Asymptotic Normality of an Entropy Estimator with Exponentially Decaying Bias [*]

Zhiyi Zhang[†]
Department of Mathematics and Statistics
University of North Carolina at Charlotte
Charlotte, NC 28223

Abstract

This paper establishes the asymptotic normality of an entropy estimator with an exponentially decaying bias on any finite alphabet. Furthermore it is shown that the nonparametric estimator is asymptotically efficient.

1 Introduction.

Let $\{p_k\}$ be a probability distribution on a finite alphabet, $\mathscr{X} = \{\ell_k; 1 \le k \le K\}$, where $K \ge 2$ is a finite integer. Let p_X be a random variable such that $P(p_X = p_k) = p_k$. Entropy in the form of

$$H = E[-\ln(p_X)] = -\sum_{k=1}^{K} p_k \ln(p_k), \tag{1}$$

was introduced by Shannon (1948), and is often referred to as Shannon's entropy. Nonparametric estimation of H has been a subject of much research for many decades. Miller (1955) and Basharin (1959) were perhaps among the first who studied the intuitive general nonparametric estimator, $\hat{H} = -\sum_{k=1}^{K} \hat{p}_k \ln(\hat{p}_k)$ where $\hat{p}_k$ is the sample relative frequency of the k th letter ℓ_k, also known as the plug-in estimator. Others have investigated the topic in various forms and directions over the

[*] *AMS 2000 Subject Classifications.* Primary 62f10, 62F12, 62G05, 62G20; secondary 62F15. *Keywords and phrases.* Turing's formula, nonparametric entropy estimation, asymptotic normality.
[†] Research partially supported by NSF Grants DMS 1004769

years. Many important references can be found in Antos and Kontoyiannis (2001) and Paninski (2003). Among many difficult issues of nonparametric entropy estimation, much research effort in the literature seems to be placed on reducing the bias of the estimators. The main reference point of such discussion is the $O(n^{-1})$ decaying bias of the plug-in $\hat{H}$ whose form may be found in Harris (1975). Many bias-adjusted nonparametric estimators have been proposed. All of them have been shown to reduce bias in certain numerical studies. However the rates of bias decay for most of the bias-adjusted estimators are largely unknown, and there is no clear theoretical evidence why any of these proposed estimators should improve the bias decay to a rate faster than $O(n^{-1})$.

Zhang (2012) proposed an estimator $\hat{H}_z$, as given in (2) below, and showed that the associated bias decays at a rate no slower than $O(n^{-1}(1-p_0)^n)$ where $p_0 = \min\{p_k > 0; k = 1, \cdots, K\}$. In addition, Zhang (2012) established that a uniform variance upper bound for the entire class of distributions with finite entropy that decays at a rate of $O(\ln(n)/n)$ compared to $O([\ln(n)]^2/n)$ for the plug-in, that in a wide range of subclasses, the variance of the proposed estimator converges at a rate of $O(1/n)$, and that the aforementioned rate of convergence carries over to the convergence rates in mean squared errors in many subclasses. The computational performances of $\hat{H}_z$, and of its variants, were compared favorably with several other commonly known estimators, such as the jackknife estimator by Zahl (1977) and Strong, Koberle, de Ruyter van Steveninck and Bialek (1998), and the NSB estimator by Nemenman, Shafee and Bialek (2002).

Let $\{y_k\}$ be the sequence of observed counts of letters in the alphabet in an independently and identically distributed (*iid*) sample of size n and $\{\hat{p}_k = y_k/n\}$. The general nonparametric estimator of entropy proposed by Zhang (2012) is

$$\hat{H}_z = \sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1}[n-(v+1)]!}{n!} \sum_{k=1}^{K} \left[\hat{p}_k \prod_{j=0}^{v-1} \left(1 - \hat{p}_k - \frac{j}{n}\right) \right] \right\}. \tag{2}$$

This paper establishes two normal laws of $\hat{H}_z$ as stated in Theorem 1 and Corollary 1 below, and the asymptotic efficiency of $\hat{H}_z$ is given in Theorem 2.

Let $H^{(2)} = E[-\ln(p_X)]^2 = \sum_{k=1}^{K} p_k \ln^2(p_k)$.

