Statistics and Probability Letters 176 (2021) 109154

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A correction to make Chao estimator conservative when the number of sampling occasions is finite

Alessio Farcomeni [a,*], Francesco Dotto [b]

[a] University of Rome "Tor Vergata", Department of Economics and Finance, Italy
[b] University of Salerno, Department of Economics and Statistics, Italy

ARTICLE INFO

Article history:
Received 9 March 2021
Received in revised form 23 April 2021
Accepted 6 May 2021
Available online 12 May 2021

Keywords:
Chao estimator
Conservative estimation
Sampling efforts

ABSTRACT

Chao estimator is not guaranteed to be asymptotically conservative with finite sampling efforts. A simple correction solves the issue. We illustrate with simulations and examples that the corrected Chao estimator is asymptotically conservative, and has lower standard error.

1. Introduction

Chao (1987) estimator is used in a wide variety of contexts to estimate population sizes after repeated sampling. Chao estimator has several advantages, including ease of computation. The most attractive property is that it is guaranteed to be asymptotically conservative under any form of unobserved heterogeneity. Much interest has been devoted to Chao and Chao-like estimators, e.g., see Chao (1989), Mao (2006), Mao and Lindsay (2007), Rivest and Baillargeon (2007), Böhning and van der Heijden (2009), Böhning (2010), Lanumteang and Böhning (2011), Böhning et al. (2013), Farcomeni (2018), Dotto and Farcomeni (2018), Puig and Kokonendji (2018), and Chao and Colwell (2017) for a review.

The premise of this work is the fact that one approximation in the original Chao (1987) proof might not be accurate in several cases. Indeed, a step in the proof involves approximating the binomial distribution through a Poisson distribution, which is acceptable only with low sampling probability and large number of sampling occasions S. This makes Chao estimator conservative only, in some sense, after a double asymptotics argument according to which both the sample size and number of sampling occasions diverge. In applied scenarios, nevertheless, the number of sampling occasions is fixed and can be as low as $S = 2$. We put forward a different proof of the same result, which does not rely on the Poisson approximation, and that results in a simple correction factor. Our corrected Chao estimator has the same advantages of the original Chao estimator, and it is conservative under any form of unobserved heterogeneity for large sample sizes regardless of the number of sampling occasions.

It shall be noted that our argument is mostly based on the Cauchy–Schwarz inequality for a binomial mixture. It more generally holds for members of the power series mixtures family, see Böhning (2015) and Böhning et al. (2019) for a detailed discussion. Interestingly, the very same correction factor can be derived as a special case (when there is no time effect) of the log-linear formulation in Rivest and Baillargeon (2007), and indeed it is implemented in function `closedpCI.t` in R library `Rcapture` (Baillargeon and Rivest, 2007).

* Corresponding author.
E-mail address: alessio.farcomeni@uniroma2.it (A. Farcomeni).

https://doi.org/10.1016/j.spl.2021.109154

The rest of the paper is as follows: in the next section we briefly set up our setting, and introduce our correction for Chao estimator. We confirm our theoretical findings in a brief simulation study in Section 3, where it is shown that classical Chao estimator is indeed not conservative even with large sample sizes. In Section 4 we discuss two real data examples. Concluding remarks are given in Section 5.

2. Finite efforts correction

Let f_j, $j = 1, \dots, S$, denote the number of subjects observed exactly j times in a discrete-time closed-population capture–recapture experiment based on S sampling occasions. We assume each of the $n = \sum_j f_j$ subjects is sampled from a finite population of unknown size N, with heterogeneous subject-specific probabilities. Consequently, for the ith subject the number of sightings n_i follow a binomial distribution with

$$p_x(\theta_i) = \Pr(n_i = x|\theta = \theta_i) = \binom{S}{x}\theta_i^x(1-\theta_i)^{S-x}.$$

Random subject-specific parameters can arise from any distribution $F(\cdot)$ with support in the unit interval. Clearly, then,

$$\Pr(n_i = x) = \binom{S}{x}\int_0^1 \theta^x(1-\theta)^{S-x}\, dF(\theta). \tag{1}$$

