Entropy Estimation in Turing's Perspective*

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

Abstract

A new nonparametric estimator of Shannon's entropy on a countable alphabet is proposed and analyzed against the well-known plug-in estimator. The proposed estimator is developed based on Turing's formula which recovers distributional characteristics on the subset of the alphabet not covered by a size-n sample. The fundamental switch in perspective brings about substantial gain in estimation accuracy for every distribution with finite entropy. In general, a uniform variance upper bound is established 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; in a wide range of subclasses, the variance of the proposed estimator converges at a rate of $O(1/n)$; and this rate of convergence carries over to the convergence rates in mean squared errors in many subclasses. Specifically for any finite alphabet, the proposed estimator has a bias decaying exponentially in n. Several new bias-adjusted estimators are also discussed.

* *AMS 2000 Subject Classifications.* Primary 62G05; secondary 62G32. *Keywords and phrases.* Entropy, entropy estimation, convergence rates, Turing's formula.

1 Introduction and Summary

Let $\{p_k\}$ be a probability distribution on a countable alphabet indexed by $k \in \mathbb{N}$. Entropy in the form of

$$H = -\sum_k p_k \ln(p_k), \tag{1}$$

was introduced by Shannon (1948), and is often referred to as Shannon's entropy. H plays a central role in many branches of mathematics. The problem of estimating entropy has naturally co-existed as long as the index itself. However in recent decades as entropy finds its way into a wider range of applications, the estimation problem also becomes increasingly important. The volume of literature on the general topic of entropy estimation is quite large and covers a wide range of models, and lately with different stochastic structures. The two most commonly cited review articles are by Beirlant, Dudewicz, Györfi & Meulen (2001) and Paninski (2003). While Beirlant, *et al.* (2001) includes results under a broader range of models, Paninski (2003) mainly focuses on the problem of entropy estimation on a countable alphabet, the most basic and therefore the most fundamental model for entropy estimation.

Entropy estimation is generally considered a difficult problem regardless the type of domain for the underlying distribution. The difficulty lies with a twofold fact. First, Shannon's entropy is an ultra sensitive measure to small perturbations on the tail probabilities. An infinitesimal perturbation in the tail of a probability distribution with finite entropy could cause the entropy to diverge to infinity. Second, a sample of size n, however large it may be, could have a hopelessly poor coverage of the alphabet. The cardinality of the not-covered subset in the alphabet is always infinite if the alphabet itself is, and even for a finite alphabet is often much larger than that of the subset covered by the sample. In light of this fact, it is clear that a good entropy estimator must find and utilize certain characteristics of the probability distribution over the not-covered subset of the alphabet. Two fundamental questions are immediately raised. The first question is whether such information exists in a finite sample; and the second question is, if it does, whether it can be extracted nonparametrically. By a perspective induced by the remarkable Turing's formula, the answer to each of the two questions is an affirmative. The proposed estimator, given in (7) below, is constructed in Turing's perspective and is shown in subsequent text to have much improved performance over a much-studied nonparametric entropy estimator, given in (2).

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 that plays a central role in the literature is commonly known as the plug-in estimator and is given by

$$\hat{H} = -\sum_k \hat{p}_k \ln(\hat{p}_k). \tag{2}$$

The plug-in estimator is appealing to practitioners since it is simple, intuitive, easily calculated and there are no known faster convergent nonparametric estimators in general. On any countable alphabet, among several key interesting results, Antos & Kontoyiannis (2001) shows that, for all distributions with finite entropy, the variance of $\hat{H}$ has an upper bound decaying at a rate of $O([\ln(n)]^2/n)$ and this rate persists for all distributions on any countably infinite alphabet. It is shown in the next section that the variance of the proposed estimator has an upper bound decaying at a rate of $O(\ln(n)n^{-1})$ for all distributions with finite entropy, which betters the rate for that of the plug-in estimator by a factor of $\ln(n)$. For a wide range of well-behaving distributions on countably infinite alphabets, it is shown that the variance of the proposed estimator decays at a rate of $O(1/n)$, which betters the rate of the the upper bound for that of the plug-in estimator by a factor of $[\ln(n)]^2$. It is also shown that the advantage of the new estimator in variance carries over to the convergence rate in mean squared errors for a wide range of subclasses.

Shannon's entropy was originally defined on a finite alphabet and therefore the problem of entropy estimation on a finite alphabet with cardinality K is considered by many as a classic problem in the literature. In this case, the variance of $\hat{H}$ decays at a rate of $O(n^{-1})$

which cannot be improved since the plug-in is the maximum likelihood estimator (MLE) when K is known. However a long standing issue there is that of adjusting $\hat{H}$ for its large bias. The bias up to the second order can be worked out. The following form is given by Harris (1975).

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

Based on (3) it is readily seen that the first bias term is proportional to an often large K, known or unknown. If K is unknown then an estimation of K must be involved in adjusting the bias and estimating K is a well-known difficult statistical problem currently without a satisfactory nonparametric solution. Miller (1955) was perhaps the first to make an attempt to adjust for the $O(n^{-1})$ bias term and many followed over the years. Among others, notable results have been reported by Basharin (1959), and more recently by Victor (2000) and Panzeri, Senatore, Montemurro, and Petersen (2007). But even if the first term bias could be satisfactorily adjusted, adjustment for the second bias term is still troublesome because the corresponding large coefficient depends on the distribution in a way that is difficult to handle non-parametrically. Many have gone to great length in effort. For examples, see Grassberger (1988), Paninski (2003) and Schürmann (2004). Generally there seem to be serious difficulties in balancing the variance and the bias terms simultaneously in adjusting for biases. However it is shown in the next section that the proposed estimator has a variance decaying at a rate of $O(1/n)$ and, more importantly, a bias with an upper bound decaying exponentially in n, or more precisely $O((1-p_0)^n/n)$ where $p_0 = \min\{p_k > 0; k \in \mathbb{N}_K\}$ and $\mathbb{N}_K$ is the set of natural numbers up to a finite integer $K \geq 2$. The extremely rapid decay of the bias brings about significant reduction of bias even for smaller samples.

In light of the claimed properties of the proposed estimator, it would seem that there is something fundamentally deficient in estimating entropy via the perspective of $\hat{H}$. As alluded to earlier, there is. The deficiency lies with the fact that entropy is essentially a tail property of the underlying distribution and that the tail information in the sample is under utilized by $\hat{H}$. Here and in the subsequent text, the word "tail" loosely refers to a subset of letters in the alphabet with low probabilities. In that sense, a "distant tail" would be a subset of letters with very small probabilities, and a distribution on a finite alphabet would have essentially "no tails". The term "tail information in the sample" refers to any useful information in the sample that can be used to make inferences about certain collective characteristics of the probabilities associated with the tail, particularly the tail that is not covered by a given sample. As it turns out, one could say quite a bit about certain characteristics of the probabilities on the not-covered set using the sample frequency distribution on the covered set. This is an area of Statistics developed around a formula introduced by Good (1953) but largely credited to Alan Turing, and hence often known as Turing's formula. Research on the statistical properties of Turing's formula has been reported only very sporadically in the history until recently. The authors of notable publications on the distributional characteristics of Turing's formula include Robbins (1968), Esty (1983), and more recently Zhang & Huang (2008) and Zhang & Zhang (2009), among others. Turing's formula does not say everything about the tail of a distribution on an alphabet, but it says exactly the right things for estimating entropy and its likes. In short, Turing's formula recovers much of the tail information in the sample that is largely missed by $\hat{H}$. In terms of entropy estimation, Chao & Shen (2003) recognized the potential of sample information from the non-covered letters of the alphabet, they proposed an adjustment for the plug-in estimator. Vu, Yu & Kass (2007) derived several convergence properties of the coverage-adjusted estimator proposed by Chao & Shen. The general theoretical results on the convergence rate were largely discouraging.

