Biometrical Journal **xx** (2022) xx, zzz–zzz / DOI: 10.1002/bimj.200100000

# A likelihood ratio test for completed sampling in population size estimation studies

**Alessio Farcomeni** *,[1]

[1] Department of Economics and Finance *(University of Rome "Tor Vergata")*

Received zzz, revised zzz, accepted zzz

We propose a likelihood ratio test to assess that sampling has been completed in closed population size estimation studies. More precisely, we assess if the expected number of subjects that have never been sampled is below a user-specified threshold. The likelihood ratio test statistic has a non-standard distribution under the null hypothesis. Critical values can be easily approximated and tabulated, and they do not depend on model specification. We illustrate in a simulation study and three real data examples, one of which involves ascertainment bias of amyothrophic lateral sclerosis in Gulf War veterans.

*Key words:* Ascertainment Bias; Capture-recapture; Constrained Optimisation; Chi-bar squared.

## 1 Introduction

Population size estimation studies involve repeatedly sampling $n$ subjects from a population of unknown finite size $N$ (Otis et al., 1978; McCrea and Morgan, 2014; Bohning et al., 2018; Silverman, 2020). We focus here on closed population experiments, in which the population size is assumed to be fixed during the observation period.

In certain contexts it is natural to wonder whether $n = N$, that is, if sampling has been completed and all subjects in the population have been observed. This is of interest in several areas. For planning of the experiment, if the test is rejected researchers might decide to proceed with further sampling (see also Alunni Fegatelli and Farcomeni (2016)). In epidemiology subjects might be of interest for further assessment and neglecting to sample some of them might put them at risk; e.g., some people with a disease might not be diagnosed and treated. Additionally, in epidemiology one might want to assess ascertainment bias: is there a subpopulation for which sampling was clearly more (or less) likely? In engineering one might investigate the abundance of defective products in a batch, or bugs in a software; and the question arises whether all items were identified. In genetics researchers might wonder whether all expressed genes have been identified based on sequence tags (e.g., Mao (2004)). Investigation of biodiversity in microbiology (or ecology) often involves wondering whether all species in a sample (or area, respectively) have been observed. Some other motivating applications are referenced in publications in the related area of species sampling (e.g., Chao (1981); Esty (1983); Mao (2004); Favaro et al. (2012)).

In this work we put forward a testing procedure to assess whether sampling has been completed. The procedure involves a possibly complex inequality constraint, leading to a chi-bar-squared null distribution.

To the best of our knowledge this problem has been tackled so far only by checking whether the observed sample size $n$ was included in a confidence/credibility interval for the population size $N$; and hence our proposal is the first formal test for completed sampling in this context. It shall be commented that the theoretical properties of inversion of a confidence interval are not straightforward, as the procedure is conditional on $n$, which should be treated instead as a random variable (e.g., Böhning (2008)). We will see in a brief simulation study that inversion of the confidence interval often exceeds the nominal level. We will also revisit three examples in which inversion of confidence intervals have been used in the literature,

*e-mail: alessio.farcomeni@uniroma2.it

 

and find that our test leads to different conclusions in some cases. A related but different problem is the one tackled in the area of species sampling, where methods exist to estimate the probability of sampling a new species in $s$ additional sampling actions. A negligibile probability of sampling a new species for large $s$ implies that all species have been sampled, but this is not a formal testing/decision procedure.

The rest of the paper is as follows: in the next section we specify our general framework and derive a likelihood ratio test. A simulation study for four specific models is reported in Section 3. Two benchmark data sets and a data set about ascertainment bias of amyothrophic lateral sclerosis in Gulf War veterans are analysed in Section 4. In Section 5 we give some concluding remarks. `R` code with an implementation of our method, data, and code for reproducing the simulation study and real data examples are available as supplementary material.

