Statistica Sinica **27** (2017), 1193-1203
doi:https://doi.org/10.5705/ss.202015.0332

ON THE ASYMPTOTIC VARIANCE OF THE CHAO ESTIMATOR FOR SPECIES RICHNESS ESTIMATION

Chang Xuan Mao, Sijia Zhang and Zhilin Liao

Shanghai University of Finance and Economics

Abstract: The Chao estimator for the number of species plays an important role in conservation biology. We show that its asymptotic variance is not estimable but admits estimable lower and upper bounds. We observe that the pre-existing variance estimator is for the variance lower bound. We propose a bias-adjusted estimator for the variance lower bound. We show that the adjusted Chao estimator in the literature and the proposed adjusted estimator for the variance lower bound are both unbiased in the limit. These findings reinforce the attractiveness of the Chao estimator. Simulation studies and applications are reported.

Key words and phrases: Bio-diversity, population size, species richness.

1. Introduction

The number of species in a species assemblage is an important parameter in conservational biology, and estimating the number of species from empirical data is an everpresent problem (Corbet, Fisher and Williams (1943); Goodman (1949); Colwell and Coddington (1994); Colwell et al. (2012)). Because observed data may empirically rule out simple models but cannot preclude complicated ones (Donoho (1988)), it is understandable that, the number of species cannot be bounded from above but can be bounded from below (Harris (1959); Bunge and Fitzpatrick (1993); Mao and Lindsay (2007)).

Suppose that there are s species and species i is observed X_i times. If X_i is a Poisson random variable with mean parameter λ_i, and the λ_i follow a mixing distribution G, then the X_i arise as a random sample from a Poisson mixture $p_j = \int (e^{-\lambda}\lambda^j/j!)\, dG(\lambda)$, $j \geqslant 0$. The number of species that appear exactly j times is $n_j = \sum_{i=1}^{s} I(X_i = j)$. The number of observed species is $n = \sum_{x=1}^{\infty} n_x$. We consider estimating the unknown number of species s from the n_j, $j \geqslant 1$. For instance, Chao and Shen (2003) studied an application about beetles. There are $n = 78$ species observed with nonzero counts $n_1 = 59$, $n_2 = 9$, $n_3 = 3$, $n_4 = 2$, $n_5 = 2$, $n_6 = 2$, and $n_{11} = 1$.

The Poisson mixture model can be used in a variety of applications (Efron and Thisted (1976); Bunge and Fitzpatrick (1993); Chao (2001)). In particular, estimating the unknown size of a population from repeated encounters can also be done in the Poisson mixture model (e.g., van der Heijden, Cruyff and van Houwelingen (2003)). Among pre-existing estimators, the Chao (1984), developed for a lower bound of the number of species, has been widely accepted in practice, see e.g., Van Hest et al. (2007); Böhning et al. (2013); Magurran (2013); Reva et al. (2015). The Zelterman estimator, once a competitor, has been shown to be inferior (Mao, Yang and Zhong (2013)). There are lower bound estimators that are less biased but demand more computational resources (Mao (2006); Mao and Lindsay (2007)).

The Chao estimator admits a bias adjusted version (Chao (2005)). We show that its bias adjusted version is unbiased in the limit. We also study the total asymptotic variance of the Chao estimator. It is a sum of an estimable component and non-estimable component. Since the non-estimable component admits an upper bound and a trivial lower bound zero, the total variance is bounded from above and below. The estimated variance in the literature is an estimator for the variance lower bound. We provide some bias-adjusted estimators for the estimable component that are also shown to be unbiased in the limit.

The rest of this article is arranged as follows. The methods are presented in Section 2. Simulation experiments are reported in Section 3. Applications are studied in Section 4. The proofs are provided as supplementary materials.