Theorem 1. *Let $\{p_k; 1 \le k \le K\}$ be a non-uniform probability distribution on a finite alphabet $\mathscr{X}$ and $\hat{H}_z$ be as in (2). Then*

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

where $\sigma^2 = Var\left[-\ln(p_X)\right] = H^{(2)} - H^2$.

Let

$$\hat{H}_z^{(2)} = \sum_{v=1}^{n-1}\left\{\left(\sum_{i=1}^{v-1}\frac{1}{i(v-i)}\right)\left\{\frac{n^{v+1}[n-(v+1)]!}{n!}\sum_{k=1}^{K}\left[\hat{p}_k\prod_{m=0}^{v-1}\left(1-\hat{p}_k-\frac{m}{n}\right)\right]\right\}\right\}. \tag{3}$$

Corollary 1. *Let $\{p_k; 1 \le k \le K\}$ be a non-uniform probability distribution on a finite alphabet, $\hat{H}_z$ be as in (2), and $\hat{H}_z^{(2)}$ be as in (3). Then*

$$\sqrt{n}\left(\frac{\hat{H}_z - H}{\sqrt{\hat{H}_z^{(2)} - \hat{H}_z^2}}\right) \xrightarrow{L} N(0,1).$$

Theorem 2. *Let $\{p_k; 1 \le k \le K\}$ be a non-uniform probability distribution on a finite alphabet $\mathscr{X}$. Then $\hat{H}_z$ is asymptotically efficient.*

2 Proofs

$\hat{H}_z$ in (2) may be re-expressed as

$$\hat{H}_z = \sum_{k=1}^{K}\left\{\hat{p}_k\sum_{v=1}^{n-1}\frac{1}{v}\left\{\frac{n^{v+1}[n-(v+1)]!}{n!}\right\}\left[\prod_{j=0}^{v-1}\left(1-\hat{p}_k-\frac{j}{n}\right)\right]\right\} \stackrel{def}{=} \sum_{k=1}^{K}\hat{p}_k\hat{g}_{k,n}. \tag{4}$$

Of first interest is an asymptotic normal law of $\hat{p}_k\hat{g}_{k,n}$. For simplicity, consider first a binomial distribution with parameters n and $p \in (0,1)$, and functions

$$g_n(p) = \sum_{v=1}^{n-1}\left\{\frac{1}{v}\left\{\frac{n^{v+1}[n-(v+1)]!}{n!}\right\}\left[\prod_{j=0}^{v-1}\left(1-p-\frac{j}{n}\right)\right]1_{[v\le n(1-p)+1]}\right\},$$

and $h_n(p) = pg_n(p)$. Let $h(p) = -p\ln(p)$.

Lemma 1 below is easily proved by induction.

Lemma 1. *Let a_j, $j = 1, \cdots, n$, be complex numbers satisfying $|a_j| \le 1$ for every j. Then* $\left|\prod_{j=1}^{n} a_j - 1\right| \le \sum_{j=1}^{n} |a_j - 1|$.

Lemma 2. *Let $\hat{p} = X/n$ where X is a binomial random variable with parameters n and p.*

1. $\sqrt{n}[h_n(p) - h(p)] \to 0$ *uniformly in $p \in (c, 1)$ for any c, $0 < c < 1$.*
2. $\sqrt{n}|h_n(p) - h(p)| < A(n) = O(n^{3/2})$ *uniformly in $p \in [1/n, c]$ for any c, $0 < c < p$.*
3. $P(\hat{p} \le c) < B(n) = O(n^{-1/2}\exp\{-nC\})$ *where $C = \frac{(p-c)^2}{p(1-p)}$ for any $c \in (0, p)$.*

