



RESEARCH ARTICLE

A Non-Parametric Estimation Method of the Population Size in Capture-Recapture Experiments With Right Censored Data

Anabel Blasco-Moreno¹  | Pedro Puig^{1,2} 

¹Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain | ²Centre de Recerca Matemàtica, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain

Correspondence: Anabel Blasco-Moreno (anabel.blasco@uab.cat)

Received: 20 August 2024 | **Revised:** 27 February 2025 | **Accepted:** 20 March 2025

Funding: This work was supported by Agencia Estatal de Investigación and Ministerio de Ciencia, Tecnología e Innovación, Grant/Award Number: PID2022-137414OB-I00; Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D, Grant/Award Number: CEX2020-001084-M.

Keywords: abundance data | censored data | Chao's estimator | incidence data | mixed-binomial distributions | mixed-Poisson distributions

ABSTRACT

We present a new non-parametric approach for estimating the total number of animals or species when we only have information on the number of animals or species that have been observed once, twice, ..., and the number of animals or species that have been observed r and more than r times. The approach, like the Chao estimator, gives a lower bound on population size while also providing bootstrap confidence intervals. We conducted simulations to compare our estimator to other competing ones in special scenarios with $r = 2$ and 3 and found that it performed quite well. In the case of uncensored samples, we analyze which censoring point is preferable in specific examples, as well as when censoring at $r = 3$ is superior to censoring at $r = 2$.

1 | Introduction

Capture-recapture methods are commonly used in ecology to estimate animal population sizes and species richness, see for instance McCrea and Morgan (2014) and the references therein. These methods have become popular, not only in ecology but also in social and medical sciences, to estimate the size of elusive populations such as illegal immigrants, illicit drug users, or people having a drinking problem (Böhning et al. 2018).

The aim of capture-recapture models is to provide an estimation of the population size N or, equivalently, of the frequency of unobserved individuals. Let $f_1, f_2, f_3, \dots, f_m$ be the frequencies of distinct individuals (or species) identified exactly $1, 2, 3, \dots, m$ times during the study period, after T trapping occasions, and f_0 be the frequency of individuals who were never found during the

study period and are therefore not being observed. As a result, the population size N can be expressed as $N = f_0 + f_1 + f_2 + \dots + f_m = f_0 + n$, where $n = \sum_{i=1}^m f_i$ is the overall number of distinct individuals seen. This sampling pattern is called *incidence data*, because it is based on the incidence of the individuals or species (detection/non-detection) in T sampling units. When only one sample of individuals is considered, this is referred to as *abundance data*. In practice, it usually indicates that T is very large or unknown.

To estimate the population size N , several non-parametric approaches have been developed in the literature. Chao (1984, 1987) estimator, Zelterman (1988), Burnham and Overton (1978), first- and second-order jackknife estimators are some of the most well-known methods. Many extensions, bias corrections, and modifications to these estimators have been detailed by

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Environmetrics* published by John Wiley & Sons Ltd.

Chao (1987), Lanumteang and Böhning (2011), and Chiu et al. (2014). Puig and Kokonendji (2018) and Jiménez-Gamero and Puig (2020) developed some non-parametric methods for abundance data for a wide range of count distributions known as the LC class.

The main goal of the present research is the development of a new non-parametric approach for estimating the total number of animals or species when we only have information on the number of animals or species that have been observed once (f_1), twice (f_2), ..., and the number of animals or species that have been observed r and more than r times (f_{r+}). If this happens, the frequency distribution of the observations is said to be right censored at r . A motivating example is the study of the number of different fish species in the coral reef of Tunku Abdul Rahman Marine Park presented by Chao et al. (2017). The data was collected in the framework of a citizen project from $T = 116$ scuba diving and snorkelling sessions. Due to the data collection procedure, only the number of unique species, $f_1 = 101$, and the number of species seen more than one (*super-doubletons*), $f_{2+} = 340$, were recorded. Therefore, the number of different fish species observed was $n = f_1 + f_{2+} = 441$. In this case, the frequency distribution of the observations is right censored at $r = 2$. This pattern of censorship is frequent in other citizen science experiences such as the Participatory Guide of the Marine Species in the Barcelona Metropolitan Area project (URBAMAR) (Fraisl et al. 2022).

This type of model has appeal outside of ecology. Press news is a common source for this kind of censored data. For instance, one in ten convicted sex offenders in North Wales were repeat offenders, according to the following information that was reported in the Daily Post on August 13, 2018: "The 34 repeat offenders convicted since 2011 represent almost 10% of the total of 342 convictions for sexual assault". In this case, $n = 342$, $f_{2+} = 34$ and $f_1 = 308$. The methods that will be discussed in this paper will make it possible to estimate the number of not detected (not reported) offenders who may be found convicted.

Even though the researcher will typically have uncensored data, the data appears to be censored in some papers because of the information provided by the authors. This is typically the case when authors employ jackknife-2 and the traditional Chao estimator, which requires just T , n , f_1 , and f_2 , that is, $r = 3$. In fact, several non-parametric estimators are suitable for censored data at specific r values. This is the case with the estimator developed by Lanumteang and Böhning (2011), which uses the first three f_i , or $r = 4$, and the estimator developed by Chiu et al. (2014), which uses the first four f_i ($r = 5$).

2 | The Statistical Model

Let Y_i be the random variable that represents the frequency with which the species or animal i has been observed in T samples or trapping occasions, $T \geq 1$. The probability of being observed k times is provided by the binomial distribution, assuming that the different samples are independent of one another, $P(Y_i = k | \lambda_i) = \binom{T}{k} \lambda_i^k (1 - \lambda_i)^{T-k}$, $i = 1, 2, \dots, N$, where λ_i denotes the probabil-

ity that the i th animal or species will be captured or observed, and therefore $0 < \lambda_i < 1$. We assume that there is only individual variation and no sample dependence. The probability of catchability or identifiability λ_i varies across target population members in most application studies. Heterogeneity is the term used in this situation. As a result, we assume that the λ_i are taken from a probability distribution with density function $f(\lambda)$, for λ belonging to the interval $(0, 1)$. Suppose the random variable Y counts the number of times in which any animal or species is observed. Therefore, the probability of being observed exactly k times is given by a mixed-binomial distribution, that is, $p_k = P(Y = k) = \int_0^1 \binom{T}{k} \lambda^k (1 - \lambda)^{T-k} f(\lambda) d\lambda$, for $k = 0, 1, 2, \dots, T$. The family of the so-called mixed-binomial distributions consists of the set of distributions produced by the various densities $f(\lambda)$. The density $f(\lambda)$ is referred to as the mixing density. The mixing distribution can be continuous, discrete, or have positive probability at a finite number of points, resulting in a finite mixture of binomial distributions.

We have to estimate f_0 , or alternatively, p_0 , in order to estimate the population size N . It is achievable if the mixing density $f(\lambda)$ is known, resulting in a parametric model, and utilizing maximum likelihood methods or other traditional statistical techniques. Morgan and Ridout (2008), for example, analyzed numerous models, including the mixing of two binomial distributions, considering $f(\lambda)$ as the density of a beta distribution (model also studied by Chiu (2022)), and the density of a logistic-normal distribution. The chosen density, as well as the estimation method, might result in significant discrepancies in the population size estimations (Schofield et al. 2023).

However, if $f(\lambda)$ is unknown, p_0 cannot be estimated because it is not identifiable (Link 2003; Holzmann et al. 2006; Aleshin-Guendel et al. 2024). Nevertheless, if we restrict to mixed-binomial distributions, then certain lower bounds on p_0 can be found independent of the knowledge of $f(\lambda)$. Chao inequality (Chao 1984, 1987) is a well-known example of this: $p_0 \geq (T - 1)p_1^2 / (2Tp_2)$. The following results will be applied to present a new bound of p_0 that will be helpful for our purposes:

Theorem 1. *Let Y be a random variable mixed-binomial distributed such that, $p_i = P(Y = i)$ and cumulative distribution function $F(k) = P(Y \leq k) = \sum_{i=0}^k p_i$, for $k = 0, 1, \dots, T$, and $T \geq 1$. Then,*

$$F(k) \leq \sum_{i=0}^k \binom{T}{i} (1 - p_0^{1/T})^i p_0^{1-i/T}, \quad k = 0, 1, 2, \dots, T \quad (1)$$

The proof establishing this result is contained in Appendix A.