2. Methods

2.1. The Chao estimator and its asymptotic limit

The vector of counts $(n_0, n_1, n_2, \dots)^\top$ follows a multinomial distribution with infinitely many cells,

$$(n_0, n_1, n_2, \dots)^\top \sim \frac{s!}{\prod_{j=0}^{\infty} n_j!} \prod_{j=0}^{\infty} \{p_j\}^{n_j}. \tag{2.1}$$

It is clear that $E(n_j) = sp_j$ and $E(n) = s(1 - p_0)$. By the Cauchy-Schwarz inequality, $\int e^{-\lambda} dG(\lambda) \int e^{-\lambda}\lambda^2 dG(\lambda) \geqslant \{\int e^{-\lambda}\lambda dG(\lambda)\}^2$ or

$$p_0(2p_2) = \int \frac{e^{-\lambda}\lambda^0}{0!} dG(\lambda) \cdot \left\{2\int \frac{e^{-\lambda}\lambda^2}{2!} dG(\lambda)\right\} \geqslant \left\{\int \frac{e^{-\lambda}\lambda^1}{1!} dG(\lambda)\right\}^2 = p_1^2,$$

where the equality holds if and only if G is degenerate. This means that $p_1^2/(2p_2) \leqslant$

p_0. Consider the parameter

$$\nu = s(1-p_0) + \frac{sp_1^2}{2p_2}. \tag{2.2}$$

It is a lower bound of s because

$$s(1-p_0) + \frac{sp_1^2}{2p_2} \leqslant s(1-p_0) + sp_0 = s, \tag{2.3}$$

and so Chao estimator can be written as

$$\widehat{\nu} = n + \frac{n_1^2}{2n_2}. \tag{2.4}$$

2.2. The total asymptotic variance and its components

By the delta method, Chao (1989) gave the variance estimator as

$$\widehat{\sigma}^2 = n_1 \left\{ \frac{n_1}{2n_2} + \frac{n_1^2}{n_2^2} + \frac{n_1^3}{4n_2^3} \right\}. \tag{2.5}$$

Proposition 1. *As s goes to infinity, $(\widehat{\nu}-\nu)/\sqrt{s}$ converges weakly to a normal distribution with mean zero and variance $\gamma^2 = \gamma_1^2 + \gamma_2^2$, where*

$$\gamma_1^2 = \frac{p_1^2}{2p_2} + \frac{p_1^3}{p_2^2} + \frac{p_1^4}{4p_2^3}, \qquad \gamma_2^2 = \left\{ 1 - p_0 + \frac{p_1^2}{2p_2} \right\} \cdot \left\{ p_0 - \frac{p_1^2}{2p_2} \right\}.$$

The total asymptotic variance of $\widehat{\nu}$ is $\sigma_T^2 = s\gamma^2 = \sigma^2 + \omega^2$ with $\sigma^2 = s\gamma_1^2$ and $\omega^2 = s\gamma_2^2$. The component σ^2 admits an estimator $\widehat{\sigma}^2$ in (2.5). The variance component $\omega^2 = \nu(1-\nu/s)$, as a function in s, is strictly increasing with a supremum over $[\nu, \infty)$ of ν and a minimum of zero. Thus $0 \leqslant \omega^2 < \nu$ and $\sigma^2 \leqslant \sigma_T^2 = \sigma^2 + \omega^2 < \sigma_U^2 = \sigma^2 + \nu$. The lower bound σ^2 and upper bound σ_U^2 of the total asymptotic variance σ_T^2 can be easily estimated. If s and ν are replaced by $\widehat{\nu}$, then $\widehat{\omega}^2 = 0$ and $\widehat{\sigma}_T^2 = \widehat{\sigma}^2$, the choice made in the literature. Another choice replaces ω^2 with $\widehat{\nu}$ and σ_T^2 with $\widehat{\sigma}_U^2 = \widehat{\sigma}^2 + \widehat{\nu}$.

