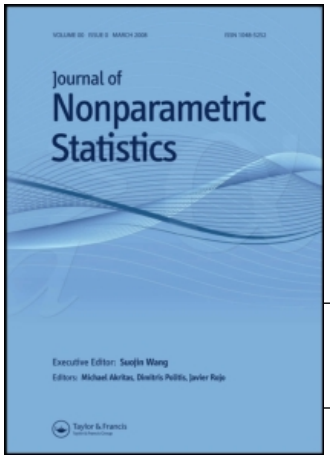

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645758

A sufficient normality condition for Turing's formula

Zhiyi Zhang [a]; Hongwei Huang [a]
[a] Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, USA

Online Publication Date: 01 July 2008

To cite this Article: Zhang, Zhiyi and Huang, Hongwei (2008) 'A sufficient normality condition for Turing's formula', Journal of Nonparametric Statistics, 20:5, 431 — 446

To link to this article: DOI: 10.1080/10485250802172126
URL: http://dx.doi.org/10.1080/10485250802172126

Journal of Nonparametric Statistics
Vol. 20, No. 5, July 2008, 431–446

A sufficient normality condition for Turing's formula

Zhiyi Zhang* and Hongwei Huang

Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, USA

(*Received 26 October 2007; final version received 25 April 2008*)

This paper establishes a previously unknown sufficient condition for the asymptotic normality of the non-parametric sample coverage estimate based on Good under a fixed underlying probability distribution $\{p_k; k = 1, \ldots\}$ where all $p_k > 0$. The sufficient condition of this paper supports a non-empty class of distributions and excludes the condition of Esty as a marginal case in which it is shown that the $\sqrt{n}$-normalised sample coverage estimate proposed by Esty necessarily degenerates under a fixed $\{p_k\}$. The convergent statistic in the newly established normality law and the resulting relevant confidence intervals are all of new forms, and specifically are different from those suggested by Esty.

Keywords: turing's formula; asymptotic normality

AMS 2000 Subject Classifications: Primary: 62f10, 62F12, 62G05, 62G20; Secondary: 62F15

1. Introduction

Consider a multinomial distribution with its countably infinite number of categories indexed by $K = \{k; k = 1, \ldots\}$ and its category probabilities denoted by $\{p_k\}$, satisfying $0 < p_k < 1$ for all k and $\sum p_k = 1$, where the sum without index is over all k as in all subsequent text unless otherwise stated. (In fact, in the subsequent text, we should observe the convention that $\prod = \prod_{k=1}^{\infty}$, $\lim = \lim_{n\to\infty}$ and that $\int = \int_{-\infty}^{+\infty}$, unless otherwise indicated. We also use '$\sim$' to indicate equality in the limit.) We denote the category counts in an *iid* sample of size n from the underlying population by $(X_1, \ldots)$ and its observed values by $(x_1, \ldots)$. For a given sample, there are at most n non-zero x_ks. Suppose the target of estimation is the 'total probability of the categories not represented in the sample', or equivalently

$$\pi_0 = \sum p_k 1[x_k = 0],$$

where $1[\cdot]$ is the indicator function. It may be interesting to note that π_0 is not a fixed constant nor is it an observable random variable. This target is interesting because it represents the probability that the $(n + 1)$th observation is from a previously unobserved category. The problem of estimating π_0 has many interesting applications. For example, Efron and Thisted [1] and Thisted and Efron [2] discuss two applications related to Shakespeare's general vocabulary and authorship

*Corresponding author. Email: zzhang@uncc.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250802172126
http://www.informaworld.com

to a poem; Good and Toulmin [3] and Chao [4], among many others, discuss the probability of discovering new species of animals in a population; more recently, Mao and Lindsay [5] study a genomic application in gene-categorisation; and Zhang [6] discuss several related issues in data confidentiality.

An estimator of π_0 described by Good [7], but largely credited to Turing and hence sometimes known as Turing's formula, is given by

$$T = \frac{N_1}{n}, \tag{1.1}$$

where N_1 is the number of categories represented exactly once in the sample, *i.e.*, $N_1 = \sum 1[X_k = 1]$. This simple formula has been used widely across many fields of study, frequently in the form of $C' = 1 - T$ estimating $C = 1 - \pi_0$. C is often referred to as the population coverage in a random sample, and C' is an estimate of C. The general problem is often known as the 'coverage' problem.

The list of authors having discussed issues related to Equation (1.1) in various settings is a long one. In addition to those mentioned above, Harris [8,9], Robbins [10], Starr [11], Holst [12], Chao [13], Esty [14–18], and Chao and Lee [19] are among those frequently referenced. However, of special relevance to the issue of this paper is Esty [15] in which the asymptotic distributional behaviour of the coverage estimate under an infinite dimensional $\{p_k\}$ is discussed. Esty [15] gives a sufficient condition for the asymptotic normality of a $\sqrt{n}$-normalised coverage estimate. Unfortunately, Esty's normality law is established not for a fixed $\{p_k\}$ but for a $\{p_k\}_n$ which is allowed to vary, as n increases, along an arbitrary path. If $\{p_k\}$ is held fixed, as seen in Section 2, the sufficient condition offered by Esty [15] never holds, and therefore the $\sqrt{n}$-normalised coverage estimate necessarily degenerates at zero.

We are motivated to find a sufficient condition for a non-degenerated asymptotic normality law for a fixed $\{p_k\}$. Toward that end, the asymptotic distributional behaviour of the coverage estimate is examined with a normalising factor, $g(n)$, that increases faster than $\sqrt{n}$. Consequently a sufficient condition of the asymptotic normality for a properly normalised coverage estimate is established. In addition, the following three facts are also established:

(1) The class of distributions satisfying the sufficient conditions, under fixed $\{p_k\}$, is non-empty.
(2) The convergent normalised coverage estimate does not depend on the choice of $g(n)$.
(3) The confidence intervals for C and π_0 respectively are of different forms from those suggested by Esty [15].

There are four sections in the remainder of the paper. A note on the result of Esty [15] is given in Section 2; most of the preliminary lemmas are summarised in Section 3; the main results are offered in Section 4; and the paper ends with several concluding remarks in Section 5.

2. A note on Esty's normality law

Esty [15] establishes a $\sqrt{n}$-normality law for $C' - C$ but allowing the underlying probability distribution $\{p_k\}$ to vary within a family $\{\{p_k\}_m; m = 1, \ldots\}$ as the sample size n changes to ensure that the imposed condition would hold:

$$\text{(a)} \quad E(N_1/n) \longrightarrow c_1, \quad 0 < c_1 < 1 \quad \text{and} \quad \text{(b)} \quad E(N_2/n) \longrightarrow c_2 \geq 0, \tag{2.1}$$

where $N_2 = \sum 1[X_k = 2]$.

Unfortunately, if the underlying distribution $\{p_k\}$ is held fixed, Equation (2.1) never holds. The following Lemma establishes that fact.

LEMMA 2.1 *Let* $\{p_k; k = 1, \ldots\}$ *be any multinomial probability distribution. Then*

$$(a) \lim_{n\to\infty} E(N_1/n) = 0 \quad \textit{and} \quad (b) \lim_{n\to\infty} E(N_2/n) = 0.$$

Proof (a) $E(N_1/n) = \sum p_k(1-p_k)^{n-1}$. Since $p_k(1-p_k)^{n-1} \le p_k$ and $\sum p_k = 1$, $\lim E(N_1/n) = \sum p_k \lim(1-p_k)^{n-1} = 0$.