The species abundance distribution is commonly modeled as Mixed-Poisson distributions. Let Y be the random variable counting the number of cases in which any animal or species is observed. Thus, Y is mixed-Poisson distributed if its probabilities can be described as, $p_k = P(Y = k) = \int_0^\infty \frac{e^{-\lambda} \lambda^k}{k!} f(\lambda) d\lambda$, for a given mixed density $f(\lambda)$. The following result is similar to Theorem 1.

Theorem 2. *Let Y be a random variable mixed-Poisson distributed such that, $p_i = P(Y = i)$ and cumulative distribution function $F(k) = P(Y \leq k) = \sum_{i=0}^k p_i$, for $k = 0, 1, 2, \dots$. Then,*

$$F(k) \leq p_0 \sum_{i=0}^k \frac{(-\log(p_0))^i}{i!}, \quad k = 0, 1, 2, \dots \quad (2)$$

The proof of this theorem is also contained in Appendix A.

Remark 1. Note that the right part of the inequality (1) corresponds to the distribution function of a binomial with probability of success $q = 1 - p_0^{1/T}$, and the right part of the inequality (2) corresponds to the distribution function of a Poisson with $\lambda = -\log(p_0)$. Then, Theorem 1 (or 2) means, in other words, that the distribution function of any mixed-binomial (or mixed-Poisson) random variable is always upper bounded by the distribution function of a binomial (or Poisson) with the same proportion of zeros.

It is important to point out that when we observe the frequencies f_1, f_2, \dots, f_{r+} , we really have a sample of size n of the random variable Y truncated at zero, because the zero is not observable. It can be expressed by means of the random variable $Y_{zi} = Y|Y > 0$, whose probabilities are $P(Y_{zi} = k) = p_k/(1 - p_0)$, for $k = 1, 2, \dots, T$. Moreover, $F_{zi}(k) = P(Y_{zi} \leq k) = (F(k) - p_0)/(1 - p_0)$. Consequently, from (1), we can obtain the following inequality for $F_{zi}(k)$:

$$F_{zi}(k) \leq \frac{\sum_{i=1}^k \binom{T}{i} (1 - p_0^{1/T})^i p_0^{1-i/T}}{1 - p_0}, \quad k = 1, 2, \dots, T \quad (3)$$

When T is very large, that is T tends to infinity, inequality (3) remains,

$$F_{zi}(k) \leq \frac{p_0}{1 - p_0} \sum_{i=1}^k \frac{(-\log(p_0))^i}{i!}, \quad k = 1, 2, \dots \quad (4)$$

The right part of (3) or (4) is an increasing function of p_0 , for each value of k , that it will be referred to, as $G_k(p_0)$, as it is stated in the following lemma.

Lemma 1. Let $G_k(p_0)$ be defined as the right part of (3) or (4). Then, for both cases $G_k(p_0)$ is an increasing function of p_0 for each value of k .

The proof (see Appendix A) is a consequence of the fact that likelihood-ratio ordering implies stochastic ordering (see Theorem 1.C.1. in Shaked and Shanthikumar (2007)).

Therefore, $G_k^{-1}(p_0)$ is also an increasing function and it leads to our main inequality:

$$p_0 \geq G_k^{-1}(F_{zi}(k)), \quad k = 1, 2, \dots, T$$

where $G_k^{-1}()$ can be numerically calculated. Note that we have an inequality for each value of k and, therefore, a sharper lower bound of p_0 can be obtained taking the maximum of all them, that is,

$$p_0 \geq \max_{1 \leq k \leq T} \{G_k^{-1}(F_{zi}(k))\} \quad (5)$$

2.1 | A Lower Bound Estimator of the Population Size N

Given the observed frequencies f_1, f_2, \dots, f_{r+} , we are able to estimate $F_{zi}(k)$, for $k = 1, 2, \dots, r - 1$, using the empirical distribution function, that is,

$$\hat{F}_{zi}(k) = \frac{\sum_{i=1}^k f_i}{n}$$

Therefore, combining this estimate with inequality (5), we obtain the following lower bound estimator of p_0 :

$$\hat{p}_0 = \max_{1 \leq k \leq r-1} \{G_k^{-1}(\hat{F}_{zi}(k))\} \quad (6)$$

For estimating p_0 , function $G_k^{-1}()$ has to be numerically evaluated. N can be estimated through the following argument: $1 - p_0 \approx n/N$, leading to the so-called Horvitz–Thompson estimator,

$$\hat{N} = \frac{n}{1 - \hat{p}_0} \quad (7)$$

The precision with which the distribution function $F_{zi}(k)$ is estimated determines the variability of the estimator \hat{p}_0 given in (6). The variance of the empirical distribution function is $V(\hat{F}_{zi}(k)) = F_{zi}(k)(1 - F_{zi}(k))/n$. A direct application of delta method shows that,

$$V(G_k^{-1}(\hat{F}_{zi}(k))) \approx \frac{F_{zi}(k)(1 - F_{zi}(k))}{n[G'_k(G_k^{-1}(F_{zi}(k)))]^2} \approx \frac{\hat{F}_{zi}(k)(1 - \hat{F}_{zi}(k))}{n[G'_k(\hat{p}_{0k})]^2} \quad (8)$$

where $\hat{p}_{0k} = G_k^{-1}(\hat{F}_{zi}(k))$, and $G'_k()$ indicates the first derivative of $G_k()$. Since $G'_k(x)$ tends to 0 when x tends to 1 for $k \geq 2$, the variance of \hat{p}_{0k} could be large when \hat{p}_{0k} is close to 1. The calculation of the variance of $\hat{p}_0 = \max_{1 \leq k \leq r-1} \hat{p}_{0k}$ is very complicated, so expression (8) will only be useful for the case $r = 2$.

In order to estimate the variance of \hat{N} and to compute a confidence interval, we suggest using the following non-parametric bootstrap method:

1. From the original sample, p_0 is estimated from (6), the population size is estimated by \hat{N} (expression (7)), and the frequency of zeros is estimated by $\hat{f}_0 = \hat{N} - n$.
2. A sample of size \hat{N} is generated from a multinomial distribution with $r + 1$ categories and cell probabilities $(\frac{\hat{f}_0}{\hat{N}}, \frac{f_1}{\hat{N}}, \dots, \frac{f_{r+}}{\hat{N}})$, obtaining $f_0^b, f_1^b, \dots, f_{r+}^b$, such that $f_0^b + f_1^b + \dots + f_{r+}^b = \hat{N}$.
3. With the zero-truncated bootstrap sample (f_1^b, \dots, f_{r+}^b) , the proportion of zeros is estimated again using (6), say \hat{p}_0^b , and the population size using (7), say \hat{N}^b , with a number of observed individuals $n_b = f_1^b + \dots + f_{r+}^b$.
4. Steps (2) and (3) are repeated 10,000 times obtaining $\hat{N}_1^b, \hat{N}_2^b, \dots, \hat{N}_{10,000}^b$. If one of the \hat{N}^b values is lower than n , the value is taken as n .

- The statistics of interest (mean and standard deviation) are computed from the 10,000 values of \hat{N}^b , and the limits of the 95% confidence interval are determined from the 2.5th and 97.5th percentiles.

The Supporting Information includes an R script for calculating \hat{p}_0 , \hat{N} , and a confidence interval for incidence and abundance data (where T is very large or unknown).

3 | Some Common Censoring Patterns

In this section, we will look at the most common censoring patterns observed in practice or in the literature, and also provide some application examples. They are represented by $r = 2$ and $r = 3$.

3.1 | $r=2$: Seeing One and More Than One

This is the case analyzed by Chao et al. (2017). When $r = 2$, only $F_{zt}(k)$ can be estimated for $k = 1$. From expression (3), we find

$$F_{zt}(1) \leq \frac{T(1 - p_0^{1/T})p_0^{1-1/T}}{1 - p_0} = G_1(p_0)$$

Using the empirical distribution function, we obtain $\hat{F}_{zt}(1) = f_1/n$, and then the lower bound estimator of p_0 given by (6) is, $\hat{p}_0 = G_1^{-1}(f_1/n)$. Therefore, the value of \hat{p}_0 can be obtained solving numerically the equation,

$$\frac{T(1 - p_0^{1/T})p_0^{1-1/T}}{1 - p_0} = \frac{f_1}{n} \quad (9)$$

or solving the following equation in case T is large and, most of the time, unknown (abundance data):

$$\frac{-\log(p_0)p_0}{1 - p_0} = \frac{f_1}{n} \quad (10)$$

Figure 1 shows the solution of Equation (10) for different values of f_1/n .