2.3. Confidence inference

The asymptotical normality of the Chao estimator $\widehat{\nu}$ can be used to construct approximate confidence intervals for the number of species s. We consider one-sided confidence intervals like $(\widehat{\ell}_{1-\alpha}, \infty)$ (Mao and Lindsay (2007)), where $\widehat{\ell}_{1-\alpha}$ is a lower confidence limit with confidence level $1-\alpha$. With $q_{0.95}$ as the 95% quantile of the standard normal distribution, one could take

$$\widehat{\ell}_{0.95} = \widehat{\nu} - q_{0.95}\widehat{\sigma}, \tag{2.6}$$

using the asymptotic normality of $\widehat{\nu}$, albeit σ_T^2 is underestimated by $\widehat{\sigma}^2$.

We consider treating $\widehat{\ell}_{0.95}$ as an approximate 95% lower confidence limit for

s. We argue that the coverage probability of $(\widehat{\ell}_{0.95}, \infty)$ is no smaller than 0.95, approximately. Take $\widehat{\ell}^{\star}_{0.95} = \widehat{\nu} - q_{0.95}\{\widehat{\sigma}^2 + (s-\nu)\nu/s\}^{1/2}$, understood as an approximate 95% "lower confidence limit" for ν. Then

$$\begin{aligned}
\widehat{\ell}_{0.95} - \widehat{\ell}^{\star}_{0.95} &= q_{0.95}\{\widehat{\sigma}^2 + \frac{(s-\nu)\nu}{s}\}^{1/2}) - q_{0.95}\widehat{\sigma} \\
&= q_{0.95}\frac{\widehat{\sigma}^2 + (s-\nu)\nu/s - \widehat{\sigma}^2}{\{\widehat{\sigma}^2 + (s-\nu)\nu/s\}^{1/2} + \widehat{\sigma}} \\
&= \left\{\frac{\nu}{s} \cdot \frac{2}{1 + \{1 + (s-\nu)\nu/(s\widehat{\sigma}^2)\}^{1/2}}\right\} \frac{q_{0.95}}{2\widehat{\sigma}}(s-\nu) \\
&\leqslant \frac{q_{0.95}}{2\widehat{\sigma}}(s-\nu).
\end{aligned}$$

The last inequality holds because $\nu/s \leqslant 1$ and $(s-\nu)\nu/(s\widehat{\sigma}^2) > 0$. If $\widehat{\sigma} > q_{0.95}/2 = 0.822$, then $(\widehat{\ell}_{0.95} - \widehat{\ell}^{\star}_{0.95}) \leqslant (s-\nu)$ or $(\widehat{\ell}_{0.95} - s) \leqslant (\widehat{\ell}^{\star}_{0.95} - \nu)$. This implies that $\Pr(\widehat{\ell}_{0.95} - s \leqslant 0) \geqslant \Pr(\widehat{\ell}^{\star}_{0.95} - \nu \leqslant 0) \approx 0.95$.

2.4. Bias adjustment

If $n_2 = 0$, then $\widehat{\nu} = \infty$ from (2.4). Chao (2005) proposes an adjusted estimator

$$\widetilde{\nu} = n + \frac{n_1(n_1-1)}{2(n_2+1)}. \tag{2.7}$$

The probability that of $n_2 = 0$ is $(1-p_2)^s$, which may be noticeable if neither s nor p_2 is large. Even if such a probability is small, one has $E(\widehat{\nu}) = \infty$. To avoid the issue of $\widehat{\nu} = \infty$ when $n_2 = 0$, we add one to the denominator of n_1^2/n_2. We provide an explicit justification on the modification in (2.7) of the numerator of n_1^2/n_2. It can be shown that

$$E(\frac{\widetilde{\nu}}{s}) = 1 - p_0 + \frac{p_1^2}{2p_2} - \frac{p_1^2}{2p_2(1-p_2)} \cdot (1-p_2)^s. \tag{2.8}$$

We obtain (2.8) using the multinomial distribution of (n_0, n_1, n_2, n_{2+}) with $n_{2+} = \sum_{x=3}^{\infty}$ derived from (2.1), and the modification from n_1^2 in (2.4) to $n_1(n_1-1)$ in (2.7).