Proof of Part 1. As the notation in $g_n(p)$ suggests, the range for v is from 1 to $\min\{n-1, n(1-p)+1\}$. For any v in that range, let $W_{n,v+1} = \frac{n^{v+1}[n-(v+1)]!}{n!}$. Noting $0 \le \frac{1-j/[n(1-p)]}{1-j/n} \le 1$ subject to $j \le n(1-p)$, by Lemma 1,

$$\left|W_{n,v+1}\prod_{j=0}^{v-1}\left(1-p-\frac{j}{n}\right) - (1-p)^v\right| = (1-p)^v\left|\prod_{j=0}^{v-1}\left(\frac{1-\frac{j}{n(1-p)}}{1-\frac{j+1}{n}}\right) - 1\right|$$

$$= (1-p)^v\left|\left(\frac{n}{n-v}\right)\prod_{j=0}^{v-1}\left(\frac{1-\frac{j}{n(1-p)}}{1-\frac{j}{n}}\right) - 1\right|$$

$$= (1-p)^v\left|\left(1+\frac{v}{n-v}\right)\prod_{j=0}^{v-1}\left(\frac{1-\frac{j}{n(1-p)}}{1-\frac{j}{n}}\right) - 1\right|$$

$$\le (1-p)^v\left(\frac{v}{n-v}\right) + (1-p)^v\left|\prod_{j=0}^{v-1}\left(\frac{1-\frac{j}{n(1-p)}}{1-\frac{j}{n}}\right) - 1\right|$$

$$\le (1-p)^v\left(\frac{v}{n-v}\right) + (1-p)^v\sum_{j=0}^{v-1}\left|\left(\frac{1-\frac{j}{n(1-p)}}{1-\frac{j}{n}}\right) - 1\right|$$

$$= (1-p)^v\left(\frac{v}{n-v}\right) + (1-p)^v\frac{p}{1-p}\sum_{j=1}^{v-1}\frac{j}{n-j}$$

$$\le (1-p)^{v-1}\left(\frac{v}{n-v}\right) + (1-p)^{v-1}\sum_{j=1}^{v-1}\frac{j}{n-j}$$

$$= (1-p)^{v-1}\sum_{j=1}^{v}\frac{j}{n-j} \le (1-p)^{v-1}\frac{v^2}{n-v}.$$

For a sufficiently large n, let $V_n = \lfloor n^{1/8} \rfloor$.

$$\sqrt{n}|h_n(p) - h(p)| \le \sqrt{n}p \sum_{v=1}^{\lfloor n(1-p)+1 \rfloor} \frac{1}{v} \left| W_{n,v+1} \prod_{j=0}^{v-1} \left(1 - p - \frac{j}{n}\right) - (1-p)^v \right|$$

$$+\sqrt{n}p \sum_{v=\lfloor n(1-p)+2 \rfloor}^{\infty} \frac{1}{v}(1-p)^v$$

$$= \sqrt{n}p \sum_{v=1}^{V_n} \frac{1}{v} \left| W_{n,v+1} \prod_{j=0}^{v-1} \left(1 - p - \frac{j}{n}\right) - (1-p)^v \right|$$

$$+\sqrt{n}p \sum_{v=V_n+1}^{\lfloor n(1-p)+1 \rfloor} \frac{1}{v} \left| W_{n,v+1} \prod_{j=0}^{v-1} \left(1 - p - \frac{j}{n}\right) - (1-p)^v \right|$$

$$+\sqrt{n}p \sum_{v=\lfloor n(1-p)+2 \rfloor}^{\infty} \frac{1}{v}(1-p)^v$$

$$\stackrel{def}{=\!=} \Delta_1 + \Delta_2 + \Delta_3.$$

$$\Delta_1 \le \sqrt{n}p \sum_{v=1}^{V_n} \frac{v}{n-v}(1-p)^{v-1} \le \frac{n^{5/8}}{n - n^{1/8}} \to 0.$$

$$\begin{aligned}\Delta_2 &\le p\sqrt{n} \sum_{v=V_n+1}^{\lfloor n(1-p)+1 \rfloor} \frac{v}{n-v}(1-p)^{v-1} \le \frac{p\sqrt{n}[n(1-p)+1]}{np-1} \sum_{v=V_n+1}^{\lfloor n(1-p)+1 \rfloor} (1-p)^{v-1} \\ &\le \frac{\sqrt{n}[n(1-p)+1]}{np-1}(1-p)^{\lfloor n^{1/8} \rfloor} \le \frac{\sqrt{n}[n(1-c)+1]}{nc-1}(1-c)^{\lfloor n^{1/8} \rfloor} \to 0.\end{aligned}$$

$$\Delta_3 \le \frac{\sqrt{n}}{n(1-p)}(1-p)^{\lfloor n(1-p)+2 \rfloor} = \frac{1}{\sqrt{n}}(1-p)^{\lfloor n(1-p)+1 \rfloor} \le \frac{1}{\sqrt{n}} \to 0.$$

Hence $\sup_{p \in (c,1)} \sqrt{n}|h_n(p) - h(p)| \to 0$.

