A Normal Law for the Plug-in Estimator of Entropy

Zhiyi Zhang and Xing Zhang

***Abstract*—This paper establishes a sufficient condition for the asymptotic normality of the plug-in estimator of Shannon's entropy defined on a countable alphabet. The sufficient condition covers a range of cases with countably infinite alphabets, for which no normality results were previously known.**

***Index Terms*—Entropy, plug-in estimator, asymptotic normality.**

I. Introduction

Let $\{p_k\}$ be a probability distribution on an alphabet $\mathscr{X} = \{\ell_k; 1 \le k \le K\}$, where K denotes either a finite integer or ∞. Let P_X be a random variable such that $P(P_X = p_k) = p_k$. Entropy in the form of

$$H = E(-\ln P_X) = -\sum_k p_k \ln p_k$$

was introduced by Shannon (1948) and is often referred to as Shannon's Entropy. Let $X_1, \cdots, X_n$ be an iid sample from $\mathscr{X}$ according to $\{p_k\}$, $\{y_{k,n} = \sum_{i=1}^n 1[X_i = \ell_k]\}$ be the sequence of observed counts of letters, and $\{\hat{p}_{k,n} = y_{k,n}/n\}$. The plug-in estimator for H, given by

$$\hat{H}_n = -\sum_k \hat{p}_{k,n} \ln \hat{p}_{k,n}$$

plays a central role in the literature. $\hat{H}_n$ is simple and intuitive; and it often serves as a reference estimator for other estimators, many of which were derived based on $\hat{H}_n$.

When K is fixed and finite,

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

where $\sigma^2 = Var(-\ln P_X) > 0$ has long been known. See Miller and Madow (1954) and Basharin (1959). In this case, it is also known that

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

$$Var(\hat{H}_n) = \tfrac{1}{n}\left(\sum_{k=1}^K p_k \ln^2 p_k - H^2\right) + \tfrac{K-1}{2n^2} + O(n^{-3}).$$

See Miller (1955), Basharin (1959) and Harris (1975).

When $K = K(n)$ is assumed to dynamically vary as the sample size n increases, *i.e.*, $\{p_{k,n}; k = 1, 2, \cdots, K(n)\}$, Paninski (2003) established a normal law for $\hat{H}_n$, stated as Lemma 1 below.

When K is infinite, Antos and Kontoyiannis (2001) obtained different rates of convergence for $\hat{H}_n$ under a variety of tail conditions on $\{p_k\}$. They also showed that no universal rate of convergence exists for any sequence of estimators. However no results regarding the asymptotic normality of $\hat{H}_n$ were known. This paper seeks to lay down a pebble in that blank space by presenting a sufficient normality condition for $\hat{H}_n$ when the cardinality of $\mathscr{X}$ is countably infinite. More specifically, the sufficient condition is satisfied by distributions with tails decaying at the rate of $k^{-2}(\ln k)^{-2}$, but not by those with tails decaying at the rate of $k^{-2}(\ln k)^{-1}$.

Zhiyi Zhang and Xing Zhang are with the Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, 28223 USA e-mail: zzhang@uncc.edu.

Manuscript received XX xx, 2011; revised December 07, 2011.

II. Main Results.

Theorem 1. *For any non-uniform distribution $\{p_k; k \ge 1\}$ satisfying $E(lnP_X)^2 < \infty$, if there exists an integer valued function $K(n)$ such that, $K(n) \to \infty$, $K(n) = o(\sqrt{n})$ and $\sqrt{n}\sum_{k \ge K(n)} p_k lnp_k \to 0$, as $n \to \infty$, then*

$$\sqrt{n}(\hat{H}_n - H)/\sigma \xrightarrow{D} N(0,1)$$

where $\sigma^2 = Var(-\ln P_X)$.

A proof of Theorem 1 requires Lemmas 1 and 2 below. Lemma 1 is due to Paninski (2003).