The development of the new estimator starts with another well-known diversity index proposed by Simpson (1949), $\tau = \sum_k p_k^2$ for a population with different species, which has an equivalent form $\zeta_{1,1} = 1 - \tau = \sum_k p_k(1 - p_k)$ and assumes a value in $[0, 1)$ with a higher level of $\zeta_{1,1}$ indicating a more diverse population. Zhang & Zhou (2010) generalized Simpson's diversity index to a family of indices of the form

$$\zeta_{u,v} = \sum_k p_k^u (1 - p_k)^v \tag{4}$$

where $u \geq 1$ and $v \geq 0$ are two arbitrarily fixed integers, proposed estimators for these generalized Simpson's indices, $Z_{u,v}$ as in (11) below, and discussed their statistical properties. Among many desirable properties, $Z_{u,v}$ is a U-statistic and therefore an unbiased estimator of $\zeta_{u,v}$ provided $u + v \leq n$. A sub-family of (4), $\{\zeta_{1,v}; v \geq 1\}$, is of special relevance to Shannon's entropy since, by Taylor's expansion and Fubini's lemma, a finite entropy has the following alternative representation:

$$H = \sum_{v=1}^{\infty} \frac{1}{v} \sum_k p_k (1 - p_k)^v = \sum_{v=1}^{\infty} \frac{1}{v} \zeta_{1,v}. \tag{5}$$

An intuitively interesting comparison can be made between the two expressions of entropy, (1) and (5). (1) may be thought of as a sum of positive terms in k and (5) as a sum of positive terms in v. Each term in (1) is a function of p_k alone. Since the coverage of any given sample is finite, only finitely many terms in (1) are deliberately estimated by $\hat{H}$ in (2), and given a fixed sample, these finite terms collectively do not reflect any tail properties of the underlying distribution. The collective remaining terms in (1) are only estimated incidentally. While this observation does not necessarily constitute a rigorous argument against the plug-in estimator in terms of its performance, it certainly suggests some potential inadequacy. On the other hand, each term in (5) is a member of the generalized Simpson's indices which in its own right reflects certain properties of the infinite tail of the underlying distribution. If an unbiased estimator exists for each term in (5), then one could naturally hope that the new representation would lead to a better estimator. According to Zhang & Zhou (2010), there do exist unbiased estimators for the terms in (5) up to $v = n - 1$, namely $Z_{1,v}/v$, where $Z_{1,v}$ is implied by (11) but explicitly given by

$$Z_{1,v} = \frac{n^{1+v}[n - (1 + v)]!}{n!} \sum_k \left[\hat{p}_k \prod_{j=0}^{v-1} \left(1 - \hat{p}_k - \frac{j}{n} \right) \right]. \tag{6}$$

Therefore

$$\hat{H}_z = \sum_{v=1}^{n-1} \frac{1}{v} Z_{1,v} \tag{7}$$

is an unbiased estimator of $\sum_{v=1}^{n-1} \frac{1}{v} \zeta_{1,v}$ as in

$$H = \sum_{v=1}^{n-1} \frac{1}{v} \zeta_{1,v} + \sum_{v=n}^{\infty} \frac{1}{v} \zeta_{1,v} = E(\hat{H}_z) + |B_n|$$

where $|B_n| = \sum_{v=n}^{\infty} \frac{1}{v} \zeta_{1,v}$ is the bias of $\hat{H}_z$.

To see that $\zeta_{1,v} = \sum_k p_k(1-p_k)^v$ is a tail property of $\{p_k\}$, one needs only to consider the random variable

$$\pi_v = \sum_k p_k 1_{[y_k=0]} \tag{8}$$

which may be thought of as the total probability associated with the letters in the alphabet not covered by an *iid* sample of size $v \geq 1$. Clearly π_v reflects a characteristic of the underlying probability distribution on low probability letters in the alphabet, and hence could be reasonably considered a tail property. Yet

$$E(\pi_v) = \sum_k p_k(1-p_k)^v = \zeta_{1,v}.$$

Given an *iid* sample of size v, Turing's formula

$$T = \sum_k \frac{1}{v} 1_{[y_k=1]}$$

estimates π_v with a bias. However, as shown by Robbins (1968), T based on an *iid* sample of size $v+1$ is an unbiased estimator of π_v and therefore of $\zeta_{1,v}$. The estimator $Z_{1,v}$, as shown in Zhang & Zhou (2010), is a U-statistic based on T as its kernel of degree $m = v+1 \leq n$. It is in this sense the proposed estimator $\hat{H}_z$ in (7) is considered an estimator in Turing's perspective.

In summary, the proposed estimator $\hat{H}_z$ not only differs greatly in form from that of $\hat{H}$, but also represents a fundamental switch in perspective, from that of (1) and (2) to that of (5) and (7). This switch in perspective brings about substantial gain in estimation accuracy for every distribution with finite entropy. The above claimed and other results with respect to $\hat{H}_z$ are established in Section 2. Several remarks on how the general spirit of the proposed estimator could be useful in other cases are given in Section 3. In Section 4, two bias-adjusted estimators are proposed for the purpose of reducing bias for small samples. In Section 5, the newly proposed estimators are compared, mostly by simulated studies, to two popular estimators, the jackknife estimator by Strong, Koberle, de Ruyter van Steveninck, & Bialek (1998) and the NSB estimator by Nemenman, Shafee, & Bialek (2002).

2 Main Results

The following lemma gives a necessary and sufficient condition for entropy H being finite.

Lemma 1 *For a given $\{p_k\}$, $H = -\sum_k p_k \ln(p_k) < \infty$ if and only if there exists a strictly increasing divergent sequence of positive real numbers a_n such that $\sum_{n=1}^{\infty}(na_n)^{-1} < \infty$ and, as $n \to \infty$,*

$$a_n \sum_{k=1}^{\infty} p_k(1-p_k)^n \to 1.$$

Proof. If an a_n satisfying the conditions exists, then $\exists\, n_0$ such that $\forall n > n_0$
$a_n \sum_k p_k(1-p_k)^n < 2$.

$$\sum_{v=1}^{\infty} \frac{1}{v} \sum_k p_k(1-p_k)^{v-1} = \sum_{v=1}^{n_0} \frac{1}{v} \sum_k p_k(1-p_k)^{v-1} + \sum_{v=n_0+1}^{\infty} \frac{1}{v} \sum_k p_k(1-p_k)^{v-1}$$

$$\leq \sum_{v=1}^{n_0} \frac{1}{v} \sum_k p_k(1-p_k)^{v-1} + 2\sum_{v=n_0+1}^{\infty} \frac{1}{va_v} < \infty.$$

On the other hand, if $\sum_{v=1}^{\infty}(1/v)\zeta_{1,v} < \infty$, then letting $a_n = 1/\zeta_{1,n}$ the following arguments show that all conditions of Lemma 1 are satisfied. First, $\zeta_{1,n}$ is strictly decreasing and therefore a_n is strictly increasing. Second, $\zeta_{1,n} = \sum_k p_k(1-p_k)^n \leq \sum_k p_k = 1$; and by the dominated convergence theorem, $\zeta_{1,n} \to 0$ and hence $a_n \to \infty$. Third, $\sum_{v=1}^{\infty} 1/(va_v) = \sum_{v=1}^{\infty}(1/v)\zeta_{1,v} < \infty$ and $a_n\zeta_{1,n} = 1$ by assumption and definition. □

Lemma 1 serves two primary purposes. First it encircles all the discrete distributions that may be of interest with regard to entropy estimation, *i.e.*, distributions with finite entropies. Second it provides a characterization of the tail of a distribution in terms of a sequence a_n and its conjugative relationship to $\zeta_{1,n}$. The rate of divergence of a_n characterizes the rate of tail decay of the underlying distribution $\{p_k\}$ as k increases. A faster (slower) rate of divergence of a_n signals a thinner (thicker) probability tail.