Let now $E[f_x] = m_x = N\Pr(n_i = x)$. Under certain assumptions, Chao (1987) shows that

$$m_0 \geq \frac{m_1^2}{2m_2}. \tag{2}$$

Plug-in of f_j as an empirical estimate of m_j, $j = 1, 2$, gives Chao estimator $\hat{N}_c = n + f_1^2/(2f_2)$. The law of large numbers guarantees that Chao estimator $\hat{N}_c \leq N$, without any further assumption on $F(\cdot)$. In order to show the result, Chao (1987) replaces the binomial with the Poisson kernel in (1). It is clearly stated that the approximation is valid under two conditions: first, that S is large; and secondly, that $\int_{p^*}^1 dF(\theta)$ is small for p^* such that the Poisson approximation is not good for $p > p^*$.

In practice, it is well known that Chao estimator might not be conservative even with large sample sizes. In the following, we show that a corrected Chao estimator of the kind

$$\hat{N}_{CC} = n + \frac{(S-1)f_1^2}{2Sf_2} \tag{3}$$

is asymptotically conservative for all $S > 1$, and without any assumption on the tail of $F(\cdot)$.

To this end, note that

$$\frac{p_1(\theta)^2}{2p_2(\theta)} = \frac{S^2\theta^2(1-\theta)^{2(S-1)}}{S(S-1)\theta^2(1-\theta)^{S-2}} = \frac{S}{S-1}(1-\theta)^S = \frac{S}{S-1}p_0(\theta).$$

Consequently,

$$\frac{S}{S-1}m_0 = \frac{S}{S-1}\int_0^1 p_0(\theta)\, dF(\theta) = \int_0^1 \frac{p_1(\theta)^2}{2p_2(\theta)}\, dF(\theta) \geq \frac{m_1}{2m_2},$$

where the last inequality follows from Jensen's' inequality. Summarizing the above result,

$$m_0 \geq \frac{(S-1)m_1}{2Sm_2}.$$

The inequality above can be replaced to (2), which holds only for large S. The rest of the proof in Chao (1987) can be seen to hold without further modifications, leading to conclude that (3) is asymptotically conservative for all $S > 1$. A similar correction factor shall be used to adjust the standard error. A consequence is that the standard error of $\hat{N}_{CC}$ will always be smaller than that of $\hat{N}_C$.

Interestingly enough, a different reasoning leads (Chao, 1989) to the same factor. The estimator finally proposed in Chao (1989) involves the approximation $(S-1)/S \cong 1$. We deem it is important to recognize the factor as a *correction* factor; which is essential to guarantee that the estimator is asymptotically conservative (in N).

We conclude this section discussing bias correction. The bias-corrected version of Chao estimator $\hat{N}_B = n + \frac{f_1(f_1-1)}{2(f_2+1)}$ guarantees existence, and has different advantages. It is by definition smaller than $\hat{N}_C$, but it still is not guaranteed to be asymptotically conservative. The theory in Böhning et al. (2019) can be directly applied to obtain a corrected-bias-corrected estimator $\hat{N}_{BC} = n + \frac{Sf_1(f_1-1)}{2S(f_2+1)}$. A comparison will be shown in the next section.

Table 1
Simulation study. For Chao ($\hat{N}_C$), bias-corrected Chao ($\hat{N}_B$), the proposed corrected Chao ($\hat{N}_{CC}$) and corrected-bias-corrected ($\hat{N}_{BC}$) estimators: probability of being conservative $\Pr(\hat{N} < N)$, bias $B(\cdot)$, standard deviation (in brackets below the bias). For these settings, $\eta_i \sim Unif(-3, -1)$. Results are averaged over $B = 1000$ replicates, biases and standard errors are relative to the population size.