Lemma 1. *Let $\{p_{k,n}; k = 1, \cdots, K(n)\}$ be a probability distribution, P_X be a random variable such that $P(P_X = p_{k,n}) = p_{k,n}$, and*

$$\tau_n^2 = Var(-\ln P_X) = \sum_{k=1}^{K(n)} p_{k,n} \ln^2 p_{k,n} - \left(\sum_{k=1}^{K(n)} p_{k,n} \ln p_{k,n}\right)^2.$$

If $K(n) = o(\sqrt{n})$ and $\liminf_{n\to\infty} n^{1-\alpha}\tau_n^2 > 0$ for some $\alpha > 0$, then

$$\sqrt{n}(\hat{H}_n - H)/\tau_n \xrightarrow{D} N(0,1).$$

Lemma 2. *For any distribution $\{p_k; k \ge 1\}$, if there exists an integer valued function $K(n)$ such that as $n \to \infty$, $K(n) \to \infty$, and $\sqrt{n}\sum_{k \ge K(n)} p_k \ln p_k \to 0$, then $\sqrt{n}\ln n \sum_{k \ge K(n)} p_k \to 0$.*

Proof. Let $p_n^* = \sum_{k \ge K(n)} p_k$. Since $1 < -\ln p_n^*$ for a sufficiently large n,

$$\begin{aligned} 0 \le \sqrt{n}p_n^* &\le -\sqrt{n}p_n^* \ln p_n^* \\ &= -\sqrt{n}\textstyle\sum_{k \ge K(n)} p_k \ln p_n^* \\ &\le -\sqrt{n}\textstyle\sum_{k \ge K(n)} p_k \ln p_k \to 0. \end{aligned}$$

$\sqrt{n}p_n^* \to 0$ implies $p_n^* = \alpha_n n^{-1/2}$ where $\alpha_n = o(1)$. On the other hand, since $\alpha_n \ln \alpha_n \to 0$, $-\sqrt{n}p_n^* \ln p_n^* = \alpha_n(\ln\sqrt{n} - \ln \alpha_n) \to 0$ implies $\alpha_n = \beta_n/\ln\sqrt{n}$ where $\beta_n = o(1)$. Hence $\sqrt{n}\ln n \sum_{k \ge K(n)} p_k = 2\beta_n \to 0$. □

Proof of Theorem 1. Consider a modified probability distribution $\{p_{k,n}; k = 1, \cdots, K(n)\}$ based on $\{p_k\}$ as follows. Let

$$p_{k,n} = \begin{cases} p_k, & \text{for } 1 \le k \le K(n)-1 \\ \\ \sum_{k \ge K(n)} p_k \equiv p_n^*, & \text{for } k = K(n). \end{cases}$$

Since $E(\ln P_X)^2 = \sum_k p_k \ln^2 p_k < \infty$ implies $H = -\sum_k p_k \ln p_k < \infty$, we have

$$\begin{aligned} 0 \le p_n^* \ln^2 p_n^* &= \textstyle\sum_{k \ge K(n)} p_k \ln^2 p_n^* \\ &\le \textstyle\sum_{k \ge K(n)} p_k \ln^2 p_k \to 0, \text{ and} \end{aligned}$$

$$\begin{aligned} 0 \le -p_n^* \ln p_n^* &= \textstyle\sum_{k \ge K(n)} (-p_k \ln p_n^*) \\ &\le \textstyle\sum_{k \ge K(n)} (-p_k \ln p_k) \to 0. \end{aligned}$$

Let $\tau_n^2 = Var(-\ln P_X)$ under the modified distribution $\{p_{k,n}\}$. After a few algebraic steps,

$$\begin{aligned} \sigma^2 - \tau_n^2 = &(\textstyle\sum_{k \ge K(n)} p_k \ln^2 p_k - p_n^* \ln^2 p_n^*) \\ &-(-\textstyle\sum_{k \ge K(n)} p_k \ln p_k + p_n^* \ln p_n^*) \\ &\times(-\textstyle\sum_{k \ge K(n)} p_k \ln p_k - p_n^* \ln p_n^* \\ &-2\textstyle\sum_{k=1}^{K(n)-1} p_k \ln p_k). \end{aligned} \tag{1}$$