Let

$$M_n = \frac{5}{n}\left(\sum_{v=1}^{n-1}\frac{1}{v}\,\zeta_{1,v-1}^{1/2}\right)^2 . \tag{9}$$

The next lemma provides an upper bound for $Var(\hat{H}_z)$ under the general conditions and plays a central role in establishing many of the subsequent results.

Lemma 2 *For any probability distribution $\{p_k\}$, $Var(\hat{H}_z) \le M_n$.*

Corollary 1 *For any probability distribution $\{p_k\}$,*

$$Var(\hat{H}_z) < 5\left\{\frac{[1+\ln(n-1)]^2}{n}\right\} = O([\ln(n)]^2/n).$$

Proof of Corollary 1. Referring to (9) and noting $\zeta_{1,v} \le 1$ and that $\sum_{k=1}^{n-1} 1/v$ is the harmonic series hence with upper bound $1+\ln(n-1)$, $M_n < 5\left\{\frac{[1+\ln(n-1)]^2}{n}\right\} = O([\ln(n)]^2/n)$. □

Corollary 2 *For any probability distribution $\{p_k\}$ with finite entropy H,*

$$Var(\hat{H}_z) < 5(1+H)\left[\frac{1+\ln(n-1)}{n}\right] = O(\ln(n)/n).$$

The implications of Corollary 2 are quite significant. The uniform upper bound in Corollary 2 for the entire class of distributions with finite entropy decays faster than the upper bound for $Var(\hat{H})$, *i.e.*, $O([\ln(n)]^2/n)$, established in Antos & Kontoyiannis (2001) for the same class of distributions by a factor of $\ln(n)$. The uniform improvement in variance certainly suggests that entropy estimation in view of $\hat{H}_z$ is fundamentally more efficient than that of $\hat{H}$. Nevertheless the fact that the upper bound is proportional to $1 + H$ suggests that, for any fixed sample size n, a distribution with finite entropy can be found to have an arbitrarily large variance.

Proof of Corollary 2. Noting $\zeta_{1,v} \geq \zeta_{1,w}$ if $v \leq w$,

$$\begin{aligned} M_n &= \tfrac{5}{n}\left(\textstyle\sum_{v=1}^{n-1}\frac{1}{v}\,\zeta_{1,v-1}^{1/2}\right)^2 = \tfrac{5}{n}\left(\textstyle\sum_{v=1}^{n-1}\frac{1}{v}\,\zeta_{1,v-1}^{1/2}\right)\left(\textstyle\sum_{w=1}^{n-1}\frac{1}{w}\,\zeta_{1,w-1}^{1/2}\right) \\ &= \tfrac{5}{n}\left[\textstyle\sum_{v=1}^{n-1}\frac{1}{v^2}\zeta_{1,v-1} + 2\sum_{1\leq v<w\leq n-1}\frac{1}{vw}\zeta_{1,v-1}^{1/2}\zeta_{1,w-1}^{1/2}\right] \\ &\leq \tfrac{5}{n}\left[\textstyle\sum_{v=1}^{n-1}\frac{1}{v^2}\zeta_{1,v-1} + 2\sum_{1\leq v<w\leq n-1}\frac{1}{vw}\zeta_{1,v-1}\right] = \tfrac{5}{n}\left[\textstyle\sum_{w=1}^{n-1}\frac{1}{w}\sum_{v=1}^{n-1}\frac{1}{v}\zeta_{1,v-1}\right] \\ &= \tfrac{5}{n}\left[\textstyle\sum_{w=1}^{n-1}\frac{1}{w}\right]\left[\textstyle\sum_{v=1}^{n-1}\frac{1}{v}\zeta_{1,v-1}\right]. \end{aligned}$$

The expression in the first pair of bracket parentheses is the harmonic series which has a well known upper bound $1 + \ln(n-1)$. Consider the expression in the second pair of bracket parentheses.

$$\sum_{v-1}^{n-1}\frac{1}{v}\zeta_{1,v-1} = 1+\sum_{v=2}^{n-1}\frac{1}{v}\zeta_{1,v-1} < 1+\sum_{v=2}^{n-1}\frac{1}{v-1}\zeta_{1,v-1} = 1+\sum_{v=1}^{n}\frac{1}{v}\zeta_{1,v} < 1+\sum_{v=1}^{\infty}\frac{1}{v}\zeta_{1,v} = 1+H.$$

Therefore $M_n < 5(1+H)\frac{1+\ln(n-1)}{n} = O(\ln(n)/n)$. □

For clarity in proving Lemma 2, a few notations and two well-known lemmas in U-statistics are first given. For each i, $1 \leq i \leq n$, let X_i be a random variable such that $x_i = k$ indicates the event that the k^{th} letter of the alphabet is observed and $P(X_i = k) = p_k$. Let $x_1, \cdots, x_n$ be an *iid* sample, and let $X_1, \cdots, X_n$ denote the corresponding random observations. 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 $\psi_c(x_1, \cdots, x_c) = E[\psi(x_1, \cdots, x_c, X_{c+1}, \cdots, X_d)]$ and $\sigma_c^2 = Var[\psi_c(X_1, \cdots, X_c)]$. The following two lemmas 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 $0 \le c \le d \le m$, $\sigma_c^2/c \le \sigma_d^2/d$.*

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

$$Var(Z_{1,v}) \le \binom{n}{d}^{-1} \sum_{c=1}^{d} c \binom{d}{c} \binom{n-d}{d-c} \sigma_d^2/d = \frac{d}{n}\sigma_d^2. \tag{10}$$

Consider $\sigma_d^2 = Var[\psi(X_1, \cdots, X_d)] = E[\psi(X_1, \cdots, X_d)]^2 - [\sum_k p_k(1-p_k)^{d-1})]^2$.

$$E[\psi(X_1, \cdots, X_d)]^2 = \tfrac{1}{d^2} E\left[\left(\textstyle\sum_k 1_{[Y_k=1]}\right)\left(\textstyle\sum_j 1_{[Y_j=1]}\right)\right]$$

$$= \tfrac{1}{d^2} E\left(\textstyle\sum_k 1_{[Y_k=1]} + 2\sum_{1\le k<j<\infty} 1_{[Y_k=1]} 1_{[Y_j=1]}\right)$$

$$= \tfrac{1}{d}\textstyle\sum_k p_k(1-p_k)^{d-1} + \tfrac{2(d-1)}{d}\sum_{1\le k<j<\infty} p_k p_j (1-p_k-p_j)^{d-2}$$

$$\le \tfrac{1}{d}\textstyle\sum_k p_k(1-p_k)^{d-1} + \tfrac{2(d-1)}{d}\sum_{1\le k<j<\infty} p_k p_j (1-p_k-p_j+p_kp_j)^{d-2}$$

$$= \tfrac{1}{d}\textstyle\sum_k p_k(1-p_k)^{d-1} + \tfrac{2(d-1)}{d}\sum_{1\le k<j<\infty} \left[p_k(1-p_k)^{d-2} p_j(1-p_j)^{d-2}\right]$$

$$\le \tfrac{1}{d}\textstyle\sum_k p_k(1-p_k)^{d-1} + \tfrac{d-1}{d}\left[\sum_k p_k(1-p_k)^{d-2}\right]^2.$$

$$
\begin{aligned}
\sigma_d^2 &\le \tfrac{1}{d}\textstyle\sum_k p_k(1-p_k)^{d-1} + \tfrac{d-1}{d}\left[\sum_k p_k(1-p_k)^{d-2}\right]^2 - \left[\sum_k p_k(1-p_k)^{d-1}\right]^2 \\
&\le \tfrac{1}{d}\textstyle\sum_k p_k(1-p_k)^{d-1} + \left[\sum_k p_k(1-p_k)^{d-2}\right]^2 - \left[\sum_k p_k(1-p_k)^{d-1}\right]^2 \\
&= \tfrac{1}{d}\textstyle\sum_k p_k(1-p_k)^{d-1} + \left[\sum_k p_k(1-p_k)^{d-2} + \sum_k p_k(1-p_k)^{d-1}\right] \\
&\quad \times \left[\textstyle\sum_k p_k(1-p_k)^{d-2} - \sum_k p_k(1-p_k)^{d-1}\right] \\
&\le \tfrac{1}{d}\textstyle\sum_k p_k(1-p_k)^{d-1} + 2\left[\sum_k p_k(1-p_k)^{d-2}\right]\left[\sum_k p_k^2(1-p_k)^{d-2}\right] \\
&= \tfrac{1}{d}\zeta_{1,d-1} + 2\zeta_{1,d-2}\zeta_{2,d-2}.
\end{aligned}
$$