Proof of Part 2. The proof is identical to that of Part 1 above until the expression $\Delta_1 + \Delta_2 + \Delta_3$ where each term is to be evaluated on the interval $[1/n, c]$. It is clear that $\Delta_1 \le O(n^{-3/8})$. For Δ_2, since $n(1-p)+1$ at $p = 1/n$ is $n > n-1$, we have

$$\begin{aligned}\Delta_2 &\le p\sqrt{n} \sum_{v=V_n+1}^{\min\{n-1, \lfloor n(1-p)+1 \rfloor\}} \frac{v}{n-v}(1-p)^{v-1} \\ &\le p\sqrt{n} \sum_{v=V_n+1}^{\min\{n-1, \lfloor n(1-1/n)+1 \rfloor\}} \frac{v}{n-v}(1-p)^{v-1} \\ &= p\sqrt{n} \sum_{v=V_n+1}^{n-1} \frac{v}{n-v}(1-p)^{v-1} \\ &< p\sqrt{n}(n-1) \sum_{v=V_n+1}^{n-1} (1-p)^{v-1} \\ &< \sqrt{n}(n-1)(1-p)^{V_n} < \sqrt{n}(n-1) = O(n^{3/2}).\end{aligned}$$

$$\begin{aligned}\Delta_3 &= p\sqrt{n} \sum_{v=\min\{n-1, \lfloor n(1-p)+1 \rfloor\}+1}^{\infty} \frac{1}{v}(1-p)^v \\ &< p\sqrt{n} \sum_{v=1}^{\infty} \frac{1}{v}(1-p)^v < \sqrt{n} = O(n^{1/2}).\end{aligned}$$

Therefore $\Delta_1 + \Delta_2 + \Delta_3 = O(n^{3/2})$.

Proof of Part 3. Let Z and $\phi(z)$ be a standard normal random variable and its density function respectively, and let $\sim$ denote asymptotic equality. Since $\sqrt{n}(\hat{p}-p) \xrightarrow{L} N(0, p(1-p))$,

$$\begin{aligned} P(\hat{p} \le c) &\sim \int_{-\infty}^{\sqrt{n}(c-p)/\sqrt{p(1-p)}} \phi(z)dz = \int_{\sqrt{n}(p-c)/\sqrt{p(1-p)}}^{\infty} \phi(z)dz \\ &< \int_{\sqrt{n}(p-c)/\sqrt{p(1-p)}}^{\infty} e^{-z\left[\sqrt{n}(p-c)/\sqrt{p(1-p)}\right]} dz \\ &= \frac{\sqrt{p(1-p)}}{\sqrt{n}(p-c)} \int_{\left[\sqrt{n}(p-c)/\sqrt{p(1-p)}\right]^2}^{\infty} e^{-x} dx \\ &= \frac{\sqrt{p(1-p)}}{\sqrt{n}(p-c)} \exp\left\{-\left[\sqrt{n}(p-c)/\sqrt{p(1-p)}\right]^2\right\} \\ &= n^{-1/2}\frac{\sqrt{p(1-p)}}{(p-c)} \exp\left\{-\frac{n(p-c)^2}{p(1-p)}\right\}. \end{aligned}$$

□

Proof of Theorem 1. Without loss of generality, consider the sample proportions of the first two letters of the alphabet $\hat{p}_1$ and $\hat{p}_2$ in an *iid* sample of size n. $\sqrt{n}\,(\hat{p}_1 - p_1, \hat{p}_2 - p_2)' \xrightarrow{L} N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})$, $i, j = 1, 2$, $\sigma_{ii} = p_i(1-p_i)$ and $\sigma_{ij} = -p_i p_j$ when $i \ne j$. Write