The variance of \hat{p}_0 can be estimated using (8), obtaining for incidence and abundance data, respectively,

$$V(\hat{p}_0) \approx \frac{\frac{f_1}{n}(1 - \frac{f_1}{n})(1 - \hat{p}_0)^4 \hat{p}_0^{2/T}}{n(1 - \hat{p}_0 - T(1 - \hat{p}_0^{1/T}))^2},$$

$$V(\hat{p}_0) \approx \frac{\frac{f_1}{n}(1 - \frac{f_1}{n})(1 - \hat{p}_0)^4}{n(1 - \hat{p}_0 + \log(\hat{p}_0))^2}. \quad (11)$$

There is also a classical non-parametric estimator that uses only n and f_1 that could be used in this scenario: the estimator jackknife-1 given by Burnham and Overton (1978): $\hat{N}_{Jk1} = n + (T - 1)f_1/T$. Chao et al. (2017) introduced an estimator based in the Good-Turing theory, denoted as \hat{N}_{ChQ2} , consisting in estimating the number of non-observed doubletons and to apply the classical Chao estimator afterwards. The performance of our estimator \hat{N} is compared with that of \hat{N}_{Jk1} and \hat{N}_{ChQ2} in Figures 2 and 3, respectively, simulating scenarios for six

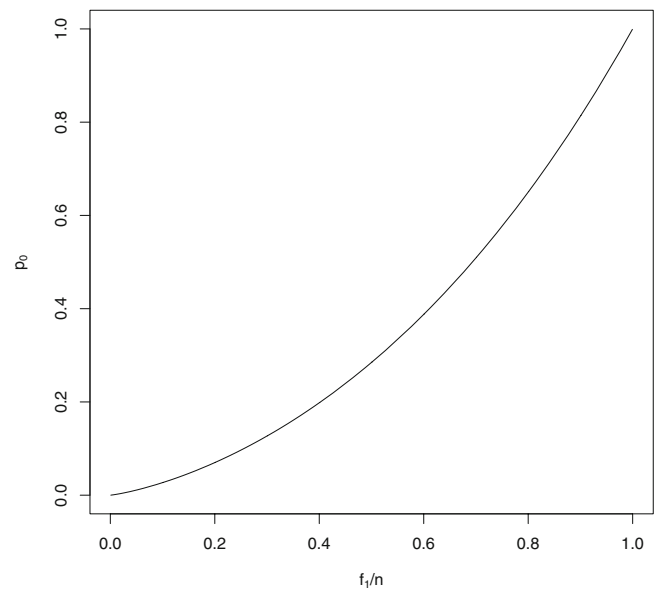


FIGURE 1 | Estimated lower-bounds of p_0 for different values of f_1/n .

alternative detection probability distributions. Because the Chao estimator (Chao 1987), $\hat{N}_{Ch} = n + (T - 1)f_1^2/(2Tf_2)$, is a common benchmark for comparison, we have also included it, even though it is inappropriate for $r = 2$ because f_2 is unknown. For each distribution, we have simulated 10,000 Binomial(T, λ_i) samples, for different trapping occasions T , and $N = 200, 500$ (the true population size).

The performance of each estimator has been evaluated computing the mean-square-error, $MSE = \sum (N - \hat{N})^2 / 10,000$. In Tables S1 and S2 of the Supporting Information, the lowest MSE value for each case has been highlighted in bold. As can be observed, for all the detection probability distributions, our estimator performs better for all trapping occasions explored: a constant detection probability equal to 0.3, a 25% mixture of the four detection probabilities (0.3, 0.4, 0.5, 0.6), a 50% mixture of the two detection probabilities (0.2, 0.4), and for the detection probabilities based on a lognormal($\mu = 0, \sigma = 0.3$), gamma($k = 12, \theta = 1$), and beta($\alpha = 8, \beta = 12$) distributions. Notice how the SE and MSE for all estimators decreases as T increases.

3.1.1 | Examples of Application

- Coral reef fish species.** Consider the data mentioned in the Introduction, about the number of different coral reef fish species presented by Chao et al. (2017). We have that $r = 2$, $T = 116$, $f_1 = 101$ and $n = 441$. Solving Equation (9) we find that $\hat{p}_0 = 0.0837$ and $\hat{N} = 481.3 \sim 481$, and the 95% bootstrap-CI is $(464.3, 498.6) \sim (464, 499)$. An approximated 95% CI for p_0 can also be computed using expression (11), as $p_0 = \hat{p}_0 \pm z_{0.95} \sqrt{V(\hat{p}_0)}$. When (7) is applied to both interval extremes of p_0 , another estimated 95% CI for N resulted, $(471, 493)$, which is quite comparable to the one obtained using bootstrap. The estimation provided by Chao et al. (2017) was $\hat{N}_{ChQ2} \sim 542$, which is clearly a higher value. It is important to note that \hat{N} estimates a lower bound

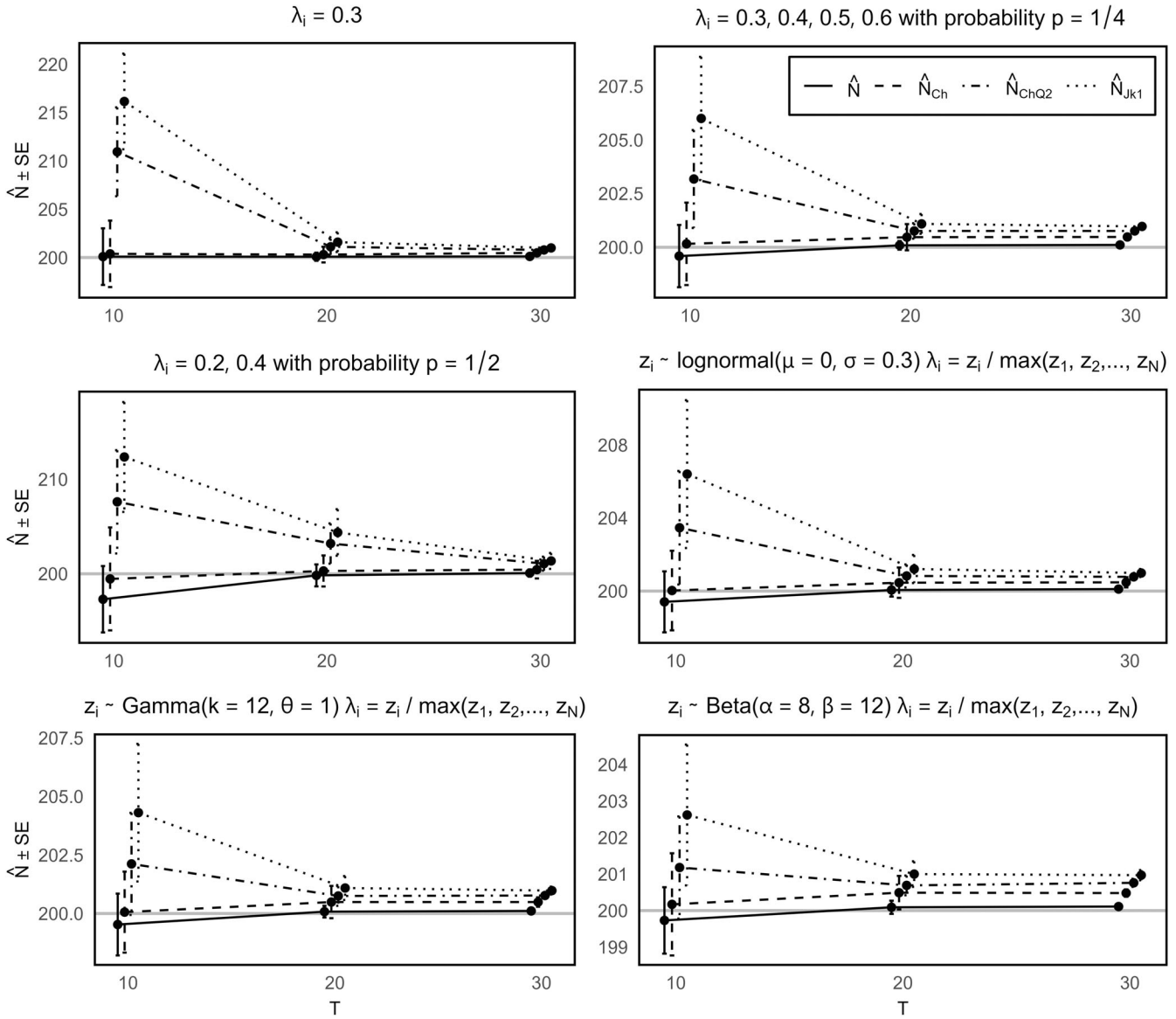


FIGURE 2 | Average values and SE for \hat{N} , \hat{N}_{ChQ2} , \hat{N}_{Ch} , and \hat{N}_{Jk1} , for 10,000 Monte Carlo simulations of six different distributions of the detection probabilities with different trapping occasions $T = 10, 20, 30$. The true population size is $N = 200, r = 2$.

on the population size, whereas \hat{N}_{ChQ2} does not because it is based on the Chao estimator, but the procedure in general overestimates the unobserved value of f_2 .