By (7) and (10)

$$
\begin{aligned}
Var(\hat{H}_z) &= \textstyle\sum_{v=1}^{n-1}\sum_{w=1}^{n-1}\frac{1}{vw}cov(Z_{1,v},Z_{1,w}) \le \sum_{v=1}^{n-1}\sum_{w=1}^{n-1}\frac{1}{vw}\sqrt{Var(Z_{1,v})Var(Z_{1,w})} \\
&\le \textstyle\sum_{v=1}^{n-1}\sum_{w=1}^{n-1}\frac{1}{vw}\sqrt{\frac{v+1}{n}\sigma_{v+1}^2\frac{w+1}{n}\sigma_{w+1}^2} = \frac{1}{n}\left(\sum_{v=1}^{n-1}\frac{\sqrt{v+1}}{v}\sigma_{v+1}\right)^2 \\
&\le \tfrac{1}{n}\left(\textstyle\sum_{v=1}^{n-1}\frac{\sqrt{v+1}}{v}\sqrt{\frac{1}{v+1}\zeta_{1,v}+2\zeta_{1,v-1}\zeta_{2,v-1}}\right)^2 = \frac{1}{n}\left[\sum_{v=1}^{n-1}\frac{1}{v}\sqrt{\zeta_{1,v}+2(v+1)\zeta_{1,v-1}\zeta_{2,v-1}}\right]^2.
\end{aligned}
$$

Finally noting that $\zeta_{1,v} \le \zeta_{1,v-1}$ and that $p(1-p)^{v-1}$ attains its maximum at $p = 1/v$ and hence $2(v+1)\zeta_{2,v-1} \le 2(v+1)(1/v)(1-1/v)^{v-1}\sum_k p_k < 4$, $Var(\hat{H}_z) < \frac{5}{n}\left(\sum_{v=1}^{n-1}\frac{1}{v}\zeta_{1,v-1}^{1/2}\right)^2 = M(n)$. □

An immediate benefit of Lemma 2 is the following statement on $\hat{H}_z$ under the condition of finite entropy.

Theorem 1 *Provided that entropy H is finite, $\hat{H}_z$ in (7) is a consistent estimator of entropy H.*

Proof. It suffices to show $E(\hat{H}_z - H)^2 = Var(\hat{H}_z) + |B_n|^2 \to 0$. The finite H implies $|B_n| \to 0$. For $Var(\hat{H}_z)$, consider its upper bound

$$
M_n = \frac{5}{n}\left(\sum_{v=1}^{n-1}\frac{1}{v}\,\zeta_{1,v-1}^{1/2}\right)^2 < \frac{5}{n^{1/2}}\left[\sum_{v=1}^{n-1}\frac{1}{v^{1+1/2}a_v^{1/2}}\left(a_v\zeta_{1,v-1}\right)^{1/2}\right]^2
$$

where a_v is as in Lemma 1. By Lemma 1, $a_v\zeta_{1,v-1}$ is bounded above, $\sum_{v=1}^{n-1} v^{-(1+1/2)}a_v^{-1/2} < \infty$, and therefore $M_n \to 0$. □

Let P be the class of all distributions on a countable alphabet with finite entropy. P contains thick tail distributions with very slowly divergent a_n's. Antos & Knotoyiannis (2001) showed that P is so rich in thick tail distributions that an estimator of H whose mean squared errors is uniformly bounded by a sequence $b_n \to 0$ for all distributions in P does not exist. The result of Antos & Knotoyiannis (2001) effectively forces the convergence rate characterization of entropy estimators into subclasses of P. Naturally P may be partitioned into layers characterized by the rate of divergence of a_n where a_n is as in Lemma 1. Consider the following conditions.

Condition 1 *For a probability sequence $\{p_k\}$, $n\,\zeta_{1,n} \to C \geq 0$.*

Condition 2 *For a probability sequence $\{p_k\}$, there exists a strictly increasing $a_n \to \infty$, such that, as $n \to \infty$, (1) $a_n/n \to 0$, (2) $n^{1/2}/a_n \to C_1 \geq 0$, and (3) $a_n\,\zeta_{1,n} \to C > 0$.*

Condition 3 *For a probability sequence $\{p_k\}$, there exist a strictly increasing $a_n \to \infty$ and a constant $\delta \in (0, 1/2)$, such that, as $n \to \infty$, (1) $a_n/n^{1/2} \to 0$, (2) $n^{\delta}/a_n \to C_1 \geq 0$, and (3) $a_n\,\zeta_{1,n} \to C > 0$.*

Conditions 1, 2 and 3 are mutually exclusive conditions. Let P_i, $i = 1, 2, 3$, be the subclasses of P under Conditions 1, 2 and 3, respectively. Let $P_4 = (P_1 \cup P_2 \cup P_3)^c$ where the complement is with respect to P.

It may be instructive to make several observations at this point. First, P_i, $1 \leq i \leq 4$, are defined in an order of the the tail decaying rate of the underlying distribution with P_1 having the fastest decaying rate. Second, it can be verified that $p_k = O(e^{-\lambda k})$ where $\lambda > 0$ satisfies $n\zeta_{1,n} \to C > 0$ and hence Condition 1. Condition 1 is therefore satisfied by all distributions with faster decaying tails than that of $p_k = O(e^{-\lambda k})$, including the distributions on any finite alphabet. Third, it can also be verified that, for $p_k = O(k^{-\lambda})$ where $\lambda > 1$, $n^{\delta}\,\zeta_{1,n} \to C > 0$ where $\delta = 1 - 1/\lambda$. It is seen here that $p_k = O(k^{-\lambda})$ where $\lambda \geq 2$ belongs to P_2 and that $p_k = O(k^{-\lambda})$ where $1 < \lambda < 2$ belongs to P_3. P_4 holds all other very thick-tailed distributions whose a_n's diverge so slow that they cannot be bounded below by a sequence diverging at a rate of $O(n^{\varepsilon})$ for any small $\varepsilon > 0$.

Theorem 2 *For any probability distribution $\{p_k\}$,*

1. *if $\{p_k\} \in P_1 \cup P_2$, then there exists $M_1(n) = O(n^{-1})$ such that $E(\hat{H}_z - H)^2 \leq M_1(n)$; and*

2. *if $\{p_k\} \in P_3$, then there exists $M_2(n) = O(n^{-2\delta})$ such that $E(\hat{H}_z - H)^2 \leq M_2(n)$.*

Proof. By Lemma 2, $E(\hat{H}_z - H)^2 \leq M(n) + |B_n|^2$ in general where $M(n)$ is as in (9) and $B_n = \sum_{v=n}^{\infty} \zeta_{1,v}/v$. For part 1, if Condition 1 holds then for any fixed $\varepsilon > 0$, there exists a sufficiently large but fixed n_0 such that for all $v \geq n_0$, $0 \leq v\,\zeta_{1,v-1} \leq C + \varepsilon$.

$$M_n = \tfrac{5}{n}\left[\sum_{v=1}^{n_0-1} \tfrac{1}{v}\,\zeta_{1,v-1}^{1/2} + \sum_{v=n_0}^{n-1} \tfrac{1}{v^{1+1/2}}(v\,\zeta_{1,v-1})^{1/2}\right]^2$$

$$\leq \tfrac{5}{n}\left[\sum_{v=1}^{n_0-1} \tfrac{1}{v} + (C+\varepsilon)^{1/2} \sum_{v=n_0}^{n-1} \tfrac{1}{v^{1+1/2}}\right]^2 \equiv V_1(n).$$

For a sufficiently large n,

$$B_n = \sum_{v=n}^{\infty} \frac{1}{v}\,\zeta_{1,v} = \sum_{v=n}^{\infty} \frac{1}{v^2}\,v\,\zeta_{1,v} \leq \sum_{v=n}^{\infty} \frac{(C+\varepsilon)}{v^2} \leq (C+\varepsilon)\int_{n-1}^{\infty} x^{-2}dx = \frac{C+\varepsilon}{n-1} \equiv B_1(n).$$

Letting $M_1(n) = V_1(n) + B_1^2(n) = O(n^{-1})$ establishes the desired result.