Proposition 2. *The estimator $\widetilde{\nu}$ is unbiased in the limit,*

$$E(\frac{\widetilde{\nu}}{s}) = 1 - p_0 + \frac{p_1^2}{2p_2} + O\big((1-p_2)^s\big). \tag{2.9}$$

We propose a bias adjusted-estimator for σ^2 as

$$\widetilde{\sigma}^2 = n_1\left\{\frac{n_1-1}{2(n_2+1)} + \prod_{j=1}^{2}\frac{n_1-j}{n_2+j} + \frac{1}{4}\prod_{j=1}^{3}\frac{n_1-j}{n_2+j}\right\}. \tag{2.10}$$

It can be shown that

$$E(\frac{\widetilde{\sigma}^2}{s}) = \gamma_1^2 - (1-p_2)^{-3} \cdot r(s, p_1, p_2)(1-p_2)^s, \tag{2.11}$$

where $r(s, p_1, p_2) = O(s^2)$ is given by

$$r(s, p_1, p_2) = \gamma_1^2(1-p_2)^2 + \binom{s-1}{1}(1-p_2)\left(\frac{p_1^3}{p_2} + \frac{p_1^4}{4p_2^2}\right) + \binom{s-1}{2}\frac{p_1^4}{4p_2}.$$

The proof of (2.11) also relies on the multinomial distribution of $(n_0, n_1, n_2, n_{2+})^\top$.

Proposition 3. *$\widetilde{\sigma}^2$ is unbiased in the limit,*

$$E(\frac{\widetilde{\sigma}^2}{s}) = \gamma_1^2 + O\big(s^2(1-p_2)^s\big). \tag{2.12}$$

There are alternative approaches to constructing bias-adjusted estimators, for example,

$$\check{\nu} = n + \frac{n_1^2}{2(n_2+1)}, \tag{2.13}$$

$$\check{\sigma}^2 = n_1\left\{\frac{n_1}{2(n_2+1)} + \frac{n_1^2}{(n_2+1)^2} + \frac{n_1^3}{4(n_2+1)^3}\right\}. \tag{2.14}$$

Colwell (2013) has provided others.

Proposition 4. *$\check{\nu}$ in (2.13) and $\check{\sigma}^2$ in (2.14) are unbiased in the limit,*

$$E(\frac{\check{\nu}}{s}) = 1 - p_0 + \frac{p_1^2}{2p_2} + \frac{b_1(p_1, p_2)}{s} + O((1-p_2)^s), \tag{2.15}$$

$$E(\frac{\check{\sigma}^2}{s}) = \gamma_1^2 + \frac{b_2(p_1, p_2)}{s} + O(s^{-3/2}), \tag{2.16}$$

where $b_1(p_1, p_2) = p_1/(2p_2)$ and

$$b_2(p_1, p_2) = \sum_{i=1}^{3} \frac{\{p_2 + (p_1+p_2)(i-1)/2\} i p_1^i}{\{1 + I(i=1) + 3I(i=3)\} p_2^{i+1}}.$$

Because $b_1(p_1, p_2) > 0$ and $b_2(p_1, p_2) > 0$, although the bias of $\check{\nu}$ in (2.13) and that of $\check{\sigma}^2$ in (2.14) are small, they are positive if s is not too small.

3. Simulation Studies

We compared the Chao estimator against the jackknife estimator $\widehat{\nu}_J = n + n_1$ with an estimand $\nu_J = s(1 - p_0 + p_1)$ (Smith and van Belle (1984); Mao, Yang and Zhong (2013)). We considered mixing distributions, $G_1 = 0.9\delta(0.3) + 0.1\delta(10)$, $G_2 = 0.5\delta(1) + 0.5\delta(2)$, $G_3 = 0.6\delta(0.3) + 0.3\delta(3) + 0.1\delta(10)$, and G_4, a gamma distribution with shape and scale 1, where $\delta(\lambda)$ is a distribution degenerate at λ.