(b) $E(N_2/n) = 1/2\sum(n-1)p_k^2(1-p_k)^{n-2}$. Since $(n-1)p(1-p)^{n-2}$ attains its maximum at $p = 1/(n-1)$, and therefore $(n-1)p_k(1-p_k)^{n-2} \le [(n-2)/(n-1)]^{n-2} < 1$. Hence $\lim E(N_2/n) = 1/2\sum \lim[(n-1)p_k^2(1-p_k)^{n-2}] = 0$. ■

On the other hand, the method of Esty [15] is instructive. The basic structure of Esty's method is based on a direct evaluation of the limit of the characteristic function of a normalised coverage estimate. The method is supported by two partitions of the index set $K = \{k; 1, \ldots\}$, denoted by $K = M \cup MC$ and $K = I \cup II$. The first partition is designed to support an exchange of a limit operator and an integral operator. The second partition is designed to control the tail probabilities of $\{p_k\}$ as n increases. The proofs of this paper are largely parallel to those of Esty [15]. However, we establish that Esty's first partition (M and MC) is not necessary and hence simplify the structure of the proofs somewhat. The second partition (I and II) adopted in this paper depends on $g(n)$ and, therefore, plays an essential role in the relevant proofs.

3. Preliminary results

Let $K_1 = \{1\}$ and $K_2 = \{2, \ldots\}$. For any $k \in K = K_1 \cup K_2$, let

$$f_k(x) = \begin{cases} p_k & x = 0, \\ -1/n & x = 1, \\ 0 & x \ge 2. \end{cases}$$

$Z = \sum f_k(X_k) = C' - C$. We are interested in the asymptotic behaviour of $Zg(n)$, where $g(n)$ is a function of n satisfying

$$g(n) = \mathrm{O}(n^{1-2\delta}) \tag{3.1}$$

for some $\delta \in (0, 1/4)$, in terms of the limit of its characteristic function, $E[\exp(\mathrm{i}sZg(n))]$. To begin, we note that $Z = Z_1 + Z_2$, where $Z_1 = \sum_{K_1} f_k(X_k)$ and $Z_2 = \sum_{K_2} f_k(X_k)$. Lemma 3.1 below is a well-known fact and Lemma 3.2 is due to Bartlett [20].

LEMMA 3.1 *Let* $\{X_k\}$ *be the counts of observations in category* k, $k = 1, 2, \ldots$, *in an iid sample under the multinomial model with probability distribution* $\{p_k\}$. *Then*

$$P(X_k = x_k; k = 1, \ldots) = P\left(Y_k = x_k; k = 1, \ldots \,\middle|\, \sum Y_k = n\right),$$

where $\{Y_k\}$ *are independent Poisson random variables with mean* np_k.

LEMMA 3.2 *Let* (U, V) *be a two-dimensional random vector with* U *integer valued. Then*

$$E(\exp(\mathrm{i}vV|U = n)) = (2\pi P(U = n))^{-1} \int_{-\pi}^{\pi} E[\exp(\mathrm{i}u(U-n) + \mathrm{i}vV)]\,\mathrm{d}u.$$

Thus, $E(\exp(\mathrm{i}sZg(n)))$ is

$$\left(2\pi P\left(\sum Y_k = n\right)\right)^{-1} \int_{-\pi}^{\pi} E\left[\exp\left(\mathrm{i}u\sum(Y_k - np_k) + \mathrm{i}sZg(n)\right)\right] \mathrm{d}u.$$

We want to evaluate $\lim E(\exp(\mathrm{i}sZg(n)))$. Toward that end, we first note that, by Stirling's formula, $(2\pi n)^{1/2} P(\sum Y_k = n) \to 1$. Therefore, we need only to evaluate the limit of

$$H_n(s) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\pi}^{\pi} E\left[\exp\left(\mathrm{i}u\sum(Y_k - np_k) + \mathrm{i}sZg(n)\right)\right] \mathrm{d}u,$$

or letting $t = un^{1/2}$,

$$H_n(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} 1[|t| < \pi\sqrt{n}] E\left[\exp\left(\mathrm{i}(n)^{-1/2}t\sum(Y_k - np_k) + \mathrm{i}sZg(n)\right)\right] \mathrm{d}t. \quad (3.2)$$

Our first task is to allow the limit operator to exchange with the integral operator. The key element to support this exchange is Equation (3.3). The proof of Equation (3.3) is given in Appendix 1.

Now that we have established

$$\lim \int |\overline{h}_{n1}| \, \mathrm{d}t = \int \lim |\overline{h}_{n1}| \, \mathrm{d}t, \quad (3.3)$$

by the Dominated Convergence Theorem, we have the following lemma.

LEMMA 3.3 *Let h_n and H_n be as defined in Equations* (3.2) *and* (A1)*, respectively. Then*

$$\lim H_n = \frac{1}{\sqrt{2\pi}} \int \lim h_n \, \mathrm{d}t.$$

For each k, it can be verified that, letting

$$B_k = \exp(-\mathrm{i}tp_k n^{1/2})[\exp(np_k(\exp(\mathrm{i}tn^{-1/2}) - 1))]$$
$$C_k = \exp(-\mathrm{i}tp_k n^{1/2})[\exp(\mathrm{i}sp_k g(n)) - 1]\exp(-np_k)$$
$$D_k = \exp(-\mathrm{i}tp_k n^{1/2})\exp(\mathrm{i}tn^{-1/2})[\exp(-\mathrm{i}sn^{-1}g(n)) - 1]np_k\exp(-np_k),$$

and $E_k = C_k + D_k$, $h_n \sim \prod(B_k + E_k)$. We are interested in evaluating $\lim \prod(B_k + E_k)$.

The facts of the following two lemmas are given by Esty [15].

LEMMA 3.4 *Let $\{\beta_k\}$ and $\{\epsilon_k\}$ be two sequences of complex numbers, and M_n be a sequence of subsets of K, indexed by n. If*

(1) $\prod_{M_n} \beta_k \sim \beta$,
(2) $(\sum_{M_n} \epsilon_k) \sim \epsilon$,
(3) $\beta_k \sim 1$ *uniformly*,
(4) $\epsilon_k \sim 0$ *uniformly*,
(5) *there exists a constant*, δ_1 *such that*, $\sum_{M_n} |\beta_k - 1| \le \delta_1$, *and*
(6) *there exists a constant*, δ_2 *such that*, $\sum_{M_n} |\epsilon_k| \le \delta_2$;

then

$$\prod_{M_n} (\beta_k + \epsilon_k) \sim \beta e^{\epsilon}$$

where β and ϵ may also depend on n.

LEMMA 3.5 *For all* $k \in K$,

$$B_k = \exp[(-t^2/2)p_k + \mathrm{O}(t^3 p_k n^{-1/2})].$$

The next lemma includes three useful facts.

LEMMA 3.6 (1) *For any complex number* x *satisfying* $|x| < 1$, $|\ln(1+x)| \le |x|/(1-|x|)$.
(2) *For any real number* $x \in [0, 1)$, $1 - x \ge \exp(-x/(1-x))$.
(3) *For any real number* $x \in (0, 1/2)$, $1/(1-x) < 1 + 2x$.