For part 1, if Condition 2 holds, for the corresponding a_n and C and any fixed $\varepsilon > 0$, there exists a sufficiently large but fixed n_0 such that for all $v \geq n_0$, $0 \leq a_v\zeta_{1,v-1} \leq C + \varepsilon$. Noting (2) of Condition 2,

$$M_n = \tfrac{5}{n}\left[\sum_{v=1}^{n_0-1} \tfrac{1}{v}\zeta_{1,v-1}^{1/2} + \sum_{v=n_0}^{n-1} \tfrac{1}{v a_v^{1/2}}(a_v\zeta_{1,v-1})^{1/2}\right]^2$$

$$\leq \tfrac{5}{n}\left[\sum_{v=1}^{n_0-1} \tfrac{1}{v} + (C+\varepsilon)^{1/2} \sum_{v=n_0}^{n-1} \tfrac{1}{v^{1+1/4}}\left(\tfrac{v^{1/2}}{a_v}\right)^{1/2}\right]^2$$

$$\leq \tfrac{5}{n}\left[\sum_{v=1}^{n_0-1} \tfrac{1}{v} + (C+\varepsilon)^{1/2}(C_1+\varepsilon)^{1/2} \sum_{v=n_0}^{n-1} \tfrac{1}{v^{1+1/4}}\right]^2 \equiv V_1(n).$$

For a sufficiently large n, and also noting (2) of Condition 2,

$$B_n = \sum_{v=n}^{\infty} \frac{1}{v}\zeta_{1,v} = \sum_{v=n}^{\infty} \frac{1}{v a_v} a_v \zeta_{1,v} \le \sum_{v=n}^{\infty} \frac{(C+\varepsilon)}{v^{1+1/2}} \left(\frac{v^{1/2}}{a_v}\right) \le \sum_{v=n}^{\infty} \frac{(C+\varepsilon)(C_1+\varepsilon)}{v^{1+1/2}} \equiv B_1(n).$$

Letting $M_1(n) = V_1(n) + B_1^2(n) = O(n^{-1})$ establishes the desired result.

For part 2, Condition 3 holds, for the corresponding a_n, δ, C, C_1 and any fixed $\varepsilon > 0$, there exists a sufficiently large but fixed n_0 such that for all $v \ge n_0$, $a_v > \frac{v^\delta}{C_1+\varepsilon}$ and

$$\begin{aligned} M_n &= \tfrac{5}{n}\left[\textstyle\sum_{v=1}^{n_0-1} \frac{1}{v}\zeta_{1,v-1}^{1/2} + \sum_{v=n_0}^{n-1} \frac{1}{v a_v^{1/2}} (a_v \zeta_{1,v-1})^{1/2}\right]^2 \\ &\le \tfrac{5}{n}\left[\textstyle\sum_{v=1}^{n_0-1} \frac{1}{v} + (C+\varepsilon)^{1/2} \sum_{v=n_0}^{n-1} \frac{1}{v a_v^{1/2}}\right]^2 \\ &= \tfrac{5}{n}\left[\textstyle\sum_{v=1}^{n_0-1} \frac{1}{v} + (C+\varepsilon)^{1/2} \sum_{v=n_0}^{n-1} \frac{1}{v^{1+\delta/2}} \left(\frac{v^\delta}{a_v}\right)^{1/2}\right]^2 \\ &\le \tfrac{5}{n}\left[\textstyle\sum_{v=1}^{n_0-1} \frac{1}{v} + (C+\varepsilon)^{1/2}(C_1+\varepsilon)^{1/2} \sum_{v=n_0}^{n-1} \frac{1}{v^{1+\delta/2}}\right]^2 \equiv V_2(n). \end{aligned}$$

For a sufficiently large n, and also noting (2) of Condition 3,

$$\begin{aligned} B_n &= \textstyle\sum_{v=n}^{\infty} \frac{1}{v}\zeta_{1,v} = \sum_{v=n}^{\infty} \frac{1}{v a_v} a_v \zeta_{1,v} \le \sum_{v=n}^{\infty} \frac{(C+\varepsilon)}{v a_v} \le \sum_{v=n}^{\infty} \frac{(C+\varepsilon)(C_1+\varepsilon)}{v^{1+\delta}} \\ &= \tfrac{(C+\varepsilon)(C_1+\varepsilon)}{\delta(n-1)^\delta} \equiv B_2(n). \end{aligned}$$

Letting $M_2(n) = V_2(n) + B_2^2(n) = O(n^{-2\delta})$ establishes the desired result. □

The statements of Theorem 2 give the convergence rates of upper bounds in mean squared errors for various types of distributions. Statement 1 says that, for all distributions with fast decaying tails, the bias of $\hat{H}_z$ decays sufficiently fast so that $|B_n|^2$ is dominated by $Var(\hat{H}_z)$ which converges at a rate of $O(1/n)$. It may be interesting to note that the so-called "fast decaying" distributions here include those with power decaying tails down to a threshold $\delta = 1/2$. Statement 2 says that, for each of the thick tailed distributions, the squared bias dominates the convergence in mean squared errors.

Example 1 *Suppose $p_k = O(k^{-\lambda})$ for some $\lambda > 1$. It can be verified that $n^\delta \sum_k p_k(1 - p_k)^n \to C > 0$ where $\delta = 1 - 1/\lambda$ for some constant C. By Theorem 2*

1. *if* $\lambda \geq 2$, *then* $E(\hat{H}_z - H)^2 \leq O\left(n^{-1}\right)$, *and*

2. *if* $1 < \lambda < 2$, *then* $E(\hat{H}_z - H)^2 = O\left(n^{-\frac{2(\lambda-1)}{\lambda}}\right)$.

Example 2 *Suppose* $p_k = O(e^{-\lambda k})$ *for some* $\lambda > 0$. *It can be verified that* $n \sum_k p_k(1 - p_k)^n \to C > 0$ *for some constant* C. *By Theorem 2,* $E(\hat{H}_z - H)^2 = O\left(n^{-1}\right)$.

Thus far, the advantage of $\hat{H}_z$ over $\hat{H}$ has been demonstrated mainly in a faster rate of convergence for the variance by a factor of $\ln(n)$ in the uniform upper bound in P in general and in P_4 in specific, and by a factor of $[\ln(n)]^2$ in P_1, P_2 and P_3. This advantage carries over in P_1 and P_2, but gets lost in P_3 in terms of convergence in mean squared errors. However $\hat{H}_z$ brings about a very different type of advantage to the subclass of distributions on any finite alphabet, denoted by $P_0 \subset P_1$.