(a) : $N = 1000$								
S	$B(N_C)$	$B(N_B)$	$B(N_{BC})$	$B(N_{CC})$	$P(N_C < N)$	$P(N_B < N)$	$P(N_{BC} < N)$	$P(N_{CC} < N)$
2	0.644 (0.72)	0.587 (0.66)	−0.061 (0.16)	−0.03 (0.17)	0	0	0.7	0.62
3	0.3 (0.37)	0.267 (0.34)	−0.056 (0.15)	−0.03 (0.15)	0.07	0.09	0.69	0.63
5	0.129 (0.22)	0.106 (0.2)	−0.054 (0.14)	−0.04 (0.14)	0.22	0.26	0.7	0.63
10	0.037 (0.15)	0.019 (0.14)	−0.052 (0.14)	−0.04 (0.14)	0.42	0.48	0.69	0.64
(b) : $N = 5000$								
S	$B(N_C)$	$B(N_B)$	$B(N_{BC})$	$B(N_{CC})$	$P(N_C < N)$	$P(N_B < N)$	$P(N_{BC} < N)$	$P(N_{CC} < N)$
2	0.604 (0.62)	0.593 (0.61)	−0.058 (0.09)	−0.05 (0.08)	0	0	0.81	0.78
3	0.273 (0.29)	0.267 (0.28)	−0.056 (0.08)	−0.05 (0.08)	0	0	0.82	0.81
5	0.11 (0.13)	0.105 (0.13)	−0.055 (0.08)	−0.05 (0.08)	0.05	0.06	0.83	0.81
10	0.021 (0.07)	0.017 (0.06)	−0.054 (0.08)	−0.05 (0.08)	0.39	0.41	0.84	0.83
(c) : $N = 10\,000$								
S	$B(N_C)$	$B(N_B)$	$B(N_{BC})$	$B(N_{CC})$	$P(N_C < N)$	$P(N_B < N)$	$P(N_{BC} < N)$	$P(N_{CC} < N)$
2	0.608 (0.62)	0.603 (0.61)	−0.053 (0.07)	−0.05 (0.07)	0	0	0.85	0.83
3	0.275 (0.28)	0.271 (0.28)	−0.053 (0.07)	−0.05 (0.07)	0	0	0.88	0.86
5	0.11 (0.12)	0.108 (0.12)	−0.053 (0.07)	−0.05 (0.07)	0.01	0.01	0.9	0.89
10	0.02 (0.05)	0.018 (0.05)	−0.052 (0.07)	−0.05 (0.07)	0.34	0.36	0.89	0.89
(d) : $N = 100\,000$								
S	$B(N_C)$	$B(N_B)$	$B(N_{BC})$	$B(N_{CC})$	$P(N_C < N)$	$P(N_B < N)$	$P(N_{BC} < N)$	$P(N_{CC} < N)$
2	0.598 (0.6)	0.597 (0.6)	−0.056 (0.06)	−0.06 (0.06)	0	0	1	1
3	0.268 (0.27)	0.268 (0.27)	−0.055 (0.06)	−0.06 (0.06)	0	0	1	1
5	0.106 (0.11)	0.106 (0.11)	−0.055 (0.06)	−0.05 (0.06)	0	0	1	1
10	0.017 (0.02)	0.017 (0.02)	−0.054 (0.06)	−0.05 (0.06)	0.11	0.12	1	1

3. Simulation study

We conduct a simulation study to empirically compare the corrected $\hat{N}_{CC}$ and corrected-bias-corrected $\hat{N}_{BC}$ estimators with classical Chao estimator $\hat{N}_C$ and its bias-corrected version $\hat{N}_B$. We generate data for N subjects, with $N \in \{1000, 5000, 10\,000, 100\,000\}$, with $S \in \{2, 3, 5, 10\}$ capture occasions. Subject specific capture probabilities p_i, $i = 1, 2, \ldots, N$, are defined as $p_i = \exp\{-\log(S) + .5\eta_i\}/(1 + \exp\{-\log(S) + .5\eta_i\})$. We let p_i depend on S in order to have a comparable sampling ratio $.2 \leq n/N \leq .4$ across difference scenarios. As far as unobserved heterogeneity is concerned, we let $\eta_i \sim N(-4, 1)$ in one setting, and $\eta_i \sim U(-3, -1)$ in another one. Note that there is a one-to-one relationship between η_i and θ_i in (1), due to the invertible logistic transformation we use to specify p_i in this section. We finally obtain 32 scenarios, and, for each scenario, generate the data $B = 1000$ times. In Table 1 we report probability of being conservative, bias, and standard deviation of the estimates relative to the population size; with Gaussian latent variables. In Table 2 the same is shown for uniform latent variables. A visual account is also given in Fig. 1, where we report the distributions of $\hat{N}_C$, and $\hat{N}_{CC}$ for different values of N, with therefore 8 scenarios collapsed in each panel. $\hat{N}_B$ and $\hat{N}_{BC}$ are not reported in Fig. 1 as their distribution almost overlaps with that of $\hat{N}_C$ and $\hat{N}_{CC}$, respectively.