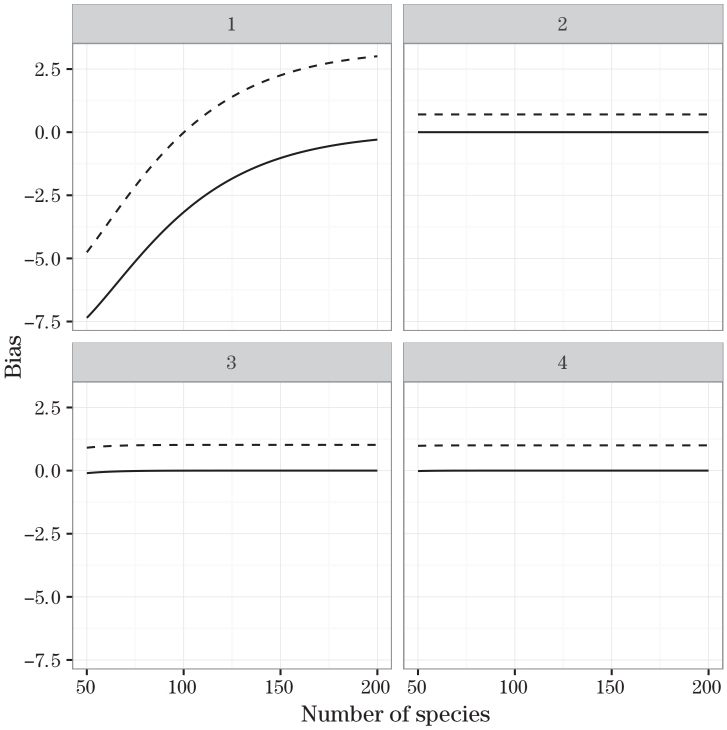

Figure 1. The biases $E(\widetilde{\nu}) - \nu$ (solid) and $E(\check{\nu}) - \nu$ (dashed) as functions of s given different mixing distributions (G_i in panel i, $i = 1$, 2, 3 and 4).

We calculated the biases of $\widetilde{\nu}$ and $\check{\nu}$ over a grid of s ($50 \leqslant s \leqslant 200$), given each of the four mixing distributions. Figure 1 presents these biases against the number of species. Under G_1, the bias of $\check{\nu}$ is negative for small s, and is positive for $s > 100$, while the bias of $\widetilde{\nu}$ is always negative and is increasing in s. Under G_2, G_3, or G_4, the biases of $\widetilde{\nu}$ and $\check{\nu}$ vary little over s, the bias of $\check{\nu}$ is increasing in s and approaches $p_1/(2p_2)$, and the bias of $\widetilde{\nu}$ is increasing in s and approaches zero.

In our simulation experiment, there were 12 settings labelled by (s, G), with $s \in \{50, 500, 5,000\}$ and $G \in \{G_1, G_2, G_3, G_4\}$. For each setting, 5,000 samples were generated and the lower bound ν was calculated. When $s = 50$, the probability of $n_2 = 0$ was 0.2155 (G_1), 0.0103 (G_3), and 0.0013 (G_4), and the proportion of samples in which $n_2 = 0$ was 0.2230 (G_1), 0.0128 (G_3) and 0.0014 (G_4). Under $(50, G_2)$, the probability of $n_2 = 0$ was 2.5×10^{-6} and no sample had $n_2 = 0$. When $s = 500$ or $5,000$, no sample had $n_2 = 0$.

For each sample, we calculated the estimate $\widehat{\nu}$, the standard error $\widehat{\sigma}$, together with their adjusted versions $\widetilde{\nu}$, $\check{\nu}$, $\widetilde{\sigma}$ and $\check{\sigma}$. Table 1 presents the true values of

ν and σ, the sample means, and the root mean square errors of their estimators. Some statistics in settings $(50, G_1)$, $(50, G_3)$, and $(50, G_4)$ were unavailable due to $n_2 = 0$. The adjustment eliminates the value of infinity and the sample means were close to the corresponding estimands. Under $(500, G_2)$, $(5,000, G_2)$, $(500, G_3)$, $(5,000, G_3)$, $(500, G_4)$, and $(5,000, G_4)$, the sample means of the adjusted and unadjusted estimators differ little. Under $(500, G_1)$ and $(5,000, G_1)$, the sample means of $\widehat{\nu}$ is larger than those of $\widetilde{\nu}$ and $\check{\nu}$. The adjusted estimators generally perform better, and indispensable on occasion.