If $\{p_k\}$ is defined on a finite alphabet with K letters, then letting $a_n = (1-p_0)^{-n}$ where $p_0 = \min\{p_k; k = 1, \cdots K\} > 0$,

$$a_n \zeta_{1,n} = \sum_{k=1}^{K} p_k[(1-p_k)/(1-p_0)]^n \to m_0 p_0$$

where m_0 is the multiplicity of letters in the alphabet with the same probability p_0. The fast divergence of a_n puts the distribution in P_1 and therefore by Theorem 2 $E(\hat{H}_z - H)^2 = O(1/n)$ which is shared by $E(\hat{H} - H)^2$. However the more interesting issue here is the bias of the estimator. The bias, for an arbitrary small $\varepsilon > 0$ and a sufficiently large n, noting that $(va_v)^{-1}$ is monotonically decreasing in v and the Euler-Maclaurin lemma,

$$|B_n| = \sum_{v=n}^{\infty} \frac{\zeta_{1,v}}{v} = \sum_{v=n}^{\infty} \frac{a_v \zeta_{1,v}}{v a_v} \leq (mp_0 + \varepsilon) \sum_{v=n}^{\infty} \frac{1}{v a_v} \leq (mp_0 + \varepsilon) \int_{n-1}^{\infty} \frac{1}{x a_x} dx$$

$$\leq -\frac{(mp_0+\varepsilon)}{\ln(1-p_0)} \frac{(1-p_0)^{n-1}}{n-1} = O\left(\frac{(1-p_0)^n}{n}\right).$$

The following corollary is given for completeness.

Corollary 3 *Let* $\{p_k\}$ *be a probability distribution defined on a finite alphabet. Then* $|B_n| = O\left(\frac{(1-p_0)^n}{n}\right)$ *where* $p_0 = \min\{p_k; k \in \mathbb{N}_K\}$.

3 Remarks

Section 2 above has made it clear that $\{\zeta_{1,v}; v \geq 1\}$ is of central importance to Shannon's entropy and to the proposed estimator. By representing a finite entropy in terms of $\zeta_{1,v}$ as in (5), it allows information of the tail beyond the observed data range to be recovered and utilized via Turing's formula. $\hat{H}_z$ offers a fundamentally different perspective in estimating entropy. However it must also be mentioned that the applicability of the methodology described is not limited to Shannon's entropy pe se. Any general index of the form $\sum_k p_k f(p_k)$ can be similarly treated provided that $f(p)$ is a positive analytic function over an interval containing $[0,1]$. The construction of the corresponding estimator starts with the Taylor expansion of $f(p)$ at $p=1$ and therefore

$$\sum_k p_k f(p_k) = \sum_{v\geq 1} b_v \zeta_{1,v} = \sum_{v=1}^{n-1} b_v \zeta_{1,v} + \sum_{v=n}^{\infty} b_v \zeta_{1,v}$$

provided that the index is finite. Let $\hat{H}_z = \sum_{v=1}^{n-1} b_v Z_{1,v}$. $\hat{H}_z$ has bias $|B(n)| = \sum_{v=n}^{\infty} b_v \zeta_{1,v}$.

For example, Rényi (1961) defined the entropy of order r, $H_r = (1-r)^{-1}\ln(\sum_k p_k^r)$ for $r \in (0,1)$. This index is considered a generalization of H since $\lim_{r\to 1} H_r = H$. Rényi's indices have an equivalent form $h_r = \sum_k p_k^r = \sum_k p_k \left[p_k^{-(1-r)}\right]$. To estimate a finite Rényi's entropy index of order r, first it can be verified that

$$h_r = \sum_{v=1}^{\infty}\left\{\prod_{i=1}^{v}\left[\frac{i-r}{i}\right]\right\}\zeta_{1,v} = \sum_{v=1}^{n-1}\left\{\prod_{i=1}^{v}\left[\frac{i-r}{i}\right]\right\}\zeta_{1,v} + \sum_{v=n}^{\infty}\left\{\prod_{i=1}^{v}\left[\frac{i-r}{i}\right]\right\}\zeta_{1,v}$$

and the corresponding estimator is

$$\hat{h}_r = \sum_{v=1}^{n-1}\left\{\prod_{i=1}^{v}\left[\frac{i-r}{i}\right]\right\} Z_{1,v}.$$

The convergence rate and other statistical properties of $\hat{h}_r$ can be passed on via transformation $\hat{H}_r = (1-r)^{-1}\ln(\hat{h}_r)$.

For indices of the form $H_u = \sum_k p_k^u f(p_k)$ where $u \geq 2$, an alternative representation would be $H_u = \sum_k p_k^u f(p_k) = \sum_{v=1}^{\infty} b_v \zeta_{u,v}$. The corresponding estimator would be $\hat{H}_u =$

$\sum_{v=1}^{n-u} b_v Z_{u,v}$ where, given in Zhang & Zhou (2010),

$$Z_{u,v} = \frac{n^{u+v}[n-(u+v)]!}{n!} \sum_k \left\{ \left[1_{[\hat{p}_k \geq \frac{u}{n}]} \prod_{i=0}^{u-1} \left(\hat{p}_k - \frac{i}{n} \right) \right] \left[\prod_{j=0}^{v-1} \left(1 - \hat{p}_k - \frac{j}{n} \right) \right] \right\}. \tag{11}$$

It may also be of some interest to mention that the mutual information of two discrete random variables X and Y can be written as

$$I(X,Y) = \sum_y \sum_x p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right) = -H_{X,Y} + H_X + H_Y$$

where $H_{X,Y}$, H_X and H_Y are entropies under the joint distribution of X and X, and the marginal distributions of X and Y respectively. Assuming the joint distribution of X and Y is defined on a discrete alphabet, the proposed methodology can be readily applied to each of the three entropies using the joint and the marginal samples. However the properties of the resulting estimator may not necessarily be easily derived as they depend on the nature of the correlation between X and Y.

Finally it is to be highlighted that $\hat{H}_z$ has an exponentially decaying bias as claimed in Corollary 3 is only true when K is finite in which case, the smallest non-zero probability is bounded away from zero. For the cases of infinite alphabet, the decaying rate of the bias depends on the tail of the distribution.

4 Bias Correction

While this paper mainly focuses on the consideration of convergence rates of $\hat{H}_z$, the practical consideration of bias in small sample cases is also important. The setup of $\hat{H}_z$ offers a natural way to correct its bias. Since the bias of $\hat{H}_z$ is explicitly expressed as $B_n = \sum_{v=n}^{\infty} \frac{1}{v} \zeta_{1,v} = \sum_{v=n}^{\infty} \frac{1}{v} \sum_k p_k (1-p_k)^v$. A correction is readily available by adding to $\hat{H}_z$ a correction term with p_k replaced by $\hat{p}_k$, $\hat{B}_n = \sum_{v=n}^{\infty} \frac{1}{v} \sum_k \hat{p}_k (1-\hat{p}_k)^v$. Let n_r, $r = 1, \cdots, n$, be the number of letters in the alphabet that are observed exactly r times in the sample of size n. Let

$$G(r,n) = \sum_{v=n}^{\infty} \frac{1}{v} \left(1 - \frac{r}{n}\right)^v .$$

Given a sample of size n, $\hat{B}_n$ may be re-expressed as,

$$\hat{B}_n = \sum_{v=n}^{\infty} \frac{1}{v} \sum_k \hat{p}_k (1-\hat{p}_k)^v = \sum_{v=n}^{\infty} \frac{1}{v} \sum_{r=1}^{n} \frac{r n_r}{n} \left(1 - \frac{r}{n}\right)^v = \sum_{r=1}^{n} \frac{r n_r}{n} G(r,n).$$

The resulting augmented estimator is

$$\hat{H}_z^{\star} = \hat{H}_z + \sum_{r=1}^{n} \frac{r n_r}{n} G(r,n). \tag{12}$$

Following Zahl (1977) and Strong, *et al.* (1998), a jackknife correction to $\hat{H}_z$ can also be made in the exactly the same way as is made to the plug-in estimator $\hat{H}$. Let the resulting jackknifed estimator be denoted as $\hat{H}_z^{JK}$. In the next section, these bias-corrected variants of the estimator, $\hat{H}_z^{JK}$ and $\hat{H}_z^{\star}$, will be compared to other estimators.