Proof (1) By Taylor's formula, $|\ln(1+x)| = |\sum_{j=1}^{\infty}(-1)^{j+1}x^j/j| \le \sum_{j=1}^{\infty}|x|^j = |x|/(1-|x|)$.

(2) The function $y = 1/(1+t)\mathrm{e}^t$ is strictly increasing over $[0, \infty)$ and has value 1 at $t = 0$. Therefore, $1/(1+t)\mathrm{e}^t \ge 1$ for $t \in [0, \infty)$. The desired inequality follows the change of variable $x = t/(1+t)$.

(3) The proof is trivial. ■

Let us consider a partition of the index set $K = I \cup II$, where

$$I = \{k;\ p_k g(n) \le n^{-\delta}\} \quad \text{and} \quad II = \{k;\ p_k g(n) > n^{-\delta}\},$$

where δ is as in Equation (3.1).

LEMMA 3.7 (a) $\sum_{II}|E_k| \longrightarrow 0$; *and* (b) $\prod_{II}(B_k + E_k)/\prod_{II} B_k \longrightarrow 1$.

Proof (a) $\sum_{II}|E_k| \le 2\sum_{II}(\mathrm{e}^{-np_k} + np_k\mathrm{e}^{-np_k})$. Since the derivative of $(\mathrm{e}^{-np_k} + np_k\mathrm{e}^{-np_k})$ for any $k \in II$, with respect to p_k is negative in $(0, 1)$, $(\mathrm{e}^{-np_k} + np_k\mathrm{e}^{-np_k})$ attains its maximum at $p_k = 1/(g(n)n^{\delta})$, with value $\mathrm{e}^{-n^{1-\delta}/g(n)}(1 + n^{1-\delta}/g(n))$. The total number of indices in II is less than or equal to $g(n)n^{\delta}$. Therefore,

$$\sum_{II}|E_k| \le 2[g(n)n^{\delta}][\mathrm{e}^{-n^{1-\delta}/g(n)}(1 + n^{1-\delta}/g(n))] = 2\mathrm{e}^{-\mathrm{O}(n^{\delta})}\mathrm{O}(n) \longrightarrow 0. \tag{3.4}$$

(b) By Lemma 3.5, $|B_k|$ is bounded away from zero, and by the fact that $\lim|E_k| = 0$ (and hence $\lim|E_k|/|B_k| = 0$) and by applying the first part of Lemma 3.6 with $x = E_k/B_k$, we have

$$\left|\ln\left[\prod_{II}(B_k + E_k)/\prod_{II}B_k\right]\right| = \left|\sum_{II}\ln\left(1 + \frac{E_k}{B_k}\right)\right| \le \sum_{II}\left|\ln\left(1 + \frac{E_k}{B_k}\right)\right|$$
$$\le \sum_{II}\left(\frac{|E_k|}{|B_k| - |E_k|}\right) = \mathrm{O}\Big(\sum_{II}|E_k|\Big) \longrightarrow 0. \quad ■$$

Now let us state the condition under which many of the subsequent results are established.

CONDITION 3.1 *As* $n \longrightarrow \infty$,

(1) $\sum(g^2(n)/n)p_k\mathrm{e}^{-np_k} \longrightarrow c_1 \ge 0$,
(2) $\sum g^2(n)p_k^2\mathrm{e}^{-np_k} \longrightarrow c_2 \ge 0$, *and*
(3) $c_1 + c_2 > 0$.

LEMMA 3.8 *Under Condition* 3.1, *all the conditions of Lemma* 3.4 *are satisfied with* $M_n = I$, $\beta_k = B_k$, $\beta = B$, $\epsilon_k = E_k$, *and* $\epsilon = E$.

The proof of Lemma 3.8 is given in Appendix 1.

Remark 3.1 It may be interesting to note that, the third term within the curly brackets in Equation (A2), $st\frac{g(n)}{\sqrt{n}}p_k$, also satisfies $\sum_I \mathrm{e}^{-np_k} st\frac{g(n)}{\sqrt{n}}p_k \to 0$. However if $g(n) = \sqrt{n}$ as in Esty [15], this term does not vanish, and, as a result, shows up as an extra term in the asymptotic variance of the normalised coverage estimator in Esty's results.

Lemma 3.4 and Lemma 3.8 give immediately the following corollary.

COROLLARY 3.1 *Under Condition* 3.1, $\prod_I (B_k + E_k) \sim \prod_I B_k \exp(\sum_I E_k)$.

LEMMA 3.9 *Under Condition* 3.1, $\prod(B_k + E_k) \to B\mathrm{e}^E$, *where* $B = \lim \prod B_k$ *and* $E = \lim \sum E_k$.

Proof

$$\begin{aligned}\prod(B_k + E_k) &= \prod_I (B_k + E_k) \prod_{II}(B_k + E_k) \sim \prod_I (B_k + E_k) \prod_{II} B_k \qquad \text{(by Lemma 3.7)}\\ &\sim \prod_I B_k \left(\exp \sum_I E_k\right) \prod_{II} B_k \qquad \text{(by Lemma 3.8)}\\ &\sim \prod B_k \left(\exp \sum E_k\right) \qquad \text{(by Lemma 3.7).}\end{aligned}$$

■

Remark 3.2 At this point, one may see the reason why it is imposed that $g(n) = \mathrm{O}(n^{1-2\delta})$ for some small positive δ. If $g(n)$ is let to be a sequence increasing to infinity in the order of n or faster, $\sum_{II} E_k \to 0$ cannot be established using the current method. The proof for (a) of Lemma 3.7 will break down. Consequently, the partition of $K = I \cup II$ will not effectively support the subsequent proofs.

4. Main results

THEOREM 4.1 *Let* $g(n)$ *be as in Equation* (3.1). *Under Condition* 3.1,

$$g(n)(C' - C) \xrightarrow{D} N(0, c_1 + c_2).$$

Proof Since $\lim \prod B_k = \mathrm{e}^{-\frac{t^2}{2}}$ and

$$\lim \sum E_k = -\frac{s^2}{2}\left(\lim \sum \frac{g^2(n)}{n} p_k \mathrm{e}^{-np_k} + \lim \sum g^2(n) p_k \mathrm{e}^{-np_k}\right),$$

$$\lim H_n = \left(\frac{1}{\sqrt{2\pi}} \int \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t\right) \mathrm{e}^{-\frac{s^2}{2}\left(\lim \sum \frac{g^2(n)}{n} p_k \mathrm{e}^{-np_k} + \lim \sum g^2(n) p_k \mathrm{e}^{-np_k}\right)} = \mathrm{e}^{-\frac{s^2}{2}(c_1 + c_2)}$$

which is the characteristic function of a normal distribution with mean zero and variance $c_1 + c_2$. ■

Given a $g(n)$ satisfying Equation (3.1), Condition 3.1 imposes a rate of convergence for $\{p_k\}$. To see that and that the condition of Theorem 4.1 describes a non-empty class of distribution, we consider the following example.

Example 4.1 Let $p_k = 2/(k+1)^2, k = 1, \ldots$. Then $g(n)$ must be of order $O(n^{3/4})$ for Condition 3.1 to hold.