$$\begin{aligned} &\sqrt{n}\,\{[h_n(\hat{p}_1) + h_n(\hat{p}_2)] - [-p_1 \ln(p_1) - p_2 \ln(p_2)]\} \\ &= \sqrt{n}\,\{[h_n(\hat{p}_1) + h_n(\hat{p}_2)] - [h(\hat{p}_1) + h(\hat{p}_2)]\} \\ &\quad + \sqrt{n}\,\{[h(\hat{p}_1) + h(\hat{p}_2)] - [-p_1 \ln(p_1) - p_2 \ln(p_2)]\} \\ &= \sqrt{n}\,[h_n(\hat{p}_1) - h(\hat{p}_1)] + \sqrt{n}\,[h_n(\hat{p}_2) - h(\hat{p}_2)] \\ &\quad + \sqrt{n}\,\{[h(\hat{p}_1) + h(\hat{p}_2)] - [-p_1 \ln(p_1) - p_2 \ln(p_2)]\} \\ &= \sqrt{n}\,[h_n(\hat{p}_1) - h(\hat{p}_1)]\,1_{[\hat{p}_1 \le p_1/2]} + \sqrt{n}\,[h_n(\hat{p}_2) - h(\hat{p}_2)]\,1_{[\hat{p}_2 \le p_2/2]} \\ &\quad + \sqrt{n}\,[h_n(\hat{p}_1) - h(\hat{p}_1)]\,1_{[\hat{p}_1 > p_1/2]} + \sqrt{n}\,[h_n(\hat{p}_2) - h(\hat{p}_2)]\,1_{[\hat{p}_2 > p_2/2]} \\ &\quad + \sqrt{n}\,\{[h(\hat{p}_1) + h(\hat{p}_2)] - [-p_1 \ln(p_1) - p_2 \ln(p_2)]\}. \end{aligned}$$

The third and fourth terms above converge to zero almost surely by Part 1 of Lemma 2. The last term, by the delta method, converges in law to $N(0,\tau^2)$ where after a few algebraic steps

$$\begin{aligned}\tau^2 &= [\ln(p_1)+1]^2 p_1(1-p_1) + [\ln(p_2)+1]^2 p_2(1-p_2)\\ &\quad -2[\ln(p_1)+1][\ln(p_2)+1]p_1p_2\\ &= [\ln(p_1)+1]^2 p_1 + [\ln(p_2)+1]^2 p_2 - \{[\ln(p_1)+1]p_1 + [\ln(p_1)+1]p_1\}^2.\end{aligned}$$

It remains to show that the first term (the second term will admit the same argument) converges to zero in probability. However this fact can be established by the following argument. By Part 2 and then Part 3 of Lemma 2,

$$\begin{aligned}E\{\sqrt{n}\,|h_n(\hat{p}_1)-h(\hat{p}_1)|\,1_{[\hat{p}_1\le p_1/2]}\} &\le A(n)P(\hat{p}_1 \le p_1/2)\\ &\le A(n)B(n) = O(n^{3/2})O(n^{-1/2}\exp\{-nC\}) \to 0\end{aligned}$$

for some positive constant C. This fact, noting that $\sqrt{n}\,|h_n(\hat{p}_1)-h(\hat{p}_1)| \ge 0$, gives immediately the desired convergence in probability, that is, $\sqrt{n}\,|h_n(\hat{p}_1)-h(\hat{p}_1)|\,1_{[\hat{p}_1\le p_1/2]} \xrightarrow{P} 0$. In turn, it gives the desired weak convergence for $\sqrt{n}\,\{[h_n(\hat{p}_1)+h_n(\hat{p}_2)] - [-p_1\ln(p_1) - p_2\ln(p_2)]\}$.

By generalization for K terms, $\sqrt{n}(\hat{H}_z - H) \xrightarrow{L} N(0,\sigma^2)$ where, letting p_X denote the random variable that assumes the value p_k when X assumes ℓ_k,

$$\begin{aligned}\sigma^2 &= \textstyle\sum_{k=1}^K \{-[\ln(p_k)+1]\}^2 p_k - \{\sum_{k=1}^K \{-[\ln(p_k)+1]\}p_k\}^2\\ &= Var\,[-\ln(p_X)-1] = Var\,[-\ln(p_X)]\,.\end{aligned}$$

□

Remark 1. *It may be interesting to note that the asymptotic variance of* $\sqrt{n}(\hat{H}_z - H)$ *is identical to that of* $\sqrt{n}(\hat{H} - H)$ *where* $\hat{H}$ *is the plug-in.*