It is clear that the first term in (1) converges to zero, that the first factor of the second term converges to zero, and that the second factor of the second term converges to $2H < \infty$. Therefore $\tau_n \to \sigma$, and hence by Lemma 1, $\sqrt{n}\sum_{k=1}^{K(n)}(-\hat{p}_{k,n} \ln \hat{p}_{k,n} + p_{k,n} \ln p_{k,n}) \xrightarrow{D} N(0, \sigma^2)$. However

$$\begin{aligned} &\sqrt{n}(\hat{H}_n - H) \\ &-\sqrt{n}\textstyle\sum_{k=1}^{K(n)}(-\hat{p}_{k,n} \ln \hat{p}_{k,n} + p_{k,n} \ln p_{k,n}) \\ &= \sqrt{n}\textstyle\sum_{k \ge K(n)}(-\hat{p}_{k,n} \ln \hat{p}_{k,n}) \\ &-\sqrt{n}\textstyle\sum_{k \ge K(n)}(-p_k \ln p_k) + \sqrt{n}\hat{p}_n^* \ln \hat{p}_n^* \\ &-\sqrt{n}\ p_n^* \ln p_n^* \end{aligned} \tag{2}$$

where $\hat{p}_n^* = \sum_{k \ge K(n)} y_{k,n}/n$. The proof is complete if it is shown that the right hand side of (2) is $o_p(1)$. Toward that end, it is to show that each of the four terms in the last expression of (2) is $o_p(1)$.

The second term converges to zero in probability by the condition of Theorem 1. The fact that the fourth term converges to zero is established in the proof of Lemma 2. For the first and third terms, we first observe $-\hat{p}_{k,n} \ln \hat{p}_{k,n} \le \hat{p}_{k,n} \ln n$ and $-\hat{p}_n^* \ln \hat{p}_n^* \le \hat{p}_n^* \ln n$, and then observe the following two inequalities

$$\begin{gathered} 0 \le \sqrt{n}\textstyle\sum_{k \ge K(n)}(-\hat{p}_{k,n} \ln \hat{p}_{k,n}) \le \sqrt{n}(\ln n)\hat{p}_n^*, \\ \text{and} \\ 0 \le -\sqrt{n}\hat{p}_n^* \ln \hat{p}_n^* \le \sqrt{n}(\ln n)\hat{p}_n^*. \end{gathered} \tag{3}$$

Since, by Lemma 2, $E[\sqrt{n}(\ln n)\hat{p}_n^*] = \sqrt{n}(\ln n)p_n^* \to 0$ and, noting $\sqrt{n}(\ln n)\hat{p}_n^* \ge 0$, $\sqrt{n}(\ln n)\hat{p}_n^* = o_p(1)$. By (3), both the first and the third terms converge to zero in probability. The theorem follows by Slutsky's lemma. □

Let $\hat{\sigma}_n^2 = \sum_k \hat{p}_{k,n} \ln^2 \hat{p}_{k,n} - (\sum_k -\hat{p}_{k,n} \ln \hat{p}_{k,n})^2$.

Corollary 1. *Under the condition of Theorem 1,*

$$\sqrt{n}(\hat{H}_n - H)/\hat{\sigma}_n \xrightarrow{D} N(0,1).$$

Corollary 1 provides a means of large sample inference on H. A proof of Corollary 1 requires the following lemma due to Devroye (1991).

Lemma 3. *Let $X_1, \cdots, X_n$ be independent random variables on $\mathscr{X}$, and assume that $\hat{F}_n$: $\mathscr{X}^n \to \mathscr{R}$ satisfies, for $1 \le i \le n$,*

$$\begin{aligned} sup_{x_1, \cdots, x_n, x_i' \in \mathscr{X}} &|\hat{F}_n(x_1, \cdots, x_n) \\ &-\hat{F}_n(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)| \le c_i. \end{aligned}$$

Then $Var\{\hat{F}_n(X_1, \cdots, X_n)\} \le \frac{1}{4}\sum_{i=1}^n c_i^2$.

Proof of Corollary 1. Let

$$\begin{aligned} \hat{F}_n \equiv \hat{F}_n(X_1, \cdots, X_n) &= \textstyle\sum_k f(\hat{p}_{k,n}) \\ &= \textstyle\sum_k \hat{p}_{k,n} \ln^2 \hat{p}_{k,n}. \end{aligned}$$

We first want to show $\lim_{n\to\infty} E(\hat{F}_n - F)^2 = 0$ for $F \equiv \sum_k p_k \ln^2 p_k < \infty$.

