Noname manuscript No.
(will be inserted by the editor)

Chao's Lower Bound Estimator and the Size of the Pleiades

Letter to the Editor

Dankmar Böhning

Received: date / Accepted: date

Abstract In this note we would like to point out that the lower bound estimator of the frequency of hidden units in a target population, developed by Chao in ecology, was developed independently in astro-physics and has been used to estimate the size of flare stars in the Pleiades.

Keywords Chao's lower bound estimator · estimating the number of flare stars · dark number estimation in astro-physics

One of the most popular and used estimators of population size using a capture-recapture approach is the estimator suggested by Chao [5], [6], [7]. Shortly after the appearance of two recent books on capture-recapture methods by McCrea and Morgan [11] and Böhning, van der Heijden and Bunge [3] an email by Dr Ashot Akopian [1] reached me referring to the two books and mentioning the following:

> I am astronomer and very interested in "capture-recapture methods". I hope that it will be interesting for you to know that such methods are used in astronomy since 1968. At this year the famous astronomer Ambartsumian had suggested and applied an estimator, which now is known as Chao's estimator. In 1970, Ambartsumian (Astrophysics, 1970, Volume 6, Issue 1, pp.1-10) had proved that this estimator gives only a lower bound. Unfortunately this fact was missed in your books.

I think that this fact is worth mentioning to a wider community in ecology, social and life sciences, public health and software engineering, criminology and text mining (and wherever the estimator of Chao is used). In the original

D. Böhning
Southampton Statistical Sciences Research Institute, University of Southampton, SO17 1BJ, UK
Tel.: +123-45-678910
E-mail: d.a.bohning@soton.ac.uk

development, Chao uses the Poisson distribution $p(y;\theta) = \exp(-\theta)\theta^y/y!$ and exploit that $p(0;\theta) = \frac{1}{2}p(1;\theta)^2/p(2;\theta)$ leading to the estimate $\hat{f}_0 = \frac{1}{2}f_1^2/f_2$ where f_x denotes the frequency of units with x identifications. More importantly, if population heterogeneity arises which can be described by a Poisson mixture

$$p(y;F) = \int_\theta k(y;\theta)dF(\theta)$$

with some mixing distribution F and Poisson mixture kernel $k(y;\theta) = \exp(-\theta)$ $\theta^y/y!$, using the Cauchy-Schwarz inequality, we find that

$$p(0;F) \geq \frac{1}{2}p(1;F)^2/p(2;F)$$

which establishes that $\hat{f}_0$ is a lower bound estimator for the number of unobserved units. The estimator was then developed for a binomial mixture kernel (Chao 1989) and, in fact, can be easily extended to any Power series as mixture kernel $k(y;\theta) = a_y\theta^y/\eta(\theta)$ where $\eta(\theta) = \sum_{y=0}^\infty a_y\theta^y$ (Böhning *et al.* 2019).

Ambartsumyan *et al.* [2] raised a problem in astrophysics, namely, how many flare stars exist in the region of the Pleiades? The Pleiades is a star cluster, 444 light years distance relatively close to planet earth. Some of its members are only visible occasionally for short periods as they burst out intense light for a brief period - a flare; these are the flare stars. For some flare stars they get identified only once, others more than once. At the time of the publication of the paper, 145 flare stars had been identified. Of these, 123 showed one flare, 16 showed two flares, and 6 showed more than two flares. The full distribution is given in Table 1. This repeated observational process establishes the analogy to capture-recapture data: the flare stars are the units that are 'recaptured' by observing their flares. Stars with flares are visible only during a short observational window. Most flare stars are observed only once and very few are seen several times. Clearly, not every flare star is observed and Ambartsumyan *et al.* (1970) asked the question: how many additional and hidden stars are out there? Indeed, Ambartsumyan *et al.* (1970) used also the Poisson distribution $p(y;\theta) = \exp(-\theta)\theta^y/y!$ and also exploited that $p(0;\theta) = \frac{1}{2}p(1;\theta)^2/p(2;\theta)$ leading to the estimate $\hat{f}_0 = \frac{1}{2}f_1^2/f_2$ where f_x denotes the frequency of stars with x flares in this case. They also argued, using the Cauchy-Schwarz inequality, that the estimator has a lower bound property. As mentioned above, this estimator was later developed independently and evidently without knowledge of the paper by Ambartsumyan *et al.* (1970) in ecology (Chao 1984, 1987, 1989) and it is now one of the most

Table 1 The distribution of flare counts per star in the Pleiades

Number of flares	0	1	2	3	4	5	6	9
Frequency		123	16	2	1	1	1	1

popular estimators in ecology (Chao and Colwell 2017). For the data of Table 1, Ambartsumyan *et al.* estimated the number of hidden flare stars as 473, at the time of the appearance of the paper. They also pointed out that the lower bound estimator will depend on the mixture kernel. So, for example, if a geometric distribution is used for the mixture kernel instead of the Poisson, the estimator changes to $\hat{f}_0 = f_1^2/f_2 = 946$. It is clear that the lower bound estimator will depend on the mixture kernel used. Hence it is advisable to compare distributions using model selection criteria. Ambartsumyan *et al.* also consider two component mixtures of Poisson distributions to get an estimate of F. However, in contrast to Chao(1978, 1989), Ambartsumyan *et al.* did not consider any uncertainty assessment such as standard errors of estimate od confidence intervals. The paper by Ambartsumyan *et al.* is remarkable to read and provides an interesting application of the capture-recapture approach in astro-physics, although it does not occur under the capture-recapture umbrella. It is not uncommon that scientific findings are developed independently by different research groups, in particular, if the application fields, like in this case, are far apart. This has been exemplified for the Petersen estimator in the area of two-sources capture-recapture methods by Goudie and Goudie (2007).

A feature of the data is the large number of ones in Table 1, which will have large impact on the estimator and could lead to a largely inflated estimate of the number of hidden flare stars if the spike at one is ignored. This point has recently attained some attention, see for example Godwin (2017) or Böhning *et al.* (2018, 2019).

References

1. Akopian, A. (2018). Personal communication.
2. Ambartsumyan,V. A., Mirzoyan,L. V., Parsamyan, E. S., Chavushyan, O.S., and Erastova, L. K. (1970). Flare stars in the Pleiades. *Astrofizika* **6**(1), 7–30.
3. Böhning, D., van der Heijden, P.G.M., and Bunge, J. (2018). *Capture-Recapture Methods for the Social and Medical Sciences.* Chapman & Hall/ CRC: Boca Raton.
4. Böhning, D., Kaskasamkul, P., and van der Heijden, P.G.M. (2019). A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika* **82**, 361-384.
5. Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**, 265–270.
6. Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–791.
7. Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics* **45**, 427–438.
8. Chao, A. and Colwell, R.K. (2017). Thirty years of progeny from Chaos inequality: Estimating and comparing richness with incidence data and incomplete sampling. *SORT* **43**, 3–54.
9. Godwin, R. (2017). One-inflation and unobserved heterogeneity in population size estimation. *Biometrical Journal* **59**, 79–93.
10. Goudie, I.B.J. and Goudie, M. (2007). Who captures the marks for the Petersen estimator? *Journal of the Royal Statistical Society* **A 170**, 825–839.
11. McCrea, R.S. and Morgan, B.J.T. (2015). *Analysis of Capture-Recapture Data.* Chapman&Hall/CRC: Boca Raton.