Remark 2. *When* $\{p_k\}$ *is a uniform distribution,* $-\ln(p_X)$ *is constant,* $Var\,[-\ln(p_X)] = 0$ *and therefore* $\sqrt{n}(\hat{H}_z - H)$ *asymptotically degenerates.*

Let $\zeta_{1,v} = \sum_k p_k(1-p_k)^v$, $C_v = \sum_{i=1}^{v-1}\frac{1}{i(v-i)}$ for $v \ge 2$ (and define $C_1 = 0$),

$$Z_{1,v} = \frac{n^{1+v}[n-(1+v)]!}{n!}\sum_{k=1}^K \left[\hat{p}_k \prod_{j=0}^{v-1}\left(1-\hat{p}_k-\frac{j}{n}\right)\right],$$

and therefore $\hat{H}_z^{(2)} = \sum_{v=1}^{n-1} C_v Z_{1,v}$.

For clarity in proving Corollary 1, a few notations and two well-known lemmas in U-statistics are first given. For each i, $1 \le i \le n$, let X_i be a random variable such that $X_i = \ell_k$ indicates the event that the k^{th} letter of the alphabet is observed and $P(X_i = \ell_k) = p_k$. Let $X_1, \cdots, X_n$ be an *iid* sample, and denote $x_1, \cdots, x_n$ as the corresponding sample realization. A U-statistic is an n-variable function obtained by averaging the values of an m-variable function (kernel of degree m, often denoted by ψ) over all $n!/[m!(n-m)!]$ possible subsets of m variables from the set of n variables. Interested readers may refer to Lee (1990) for an introduction. Turing's formula, also known as the Good-Turing estimator, is a nonparametric estimator introduced by Good (1953), but largely credited to Alan Turing, as a means of estimate the total probability associated with letters in the alphabet that are not represented in a random sample. In Zhang & Zhou (2010), it is shown that $Z_{1,v}$ is a U-statistic with kernel ψ being Turing's formula with degree $m = v + 1$. Let $\psi_c(x_1, \cdots, x_c) = E[\psi(x_1, \cdots, x_c, X_{c+1}, \cdots, X_m)]$ and $\sigma_c^2 = Var[\psi_c(X_1, \cdots, X_c)]$. Lemmas 3 and 4 below are due to Hoeffding (1948).

Lemma 3. *Let U_n be a U-statistic with kernel ψ of degree m.*

$$Var(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2.$$

Lemma 4. *Let U_n be a U-statistic with kernel ψ of degree m. For $0 \le c \le d \le m$, $\sigma_c^2/c \le \sigma_d^2/d$.*

Lemma 5. $Var(Z_{1,v}) \le \frac{1}{n}\zeta_{1,v} + \frac{v+1}{n}\zeta_{1,v-1}^2$.

Proof. Let $m = v + 1$. By Lemmas 3, 4, and identity $\binom{n}{m}^{-1} \sum_{c=1}^{m} c\binom{m}{c}\binom{n-m}{m-c} = \frac{m^2}{n}$,

$$Var(Z_{1,v}) \le \binom{n}{m}^{-1} \sum_{c=1}^{m} c \binom{m}{c} \binom{n-m}{m-c} \sigma_m^2/m = \frac{m}{n}\sigma_m^2. \tag{5}$$

Consider $\sigma_m^2 = Var[\psi(X_1, \cdots, X_m)] = E[\psi(X_1, \cdots, X_m)]^2 - [\sum_{k=1}^{K} p_k(1-p_k)^{m-1})]^2$. Let $y_k^{(m)}$

denote the frequency of the k th letter in the sample of size m.