As expected, it can be seen that as N grows the corrected Chao estimator is more and more likely to be conservative. This does not necessarily apply to the classical Chao estimator and to its bias-corrected version, which would also require S to be large and/or the unobserved latent variable to concentrate probability mass around zero. There is not much difference in these settings between $\hat{N}_{cc}$ and $\hat{N}_{bc}$, with the latter being slightly more likely to be conservative; at the price of slightly larger bias.

Table 2
Simulation study. For Chao ($\hat{N}_C$), bias-corrected Chao ($\hat{N}_B$), the proposed corrected Chao ($\hat{N}_{CC}$) and corrected-bias-corrected ($\hat{N}_{BC}$) estimators: probability of being conservative $\Pr(\hat{N} < N)$, bias $B(\cdot)$, standard deviation (in brackets below the bias). For these settings, $\eta_i \sim N(-3, 2)$. Results are averaged over $B = 1000$ replicates, biases and standard errors are relative to the population size.

S	$B(N_C)$	$B(N_B)$	$B(N_{BC})$	$B(N_{CC})$	$P(N_C < N)$	$P(N_B < N)$	$P(N_{BC} < N)$	$P(N_{CC} < N)$
(a) : $N = 1000$								
2	0.275	0.267	−0.089	−0.08	0	0	0.95	0.95
	(0.29)	(0.28)	(0.1)	(0.1)				
3	0.094	0.09	−0.081	−0.08	0.07	0.09	0.94	0.93
	(0.12)	(0.11)	(0.09)	(0.09)				
5	0.011	0.007	−0.075	−0.07	0.46	0.48	0.94	0.93
	(0.06)	(0.06)	(0.09)	(0.09)				
10	−0.032	−0.035	−0.07	−0.07	0.74	0.76	0.92	0.92
	(0.06)	(0.06)	(0.08)	(0.08)				
(b) : $N = 5000$								
S	$B(N_C)$	$B(N_B)$	$B(N_{BC})$	$B(N_{CC})$	$P(N_C < N)$	$P(N_B < N)$	$P(N_{BC} < N)$	$P(N_{CC} < N)$
2	0.267	0.266	−0.089	−0.09	0	0	1	1
	(0.27)	(0.27)	(0.09)	(0.09)				
3	0.088	0.087	−0.083	−0.08	0	0	1	1
	(0.09)	(0.09)	(0.09)	(0.09)				
5	0.006	0.005	−0.077	−0.08	0.42	0.43	1	1
	(0.03)	(0.03)	(0.08)	(0.08)				
10	−0.036	−0.037	−0.072	−0.07	0.94	0.95	1	1
	(0.04)	(0.04)	(0.07)	(0.07)				
(c) : $N = 10\,000$								
S	$B(N_C)$	$B(N_B)$	$B(N_{BC})$	$B(N_{CC})$	$P(N_C < N)$	$P(N_B < N)$	$P(N_{BC} < N)$	$P(N_{CC} < N)$
2	0.267	0.266	−0.089	−0.09	0	0	1	1
	(0.27)	(0.27)	(0.09)	(0.09)				
3	0.088	0.087	−0.083	−0.08	0	0	1	1
	(0.09)	(0.09)	(0.08)	(0.08)				
5	0.005	0.005	−0.077	−0.08	0.4	0.4	1	1
	(0.02)	(0.02)	(0.08)	(0.08)				
10	−0.037	−0.037	−0.072	−0.07	0.99	0.99	1	1
	(0.04)	(0.04)	(0.07)	(0.07)				
(d) : $N = 100\,000$								
S	$B(N_C)$	$B(N_B)$	$B(N_{BC})$	$B(N_{CC})$	$P(N_C < N)$	$P(N_B < N)$	$P(N_{BC} < N)$	$P(N_{CC} < N)$
2	0.265	0.265	−0.09	−0.09	0	0	1	1
	(0.27)	(0.26)	(0.09)	(0.09)				
3	0.086	0.086	−0.083	−0.08	0	0	1	1
	(0.09)	(0.09)	(0.08)	(0.08)				
5	0.004	0.004	−0.077	−0.08	0.23	0.23	1	1
	(0.01)	(0.01)	(0.08)	(0.08)				
10	−0.037	−0.037	−0.072	−0.07	1	1	1	1
	(0.04)	(0.04)	(0.07)	(0.07)				