5 Comparison to Other Estimators

One of the most used general purpose estimators in practice is perhaps the jackknifed version of the plug-in estimator proposed by Strong *et al.* (1998) (which will be referred to as $\hat{H}^{JK}$ in the text below). $\hat{H}^{JK}$ reduces the bias of $\hat{H}$ significantly for small samples. Another intriguing estimator frequently referred to in the literature is the estimator proposed by Nemenman, *et al.* (2002), also known as the NSB estimator. The NSB estimator relies on a prior, and therefore a posterior distribution to reduce the effect of sample size effect on bias. However the NSB requires the knowledge of the finite cardinality K. When $K < \infty$ is known, the NSB generally works well in terms of bias for small samples across a range of distributions provided that the distributions do not have very long and thin tails.

In this section, an effort is made to compare the relative performances of six estimators, namely $\hat{H}$, $\hat{H}_z$, $\hat{H}^{JK}$, $\hat{H}_z^{JK}$, $\hat{H}_z^{\star}$, and the NSB, mainly by simulations. The main issue of interest is the bias in small sample cases. The comparison of the estimators is made in three different categories, 1) finite alphabets with known cardinalities, 2) finite alphabets with unknown cardinalities, and 3) infinite alphabets. As it turned out, $\hat{H}$ and $\hat{H}^{JK}$ are quite easily dominated by the other estimators across board. The remaining four estimators, $\hat{H}_z$, $\hat{H}_z^{JK}$, $\hat{H}_z^{\star}$, and the NSB, turned out to have a mixed picture.

To make the comparison simpler, the first effort is to eliminate $\hat{H}$ and $\hat{H}^{JK}$ from the list for further comparison. $\hat{H}$, $\hat{H}^{JK}$, $\hat{H}_z$, and $\hat{H}_z^{JK}$ are compared under the following six distributions:

1. $p_k = 1/100$, $k = 1, \cdots, 100$. Uniform Distribution. $H = 4.605170$.

2. $p_k = C/k$, $k = 1, \cdots, 100$, (Zipf's,) where $C = 0.192776$. $H = 3.680763$.

3. $p_k = C/k^2$; $k = 1, \cdots, 100$, where $C = 0.611627$. $H = 1.570208$.

4. $p_k = C/k^{2.0}$, $k = 1, \cdots$, where $C = 0.607927$. $H = 1.637621$.

5. $p_k = e^{-\lambda}\lambda^k/k!$, $k = 1, \cdots$, where $\lambda = e$. $H = 1.877220$.

6. $p_k = Ce^{-k^2}$, $k = 1, \cdots$, where $C = 1.718282$. $H = 1.040652$.

For each case in the simulation, the bias reported is based on the average of 5000 simulated samples. The sample size varies from 10 to several hundreds. The jackknifing is done using all sub-samples of sizes m, where $5 \leq m < n$. The comparative results of these four estimators turned out to be nearly identical in each of the above six cases: $\hat{H}$ is clearly dominated by $\hat{H}_z$, and $\hat{H}^{JK}$ is clearly dominated by $\hat{H}_z^{JK}$. Only the results for the Zipf's distribution ($p_k = O(k^{-1})$, $k = 1, \cdots, 100$) are given in Figure 1, in which, the darker solid curve represents the bias of $\hat{H}_z$; the lighter solid curve does that of $\hat{H}$; the darker dashed curves does that of $\hat{H}_z^{JK}$; and the lighter dashed curve does that of $\hat{H}^{JK}$.

Next $\hat{H}_z^{JK}$, $\hat{H}_z^{\star}$, and the NSB, are compared.

For an alphabet with a known finite cardinality K, among these estimators, only the NSB requires the knowledge of K, and when available not surprisingly performs better than the rest in many simulated small sample cases across a range of distributions. Generally the BSN performs much better than $\hat{H}_z^\star$ but somewhat better than $\hat{H}_z^{JK}$. Only the results for the Zipf's distribution ($p_k = O(k^{-1})$, $k = 1, \cdots, 100$) are graphed in Figure 2, where the darker solid curve represents the bias of the NSB; the lighter solid curve does that of $\hat{H}_z^{JK}$; and the dashed curves does that of $\hat{H}_z^\star$. The qualitative comparison of the relative performances observed in Figure 2 is quite typical in a range of distributions simulated. However a general claim of the NSB's superiority cannot be made even for cases of a known finite K. The NSB has a tendency to overestimate entropy when the underlying distribution has a very long and thin tail, for examples, a Poisson distribution or a one-sided discrete Gaussian distribution truncated at a very large value of k, say $K = 10^{10}$. Both of these cases are alluded to in Figures 3 and 4, although these figures are for un-truncated cases. With larger samples however, also not surprisingly $\hat{H}_z$ and almost equivalently $\hat{H}_z^\star$ dominate the field since the exponentially decaying rate of the bias takes effect.

For an alphabet with an unknown finite cardinality K, the NSB encounters an added difficulty of fixing the value of K. Strictly speaking, the NSB requires a known K. When K is unknown, it needs to be estimated or assigned before an estimate could be obtained. However in many situations, the NSB proves to be quite robust against errors in estimated or assigned K, and resiliently gives smaller biases for small samples. The NSB does not perform as well when the estimated or assigned value of K is grossly wrong, particularly when the underlying distributions have very long and thin tails. With larger samples, $\hat{H}_z$ and $\hat{H}_z^\star$ continue to dominate the field.

For a countably infinite alphabet, the cardinality K required by the NSB loses its meaning and becomes a pure computational parameter. However for some distributions, such as $p_k = O(k^{-2})$ $k \geq 1$, letting $K = 10^{10}$, the NSB continues to outperform $\hat{H}_z^{JK}$ and $\hat{H}_z^\star$. On the other hand, for thin tailed distributions, such as Poisson and Gaussian-like distributions, the NSB overestimates H with small samples and has much larger bias than those of $\hat{H}_z$ and $\hat{H}_z^\star$. Figure 3 demonstrates that fact under the Poisson distribution with $\lambda = e$; and Figure 4 demonstrates the same under the distribution $p_k = O(e^{-k^2})$, $k \geq 1$. In both Figures 3 and 4, the darker solid curves represent the biases of NSB, the lighter solid curves do that of $\hat{H}_z^\star$, and the lighter dashed curves do that of $\hat{H}_z$.

In summary, it would seem to be reasonable to make the following three observations based on the numerical studies:

1. $\hat{H}_z$ outperforms $\hat{H}$ and $\hat{H}_z^{JK}$ outperforms $\hat{H}^{JK}$ under a wide range of distributions and sample sizes.

2. $\hat{H}_z^{JK}$, $\hat{H}_z^\star$ and the NSB have mixed relative performances for small samples.

3. $\hat{H}_z$, $\hat{H}_z^{JK}$, and $\hat{H}_z^\star$ outperform the others under a wide range of distributions for large samples.

References

[1] Antos, A. & Kontoyiannis, I. (2001). *Convergence properties of functional estimates for discrete distributions*, Random Structures & Algorithms, 19, 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, 333-336.

[3] Beirlant, J., Dudewicz, E.J., Györfi, L. & Meulen, E.C. (2001). *Nonparametric Entropy Estimation: An Overview*, International Journal of the Mathematical Statistics Sciences, 6, 17-39.

[4] Chao, A. and Shen, T.J. (2003) *Nonparametric estimation of Shannons index of diversity when there are unseen species in sample*, Environmental and Ecological Statistics, 10, 429443.

[5] Esty, W.W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. Annal of Statistics, 11, 905-912.

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

[7] Grassberger, P. (1988). *Finite sample corrections to entropy and dimension estimates*, Physics Letters A, 128 (6-7), 369-373.

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

[9] Hoeffding, W. (1948). *A class of statistics with asymptotically normal distribution*, Annals of Mathematical Statistics, 19 (3), 293-325.

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

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

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

[13] Panzeri, S., Senatore, R., Montemurro, M.A., and Petersen, R.S. (2007). *Correcting for the sampling bias problem in spike train information measures*, J. Neurophysiol 98, 10641072.

[14] Rényi, A. (1961). *On Measures of Entropy and Information*, Proc. 4th Berk. Symp. Math. Statist. and Prob., University of California Press, 1, 547-461.