To see this, we have

$$\begin{aligned}\frac{g^2(n)}{n}\int_1^\infty \frac{2}{(x+1)^2}e^{-2n/(x+1)^2}dx &= 2\frac{g^2(n)}{n}\int_0^{1/2} e^{-2nt^2}\,dt \\ &= 2\frac{g^2(n)}{n}\sqrt{2\pi}\left(\frac{1}{2\sqrt{n}}\right)\left[\frac{1}{\sqrt{2\pi}}\int_0^{\sqrt{n}} e^{-t^2/2}\,dt\right] \\ &= O\left(\frac{g^2(n)}{n\sqrt{n}}\right).\end{aligned}$$

The last expression goes to a non-zero constant if and only if $g(n) = O(n^{3/4})$.

Similarly,

$$\begin{aligned}g^2(n)\int_1^\infty \frac{4}{(x+1)^4}e^{-2n/(x+1)^2}\,dx &= 4g^2(n)\int_0^{1/2} t^2 e^{-2nt^2}\,dt \\ &= g^2(n)\frac{4}{(2\sqrt{n})^3}\int_0^{\sqrt{n}} t^2 e^{-t^2/2}\,dt \\ &= O\left(\frac{g^2(n)}{n\sqrt{n}}\right).\end{aligned}$$

The last expression goes to a non-zero constant if and only if $g(n) = O(n^{3/4})$.

Let us consider the following condition:

CONDITION 4.1 *As* $n \longrightarrow \infty$,

(1) $\frac{g^2(n)}{n^2}E(N_1) \longrightarrow c_1 \geq 0$,

(2) $\frac{g^2(n)}{n^2}E(N_2) \longrightarrow \frac{c_2}{2} \geq 0$, *and*

(3) $c_1 + c_2 > 0$.

LEMMA 4.1 *Condition* 3.1 *and Condition* 4.1 *are equivalent.*

The proof of Lemma 4.1 is given in Appendix 1.

Lemma 4.1 allows us to re-state Theorem 4.1:

THEOREM 4.2 *If there exists a* $g(n)$ *satisfying Equation* (3.1) *and Condition* 4.1, *then*

$$\frac{n(C'-C)}{\sqrt{E(N_1)+2E(N_2)}} \xrightarrow{D} N(0,1). \tag{4.1}$$

Remark 4.1 The statement in Theorem 4.2 can be re-written as

$$\frac{\sqrt{n}(C'-C)}{\sqrt{E(N_1)/n+2E(N_2)/n}} \xrightarrow{D} N(0,1)$$

which resembles very much Theorem 4 of Esty [15] except the third term in the variance of Esty [15] is missing. However, it is to be noted that in the current context, the coverage statistic, even

though its normalised form can be expressed as above, is not normalised by $\sqrt{n}$ but by $g(n)$ satisfying $g(n)/\sqrt{n} \to \infty$.

As a consequence of Theorem 4.1, we have the following theorem.

THEOREM 4.3 *If there exists a $g(n)$ satisfying Equation* (3.1) *and Condition* 4.1, *then*

$$\frac{n(C' - C)}{\sqrt{N_1 + 2N_2}} \xrightarrow{D} N(0, 1).$$

The proof of Theorem 4.3 is given in Appendix 1.

We note that the condition of Theorems 4.2 and 4.3 requires no further knowledge of $g(n)$ other than its existence.

Theorem 4.3 leads to an approximate $(1 - \alpha)$-level confidence interval for C:

$$(1 - N_1/n) \pm z_{\alpha/2}\sqrt{N_1/n^2 + 2N_2/n^2}; \tag{4.2}$$

and, for completeness, an approximate $(1 - \alpha)$-level confidence interval for π_0:

$$\frac{N_1}{n} \pm z_{\alpha/2}\sqrt{N_1/n^2 + 2N_2/n^2}. \tag{4.3}$$

5. Concluding remarks

The sufficient condition of this paper and Example 1 together ensure the existence of a non-degenerated asymptotic normality law for a non-empty class of distributions. Prior to this paper, no such results were known. However, two papers in the existing literature dealing with asymptotic distribution of the coverage estimate are worth noting. The first is Esty [15] which is highly relevant to this paper but does not establish a non-degenerated normality law for a fixed $\{p_k\}$ as already explained above. The second is Mao and Lindsay [5] which also establishes a normality law. However, they assume a multinomial distribution with finite number of categories, N, and $(X_1, \ldots, X_N)$ to be an independent sample from N Poisson distributions with intensity parameters λ_k, $k = 1, \ldots, N$, which are themselves an *iid* random sample from an unspecified distribution. The normality law established by Mao and Lindsay [5] is a limiting characteristic as the number of categories increases to infinity. This is an interesting idea, but not exactly non-parametric and therefore essentially a different problem.

For theoretical as well as practical reasons, it is perhaps also of interest to understand what type of $\{p_k\}$ is supported by the sufficient condition established in this paper. It can be verified that, if $p_k = \mathrm{O}(k^{-\alpha})$ for some $\alpha \in (1, \infty)$, a $g(n)$ satisfying Equation (3.1) and Condition 4.1 exists. This implies that Theorems 4.1, 4.2, and 4.3, and the confidence intervals in Equations (4.2) and (4.3) are all valid for $\{p_k = \mathrm{O}(k^{-\alpha})\}$, where $\alpha \in (1, \infty)$. (In fact, even for $p_k = 1/k$, obviously not a probability sequence, $g(n) = \sqrt{n}$ satisfies Condition 4.1.) It is however not clear whether the sufficient condition supports the super slowly convergent probability sequences such as $p_k = \mathrm{O}(1/[k(\ln(k))^2])$.

On the other hand, it is clear that the sufficient condition does not support very rapidly convergent probability sequences such as $p_k = \alpha^k$ for $\alpha \in (0, 1)$. It can also be shown that, if $p_k = \mathrm{O}(\alpha^k)$ for some $\alpha \in (0, 1)$, $g(n)$ must be of the order of n which does not satisfy Equation (3.1). Obviously for the extreme case of a distribution with finite categories, Condition 3.1 fails to hold for any $g(n) = n^\alpha$, where $\alpha \in (0, \infty)$.

One of the crucial elements of this paper is the partition of $K = I \cup II$. When the proper partition is found, the total tail probabilities, $\sum_I p_k$, shrinks as $n \to \infty$ at an appropriate rate so that Equation (3.4) holds. It is not clear, if $g(n)$ is taken to be of the order of n or higher, how the tail probabilities would be controlled.

Although the sufficient condition of Esty and that of this paper describe different populations, an intuitive comparison is still possible. Esty's condition is essentially a thicker tail condition. It says that as n increases, the total probability of unrepresented categories does not converge to zero but inflates at a rate such that the total probability remains constant. Condition 4.1 on the other hand allows the total probability to converge to zero. It is therefore conceivable that admittedly in a somewhat speculative sense, the respective biases converge to zero at different rates, slower under Esty's condition, and faster under Condition 4.1. The difference is reflected by the fact that $g(n)$ increases at a higher rate than $\sqrt{n}$.