2. *Daily Post information.* The other example mentioned in the Introduction reports $f_1 = 308, n = 342$ (number of convictions for sexual assault), $r = 2$, and T is unknown and can be though very large. Therefore, solving equation (10) we find that $\hat{p}_0 = 0.8139$ and $\hat{N} = 1837.9 \sim 1838$, and the 95% bootstrap-CI is $(1381.4, 2628.7) \sim (1381, 2629)$. When applying (7) to the CI of p_0 derived from $V(\hat{p}_0)$ in (11), the approximated 95% CI in this case is $(1416, 2617)$.

3.2 | r=3: Seeing One, Two, and More Than Two

When $r = 3$, $F_{zi}(k)$ can be estimated for $k = 1$ and $k = 2$. From expression (3), we find

$$G_1(p_0) = \frac{T(1 - p_0^{1/T})p_0^{1-1/T}}{1 - p_0},$$

$$G_2(p_0) = G_1(p_0) + \frac{T(T-1)(1 - p_0^{1/T})^2 p_0^{1-2/T}}{2(1 - p_0)}.$$

Using the empirical distribution function, we obtain $\hat{F}_{zi}(1) = f_1/n$ and $\hat{F}_{zi}(2) = (f_1 + f_2)/n$. Then, the lower bound estimator of p_0 given by (6) is, $\hat{p}_0 = \max(G_1^{-1}(f_1/n), G_2^{-1}((f_1 + f_2)/n))$. It means that we have to solve the following two equations,

$$\frac{T(1 - p_0^{1/T})p_0^{1-1/T}}{1 - p_0} = \frac{f_1}{n},$$

$$\frac{T(1 - p_0^{1/T})p_0^{1-1/T}}{1 - p_0} + \frac{T(T-1)(1 - p_0^{1/T})^2 p_0^{1-2/T}}{2(1 - p_0)} = \frac{f_1 + f_2}{n}, \quad (12)$$

and \hat{p}_0 is the largest of both solutions. In this censoring pattern, two classical estimators are also suitable because they require just

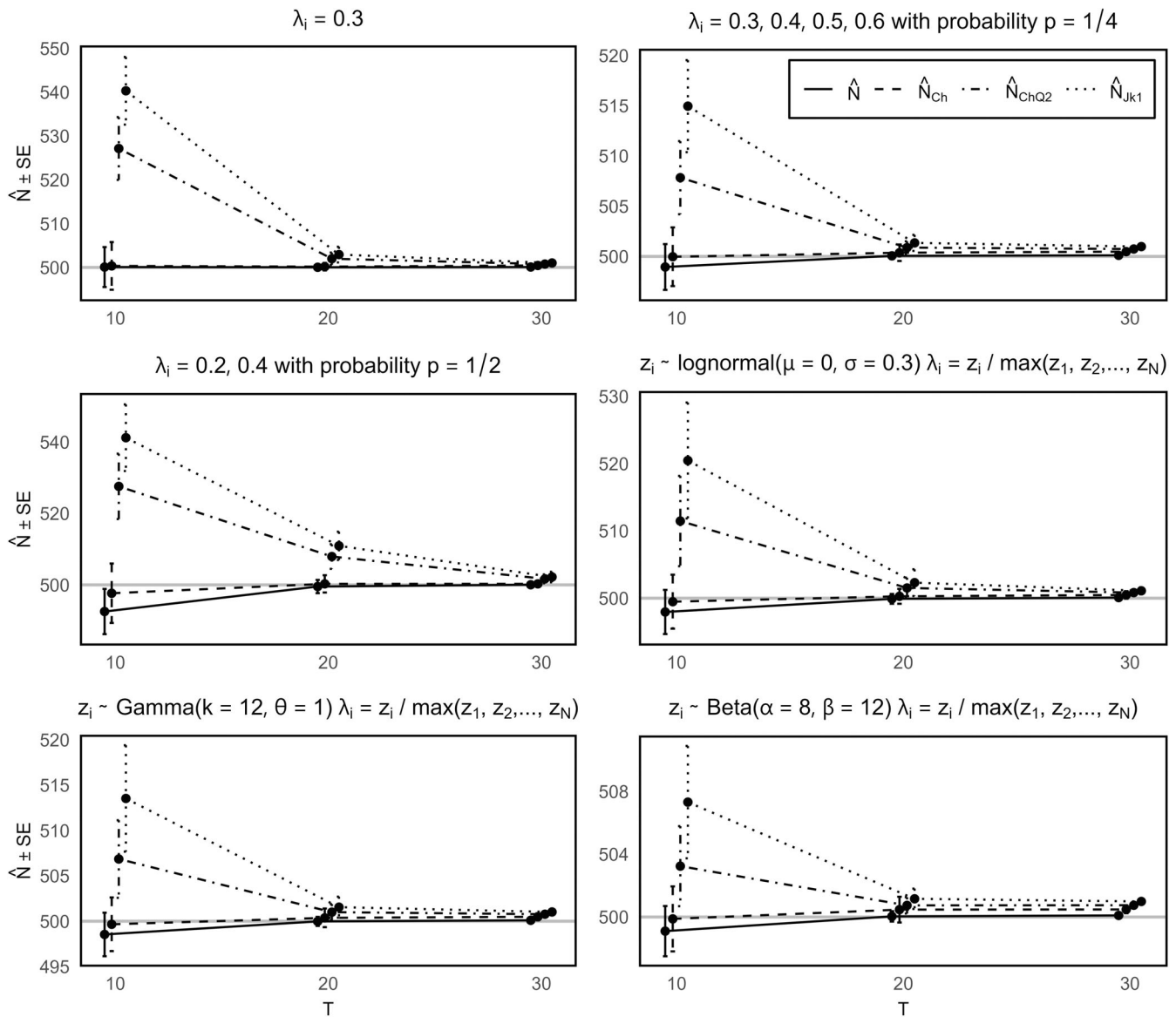


FIGURE 3 | Average values and SE for \hat{N} , \hat{N}_{ChQ2} , \hat{N}_{Ch} , and \hat{N}_{Jk1} , for 10,000 Monte Carlo simulations of six different distributions of the detection probabilities with different trapping occasions $T = 10, 20, 30$. The true population size is $N = 500$, $r = 2$.

T , n , f_1 , and f_2 . These are the previously mentioned Chao estimator, \hat{N}_{Ch} , and jackknife-2 (Burnham and Overton 1978), $\hat{N}_{Jk2} = n + (2T - 3)f_1/T + (T - 2)^2 f_2/(T(T - 1))$. Figures 4 and 5 compare the performance of our estimator \hat{N} to that of \hat{N}_{Jk2} and \hat{N}_{Ch} . More detailed information is shown in Tables S3 and S4, where the lowest MSE value for each case has been highlighted in bold. We considered the same six alternative detection probability distributions that we used in Figures 2 and 3, and the true population size is $N = 200, 500$ as well.

Similarly to the situation with $r = 2$, for all detection probability distributions and T values, our estimator outperforms the others. However, in many cases, \hat{N}_{Ch} is completely comparable to our estimator, with very similar MSE values. Moreover, as it can be seen in Table S4, for $N = 500$ we observe that \hat{N}_{Ch} is slightly better than the others for $T = 5$ for five of the considered detection probability distributions.