[15] Robbins, H.E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. Annals of Statistics, 39 (1), 256-257.

[16] Schürmann, T. (2004). *Bias analysis in entropy estimation*, J. Phys. A: Math. Gen. 37, 295-301.

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

[18] Simpsom, E.H. (1949). *Measurement of diversity.* Nature, 163, 688.

[19] 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), 197-200.

[20] Victor, J.D. (2000). *Asymptotic bias in iinformation estimates and the exponential (bell) polynomials*, Neural Computation, 12, 2797–2804.

[21] Vu, V.Q., Yu, B., and Kass, R.E. (2007). *Coverage-adjusted entropy estimation*, Statist. Med., 26, 4039-4060.

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

[23] Zhang, Z. & Huang, H. (2008). *A sufficient normality condition for Turing's Formula*, Journal of Nonparametric Statistics, 20 (5), 431-446.

[24] Zhang, C.-H. & Zhang, Z. (2009). *Asymptotic normality of a nonparametric estimator of sample coverage*, Annals of Statistics, 37 (5A), 2582-2595.

[25] Zhang, Z. & Zhou, J. (2010). *Re-parameterization of multinomial distribution and diversity indices*, J. of Statistical Planning and Inference, 140 (7), 1731-1738.

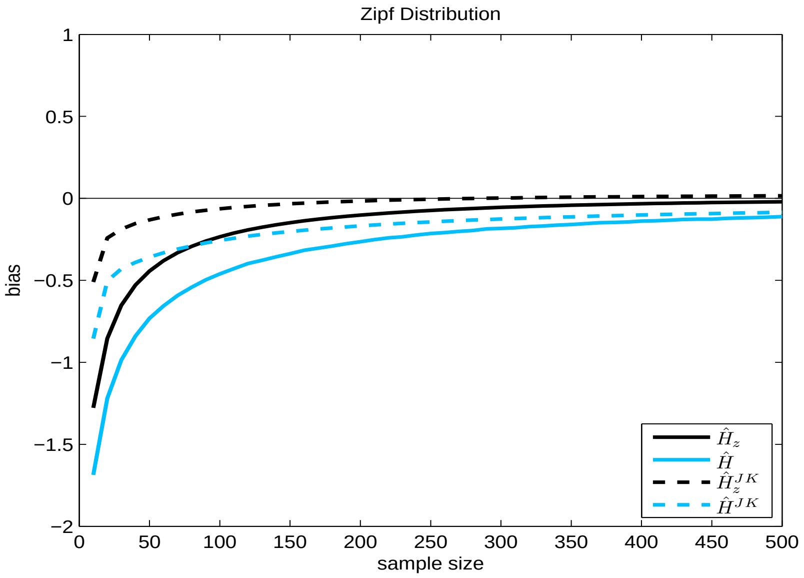

Figure 1: $\hat{H}_z$ vs. $\hat{H}$ and $\hat{H}_z^{JK}$ vs. $\hat{H}^{JK}$ under Zipf

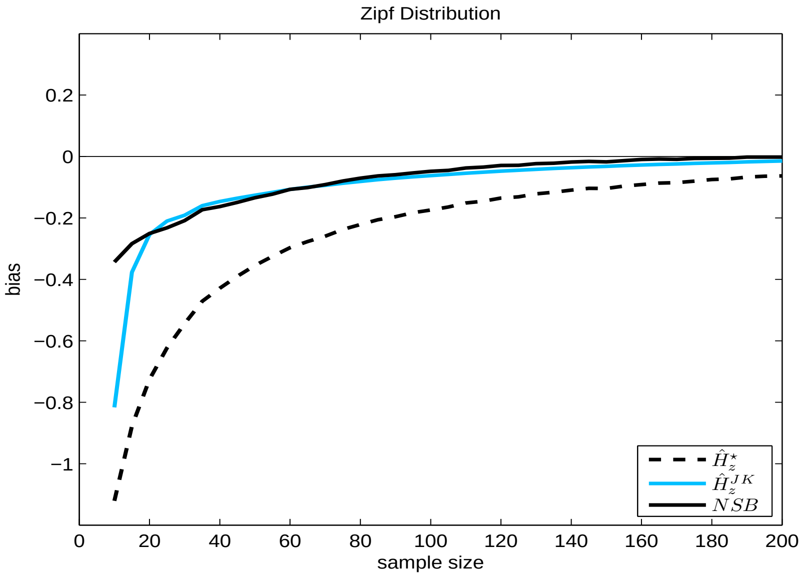

Figure 2: $\hat{H}_z^\star$, $\hat{H}_z^{JK}$, and NSB under Zipf

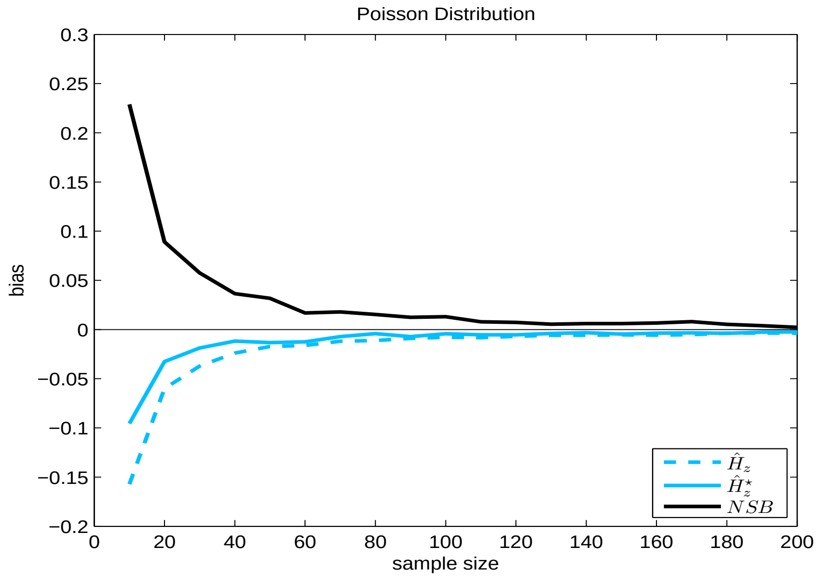

Figure 3: $\hat{H}_z$, $\hat{H}^\star_z$, and NSB under Poisson

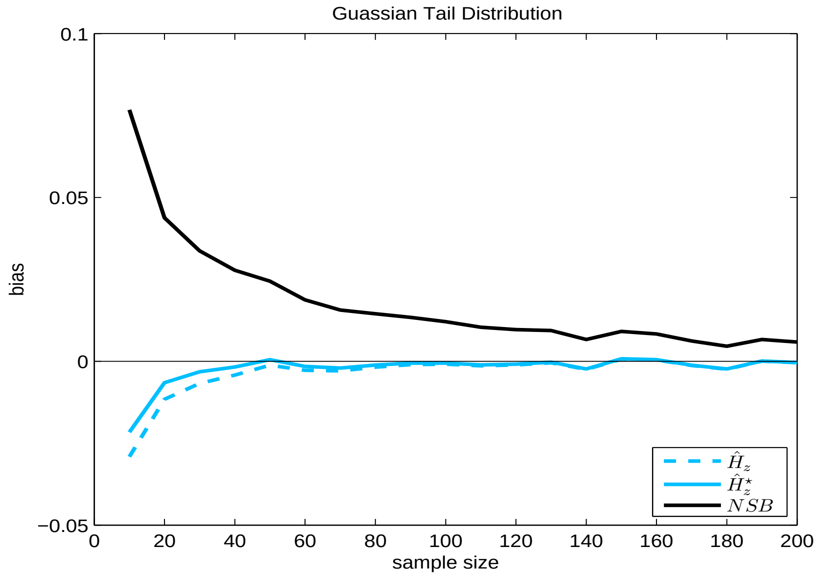

Figure 4: $\hat{H}_z$, $\hat{H}^\star_z$, and NSB under Gaussian