$$\sigma_m^2 \le E[\psi(X_1,\cdots,X_m)]^2 = \frac{1}{m^2}E\left[\left(\sum_{k=1}^K 1_{[y_k=1]}\right)\left(\sum_{k'=1}^K 1_{[y_{k'}=1]}\right)\right]$$

$$= \frac{1}{m^2}E\left(\sum_{k=1}^K 1_{[y_k=1]} + 2\sum_{1\le k<k'\le K} 1_{[y_k=1]}1_{[y_{k'}=1]}\right)$$

$$= \frac{1}{m}\sum_{k=1}^K p_k(1-p_k)^{m-1} + \frac{2(m-1)}{m}\sum_{1\le k<k'\le K} p_k p_{k'}(1-p_k-p_{k'})^{m-2}$$

$$\le \frac{1}{m}\sum_{k=1}^K p_k(1-p_k)^{m-1} + 2\sum_{1\le k<k'\le K} p_k p_{k'}(1-p_k-p_{k'}+p_kp_{k'})^{m-2}$$

$$= \frac{1}{m}\sum_{k=1}^K p_k(1-p_k)^{m-1} + 2\sum_{1\le k<k'\le K} \left[p_k(1-p_k)^{m-2}p_{k'}(1-p_{k'})^{m-2}\right]$$

$$\le \frac{1}{m}\sum_{k=1}^K p_k(1-p_k)^{m-1} + \left[\sum_{k=1}^K p_k(1-p_k)^{m-2}\right]^2 = \frac{1}{m}\zeta_{1,m-1} + \zeta_{1,m-2}^2.$$

By (5), $Var(Z_{1,v}) \le \frac{1}{n}\zeta_{1,v} + \frac{v+1}{n}\zeta_{1,v-1}^2$. □

Proof of Corollary 1. By Zhang & Zhou (2010), $E(Z_{1,v}) = \sum_{k=1}^K p_k(1-p_k)^v = \zeta_{1,v}$, and therefore

$$E(\hat{H}_z^{(2)}) = \sum_{v=1}^{n-1} C_v \sum_{k=1}^K p_k(1-p_k)^v$$

$$\to \sum_{k=1}^K p_k \sum_{v=1}^{n-1} C_v(1-p_k)^v = \sum_{k=1}^K p_k[-\ln(p_k)]^2 = \sum_{k=1}^K p_k \ln^2(p_k).$$

It only remains to show $Var(\hat{H}_z^{(2)}) \to 0$.

$$Var(\hat{H}_z^{(2)}) = \sum_{v=1}^{n-1}\sum_{w=1}^{n-1} C_vC_w cov(Z_{1,v},Z_{1,w}) \le \sum_{v=1}^{n-1}\sum_{w=1}^{n-1} C_vC_w\sqrt{Var(Z_{1,v})Var(Z_{1,w})}$$

$$= \left[\sum_{v=1}^{n-1} C_v\sqrt{Var(Z_{1,v})}\right]^2.$$

Note $C_v = \sum_{i=1}^{v-1}\frac{1}{i(v-i)} \le \sum_{i=1}^{v-1}\frac{1}{v-1} = 1$, $\zeta_{1,v} \le \zeta_{1,v-1}$, $\zeta_{1,v-1}^2 \le \zeta_{1,v-1}$,

$$\zeta_{1,v-1} = \sum_{k=1}^K p_k(1-p_k)^{v-1} \le \sum_{k=1}^K p_k(1-p_0)^{v-1} = (1-p_0)^{v-1}$$

where $p_0 = \min\{p_k > 0; k=1,\cdots,K\}$, and therefore, from Lemma 5 for $v \ge 2$,

$$\sqrt{Var(Z_{1,v})} \le \frac{1}{\sqrt{n}}\sqrt{(v+2)\zeta_{1,v-1}} \le \frac{\sqrt{2}v^{1/2}}{\sqrt{n}}(1-p_0)^{(v-1)/2}.$$

As $n \to \infty$,

$$\sum_{v=1}^{n-1} C_v \sqrt{Var(Z_{1,v})} \le \frac{\sqrt{2}}{\sqrt{n}} \sum_{v=1}^{n} v^{1/2} \left(\sqrt{1-p_0}\right)^{v-1}$$

$$= \frac{\sqrt{2}}{\sqrt{n}} \sum_{v=1}^{\lfloor n^{1/4} \rfloor} v^{1/2} \left(\sqrt{1-p_0}\right)^{v-1} + \frac{\sqrt{2}}{\sqrt{n}} \sum_{v=\lfloor n^{1/4} \rfloor+1}^{n} v^{1/2} \left(\sqrt{1-p_0}\right)^{v-1}$$