4. Examples

We revisit in this section three real data examples: two of them belong to the field of epidemiology , while the third is a benchmark dataset in ecology.

4.1. HIV prevalence

Abeni et al. (1994) collected data about HIV-1 infected in the Lazio region in Italy in 1990. Patient lists were from $S = 4$ centers in the city of Rome, with $n = 1896$, $f_1 = 1774$, $f_2 = 115$, $f_3 = 7$ and $f_4 = 0$. A general agreement, using log-linear models or more complex approaches (Wang et al., 2005; Baillargeon and Rivest, 2007; Farcomeni, 2016) is that the true population size is between 11,000 and 13,000. For this data set, $\hat{N}_C = 15\,579$, which is at odds with previous analyses and unlikely to be a lower bound for the true population size. On the other hand, $\hat{N}_{CC} = 12\,158$, which is much more reasonable.

4.2. Multiple sclerosis prevalence

Farcomeni et al. (2018), Farcomeni (2020) report data about Multiple Sclerosis (MS) in the metropolitan area of Rome. Data arise from treatment lists in $S = 6$ centers, with $n = 1007$, $f_1 = 894$, $f_2 = 105$, $f_3 = 8$. It is well known that

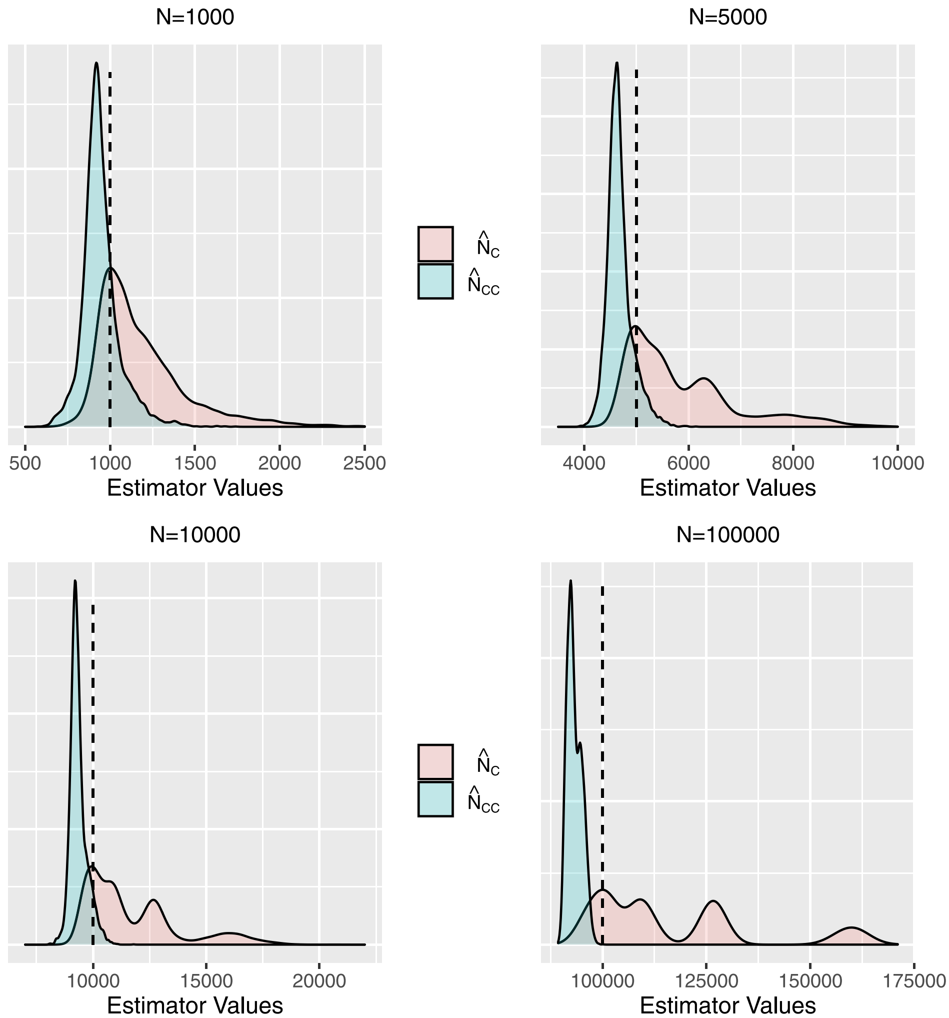

Fig. 1. Simulation study. Observed distribution of $\hat{N}_C$ and $\hat{N}_{CC}$ for different values of N. Each panel reports the 8000 estimates obtained for each estimator across settings with the same value of N.

Italy is a high risk country for MS, with an expected prevalence in Lazio region between 110 (the national average) and 140 per 100,000 inhabitants (e.g., Bargagli et al. (2016)). For the data at hand, Chao estimator gives $\hat{N}_C = 4813$, which would lead to an estimated prevalence of about 150 per 100,000 inhabitants. Once again, this is possibly slightly too large, and unlikely to be a lower bound in the face of other works dealing prevalence of MS in the area. On the other hand, $\hat{N}_{CC} = 4179$, leading to the reasonable estimated prevalence of 129 per 100,000.

4.3. Cottontail abundance