In terms of verifying the conditions of the theorem for real data sets in application, one may consider the following logic. If a value of $\delta \in (0, 1/4)$ exists (denoted by δ_0) such that $g(n) = n^{1/2+\delta_0}$ in Equation (3.1) satisfies Condition 4.1, then

$$\frac{g^2(n)}{n^2} E(N_1) + \frac{g^2(n)}{n^2} E(N_2) \longrightarrow c > 0 \tag{5.1}$$

at $\delta = \delta_0$ for some c. For any $\delta < \delta_0$, the left-hand side of Equation (5.1) converges to zero. For any $\delta > \delta_0$, the left-hand side of Equation (5.1) diverges to infinity. The sample version of the left-hand side of Equation (5.1)

$$y(\delta, n, N_1, N_2) = \frac{g^2(n)}{n^2} N_1 + \frac{g^2(n)}{n^2} N_2$$

should behave similarly. Suppose sub-samples of sizes $n_1 = [|n \times 50\%|]$, $n_2 = [|n \times 60\%|]$, $n_3 = [|n \times 70\%|]$, $n_4 = [|n \times 80\%|]$, and $n_5 = [|n \times 90\%|]$, where $[|\cdot|]$ is the greatest integer function, are randomly formed from the original sample of size $n_6 = n$. Let $N_1^{(i)}$ and $N_2^{(i)}$, where $i = 1, \ldots, 6$, be the numbers of categories represented once and twice, respectively, in the ith sub-sample. Let us consider the sequence of points

$$\left\{\left(n_i, y(\delta, n_i, N_1^{(i)}, N_2^{(i)})\right); i = 1, \ldots, 6\right\}$$

in $n \times y$ plane. For each $\delta \in (0, 1/4)$ such a sequence can be calculated and plotted. If a set of δ values ranging from 0 to 1/4 are used, a set of such sequences are obtained. If a $\delta = \delta_0$ satisfying Condition 4.1 exists, then some of these sequences with higher δ values will trend upward and those with lower δ values will trend downward. Observing both trends among the set of sequences empirically verifies the condition for the theorem.

Finally, we end the paper with two conjectures:

(1) For any probability sequence $\{p_k\}$ with convergence rate slower than any in the family $\{p_k = O(k^{-\alpha})\}$, where $\alpha \in (1, \infty)$, $\sqrt{n}(C' - C)$ degenerates asymptotically at zero.
(2) For any probability sequence $\{p_k\}$ with the convergence rate of $O(\alpha^k)$ where $\alpha \in (0, 1)$ or faster, $n(C' - C)$ either degenerates at zero or converges in distribution to a non-Gaussian law.

References

[1] B. Efron and R. Thisted, *Estimating the number of unseen species: how many words did Shakespeare know?*, Biometrika 63 (1976), pp. 435–447.

[2] R. Thisted and B. Efron, *Did Shakespeare write a newly-discovered poem?*, Biometrika 74 (1987), pp. 445–455.
[3] I.J. Good and G.H. Toulmin, *The number of new species, and the increase in population coverage, when a sample is increased*, Biometrika 43 (1956), pp. 45–63.
[4] A. Chao, *On estimating the probability of discovering a new species*, Ann. Stat. 9 (1981), pp. 1339–1342.
[5] C.X. Mao and B.G. Lindsay, *A Poisson model for the coverage problem with a genomic application*, Biometrika 89 (2002), pp. 669–681.
[6] C.-H. Zhang, *Estimation of sums of random variables: examples and information bounds*, Ann. Stat. 33 (2005), pp. 2022–2041.
[7] I.J. Good, *The population frequencies of species and the estimation of population parameters*, Biometrika 40 (1953), pp. 237–264.
[8] B. Harris, *Determining bounds on integrals with applications to cataloging problems*, Ann. Math. Stat. 30 (1959), pp. 521–548.
[9] ———, *Statistical inference in the classical occupancy problem unbiased estimation of number of classes*, J. Am. Stat. Assoc. 63 (1968), pp. 837–847.
[10] H.E. Robbins, *Estimating the total probability of the unobserved outcomes of an experiment*, Ann. Math. Stat. 39 (1968), pp. 256–257.
[11] N. Starr, *Linear estimation of probability of discovering a new species*, Ann. Stat. 7 (1979), pp. 644–652.
[12] L. Holst, *Some assymptotic results for incomplete multinomial or Poisson samples*, Scand. J. Stat. 8 (1981), pp. 243–246.
[13] A. Chao, *Nonparametric estimation of the number of the classes in a population*, Scand. J. Stat. 11 (1984), pp. 265–270.
[14] W.W. Esty, *Confidence intervals for the coverage of low coverage samples*, Ann. Stat. 10 (1982), pp. 190–196.
[15] ———, *A normal limit law for a nonparametric estimator of the coverage of a random sample*, Ann. Stat. 11 (1983), pp. 905–912.
[16] ———, *Estimation of the number of classes in a population and the coverage of a sample*, Math. Sci. 10 (1985), pp. 41–50.
[17] ———, *The size of a coinage*, Numis. Chron. 146 (1986a), pp. 185–215.
[18] ———, *The efficiency of Good's nonparametric coverage estimator*, Ann. Stat. 14 (1986b), pp. 1257–1260.
[19] A. Chao and S. Lee, *Estimating the number of classes via sample covergae*, J. Am. Stat. Assoc. 87 (1992), pp. 210–217.
[20] M.S. Bartlett, *The characteristic function of a conditional statistic*, J. Lond. Math. Soc. 13 (1938), pp. 62–67.

Appendix A

*Proof of Equation (*3.3*)* Let

$$\begin{aligned} h_n &= 1[|t| < \pi\sqrt{n}]E[\exp(\mathrm{i}(n)^{-1/2}t\sum(Y_k - np_k) + \mathrm{i}sZg(n)] \\ h_{n1} &= 1[|t| < \pi\sqrt{n}]E[\exp(\mathrm{i}(n)^{-1/2}t(Y_1 - np_1) + \mathrm{i}sZ_1g(n)] \\ h_{n2} &= 1[|t| < \pi\sqrt{n}]E[\exp(\mathrm{i}(n)^{-1/2}t\sum_{K_2}(Y_k - np_k) + \mathrm{i}sZ_2g(n)], \end{aligned} \tag{A1}$$

$$H_n(s) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} h_n\,\mathrm{d}t = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} h_{n1}h_{n2}\,\mathrm{d}t.$$

Since $|h_{n2}| \le 1$, $|h_n| \le |h_{n1}|$. On the other hand,

$$\begin{aligned} &E[\exp(\mathrm{i}u(Y_1 - np_1) + \mathrm{i}sf_1(Y_1)g(n))] \\ &\quad= \exp(\mathrm{i}u(-np_1) + \mathrm{i}sp_1g(n))\exp(-np_1) + \exp(\mathrm{i}u(1 - np_1) - \mathrm{i}sn^{-1}g(n))np_1\exp(-np_1) \\ &\qquad+ \sum_{j=2}^{\infty}\exp(\mathrm{i}u(j - np_1))P(Y_1 = j) \\ &\quad= \sum_{j=0}^{\infty}\exp(\mathrm{i}u(j - np_1))P(Y_1 = j) \end{aligned}$$

$$
\begin{aligned}
&- \exp(-\mathrm{i}unp_1)\exp(-np_1) - \exp(\mathrm{i}u(1-np_1))np_1\exp(-np_1)\\
&+ \exp(\mathrm{i}u(-np_1) + \mathrm{i}sp_1g(n))\exp(-np_1) + \exp(\mathrm{i}u(1-np_1) - n^{-1}\mathrm{i}sg(n))np_1\exp(-np_1)\\
&= \left[\exp(-\mathrm{i}unp_1)\exp(\mathrm{i}\sin(u)np_1)\exp(np_1(\cos(u)-1))\right]\\
&- \exp(-\mathrm{i}unp_1)\exp(-np_1) - \exp(\mathrm{i}u(1-np_1))np_1\exp(-np_1)\\
&+ \exp(\mathrm{i}u(-np_1) + \mathrm{i}sp_1g(n))\exp(-np_1) + \exp(\mathrm{i}u(1-np_1) - n^{-1}\mathrm{i}sg(n))np_1\exp(-np_1).
\end{aligned}
$$

