2]} = \left(\frac{\varepsilon}{n-2}\right)^{\varepsilon}\left(\frac{1-\varepsilon}{2}\right)^{1-\varepsilon} = \mathcal{O}(n^{-\varepsilon}),$$

as required. □

Proof of Proposition 1. For every k and every $m \leq n$, let

$$S_{m,k} = \sum_{i=1}^{m} 1[X_i = \ell_k] \text{ and } \hat{H}_m = -\sum_{k\geq 1}\frac{S_{m,k}}{m}\ln\left(\frac{S_{m,k}}{m}\right)$$

be the observed letter counts and the plug-in estimator of entropy based on the first m observations. Thus, $S_{m,k} = Y_k$ and $\hat{H}_n = \hat{H}$. We are interested in evaluating

$$\begin{aligned}
B_n - B_{n-1} &= \mathrm{E}\left(\hat{H}_n - \hat{H}_{n-1}\right) \\
&= \sum_{k\geq 1}\mathrm{E}\left[\frac{S_{n-1,k}}{n-1}\ln\left(\frac{S_{n-1,k}}{n-1}\right) - \frac{S_{n,k}}{n}\ln\left(\frac{S_{n,k}}{n}\right)\right] \\
&\leq \sum_{k\geq 1}\frac{2p_k}{(n-1)[(n-2)p_k+2]} \\
&= 2\sum_{k\geq 1} p_k^{1-\varepsilon}\frac{g(n,p_k,\varepsilon)}{(n-1)},
\end{aligned}$$

where the third line follows by Lemma 8. Now, applying Lemmas 5 and 9 gives

$$\begin{aligned} 0 \le n^{3/2}(B_n - B_{n-1}) &\le 2n^{3/2} \sum_{k\ge 1} p_k^{1-\varepsilon} \frac{g(n, p_k, \varepsilon)}{(n-1)} \\ &\le n^{3/2} \frac{g(n, p_0, \varepsilon)}{n-1} \sum_{k \ge 1} p_k^{1-\varepsilon} = \mathcal{O}(n^{1/2-\varepsilon}), \end{aligned}$$

which converges to zero when $\varepsilon \in (1/2, 1)$. □

Author Contributions: C.C., M.G., A.S., J.Z. and Z.Z. contributed to the proofs; C.C., M.G. and J.Z. contributed editorial input; Z.Z. wrote the paper.

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

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Son, Inc.: Hoboken, NJ, USA, 2006.
3. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
4. Zhang, Z. *Statistical Implications of Turing's Formula*; John Wiley & Sons: New York, NY, USA, 2017.
5. Zhang, Z.; Grabchak, M. Bias adjustment for a nonparametric entropy estimator. *Entropy* **2013**, *15*, 1999–2011. [CrossRef]
6. Miller, G.A.; Madow, W.G. *On the Maximum-Likelihood Estimate of the Shannon-Wiener Measure of Information*; Operational Applications Laboratory, Air Force, Cambridge Research Center, Air Research and Development Command, Report AFCRC-TR-54-75; Luce, R.D., Bush, R.R., Galanter, E., Eds.; Bolling Air Force Base: Washington, DC, USA, 1954.
7. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithm* **2001**, *19*, 163–193. [CrossRef]
8. Zhang, Z.; Zhang, X. A normal law for the plug-in estimator of entropy. *IEEE Trans. Inf. Theory* **2012**, *58*, 2745–2747. [CrossRef]
9. Zahl, S. Jackknifing an index of diversity. *Ecology* **1977**, *58*, 907–913. [CrossRef]
10. Harris, B. The statistical estimation of entropy in the non-parametric case. In *Topics in Information Theory*; Csiszár, I., Ed.; North-Holland: Amsterdam, The Netherlands, 1975; pp. 323–355.
11. Bingham, N.H.; Goldie, C.M.; Teugels, J.L. *Regular Variation*; Cambridge University Press: New York, NY, USA, 1987.
12. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions*, 10th ed.; Dover Publications: New York, NY, USA, 1972.