Table 2 presents the coverage probabilities of 95% lower confidence limits obtained from the Chao estimator and its adjusted versions, and those of the 95% lower confidence limit $\widehat{\ell}_{0.95,J} = \widehat{\nu}_J - q_{0.95}\widehat{\sigma}_J$ obtained from the jackknife estimator $\widehat{\nu}_J$, where $\widehat{\sigma}^2_J = 2n_1$ estimates $\sigma^2_J = 2sp_1$. The asymptotic variance of $\widehat{\nu}_J$ is $\sigma^2_{T,J} = s\sigma^2_J + \nu_J - \nu^2_J/s$. Table 2 presents the expected number of observed species $E(n)$, and the true values of ν and ν_J. The poor coverage probabilities of $(\widehat{\ell}_{0.95,J}, \infty)$ in settings $(50, G_2)$, $(500, G_2)$, and $(5,000, G_2)$ are due to the fact that $\nu_J > s$ under G_2. From Table 2, confidence intervals from the Chao estimator and its adjusted versions are conservative. However, it is difficult to make some improvements upon them as the lower bound can be quite sharp when the population is homogeneous or close to homogeneous. To obtain larger lower confidence limits, we can consider the possibility of using larger lower bounds such as those in Mao (2006) and Mao and Lindsay (2007).

4. Applications

In addition to the application beetle, we introduce three applications: tomato, firearm and coin. The application tomato studied in Mao and Lindsay (2002) is about expressed genes of tomato flowers. There were $n = 1,825$ expressed genes observed from 2,586 expressed sequence tags. The non-zero counts were $n_1 = 1,434$, $n_2 = 253$, $n_3 = 71$, $n_4 = 33$, $n_5 = 11$, $n_6 = 6$, $n_7 = 2$, $n_8 = 3$, $n_9 = 1$, $n_{10} = 2$, $n_{11} = 2$, $n_{12} = 1$, $n_{13} = 1$, $n_{14} = 1$, $n_{16} = 2$, $n_{23} = 1$, and $n_{27} = 1$. The application firearm studied in van der Heijden, Cruyff and van Houwelingen (2003) is about illegal possession of firearms from the Dutch police. There were $n = 2,638$ cases with $n_1 = 2,561$, $n_2 = 72$, $n_3 = 5$, and $n_x = 0$ for $x \geqslant 4$. In firearm, we are interested in estimating the size s of a population consisting of individuals that can be seen multiple times. The application coin concerns 662 coins struck from a variety of dies, with $n = 660$, $n_1 = 658$ and $n_2 = 2$ (Chao (1984); Eddy (1967)). In coin, we are interested in estimating the number of dies

Table 1. The sample mean (mean) and root mean square error (rmse) of the Chao estimator and the estimator for the lower bound of its asymptotic variance, and their adjusted versions, together with the true values in 12 settings.