Therefore (recall $t = u\sqrt{n}$),

$$
|h_{n1}| \le 1[|t| < \pi\sqrt{n}][\exp(np_1(\cos(tn^{-1/2})-1)) + 2[\exp(-np_1) + np_1\exp(-np_1)]] \quad (= \overline{h}_{n1}).
$$

It is clear that, for any t, by Taylor's formula for $\cos(x)$,

$$
\lim \overline{h}_{n1} = \lim 1[|t| < \pi\sqrt{n}]\exp(np_1(\cos(tn^{-1/2})-1)) = \exp(-p_1t^2/2) \qquad (= \overline{h}_1).
$$

$$
\begin{aligned}
\int |\overline{h}_{n1}|\,\mathrm{d}t &= \int 1[|t| < \pi\sqrt{n}]\left[\exp(np_1(\cos(tn^{-1/2})-1))\right]\mathrm{d}t\\
&\quad + 2\int 1[|t| < \pi\sqrt{n}]\exp(-np_1)\,\mathrm{d}t + 2\int 1[|t| < \pi\sqrt{n}]np_1\exp(-np_1)\,\mathrm{d}t\\
&= \int 1[|t| < \pi\sqrt{n}]\left[\exp(np_1(\cos(tn^{-1/2})-1))\right]\mathrm{d}t\\
&\quad + 2\times 2\pi\sqrt{n}\exp(-np_1) + 2\times 2\pi\sqrt{n}np_1\exp(-np_1)
\end{aligned}
$$

Since the last two terms above vanish to zero as $n \to \infty$, we have, letting θ be a constant in $(0, 1/2)$,

$$
\begin{aligned}
\lim \int |\overline{h}_{n1}|\,\mathrm{d}t &= \lim \int 1[|t| < \pi\sqrt{n}]\left[\exp(np_1(\cos(tn^{-1/2})-1))\right]\mathrm{d}t\\
&= \lim \int_{-\pi}^{+\pi} \sqrt{n}\left[\exp(np_1(\cos(u)-1))\right]\mathrm{d}u\\
&= \lim \int_{|u|<1/(n^{(1-\theta)/2})} \sqrt{n}\left[\exp(np_1(\cos(u)-1))\right]\mathrm{d}u\\
&\quad + \lim \int_{(1/n^{(1-\theta)/2})\le|u|<\pi} \sqrt{n}\left[\exp(np_1(\cos(u)-1))\right]\mathrm{d}u\\
&(= \lim \eta_1 + \lim \eta_2).
\end{aligned}
$$

The second term of the last expression above is zero. To see this, we note that for any u satisfying $1/n^{(1-\theta)/2} \le |u| < \pi$, $\cos(u) - 1 \le \cos(1/n^{(1-\theta)/2}) - 1$, and hence

$$
\begin{aligned}
\lim \eta_2 &\le \lim \int_{(1/n^{(1-\theta)/2})\le|u|<\pi} \sqrt{n}\left[\exp(np_1(\cos(1/n^{(1-\theta)/2})-1))\right]\mathrm{d}u\\
&\le \lim 2\pi\sqrt{n}\left[\exp(np_1(\cos(1/n^{(1-\theta)/2})-1))\right]\\
&= \lim 2\pi\sqrt{n}\left[\exp\left(-np_1\left(1-\cos(1/n^{(1-\theta)/2})\right)\right)\right]\\
&= \lim 2\pi\sqrt{n}\left[\exp\left(-np_1\left(\frac{\sin^2(1/n^{(1-\theta)/2})}{1+\cos(1/n^{(1-\theta)/2})}\right)\right)\right]
\end{aligned}
$$

$$= \lim 2\pi\sqrt{n}\exp\left(-np_1\mathrm{O}\left(\frac{1}{n^{1-\theta}}\right)\right)$$
$$= \lim 2\pi\sqrt{n}\exp\left(-p_1\mathrm{O}(n^{\theta})\right) = 0.$$

For u satisfying $|u| < 1/n^{(1-\theta)/2}$, consider the Taylor expansion of

$$\begin{aligned}\cos(u) - 1 &= -\frac{u^2}{2!} + \frac{u^4}{4!} - \frac{u^6}{6!} + \cdots + \frac{(-1)^m u^{2m}}{(2m)!} + \cdots \\ &\le -\frac{u^2}{2} + (u^4 + u^8 + \cdots + u^{4m} + \cdots) \\ &= -\frac{u^2}{2} + \frac{u^4}{1-u^4}.\end{aligned}$$

Therefore,

$$\begin{aligned}\lim \eta_1 &\le \lim \int_{|u|<1/n^{(1-\theta)/2}} \sqrt{n}\exp\left(np_1\left(-\frac{u^2}{2} + \frac{1/n^{2-2\theta}}{1-1/n^{2-2\theta}}\right)\right)\mathrm{d}u \\ &= \lim \int_{|u|<1/n^{(1-\theta)/2}} \sqrt{n}\exp\left(-\frac{np_1u^2}{2} + np_1\frac{1/n^{2-2\theta}}{1-1/n^{2-2\theta}}\right)\mathrm{d}u \\ &= \lim \left[\left(\int_{|u|<1/n^{(1-\theta)/2}} \sqrt{n}\exp\left(-\frac{np_1u^2}{2}\right)\mathrm{d}u\right)\exp\left(\mathrm{O}(\frac{1}{n^{1-2\theta}})\right)\right] \\ &\quad \text{(letting } t = u\sqrt{n}\text{)} \\ &= \lim \left[\left(\int_{|t|<n^{\theta/2}} \exp\left(-\frac{p_1t^2}{2}\right)\mathrm{d}t\right)\exp\left(\mathrm{O}(\frac{1}{n^{1-2\theta}})\right)\right] \\ &= \int \exp\left(-p_1t^2/2\right)\mathrm{d}t.\end{aligned}$$

Since $\cos(u) \ge -u^2/2$ for all u satisfying $|u| < 1/n^{(1-\theta)/2}$, it is easy to establish $\lim \eta_1 \ge \int \exp(-p_1t^2/2)\,\mathrm{d}t$, and hence $\lim \eta_1 = \int \exp(-p_1t^2/2)\,\mathrm{d}t$. ■

Proof of Lemma 3.8 We need to check all six conditions in Lemma 3.4.

(3) is true because

$$B_k = \exp(-(t^2/2)\,p_k)\exp(\mathrm{O}((t^3/\sqrt{n})\,p_k))),$$

and p_k and $p_k/\sqrt{n}$ are uniformly bounded by $1/g(n)n^{\delta}$ and $1/g(n)\sqrt{n}n^{\delta}$, respectively.