Edwards and Eberhardt (1967) report data about Cottontail Rabbits (*Sylvilagus floridanus*). For this data, the true population size $N = 135$ is known. A sample of $n = 76$ rabbits are observed in $S = 18$ capture occasions, with $f_1 = 43$ and $f_2 = 16$. For these data, $\hat{N}_C = 134$ with 95% confidence interval $(87 - 181)$; while $\hat{N}_{CC} = 131$ with 95% confidence interval $(86 - 175)$. As S is quite large, there is not much difference between the point estimates; with $\hat{N}_C$ being conservative and therefore automatically closer to the true population size than $\hat{N}_{CC}$. Nonetheless, the correction leads by construction to a smaller standard error, and consequently to a confidence interval that is 5 units narrower.

5. Concluding remarks

While the simple correction factor that we put forward in this work is not less than 0.5 (when $S = 2$) and linearly increases to the unity, it is important to remark that it is a *multiplicative* factor. If the term $f_1^2/(2f_2)$ is large (e.g., for large population sizes, when f_2 is small, when there is one-inflation) the corrected and original Chao estimators can differ substantially. It shall be remarked that the approximation does not apply in continuous-time experiments, which indeed one can recover by letting $S \to \infty$. In all other cases in which the number of capture occasions is finite we recommend using the proposed correction factor. We have seen in simulated and real data scenarios that the classical Chao estimator, when S is moderately small, can lead to unreasonably large estimates for the population size. These are then believed by the researcher to be lower bounds, due to theoretical properties that might not hold. Our correction guarantees that the final estimator is conservative regardless of S, as N increases. The MSE will compare variably with that of the uncorrected Chao estimator. When the original Chao estimator is not conservative, the MSE of the corrected version can be expected to be lower; but clearly the reverse might apply as soon as indeed $\hat{N}_C \leq N$.

As further work, we mention extensions of our reasoning to the case of lower-bound estimation without replacement (Chao and Lin, 2012) or using higher-order frequency counts (Chiu et al., 2014). On the other hand, we expect the correction factor to extend to the case of covariates (Böhning et al., 2013; Farcomeni, 2018) directly.