$$\le \frac{\sqrt{2}}{\sqrt{n}} n^{1/4} \left(n^{1/4}\right)^{1/2} + \frac{\sqrt{2}}{\sqrt{n}} n^{1/2} \left(\sqrt{1-p_0}\right)^{\lfloor n^{1/4} \rfloor} \frac{1}{1-\sqrt{1-p_0}}$$

$$= \sqrt{2} n^{-1/8} + \sqrt{2} \left(\sqrt{1-p_0}\right)^{\lfloor n^{1/4} \rfloor} \frac{1}{1-\sqrt{1-p_0}} \to 0,$$

and $Var(\hat{H}_z^{(2)}) \to 0$ follows. Hence $\hat{H}_z^{(2)} \xrightarrow{p} H^{(2)}$. The fact of $\hat{H}_z \xrightarrow{p} H$ is implied by Theorem 1. Finally the corollary follows Slutsky's Theorem. □

Proof of Theorem 2. First consider the plug-in estimator $\hat{H}$. It can be verified that $\sqrt{n}(\hat{H} - H) \to N(0, \sigma^2)$ where $\sigma^2 = \sigma^2(\{p_k\})$ is as in Theorem 1. We want to show first that $\hat{H}$ is asymptotically efficient in two separate cases: 1) when K is known and 2) when K is unknown. If K is known, then the underlying model $\{p_k; 1 \le k \le K\}$ is a $(K-1)$-parameter multinomial distribution and therefore $\hat{H}$ is the maximum likelihood estimator of H which implies that it is asymptotically efficient. Since the estimator $\hat{H}$ takes the same value, given a sample, regardless whether K is known or not, its asymptotic variance is the same whether K is known or not. Therefore $\hat{H}$ must be asymptotically efficient when K is finite but unknown, or else, it would contradict the fact that $\hat{H}$ is asymptotically efficient when K is known. The asymptotic efficiency of $\hat{H}_z$ follows from the fact that $\sqrt{n}(\hat{H}_z - H)$ and $\sqrt{n}(\hat{H} - H)$ have identical limiting distribution. □

References

[1] Antos, A. and Kontoyiannis, I. (2001). *Convergence properties of functional estimates for discrete distributions*, Random Structures & Algorithms, Vol. 19, pp. 163-193.

[2] Basharin, G. (1959). *On a statistical estimate for the entropy of a sequence of independent random variables*, Theory of Probability and Its Applications, 4, pp. 333-336.

[3] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. Biometrika, 40, pp. 237-264.

[4] Harris, B. (1975). *The statistical estimation of entropy in the non-parametric case*, Topics in Information Theory, edited by I. Csiszar, Amsterdam: North-Holland, pp. 323-355.

[5] Hoeffding, W. (1948). *A class of statistics with asymptotically normal distribution*, Annals of Mathematical Statistics, Vol. 19, No. 3, pp. 293-325.

[6] Lee, A.J. (1990). *U-Statistics: Theory and Practice*, Marcel Dekker, Inc. New York.

[7] Miller, G. (1955). *Note on the bias of information estimates*, Information theory in psychology II-B, ed. H. Quastler, Glencoe, IL: Free Press, pp. 95-100.

[8] Nemenman, I., Shafee, F. & Bialek, W. (2002). *Entropy and inference, revisited*. Advances in Neural Information Processing Systems 14, Cambridge, MA, 2002. MIT Press.

[9] Paninski, L. (2003). *Estimation of entropy and mutual information*, Neural Comp. 15, pp. 1191-1253.

[10] Shannon, C.E. (1948). *A Mathematical Theory of Communication*, Bell Syst. Tech. J., 27, pp. 379-423, and pp. 623-656.

[11] Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R., & Bialek, W. (1998). *Entropy and information in neural spike trains*. Physical Review Letters, 80 (1), pp. 197-200.

[12] Zahl, S. (1977). *Jackknifing an index of diversity*, Ecology, 58: pp. 907913.

[13] Zhang, Z. (2012). *Entropy estimation in Turing's perspective*, Neural Computation, Vol. 24, No. 5, pp. 1368-1389.

[14] Zhang, Z. and Zhou, J. (2010). *Re-parameterization of multinomial distribution and diversity indices*, J. of Statistical Planning and Inference, Vol. 140, No. 7, pp. 1731-1738.