For (1), since $\sum_I p_k \to 0$,

$$\prod_I B_k = \exp\left(-(t^2/2)\sum_I p_k\right)\exp\left(\mathrm{O}\left((t^3/\sqrt{n})\sum_I p_k\right)\right)) \longrightarrow 1.$$

For (2), (4), and (6),

$$
\begin{aligned}
E_k &= \mathrm{e}^{-np_k}\exp(-\mathrm{i}tp_k\sqrt{n})\left\{\mathrm{i}sg(n)p_k - \frac{s^2g^2(n)p_k^2}{2} + \mathrm{O}(s^3g^3(n)p_k^3)\right.\\
&\quad \left. + np_k\left[1 + \frac{\mathrm{i}t}{\sqrt{n}} - \frac{t^2}{2n} + \mathrm{O}\left(\frac{t^3}{n^{3/2}}\right)\right]\left[-\frac{isg(n)}{n} - \frac{s^2g^2(n)}{2n^2} + \mathrm{O}\left(\frac{s^3g^3(n)}{n^3}\right)\right]\right\}\\
&= \mathrm{e}^{-np_k}\exp(-\mathrm{i}tp_k\sqrt{n})\left\{\mathrm{i}sg(n)p_k - \frac{s^2}{2}g^2(n)p_k^2 + \mathrm{O}(s^3g^3(n)p_k^3)\right.\\
&\quad \left. + \left[np_k + \mathrm{i}tp_k\sqrt{n} - \frac{t^2}{2}p_k + np_k\mathrm{O}\left(\frac{t^3}{n^{3/2}}\right)\right]\left[-\frac{\mathrm{i}sg(n)}{n} - \frac{s^2g^2(n)}{2n^2} + \mathrm{O}\left(\frac{s^3g^3(n)}{n^3}\right)\right]\right\}\\
&= \mathrm{e}^{-np_k}\exp(-\mathrm{i}tp_k\sqrt{n})\left\{\mathrm{i}sg(n)p_k - \frac{s^2}{2}g^2(n)p_k^2 + \mathrm{O}(s^3g^3(n)p_k^3)\right.\\
&\quad - \mathrm{i}sg(n)p_k - \frac{s^2}{2}\left(\frac{g^2(n)p_k}{n}\right) + np_k\mathrm{O}\left(s^3\frac{g^3(n)}{n^3}\right)\\
&\quad + st\frac{g(n)}{\sqrt{n}}p_k - \frac{\mathrm{i}s^2t}{2n^{3/2}}g^2(n)p_k + \mathrm{i}tp_k\sqrt{n}\mathrm{O}\left(s^3\frac{g^3(n)}{n^3}\right)\\
&\quad + \frac{\mathrm{i}st^2}{2}\frac{g(n)}{n}p_k + \frac{s^2t^2}{4}\frac{g^2(n)}{n^2}p_k - \frac{t^2}{2}p_k\mathrm{O}\left(s^3\frac{g^3(n)}{n^3}\right)\\
&\quad \left. -\mathrm{i}sg(n)p_k\mathrm{O}\left(\frac{t^3}{n^{3/2}}\right) - \frac{s^2}{2}\frac{g^2(n)}{n}p_k\mathrm{O}\left(\frac{t^3}{n^{3/2}}\right) + np_k\mathrm{O}\left(\frac{t^3}{n^{3/2}}\right)\mathrm{O}(s^3\frac{g^3(n)}{n^3})\right\}\\
&= \mathrm{e}^{-np_k}\exp(-\mathrm{i}tp_k\sqrt{n})\left\{-\frac{s^2}{2}g^2(n)p_k^2 - \frac{s^2}{2}\left(\frac{g^2(n)}{n}\right)p_k\right.\\
&\quad + st\frac{g(n)}{\sqrt{n}}p_k + \frac{s^2t^2}{4}\frac{g^2(n)}{n^2}p_k - \frac{\mathrm{i}s^2t}{2n^{3/2}}g^2(n)p_k + \frac{\mathrm{i}st^2}{2}\frac{g(n)}{n}p_k\\
&\quad + \mathrm{O}(s^3g^3(n)p_k^3) + \mathrm{O}\left(s^3\frac{g^3(n)}{n^2}p_k\right) + \mathrm{iO}\left(ts^3\frac{g^3(n)}{n^{5/2}}p_k\right) - \mathrm{O}\left(\frac{s^3t^2}{2}\frac{g^3(n)}{n^3}p_k\right)\\
&\quad \left. -\mathrm{iO}\left(st^3\frac{g(n)}{n^{3/2}}p_k\right) - \mathrm{O}\left(\frac{s^2t^3}{2}\frac{g^2(n)}{n^{5/2}}p_k\right) + \mathrm{O}\left(\frac{s^3t^3}{2}\frac{g^3(n)}{n^{7/2}}p_k\right)\right\}.
\end{aligned}
\tag{A2}
$$

Now we observe the following:

(1) For all $k \in I$, $\exp(-\mathrm{i}tp_k\sqrt{n}) \to 1$ uniformly since $p_k\sqrt{n} \le \sqrt{n}/g(n)n^{\delta} \to 0$.
(2) It is easily checked that every additive term of E_k converges to zero uniformly for all $k \in I$. Therefore, (4) is checked.
(3) It is easily checked that, for every term within the curly brackets in Equation (A2), denoted by $\tau(s, t, n, p_k)$, except the first two terms,

$$\sum_I \mathrm{e}^{-np_k}|\tau(s, t, n, p_k)| \le \sum \mathrm{e}^{-np_k}|\tau(s, t, n, p_k)| \longrightarrow 0$$

uniformly by Condition 3.1.

The uniform convergence of $\sum_I \mathrm{e}^{-np_k}g^2(n)p_k^2$ and $\sum_I \mathrm{e}^{-np_k}g^2(n)/np_k$ are directly guaranteed by Condition 3.1. Therefore, (2) is checked. The uniformity of the convergence for $\sum_I E_k$ and hence for $\sum_I |E_k|$ guarantees (6).

For (5), since $B_k = \exp\left(-\frac{t^2}{2}p_k + \mathrm{O}(t^3 p_k n^{-1/2})\right)$ and $-\frac{t^2}{2}p_k + \mathrm{O}(t^3 p_k n^{-1/2}) \to 0$ uniformly, we have

$$|B_k - 1| \le \frac{|-\frac{t^2}{2}p_k + \mathrm{O}(t^3 p_k n^{-1/2})|}{1 - |-\frac{t^2}{2}p_k + \mathrm{O}(t^3 p_k n^{-1/2})|} = \mathrm{O}\left(-\frac{t^2}{2}p_k + t^3 p_k n^{-1/2}\right)$$

and hence

$$\sum_I |B_k - 1| \le \mathrm{O}\left(\frac{t^2}{2}\sum_I p_k) + \frac{|t^3|}{\sqrt{n}}\sum_I p_k\right) < \mathrm{O}(t^2 + |t^3|).$$

■

Proof of Lemma 4.1 Let us again consider the partition of $K = I \cup II$. First, we note that pe^{-np} has a negative derivative with respect to p on interval $(1/n, 1]$ and hence on $(1/(g(n)n^\delta), 1]$ for large n. Therefore, since there are at most $g(n)n^\delta$ terms in II,

$$0 \le \frac{g^2(n)}{n^2} n \sum_{II} p_k (1-p_k)^{n-1} \le \frac{g^2(n)}{n^2} n \sum_{II} p_k e^{-(n-1)p_k} \le \frac{g^2(n)}{n^2} n \sum_{II} \left(\frac{1}{g(n)n^\delta} e^{-(n-1)/g(n)n^\delta}\right)$$

$$\le \frac{g^2(n)}{n^2} n g(n) n^\delta \frac{1}{g(n)n^\delta} e^{-(n-1)/g(n)n^\delta} = \frac{g^2(n)}{n} e^{-(n-1)/g(n)n^\delta} = \mathrm{O}(n^{1-4\delta})\mathrm{O}(e^{-n^\delta}) \longrightarrow 0.$$