References

Abeni, D.A., Brancato, G., Perucci, C.A., 1994. Capture-recapture to estimate the size of the population with Human Immunodeficiency Virus type 1 infection. Epidemiology 5, 410–414.

Baillargeon, S., Rivest, L.P., 2007. Rcapture: loglinear models for capture-recapture in R. J. Stat. Softw. 19 (5).

Bargagli, A.M., Colais, P., Agabiti, N., Mayer, F., Buttari, F., Centonze, D., Di Folco, M., Filippini, G., Francia, A., Galgani, S., Gasperini, C., Giulini, M., Mirabella, M., Nociti, V., Pozzilli, C., Davoli, M., 2016. Prevalence of multiple sclerosis in the Lazio region, Italy: use of an algorithm based on health information systems. J. Neurol. 263, 751–759.

Böhning, D., 2010. Some general comparative points on Chao's and Zelterman's estimators of the population size. Scand. J. Stat. 37, 221–236.

Böhning, D., 2015. Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. Metron 73, 201–216.

Böhning, D., van der Heijden, P.G.M., 2009. A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. Ann. Appl. Stat. 3, 595–610.

Böhning, D., Kaskasamkul, P., van der Heijden, P., 2019. A modification of Chao's lower bound estimator in the case of one-inflation. Metrika 82, 361–384.

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

Chao, A., 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783–791.

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

Chao, A., Colwell, R.K., 2017. Thirty years of progeny from Chao's inequality: estimating and comparing richness with incidence data and incomplete sampling. Statist. Oper. Res. Trans. 41, 3–54.

Chao, A., Lin, C.-W., 2012. A nonparametric lower bound for species richness and shared species richness under sampling without replacement. Biometrics 68, 912–921.

Chiu, C.H., Wang, Y.T., Walther, B.A., Chao, A., 2014. An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. Biometrics 70, 671–682.

Dotto, F., Farcomeni, A., 2018. A generalized Chao estimator with measurement error and external information. Environ. Ecol. Stat. 25, 53–69.

Edwards, W.R., Eberhardt, L., 1967. Estimating cottontail abundance from livetrapping data. J. Wildl. Manag. 87–96.

Farcomeni, A., 2016. A general class of recapture models based on the conditional capture probabilities. Biometrics 72, 116–124.

Farcomeni, A., 2018. Fully general Chao and Zelterman estimators with application to a whale shark population. J. R. Statist. Soc. (Ser. C) 67, 217–229.

Farcomeni, A., 2020. Population size estimation with interval censored counts and external information: prevalence of multiple sclerosis in Rome. Biom. J. 62, 945–956.

Farcomeni, A., Cortese, A., Sgarlata, E., Alunni Fegatelli, D., Marfia, G.A., Buttari, F., Mirabella, M., De Fino, C., Prosperini, L., Pozzilli, C., Grasso, M.G., Iasevoli, L., Di Battista, G., Millefiorini, E., 2018. The prevalence of multiple sclerosis in the metropolitan area of Rome: a capture-recapture analysis. Neuroepidemiology 50, 105–110.

Lanumteang, K., Böhning, D., 2011. An extension of Chao's estimator of population size based on the first three capture frequency counts. Comput. Statist. Data Anal. 55, 2302–2311.

Mao, C.X., 2006. Inference on the number of species through geometric lower bounds. J. Amer. Statist. Assoc. 101, 1663–1670.

Mao, C.X., Lindsay, B.G., 2007. Estimating the number of classes. Ann. Statist. 35, 917–930.

Puig, P., Kokonendji, C.C., 2018. Non-parametric estimation of the number of zeros in truncated count distributions. Scand. J. Stat. 45, 347–365.

Rivest, L.P., Baillargeon, S., 2007. Applications and extensions of Chao's moment estimator for the size of a closed population. Biometrics 63, 999–1006.

Wang, X., He, C.Z., Sun, D., 2005. Bayesian inference on the patient population size given list mismatches. Stat. Med. 24, 249–267.