## 2 Setup

Capture histories of the $n$ subjects observed at least once are collected in a data matrix $X$. Let $X_{ij} = 1$ if the $i$-th subject has been observed on the $j$-th occasion, and zero otherwise; with $i = 1, \ldots, n$ and $j = 1, \ldots, S$. Without loss of generality, we assume $N - n \geq 0$ rows of the data matrix are not observed, so if $N > n$ $X_{ij} = 0$ for all $i = n+1, \ldots, N$. Let $\Pr(X_{ij} = 1) = p_{ij}(\theta)$, where $\theta$ is a short-hand notation for a vector of parameters corresponding to some model specification. Let also $p_0(\theta) = \Pr(X_{i1} = X_{i2} = \cdots = X_{iS} = 0)$. We will outline few special cases below, and keep the exposition completely general for the time being. The full likelihood can be expressed as

$$L(N, \theta) = \prod_{i=1}^{N} \prod_{j} p_{ij}(\theta)^{X_{ij}} (1 - p_{ij}(\theta))^{1-X_{ij}}$$

We rely on a common likelihood decomposition (Sanathanan, 1972; Farcomeni and Tardella, 2012), according to which $L(N, \theta) \propto L_c(\theta) L_r(N, \theta)$, with

$$L_c(\theta) \propto \prod_{i=1}^{n} \prod_{j} p_{ij}(\theta)^{X_{ij}} (1 - p_{ij}(\theta))^{1-X_{ij}} / (1 - p_0(\theta))$$

(the *conditional* likelihood) and

$$L_r(N, \theta) \propto (1 - p_0(\theta))^n p_0(\theta)^{N-n}$$

(the *residual* likelihood). The decomposition is convenient since the cumbersome optimisation of the full likelihood is asymptotically equivalent to a two-step procedure where one can optimise the conditional likelihood $L_c(\theta)$, which does not depend on $N$, and plug $\hat{\theta}$ into the residual likelihood to obtain the Horvitz-Thompson estimator

$$\hat{N} = \frac{n}{1 - p_0(\hat{\theta})}.$$

Under general conditions the two step approach has got the same inferential properties, as $N$ grows, as direct maximisation of the full likelihood (Sanathanan, 1972) and one can set $L(\hat{N}, \hat{\theta}) = L_c(\hat{\theta}) L_r(\hat{N}, \hat{\theta})$ to obtain the maximum of the full likelihood.

We are now ready to set up our procedure for testing whether all subjects can be expected to have been sampled. A natural way of proceeding would be to specify a null hypothesis of the kind $H_0 : n = N$, or $H_0 : p_0(\theta) = 0$. The first would not be well specified: $n$ is not a parameter and, more importantly, it is actually a random variable (Böhning, 2008). Inversion of the confidence interval for $N$ would informally correspond to a test for $H_0 : n = N$, and we will see below that this does not have the desired properties. The condition that $p_0(\theta) = 0$ on the other hand might be too restrictive and, in most cases, at odds with

 

the model specification. Indeed, for many models one can have $p_0(\theta) = 0$ if and only if $p_{ij}(\theta) = 1$ for all $i$ and $j = 1, \ldots, S$. In turn, $p_{ij}(\theta) = 1$ if and only if $X_{ij} = 1$ for all $i = 1, \ldots, n$; which is not at all common in real data.

We propose to test a slightly more general, albeit complex, null hypothesis:

$$H_0 : Np_0(\theta) \leq c, \tag{1}$$

for some $c \geq 0$ that is specified by the user. The left hand side of the inequality corresponds to the expected number of subjects that will not be sampled during the experiment. For the upper bound a prominent case is $c = 1$, indicating that less than 1 subject is expected to be unobserved. For several common modeling assumptions the case $c = 0$ does not correspond to any solution within the parameter space, as discussed above. For instance for model $M_0$, according to which $\Pr(X_{ij} = 1) = p$, $Np_0(\theta) = N(1-p)^S$. The latter expression can be zero only if $N = 0$ or $p = 1$. The first case is excluded by the fact that $n > 0$, and the second is compatible only with the case $\sum_j X_{ij} = S$ for all $i$.

For testing (1) we propose a likelihood-ratio statistic of the kind

$$T(c) = -2 \left( \log \left( \sup_{\theta:\ Np_0(\theta) \leq c} L(\theta) \right) - \log \left( \sup_{\theta} L(\theta) \right) \right). \tag{2}$$

The first addend shall be obtained through constrained optimisation. Since the parameter space is usually low dimensional we propose to simply use a numerical method, like the Constrained Optimization BY Linear Approximation (COBYLA) algorithm (Powell, 1994).