			mean			rmse		
G	s	ν	$\widehat{\nu}$	$\widetilde{\nu}$	$\check{\nu}$	$\widehat{\nu}$	$\widetilde{\nu}$	$\check{\nu}$
G_1	50	50	-	42	45	-	7.31	4.70
	500	498	527	498	501	29.27	0.03	3.28
	5,000	4,976	5,001	4,976	4,979	24.84	0.70	2.61
G_2	50	49	50	48	49	1.77	0.19	0.51
	500	486	488	486	487	1.76	0.04	0.75
	5,000	4,863	4,866	4,864	4,865	2.44	0.75	1.45
G_3	50	36	-	36	37	-	0.19	0.82
	500	361	364	361	362	2.98	0.24	0.78
	5,000	3,611	3,612	3,609	3,610	1.13	1.97	0.95
G_4	50	37	-	37	38	-	0.19	0.80
	500	375	378	375	376	3.47	0.37	1.37
	5,000	3,750	3,755	3,752	3,753	4.63	1.62	2.62
		σ	$\widehat{\sigma}$	$\widetilde{\sigma}$	$\check{\sigma}$	$\widehat{\sigma}$	$\widetilde{\sigma}$	$\check{\sigma}$
G_1	50	35	-	15	29	-	19.52	5.42
	500	109	125	102	112	15.16	7.60	3.10
	5,000	346	350	343	347	3.95	2.39	0.94
G_2	50	7	8	6	8	1.08	0.85	0.23
	500	23	24	23	23	0.32	0.24	0.12
	5,000	73	73	73	73	0.12	0.06	0.05
G_3	50	8	-	6	9	-	2.02	0.83
	500	25	26	25	26	0.90	0.70	0.23
	5,000	81	81	80	81	0.17	0.33	0.03
G_4	50	9	-	8	10	-	1.71	0.64
	500	30	30	29	30	0.82	0.48	0.29
	5,000	94	94	93	94	0.25	0.15	0.09

from which coins were struck.

Table 3 presents the estimate $\widehat{\nu}$ and the standard error $\widehat{\sigma}$, together with their two adjusted versions $\widetilde{\nu}$, $\check{\nu}$, $\widetilde{\sigma}$ and $\check{\sigma}$ for each application. Table 3 also presents $\widehat{\sigma}_U$, $\widetilde{\sigma}_U$ and $\check{\sigma}_U$. It is clear that the differences between $\widehat{\nu}$ and $\widehat{\sigma}$ and their adjusted versions $\widetilde{\nu}$ and $\widetilde{\sigma}$ (or $\check{\nu}$ and $\check{\sigma}$) vary over applications. In tomato, the unadjusted version and each of the adjusted versions are quite close, and in firearm, the two versions have some differences. In beetle, $\widehat{\nu}/\widetilde{\nu} - 1 = 8.9\%$, $\widehat{\sigma}/\widetilde{\sigma} - 1 = 31.0\%$, $\widehat{\sigma}_U/\widetilde{\sigma}_U - 1 = 29.6\%$, and $\widehat{\nu}/\check{\nu} - 1 = 7.7\%$, $\widehat{\sigma}/\check{\sigma} - 1 = 14.4\%$, $\widehat{\sigma}_U/\check{\sigma}_U - 1 = 14.0\%$. In coin, the unadjusted and adjusted versions differ dramatically (e.g., $\widehat{\sigma}/\widetilde{\sigma} - 1 = 172.6\%$). We calculated $(\widetilde{\nu} - q_{0.95} \cdot \widetilde{\sigma}, \infty)$ for each application: $(145, \infty)$ (beetle), $(5,317, \infty)$ (tomato), $(38,589, \infty)$ (firearm), and $(26,255, \infty)$ (coin).

Table 2. The coverage probabilities of 95% lower confidence limits obtained from the Chao estimator, its adjusted versions, and the jackknife estimator.

					coverage probability			
G	s	$E(n)$	ν	ν_J	$\widehat{\nu}$	$\widetilde{\nu}$	$\check{\nu}$	$\widehat{\nu}_J$
G_1	50	17	50	27	-	1	1	1
	500	167	498	267	1	0.99	1	1
	5,000	1,666	4,976	2,666	0.97	0.97	0.97	1
G_2	50	37	49	53	1	1	1	0.91
	500	374	486	534	0.99	0.99	0.99	0.39
	5,000	3,742	4,863	5,339	1	1	1	0
G_3	50	27	36	36	-	1	1	1
	500	270	361	359	1	1	1	1
	5,000	2,703	3,611	3,593	1	1	1	1
G_4	50	25	37	37	-	1	1	1
	500	250	375	375	1	1	1	1
	5,000	2,500	3,750	3,752	1	1	1	1

Table 3. The estimate, standard error, and their adjusted versions.