3.2.1 | Examples of Application

1. Budka et al. (2018) studied the overall macrophyte species richness in the catchment area of the lowland Wel river in northern Poland. They investigated various strategies for assessing sampling effort, including the classical Chao estimator. Although they most likely had a complete sample, the data presented in the article are censored because they only provide what is required to calculate the Chao estimate. In this example, $r = 3$, $T = 18$, $f_1 = 38$, $f_2 = 17$ and $n = 111$. Solving Equations (12) and taking the maximum, we find that $\hat{p}_0 = 0.1428$ and $\hat{N} = 129.5 \sim 130$, and the 95% bootstrap-CI is $(117.6, 143.3) \sim (118, 143)$. Budka et al. (2018) obtained $\hat{N}_{Ch} = 151.1$ and a 95% CI $(128.7, 201.7) \sim (129, 202)$. Although both CI overlap and the results are comparable, the 95% bootstrap-CI is narrower than that of the Chao estimator.

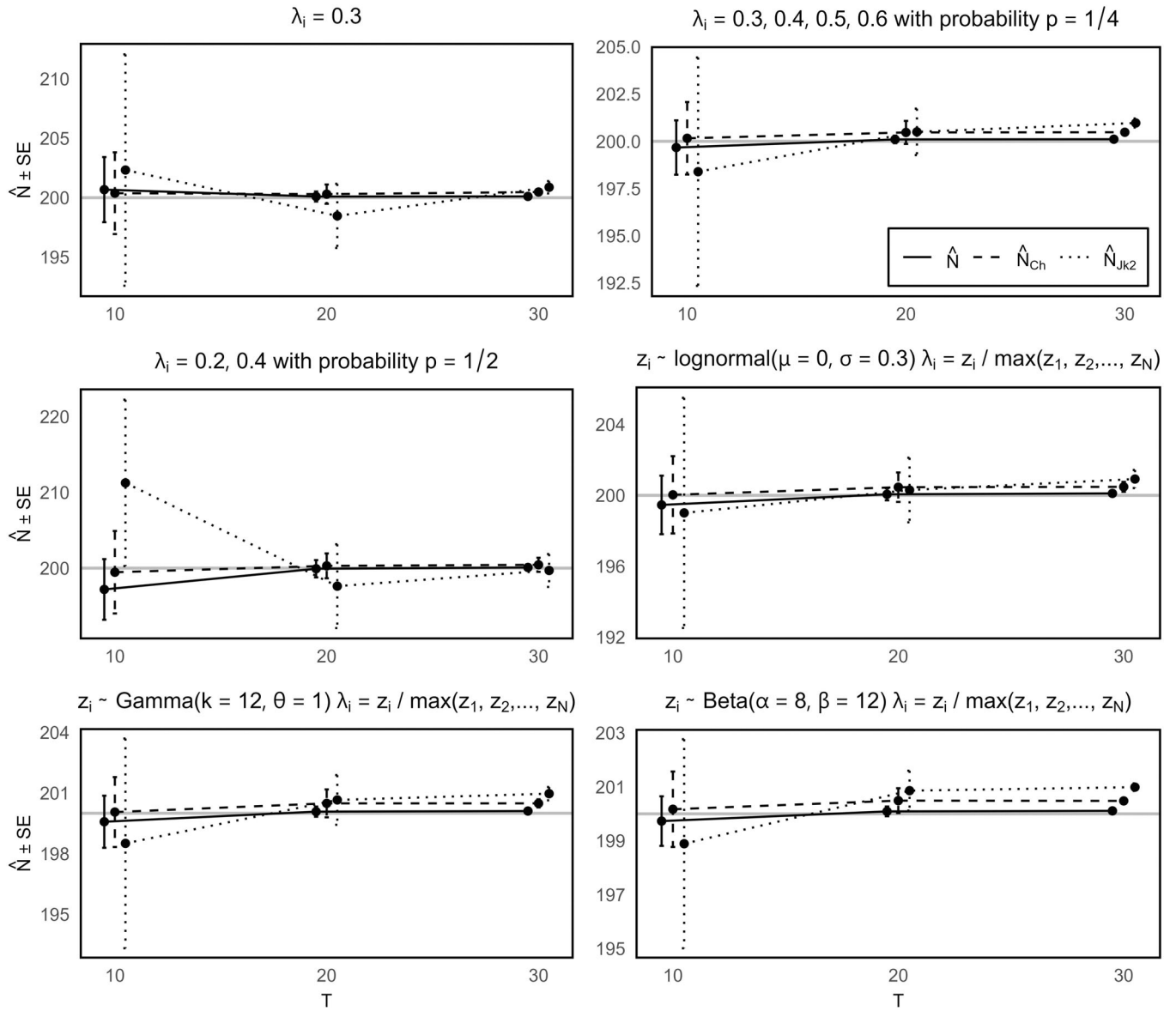


FIGURE 4 | Average values and SE for \hat{N} , \hat{N}_{Ch} , and \hat{N}_{Jk2} , for 10,000 Monte Carlo simulations of six different distributions of the detection probabilities with different trapping occasions $T = 10, 20, 30$. The true population size is $N = 200$, $r = 3$.

2. Cazzolla Gatti et al. (2022) estimated the total tree species richness at global, continental, and biome levels, based on the number of uniques (f_1) and duplicate (f_2) species. The authors adjusted the number of uniques using the observed triplicates (f_3) and quadruplicates (f_4), which are not supplied in the paper, and then applied the Chao estimator. In this example, $r = 3$, $T = 9353$, $f_1 = 24768$, $f_2 = 9426$ and $n = 64088$. Solving Equations (12) and taking the maximum, we find that $\hat{p}_0 = 0.1876$ and $\hat{N} = 78889.9 \sim 78890$, and the 95% bootstrap-CI is (78505.9, 79280.7) \sim (78506, 79281). Cazzolla Gatti et al. (2022) found that there are 73,274 tree species globally, among which 9186 tree species are yet to be discovered. However, our result shows that the number of species yet to be discovered is around 61% greater.

4 | To Censor or Not to Censor, That Is the Question

It is generally believed that the number of undetected species should be estimated mainly from data on the least frequent species (number of uniques and duplicates). This is supported by the simple reason that most common species (those found in a large number of sampling units) do not contain relevant information about undetected species; only rare/infrequent species do. Comparing the MSE of \hat{N} for $r = 2$ and $r = 3$ for the six distributions analyzed in Sections 3.1 and 3.2 (Tables S1 to S4), we can see that, depending on the distribution, the MSE can be improved when we add more information, that is, MSE is lower for $r = 3$ than for $r = 2$. For several of these distributions, the differences in MSEs are almost non-existent and even depend on whether