For all integers $0 \le i < n$ and $n \ge 21 > e^3$, $\left|\frac{i+1}{n}\ln^2(\frac{i+1}{n}) - \frac{i}{n}\ln^2(\frac{i}{n})\right| \le \frac{\ln^2 n}{n}$. Therefore

$$\begin{aligned} sup_{x_1, \cdots, x_n, x_i' \in \mathscr{X}} &\left|\hat{F}_n(x_1, \cdots, x_n)\right. \\ &\left.-\hat{F}_n(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)\right| \le \tfrac{2\ln^2 n}{n}. \end{aligned}$$

By Lemma 3, $Var(\hat{F}_n) \le \frac{\ln^4 n}{n} \to 0$. For each k, $\hat{p}_{k,n} \xrightarrow{as} p_k$, $f(\hat{p}_{k,n}) \xrightarrow{as} f(p_k)$, $f(\hat{p}_{k,n}) \le e^{-2}\ln^2 e^{-2} = 4e^{-2}$, so $Ef(\hat{p}_{k,n}) \to f(p_k)$.

Since $0 \le f(\hat{p}_{k,n}) \le 4e^{-2}$, by Fatou's Lemma,

$$\begin{aligned} &\limsup_{n\to\infty} \textstyle\sum_k Ef(\hat{p}_{k,n}) \\ &\quad\le \textstyle\sum_k \limsup_{n\to\infty} Ef(\hat{p}_{k,n}) \\ &\quad= \textstyle\sum_k f(p_k) \text{ and} \end{aligned}$$

$$\begin{aligned} &\liminf_{n\to\infty} \textstyle\sum_k Ef(\hat{p}_{k,n}) \\ &\quad\ge \textstyle\sum_k \liminf_{n\to\infty} Ef(\hat{p}_{k,n}) \\ &\quad= \textstyle\sum_k f(p_k); \text{ and therefore} \end{aligned}$$

$$\begin{aligned} &\lim_{n\to\infty} E\hat{F}_n = \lim_{n\to\infty} \textstyle\sum_k Ef(\hat{p}_{k,n}) \\ &\quad= \textstyle\sum_k f(p_k) = F. \end{aligned}$$

By Theorem 1, $\hat{H}_n^2 \xrightarrow{p} H^2$, and therefore $\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2$. Finally the corollary follows by Theorem 1 and Slusky's lemma. □

Example 1. *If $p_k = C_\lambda k^{-\lambda}$ where $\lambda > 1$, the sufficient condition of Theorem 1 holds for $\lambda > 2$ but not for $1 < \lambda \le 2$.*

Note

$$\begin{aligned} &\sqrt{n}\textstyle\sum_{k \ge K(n)}(-p_k \ln p_k) \sim \\ &\quad\sqrt{n}\textstyle\int_{K(n)}^\infty \frac{C_\lambda}{x^\lambda} \ln\left(\frac{x^\lambda}{C_\lambda}\right) dx \\ &\quad= \frac{C_\lambda \lambda}{\lambda - 1}\frac{\sqrt{n}\ln K(n)}{(K(n))^{\lambda-1}} + \left(\frac{C_\lambda \lambda}{(\lambda-1)^2} - \frac{C_\lambda \ln C_\lambda}{\lambda - 1}\right) \\ &\quad\frac{\sqrt{n}}{(K(n))^{\lambda-1}} \sim \frac{C_\lambda \lambda}{\lambda - 1}\frac{\sqrt{n}\ln K(n)}{(K(n))^{\lambda-1}}. \end{aligned}$$

If $\lambda > 2$, letting $K(n) \sim n^{1/\lambda}$, $\frac{C_\lambda \lambda}{\lambda-1}\frac{\sqrt{n}\ln K(n)}{(K(n))^{\lambda-1}} = \frac{C_\lambda}{\lambda-1}\frac{\ln n}{n^{1/2-1/\lambda}} \to 0$.

If $1 < \lambda \leq 2$, for any $K(n)$ satisfying $K(n) \sim o(\sqrt{n})$ and a sufficiently large n, $\frac{C_\lambda \lambda}{\lambda-1} \frac{\sqrt{n} \ln K(n)}{(K(n))^{\lambda-1}} \geq \frac{C_\lambda \lambda}{\lambda-1} \frac{\sqrt{n}}{n^{\lambda/2-1/2}} \geq \frac{C_\lambda \lambda}{\lambda-1} \geq 2C_\lambda > 0$.