Thus, we have

$$\lim \frac{g^2(n)}{n^2} E(N_1) = \lim \frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} \tag{A3}$$

and

$$\lim \frac{g^2(n)}{n} \sum p_k \exp(-np_k) = \lim \frac{g^2(n)}{n} \sum_I p_k \exp(-np_k). \tag{A4}$$

On the other hand,

$$\frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} \le \frac{g^2(n)}{n^2} n \sum_I p_k e^{-(n-1)p_k} \le \frac{g^2(n)}{n^2} n \exp\left(\sup_I p_k\right) \sum_I p_k e^{-np_k}.$$

Furthermore, applying (2) and (3) of Lemma 3.6 in the first and the third steps below, respectively, we have

$$\begin{aligned}
\frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} &\ge \frac{g^2(n)}{n^2} n \sum_I p_k \exp\left(-\frac{(n-1)p_k}{1-p_k}\right) \\
&\ge \frac{g^2(n)}{n^2} n \sum_I p_k \exp\left(-\frac{np_k}{1-\sup_I p_k}\right) \\
&\ge \frac{g^2(n)}{n^2} n \sum_I p_k \exp\left(-\frac{np_k}{1+\sup_I p_k}\right) \\
&\ge \frac{g^2(n)}{n^2} n \exp(-2n(\sup_I p_k)^2) \sum_I p_k e^{-np_k}.
\end{aligned}$$

Noting the fact that $\lim \exp(\sup_I p_k) = 1$ and $\lim \exp(-2n(\sup_I p_k)^2) = 1$ by the definition of I,

$$\lim \frac{g^2(n)}{n^2} n \sum_I p_k (1-p_k)^{n-1} = \lim \frac{g^2(n)}{n^2} n \sum_I p_k e^{-np_k},$$

and hence by Equations (A3) and (A4), we have the equivalence of the first parts of Condition 3.1 and Condition 4.1:

$$\lim \frac{g^2(n)}{n^2} E(N_1) = \lim \frac{g^2(n)}{n} \sum p_k \exp(-np_k).$$

The equivalence of the second parts can be established similarly. ■

Proof of Theorem 4.3 Based on Theorem 4.1, it suffices to show that

$$\hat{c}_1 = \frac{g^2(n)}{n^2} N_1 \quad \text{and} \quad \hat{c}_2 = \frac{2g^2(n)}{n^2} N_2$$

are consistent estimates of c_1 and c_2, respectively. Since $\hat{c}_1$ and $\hat{c}_2$ are unbiased estimates of c_1 and c_2. We only need to verify that their variances approach zero as n increases to infinity.

$$Var(\hat{c}_1) = \frac{g^4(n)}{n^4} Var(N_1) \le \frac{g^4(n)}{n^4} \left(E(N_1) + \sum_{j \ne k} Cov(1[X_j = 1]1[X_k = 1]) \right). \tag{A5}$$

By the first part of Condition 4.1, $g^4(n)/n^4 E(N_1) \to 0$ since $g^2/n^2 \to 0$. On the other hand, we have

$$(1-p_j-p_k)^{n-2} \le (1-p_j-p_k+p_jp_k)^{n-2} = (1-p_j)^{n-2}(1-p_k)^{n-2},$$

and therefore for the second term in Equation (A5),

$$\begin{aligned}
&\frac{g^4(n)}{n^4} \sum_{j \ne k} Cov(1[X_j = 1]1[X_k = 1])) \\
&= \frac{g^4(n)}{n^4} \sum_{j \ne k} [n(n-1)p_jp_k(1-p_j-p_k)^{n-2} - n^2 p_j (1-p_j)^{n-1} p_k (1-p_k)^{n-1}] \\
&\le \frac{g^4(n)}{n^2} \sum_{j \ne k} [p_jp_k(1-p_j)^{n-2}(1-p_k)^{n-2}(p_j+p_k)] \\
&\le \frac{2g^4(n)}{n^2} \sum_{j,k} [p_j^2 p_k (1-p_j)^{n-2}(1-p_k)^{n-2}] \\
&= \frac{2}{n} \left\{ \frac{g^2(n)}{n} \left[\sum p_k (1-p_k)^{n-2} \right] \right\} \left\{ g^2(n) \left[\sum p_k^2 (1-p_k)^{n-2} \right] \right\} \longrightarrow 0.
\end{aligned}$$

The consistency of $\hat{c}_1$ follows.

$$Var(\hat{c}_2) = \frac{4g^4(n)}{n^4} Var(N_2) \le \frac{4g^4(n)}{n^4} \left[\sum E(1[X_k = 2]) + \sum_{k \ne j} Cov(1[X_k = 2], 1[X_j = 2]) \right].$$

The first term converges to 0 since

$$\frac{4g^4(n)}{n^4}\sum E(1[X_k=2]) \le \frac{2g^4(n)}{n^2}\sum p_k^2(1-p_k)^{n-2}$$
$$= \frac{2g^2(n)}{n^2}\left[g^2(n)\sum p_k^2(1-p_k)^{n-2}\right] \longrightarrow 0.$$

For the second term, since it is easily verified that

$$1-(1-p_k)^2(1-p_j)^2 \le 4(p_k+p_j),$$

we have

$$\frac{4g^4(n)}{n^4}\sum_{j\neq k} Cov(1[X_k=2], 1[X_j=2])$$
$$= \frac{4g^4(n)}{n^4}\sum_{j\neq k}\left[\frac{n!}{4(n-4)!}p_j^2p_k^2(1-p_j-p_k)^{n-4}\right.$$
$$\left.-\left(\frac{n!}{2(n-2)!}\right)^2 p_j^2p_k^2(1-p_j)^{n-2}(1-p_k)^{n-2}\right]$$
$$\le \frac{4g^4(n)}{n^4}\sum_{j\neq k}\left\{\frac{n^2(n-1)^2}{4}p_j^2p_k^2(1-p_j)^{n-4}(1-p_k)^{n-4}\left[1-(1-p_k)^2(1-p_j)^2\right]\right\}$$
$$\le 4g^4(n)\sum_{j\neq k}\left[p_j^2p_k^2(1-p_j)^{n-4}(1-p_k)^{n-4}(p_k+p_j)\right]$$
$$= 8\left[g^2(n)\sum p_k^3(1-p_k)^{n-4}\right]\left[g^2(n)\sum p_k^2(1-p_k)^{n-4}\right].$$

The third factor converges to a non-zero constant by Condition 4.1. To see that the second factor converges to zero, we note that $p^2(1-p)^{n-4}$ attains its maximum over $(0,1)$ at $p=2/(n-2)$. Therefore,

$$g^2(n)\sum p_k^3(1-p_k)^{n-4} \le g^2(n)\left(\frac{2}{n-2}\right)^2\left(1-\frac{2}{n-2}\right)^{n-4}\sum p_k \le \frac{4g^2(n)}{(n-2)^2} \longrightarrow 0.$$

The consistency of $\hat{c}_2$ follows. ■