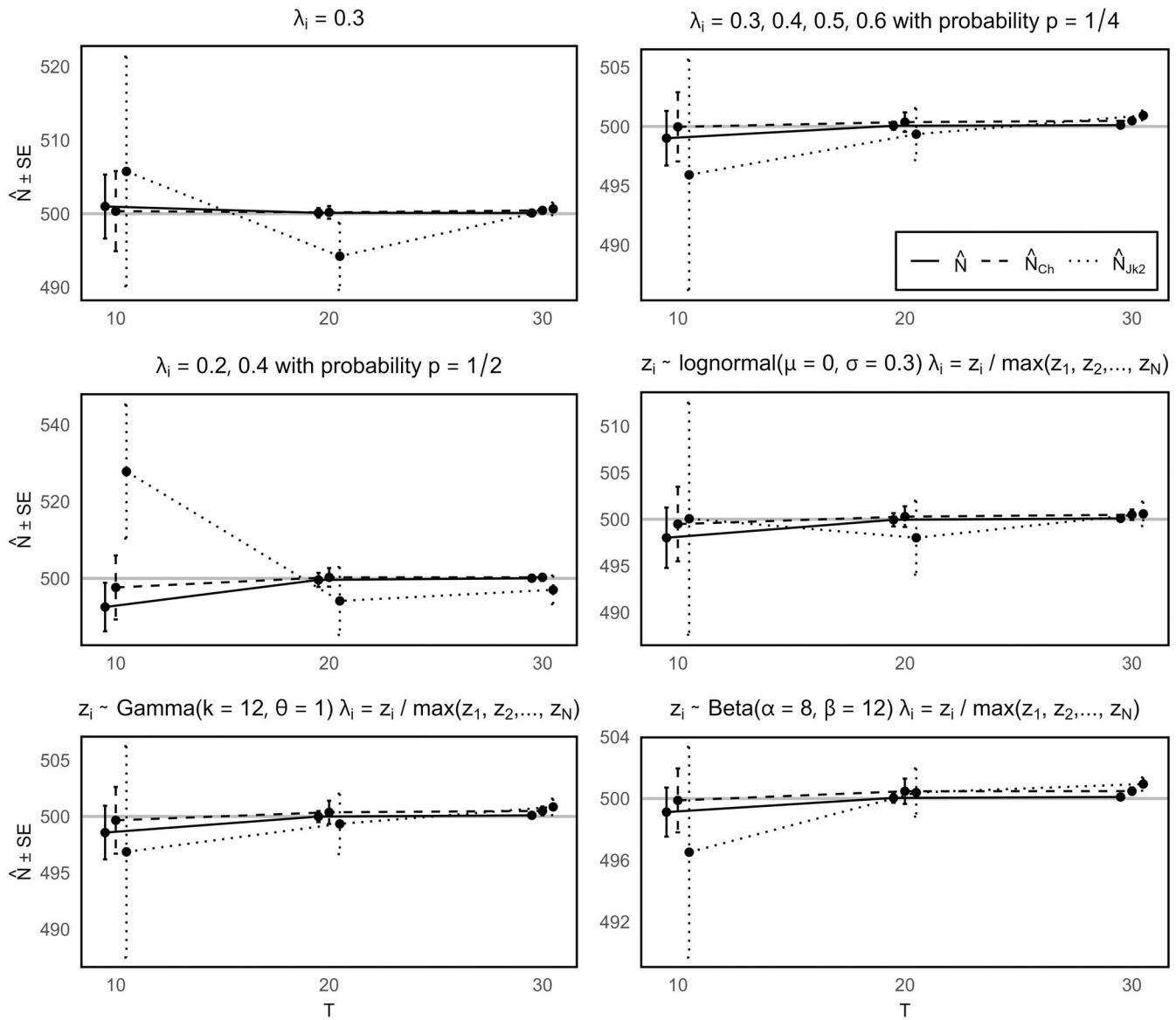


FIGURE 5 | Average values and SE for \hat{N} , \hat{N}_{Ch} , and \hat{N}_{Jk2} , for 10,000 Monte Carlo simulations of six different distributions of the detection probabilities with different trapping occasions $T = 10, 20, 30$. The true population size is $N = 500, r = 3$.

N is 200 or 500. There is a clear reduction, for example, for a 50% mixture of the two detection probabilities (0.2, 0.4). Tables S1 and S3 show that, for $T = 5$, we go from an MSE of 187.08 for $r = 2$ to an MSE of 170.14 for $r = 3$, indicating a 9.05% decrease. As the censor point rises, however, this reduction diminishes, and by simulating 10,000 further values for $r = 4$ and $r = 5$, we obtain MSEs of 164.63 and 163.55, respectively, representing a successive reduction of 3.24% and 0.66%.

As a result, even if we have a complete sample, is it better to keep only the data of the less frequent species and thus work as if we had a censored sample? Given a complete sample, what would be the best censoring point r ? These questions can be answered in practice on a case-by-case basis. A couple of examples of how this can be done are given below.

1. Consider the following toy example, where $T = 15$, $f_1 = 10$, $f_2 = 25$, $f_3 = 55$, $f_4 = 50$, $f_5 = 30$, $f_6 = 20$, $f_7 = 5$ and

$f_8 = 5$. Here, $n = \sum_{i=1}^8 f_i = 200$. Although this is a complete sample, we can analyze this example using our methodology assuming that this is a censored sample at $r = 8$, that is, $f_{8+} = 5$. The lower bound estimator of p_0 given by (6) is, $\hat{p}_0 = \max(G_1^{-1}(f_1/n), G_2^{-1}((f_1 + f_2)/n), \dots, G_7^{-1}((f_1 + f_2 + \dots + f_7)/n))$. Solving numerically the seven equations, we find that the maximum is attained at $G_6^{-1}((f_1 + f_2 + \dots + f_6)/n)$, providing $\hat{p}_0 = 0.0154$ and $\hat{N} = 203.1 \sim 203$. However, just because the maximum is attained at $G_6^{-1}(\hat{F}_{zi}(6))$, it means that the same result would be obtained considering $r = 7$, that is, $f_{7+} = 10$. We call this value the *effective censoring point* and this is provided by the R script included in the Supporting Information. But what happens if we censor at values lower than $r = 7$? Figure 6a shows the estimates \hat{N} obtained for different values of the censor point r . As can be seen, the same result is obtained by censoring at $r = 6$ rather than $r = 5$, as well as at $r = 3$ and $r = 2$. In this example, the variations in the estimate of N with the

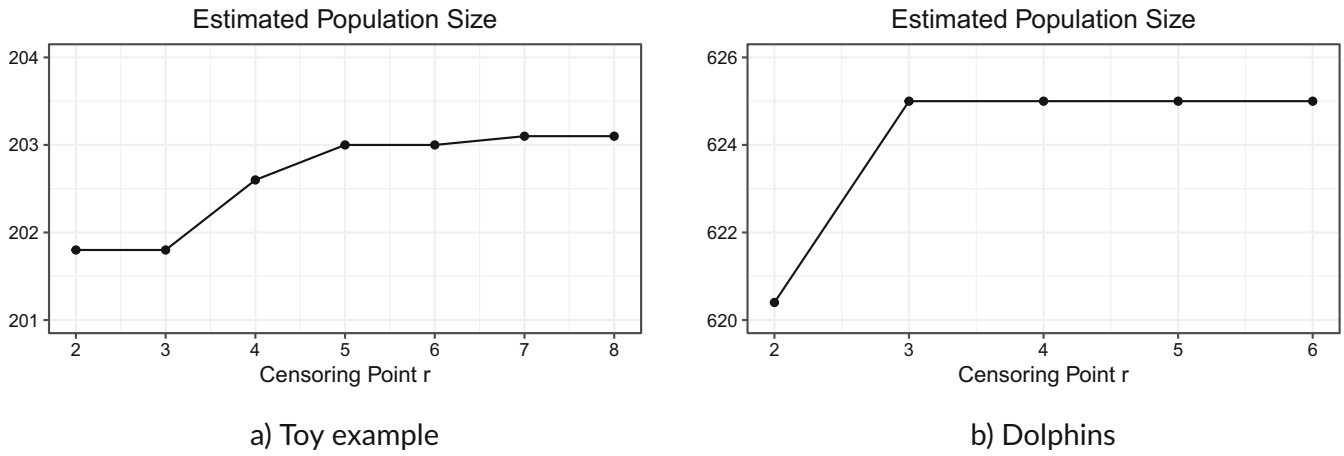


FIGURE 6 | Different values of \hat{N} according to the censor point r for the Toy example (a) and the Dolphins in Sepetiba bay (b).

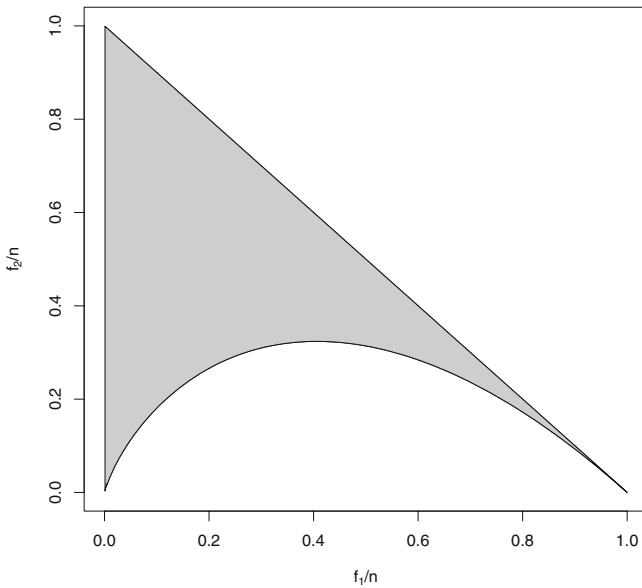


FIGURE 7 | Region (grey area) where censoring at $r = 3$ is better than censoring at $r = 2$ for abundance data.

various censoring schemes are small, confirming that the majority of the information is contained in the number of less frequent species. The effective censoring point (in this case, $r = 7$) is an inherent feature of the data. But what happens if we create similar samples using the non-parametric bootstrap method described in Section 2.1? Based on 10,000 bootstrap samples, we found that the effective censoring point was $r = 7$, occurring 43.1% of the time, followed by $r = 5$, occurring 28.3% of the time, suggesting that $r = 7$ is fairly robust.