Example 2. *If $p_k = C_\lambda e^{-\lambda k}$ for any $\lambda > 0$, then the sufficient condition of Theorem 1 holds.*

Letting $K(n) \sim \lambda^{-1} \ln n$, for a sufficiently large n,

$$\begin{aligned}&\sqrt{n} \textstyle\sum_{k \geq K(n)} (-p_k \ln p_k) \\ &\quad \sim -\sqrt{n} \textstyle\int_{\ln n^{1/\lambda}}^{\infty} C_\lambda e^{-\lambda x} \ln(C_\lambda e^{-\lambda x}) dx \\ &\quad \sim \tfrac{C_\lambda}{\lambda} (\ln n) n^{-1/2} \to 0.\end{aligned}$$

Example 3. *If $p_k = C/(k^2 \ln^2 k)$, then the sufficient condition of Theorem 1 holds.*

Letting $K(n) \sim \sqrt{n}/\ln \ln n$, for a sufficiently large n,

$$\begin{aligned}&\sqrt{n} \textstyle\sum_{k \geq K(n)} (-p_k \ln p_k) \\ &\quad \sim \sqrt{n} C \textstyle\int_{K(n)}^{\infty} \frac{2 \ln x + 2 \ln \ln x - \ln C}{x^2 \ln^2 x} dx \\ &\quad \sim 2\sqrt{n} C \textstyle\int_{K(n)}^{\infty} \frac{1}{x^2 \ln x} dx \leq \frac{2C\sqrt{n}}{K(n) \ln K(n)} \to 0.\end{aligned}$$

Example 4. *If $p_k = C/(k^2 \ln k)$, the sufficient condition of Theorem 1 does not hold.*

For any $K(n)$ satisfying with $K(n) \sim o(\sqrt{n})$, for a sufficiently large n,

$$\begin{aligned}&\sqrt{n} \textstyle\sum_{k \geq K(n)} (-p_k \ln p_k) \\ &\quad \sim \sqrt{n} C \textstyle\int_{K(n)}^{\infty} \frac{2 \ln x + \ln \ln x - \ln C}{x^2 \ln x} dx \\ &\quad \sim \sqrt{n} C \textstyle\int_{K(n)}^{\infty} \frac{2}{x^2} dx = \frac{2C\sqrt{n}}{K(n)} \to \infty.\end{aligned}$$

III. Remarks.

Under distributions $p_k = C/k^\lambda$, a necessary condition for $\sqrt{n}(\hat{H}_n - H)$ to hold asymptotic normality is $\lambda > 2$ since the bias terms $E[(\hat{H}_n - H)]$ decays at rate of $n^{-(\lambda-1)/\lambda}$, no faster than $n^{-1/2}$ if $\lambda \in (1, 2]$, *e.g.*, see Theorem 7 of Antos and Kontoyiannis (2001) . On the other hand, as shown in Example 3, $\sqrt{n}(\hat{H}_n - H)$ does have asymptotic normality when $p_k = C/(k^2 \ln^2 k)$. Even though Theorem 1 gives only a sufficient condition, the band of distributions which are not covered by the sufficient condition but may still support asymptotic normality of $\hat{H}_n$ must be, if existed, very narrow. In other words, there is a clear threshold in the space of $\{p_k\}$ which encircles a subclass of thick tail distributions for which the plug-in estimator would not hold normality; and that threshold splits the class of distributions with power decaying tails.

While it may still be of theoretical interest to find sharper sufficient normality conditions for $\hat{H}_n$, it is perhaps practically more important to develop new estimators with faster decaying bias which could support asymptotic normality under thicker tail distributions not supported by the plug-in. Such estimators, however, by no means necessarily exist.

References

[1] A. Antos and I. Kontoyiannis. *Convergence properties of functional estimates for discrete distributions*, Random Structures and Algorithm, 19, 163-193, 2001.

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

[3] L. Devroye. *Expontential inequalities in nonparametric estimation*, In Nonparametirc Functional Estimation and Related Topics, ed. G. Roussas, NATO ASI Series, Kluwer Academic, Dordrecht, 31-44, 1991.

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

[5] G.A. Miller and W.G. Madow. *On the maximum likelihood estimate of the Shannon-Wiener measure of information*, Air Force Cambridge Research Center Technical Report 54-75, 1954.

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

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

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