	$\widehat{\nu}$	$\widetilde{\nu}$	$\check{\nu}$	$\widehat{\sigma}$	$\widetilde{\sigma}$	$\check{\sigma}$	$\widehat{\sigma}_U$	$\widetilde{\sigma}_U$	$\check{\sigma}_U$
beetle	271	249	252	83	63	73	85	65	74
tomato	5,889	5,870	5,873	340	336	338	348	345	347
firearm	48,185	47,543	47,561	5,666	5,444	5,554	5,670	5,448	5,558
coin	108,901	72,711	72,821	77,003	28,243	42,041	77,003	28,244	42,042

5. Discussion

We shed some light on the Chao estimator in the Poisson mixture model. Its variance is discussed in detail. We justify the adjustment of the Chao estimator using the concept "unbiased in the limit" and propose two adjusted estimators for the lower bound of the variance of the Chao estimator. We demonstrate that, using the Chao estimator and the estimated lower bound of its variance, one can calculate a lower confidence limit for the number of species that approximately achieves its nominal confidence level.

Supplementary Materials

The proofs are provided as the Web Supplementary Materials.

Acknowledgment

The authors thank the Editor, an associate editor, and the referees for their very constructive comments.

References

Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C. and Arnold, M. (2013). A generalization of Chao's estimator for covariate information. *Biometrics* **69**, 1033–1042.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of American Statistical Association* **88**, 364–373.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**, 265–270.

Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics* **45**, 427–438.

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics* **6**, 138–155.

Chao, A. (2005). Species estimation and applications. *Encyclopedia of Statistical Sciences.*

Chao, A. and Shen, T. J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* **10**, 429–443.

Colwell, R. (2013). EstimateS: Statistical estimation of species richness and shared species from samples. `http://viceroy.eeb.uconn.edu/estimates`.

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L. and Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* **5**, 3–21.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **345**, 101–118.

Corbet, A. S., Fisher, R. A. and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* **12**, 42–58.

Donoho, D. L. (1988). One-sided inference about functionals of a density. *Annals of Statistics* **16**, 1390–1420.

Eddy, S. K. (1967). *The minting of Antoniniani AD 238-249 and the Smyrna hoard.* American Numismatic Society.

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

Goodman, L. A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics* **20**, 572–579.

Harris, B. (1959). Determining bounds on integrals with applications to cataloging problems. *Annals of Mathematical Statistics* **30**, 521–548.

Magurran, A. E. (2013). *Ecological Diversity and its Measurement.* Springer Science & Business Media.

Mao, C. X. (2006). Inference on the number of species through geometric lower bounds. *Journal of the American Statistical Association*, **101**, 1163–1170.

Mao, C. X. and Lindsay, B. G. (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669–682.

Mao, C. X. and Lindsay, B. G. (2007). Estimating the number of classes. *The Annals of Statistics* **35**, 917–930.

Mao, C. X., Yang, N. and Zhong, J. (2013). On population size estimators in the Poisson mixture model. *Biometrics* **69**, 758–765.

Reva, O. N., Zaets, I. E., Ovcharenko, L. P., Kukharenko, O. E., Shpylova, S. P., Podolich, O. V., de Vera, J.-P. and Kozyrovska, N. O. (2015). Metabarcoding of the kombucha microbial community grown in different microenvironments. *AMB Express* **5**, 35.

Smith, E. P. and van Belle, G. (1984). Nonparametric estimation of species richness. *Biometrics* **40**, 119–129.

van der Heijden, P. G. M., Cruyff, M. and van Houwelingen, H. C. (2003). Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica* **57**, 289–304.

Van Hest, N., Smit, F., Baars, H., de Vries, G., De Haas, P., Westenend, P., Nagelkerke, N. and Richardus, J. H. (2007). Completeness of notification of tuberculosis in the Netherlands: how reliable is record-linkage and capture–recapture analysis? *Epidemiology and infection* **135**, 1021–1029.

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: mao.changxuan@mail.shufe.edu.cn

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: zhangsijia19@hotmail.com

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: zhilinliao@yeah.net

(Received September 2015; accepted May 2016)