- Nery and Simão (2012) studied the abundance of Guiana dolphins (*Sotalia guianensis*) in Sepetiba Bay, southeastern Brazil, using photo-identification, and it was also analyzed by Jiménez-Gamero and Puig (2020). This is a complete sample, where $f_1 = 228$, $f_2 = 110$, $f_3 = 19$, $f_4 = 18$, $f_5 = 6$, $f_6 = 1$, and $n = 382$. Like in the previous example, we analyze these data using our methodology assuming that this is a censored sample at $r = 6$, that is, $f_{6+} = 1$. The lower

bound estimator of p_0 given by (6) is, $\hat{p}_0 = 0.3888$ and $\hat{N} = 625.04 \sim 625$. The effective censoring point is $r = 3$, indicating that only the uniques and duplicates are needed for using our method in this example. Figure 6b also illustrates this fact. Note that when $r = 6$, the proportion of censored data is very small ($1/382$). This causes bootstrap CIs to have excessively large amplitudes, a problem that disappears when $r < 6$ is used. Taking $r = 3$, the 95% bootstrap-CI is $(571.5, 714.7) \sim (572, 715)$. This result is consistent with that of Jiménez-Gamero and Puig (2020), $\hat{N} = 632$, who used the complete sample. It is worth noting that while computing the bootstrap-CI, we find that 56.8% of the time the effective censoring point was $r = 3$, while 43.2% of the time it was $r = 2$.

Given the last example of the dolphins in Sepetiba Bay, we might ask: Under what conditions does censoring at $r = 3$ improve censoring at $r = 2$? According to (6), it is clear that this will happen when $G_1^{-1}(f_1/n) < G_2^{-1}((f_1 + f_2)/n)$. Calling $x = f_1/n$ and $y = f_2/n$ this condition is equivalent to $y > G_2(G_1^{-1}(x)) - x$. On the other hand, $x + y \leq 1$ must hold. Figure 7 shows the region (grey area) of the values of $(x, y) = (f_1/n, f_2/n)$, for abundance data, where censoring at $r = 3$ improves censoring at $r = 2$. The border of the region is formed by the curve of the function $G_2(G_1^{-1}(x)) - x$ and the line $y = 1 - x$. Note that for the example of the dolphins in Sepetiba Bay, $(f_1/n, f_2/n) \sim (0.6, 0.3)$ is within the region. For incidence data, similar regions can be constructed that will depend on the value of T .

5 | Final Remarks

The methodology provided here enables us to estimate a lower bound on the size of a population, of nature or society, when the record of identified individuals (or species) only includes those who have been observed once, twice, three times, and r or more times. It is a non-parametric methodology that works with all mixed-binomial (incidence data) and mixed-Poisson (abundance data) frequency distributions. To get better bounds, we would have to restrict ourselves to subfamilies of these distributions, which would be an interesting research area. For example, in practice, it would appear quite reasonable to consider including just unimodal mixed-binomial or mixed-Poisson distributions.

There are other estimators in the literature that, although in principle they were not designed for censored data, in practice they can be used for censored data in specific cases. This is the case, for example, of the jackknife ($r = 2, 3$), \hat{N}_{ChQ2} ($r = 2$), and Chao ($r = 3$) estimators. In these specific censoring patterns, and as our simulations show, our estimator is better than the others for $r = 2$ and slightly better than the Chao estimator for $r = 3$, for trapping occasions higher than five. Nonetheless, we think that none of the estimators under consideration, including our own, is superior to the others in every situation, since there is always a particular distribution for which one estimator outperforms the others. Therefore, our recommendation would be to use all of them, also calculating confidence intervals, and compare the results.

A research topic of interest would be how to combine the results of different estimators to create a better one. In the case of Chao's estimator and ours, because both provide estimates of a lower bound of the population size, using the highest of the two values is an evident option. However, this approach should be thoroughly investigated because the variance of the new estimator built in this way may significantly rise. Furthermore, there are other theoretical aspects that should be taken into account as well.

The maximum of $r - 1$ estimators (lower bounds) of p_0 , which are the values of $G_k^{-1}(\hat{F}_{zi}(k))$, is our estimator suggested in (6). Using the inverse of the estimated variances provided in (8) as weights, another estimator may be created from these by taking the weighted average of all of them. It would be interesting to study the properties of this estimator and others constructed by combining the values of $G_k^{-1}(\hat{F}_{zi}(k))$ in other ways.

Introducing covariates to our estimation method is an interesting topic that could be studied. The work of Böhning et al. (2013), who offer a technique to introduce covariates related to the Chao estimator, constitutes a precedent for this. In our situation, how could it be done? Consider a zero-truncated sample with covariate information $\{(Y_{zi1}, W_1), \dots, (Y_{zin}, W_n)\}$ where W_i is a q -dimensional vector of covariates on individual i . If we were able to construct a non-parametric (or semi-parametric) estimator of the distribution function for each individual given the information of the covariates, say $\hat{F}_{zi}(k|W_i)$, a simple estimator of the population size based in (7) would be, $\hat{N} = \sum_{i=1}^n 1/(1 - \hat{p}_{0i})$, where $\hat{p}_{0i} = \max_{1 \leq k \leq r-1} \{G_k^{-1}(\hat{F}_{zi}(k|W_i))\}$. This is an issue that might be thoroughly researched, and the findings would be very helpful.

If we have a complete sample, does it make sense to censor it? Our opinion is that this depends on the method of estimating the population size N . We have seen that with our method there can be an optimal censoring point, in the sense that it gives us the highest estimate of N together with a confidence interval of length not too high. However, there are other estimation methods based on the complete sample where all observations, both of the less frequent and the most frequent species, are automatically involved. For example, the method by Jiménez-Gamero and Puig (2020) is based on an estimate of the probability generating function (pgf) of the observations, given by the power series $\hat{\Phi}(t) = \frac{f_1}{n}t + \frac{f_2}{n}t^2 + \frac{f_3}{n}t^3 + \dots$, where $0 \leq t \leq 1$. Note, however, that the influence on the value of $\hat{\Phi}(t)$ of the terms

corresponding to high powers of t is very small. This also agrees with the previously mentioned idea that the least frequent species are more relevant than the most frequent in order to estimate population size.

Acknowledgments

The authors would like to express their gratitude to the anonymous referee for his/her insightful comments, which helped us improve the original work.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data is included in the text of the article.

References

- Aleshin-Guendel, S., M. Sadinle, and J. Wakefield. 2024. "The Central Role of the Identifying Assumption in Population Size Estimation." *Biometrics* 80, no. 1: ujad028. <https://doi.org/10.1093/biomtc/ujad028>.
- Böhning, D., J. Bunge, and P. G. Heijden. 2018. *Capture-Recapture Methods for the Social and Medical Sciences*. CRC Press.
- Böhning, D., A. Vidal-Diez, R. Lerdsuwansri, C. Viwatwongkasem, and M. Arnold. 2013. "A Generalization of Chao's Estimator for Covariate Information." *Biometrics* 69, no. 4: 1033–1042.
- Budka, A., A. Łacka, and K. Szoszkiewicz. 2018. "Estimation of River Ecosystem Biodiversity Based on the Chao Estimator." *Biodiversity and Conservation* 27: 205–216.
- Burnham, K. P., and W. S. Overton. 1978. "Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals." *Biometrika* 65, no. 3: 625–633.
- Cazzolla Gatti, R., P. B. Reich, J. G. P. Gamarra, et al. 2022. "The Number of Tree Species on Earth." *Proceedings of the National Academy of Sciences of the United States of America* 119, no. 6: e2115329119. <https://doi.org/10.1073/pnas.2115329119>.
- Chao, A. 1984. "Nonparametric Estimation of the Number of Classes in a Population." *Scandinavian Journal of Statistics* 11: 265–270.
- Chao, A. 1987. "Estimating the Population Size for Capture-Recapture Data With Unequal Catchability." *Biometrics* 43, no. 4: 783–791.
- Chao, A., R. K. Colwell, C. H. Chiu, et al. 2017. "Seen Once or More Than Once: Applying Good–Turing Theory to Estimate Species Richness Using Only Unique Observations and a Species List." *Methods in Ecology and Evolution* 8, no. 10: 1221–1232.
- Chiu, C. H. 2022. "Incidence-Data-Based Species Richness Estimation via a Beta-Binomial Model." *Methods in Ecology and Evolution* 13: 2546–2558.
- Chiu, C. H., Y. T. Wang, B. A. Walther, and A. Chao. 2014. "An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula." *Biometrics* 70: 671–682.
- Fraisl, D., G. Hager, B. Bedessem, et al. 2022. "Citizen Science in Environmental and Ecological Sciences." *Nature Reviews Methods Primers* 2: 65.
- Holzmann, H., A. Munk, and W. Zucchini. 2006. "On Identifiability in Capture-Recapture Models." *Biometrics* 62, no. 3: 934–936.
- Jiménez-Gamero, M. D., and P. Puig. 2020. "A Nonparametric Method of Estimation of the Population Size in Capture–Recapture Experiments." *Biometrical Journal* 62: 970–988.

Lanumteang, K., and D. Böhning. 2011. "An Extension of Chao's Estimator of Population Size Based on the First Three Capture Frequency Counts." *Computational Statistics & Data Analysis* 55: 2302–2311.

Link, W. A. 2003. "Nonidentifiability of Population Size From Capture-Recapture Data With Heterogeneous Detection Probabilities." *Biometrics* 59: 1123–1130.

McCrea, R. S., and B. J. Morgan. 2014. *Analysis of Capture-Recapture Data*. CRC Press.

Morgan, B. J. T., and M. S. Ridout. 2008. "A New Mixture Model for Capture Heterogeneity." *Journal of the Royal Statistical Society: Series C: Applied Statistics* 57, no. 4: 433–446.

Nery, M. F., and S. M. Simão. 2012. "Capture-Recapture Abundance Estimate of Guiana Dolphins in Southeastern Brazil." *Ciencias Marinas* 38: 529–541.

Puig, P., and C. C. Kokonendji. 2018. "Non-Parametric Estimation of the Number of Zeros in Truncated Count Distributions." *Scandinavian Journal of Statistics* 45: 347–365.

Schofield, M. R., R. J. Barker, W. A. Link, and H. Pavanato. 2023. "Estimating Population Size: The Importance of Model and Estimator Choice." *Biometrics* 79, no. 4: 3803–3817.

Shaked, M., and J. G. Shanthikumar. 2007. *Stochastic Orders*. Springer Science & Business Media.

Zelterman, D. 1988. "Robust Estimation in Truncated Discrete Distributions With Application to Capture-Recapture Experiments." *Journal of Statistical Planning and Inference* 18: 225–237.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix A

Proof of Theorem 1. Note that,

$$F(k) = P(Y \leq k) = \int_0^1 \sum_{i=0}^k \binom{T}{i} \left(\frac{\lambda}{1-\lambda} \right)^i (1-\lambda)^T f(\lambda) d\lambda$$

Because $p_0 = \int_0^1 (1-\lambda)^T f(\lambda) d\lambda$,

$$\frac{F(k)}{p_0} = \int_0^1 \sum_{i=0}^k \binom{T}{i} \left(\frac{\lambda}{1-\lambda} \right)^i dG(\lambda) \quad (A1)$$

where $dG(\lambda) = \frac{(1-\lambda)^T f(\lambda) d\lambda}{\int_0^1 (1-\lambda)^T f(\lambda) d\lambda}$. Consider now a real valued function $h(x)$ such that, $h\left(\sum_{i=0}^k \binom{T}{i} \left(\frac{\lambda}{1-\lambda}\right)^i\right) = (1-\lambda)^{-T}$. Taking $(1-\lambda)^{-T} = u$ is clear that, $h\left(\sum_{i=0}^k \binom{T}{i} (u^{1/T} - 1)^i\right) = u$ or, in other words, $h^{-1}(u) = \sum_{i=0}^k \binom{T}{i} (u^{1/T} - 1)^i$. Function $h^{-1}(u)$ is defined for $u \in [1, \infty)$, and it is increasing and concave. Consequently, $h(x)$ is a convex increasing function defined for $x \in [0, 1)$. Therefore, applying the Jensen's inequality, to the right part of (A1), we obtain,

$$\begin{aligned} h\left(\int_0^1 \sum_{i=0}^k \binom{T}{i} \left(\frac{\lambda}{1-\lambda}\right)^i dG(\lambda)\right) &\leq \int_0^1 h\left(\sum_{i=0}^k \binom{T}{i} \left(\frac{\lambda}{1-\lambda}\right)^i\right) dG(\lambda) \\ &= \int_0^1 (1-\lambda)^{-T} dG(\lambda) \\ &= \int_0^1 (1-\lambda)^{-T} \frac{(1-\lambda)^T f(\lambda) d\lambda}{\int_0^1 (1-\lambda)^T f(\lambda) d\lambda} = \frac{1}{p_0} \end{aligned}$$

Therefore, applying $h^{-1}(u)$ to both sides of the inequality we obtain, $\frac{F(k)}{p_0} \leq h^{-1}\left(\frac{1}{p_0}\right)$, and this is equivalent to,

$$F(k) \leq p_0 \sum_{i=0}^k \binom{T}{i} \left(\frac{1}{p_0^{1/T}} - 1 \right)^i = \sum_{i=0}^k \binom{T}{i} (1 - p_0^{1/T})^i p_0^{1-i/T} \quad (A2)$$

and it concludes the proof.

Proof of Theorem 2. We can write,

$$\frac{F(k)}{p_0} = \int_0^\infty \sum_{i=0}^k \frac{\lambda^i}{i!} dG(\lambda) \quad (A3)$$

where $dG(\lambda) = \frac{e^{-\lambda} f(\lambda) d\lambda}{\int_0^\infty e^{-\lambda} f(\lambda) d\lambda}$. Define now a real valued function $h(x)$ such that, $h\left(\sum_{i=0}^k \frac{\lambda^i}{i!}\right) = e^\lambda$. Taking $u = e^\lambda$ we obtain, $h\left(\sum_{i=0}^k \frac{\log(u)^i}{i!}\right) = u$, or equivalently, $h^{-1}(u) = \sum_{i=0}^k \frac{\log(u)^i}{i!}$. Again, function $h^{-1}(u)$ is defined for $u \in [1, \infty)$, and it is increasing and concave. Therefore, $h(x)$ is a convex increasing function defined for $x \in [0, \infty)$. Now we can apply Jensen's inequality to the right part of (A3), obtaining,

$$h\left(\int_0^\infty \sum_{i=0}^k \frac{\lambda^i}{i!} dG(\lambda)\right) \leq \int_0^\infty h\left(\sum_{i=0}^k \frac{\lambda^i}{i!}\right) dG(\lambda) = \int_0^\infty e^\lambda dG(\lambda) = \frac{1}{p_0}$$

Finally, applying $h^{-1}(u)$ to both sides of the inequality we obtain, $\frac{F(k)}{p_0} \leq h^{-1}\left(\frac{1}{p_0}\right)$, and it concludes the proof.

Proof of Lemma 1. We have to prove that, for any value of k , $G_k(p) \leq G_k(q)$ when $p \leq q$. Note that $G_k(p_0)$ is the distribution function of a zero-truncated Binomial random variable with probability of success $1 - p_0^{1/T}$, that it will be denoted as X_{p_0} .

A random variable X with distribution function F is said to be stochastically greater than a random variable Y with distribution function H if $F(x) \leq H(x)$ for every x . This ordering is called *stochastic ordering* and it is wrote as $Y \leq X$. Therefore, to prove that $G_k(p_0)$ is an increasing function is equivalent to prove that $X_q \leq X_p$ when $p \leq q$. A stronger ordering, known as *likelihood-ratio ordering*, implies stochastic ordering. We say that X is greater or equal than Y according to likelihood-ratio ordering, indicated as $Y \leq X$, if $f(x)/h(x)$ is not decreasing in x , where $f(x)$ and $h(x)$ are their respective density or probability functions. Consequently, if we prove that $X_q \leq X_p$ when $p \leq q$ then we will prove that $G_k(p_0)$ is an increasing function of p_0 for all k . Note that,

$$\begin{aligned} \frac{P(X_p = x)}{P(X_q = x)} &= \frac{\binom{T}{x} (1 - p^{1/T})^x p^{1-x/T} / (1 - p)}{\binom{T}{x} (1 - q^{1/T})^x q^{1-x/T} / (1 - q)} \\ &= C \left(\frac{q^{1/T} - (pq)^{1/T}}{p^{1/T} - (pq)^{1/T}} \right)^x \end{aligned}$$

where $C = \frac{p/(1-p)}{q/(1-q)}$. Therefore, when $p \leq q$ this is a not decreasing function in x and it concludes the proof.

The proof presented here is immediately applicable to the case of abundance data, as T tends to infinity.