Once (2) has been computed, the matter arises to determine critical values based on its null distribution. To this end we rely on Silvapulle and Sen (2004). Under regularity conditions, it is shown in (Silvapulle and Sen, 2004, Chapter 4.8) that under the null hypothesis $T$ asymptotically follows a chi-bar-squared distribution, which can be explicited for $t > 0$ as $Pr(T(c) \geq t \mid H_0) = w_0 Pr(\chi^2_0 \geq t) + w_1 Pr(\chi^2_1 \geq t)$, $w_0 + w_1 = 1$. The chi-bar-squared weights $w_0$ and $w_1$ are not at all straightforward to obtain, but it is shown in Silvapulle and Sen (2004) that setting $w_0 = 0.5$ stochastically dominates the true null distribution. As we will see in Section 3 the bound seems to be rather sharp in many cases, leading to a test almost precisely of the desired size. Notably, the resulting critical values are model independent. By inverting the chi-bar squared CDF a critical value of 2.705 is obtained at the 5% level, a value of 5.412 at the 1% level, a value of 1.642 at the 10% level. Testing might occur after model selection. In this case, as in many other contexts, we are faced with a problem of post-selection-inference (e.g., Kuchibhotla et al. (2022), and references therein). Many researchers proceeed naively without performing any adjustment. A simple and effective alternative involves an adjustment for multiple testing. One can indeed simply use a Bonferroni adjustment by dividing the desired significance level by the number of models evaluated. A power-sparing approach would involve on the other hand control of the False Discovery Rate, for instance through Simes procedure. See Farcomeni (2008) for a review. We finally mention that one might also derive the null distribution after conditioning on the model selection event (the conditional selective approach). This would be cumbersome in our context, but the model-independent null distribution of the test statistic makes the naive approach promising. Indeed, in the simulation study in the next section we find evidence that the naive approach does not seem to lead to exceedance of the nominal level.

## 3 Simulation

In this section we evaluate size and power of our proposed test on simulated data, and compare with simple inversion of a confidence interval. We fix $c = 1$, $N = \{250, 1000\}$, and try different data generating distributions as follows:

- $$M_0 :\ \Pr(X_{ij} = 1) = p, \tag{3}$$

 

with $S = 4$,

- $$M_h : \Pr(X_{ij} = 1) = p_i,\ \log(p_i/(1-p_i)) \sim N(\mu, 0.25), \tag{4}$$

  with $S = 4$,

- *Poisson*: $\sum_j X_{ij} \sim Poisson(\lambda)$, with $S = \infty$
- *MixPoisson*: $\sum_j X_{ij}|\lambda \sim Poisson(\lambda_i)$, with $\log(\lambda_i) \sim N(\xi, 0.25)$, with $S = \infty$

For each scenario we set different values of $p$, $\mu$, $\lambda$, and $\xi$, respectively. Data are generated $B = 1000$ times for each setting. For each data set we compute the proposed likelihood ratio statistic and evaluate the corresponding null hypothesis at the 5% level. For comparison with the common practice we compute a 95% confidence interval for $\hat{N}$ and invert the interval, rejecting the test if the observed sample size $n$ is strictly below the lower limit of the interval. To show the consequences of post-selection-inference, for the first two data generating distributions we also compare four models, select based on the Akaike Information Criterion (AIC), and naively perform the test conditionally on the selected model. In addition to models $M_0$ and $M_h$ above we test models

$$M_t : \Pr(X_{ij} = 1) = p_j \tag{5}$$

and

$$M_{th} : \Pr(X_{ij} = 1) = p_{ij},\ \log(p_{ij}/(1-p_{ij})) \sim N(\mu_j, \sigma). \tag{6}$$

**Table 1** Simulation study. Model $M_0$. Proportion of rejections for our test ($R_T$), inversion of confidence interval ($R_{CI}$), and our test after model selection ($R_{MS}$) for different values of $N$ and $p$. For the test, $c = 1$. Results are averaged over $B = 1000$ replicates.

| $N$ | $p$ | $Np_0(\theta)$ | $R_T$ | $R_{CI}$ | $R_{MS}$ |
|---|---|---|---|---|---|
| 250 | 0.250 | 79.102 | 1.000 | 1.000 | 1.000 |
| 250 | 0.330 | 50.378 | 1.000 | 1.000 | 1.000 |
| 250 | 0.500 | 15.625 | 1.000 | 1.000 | 1.000 |
| 250 | 0.660 | 3.341 | 1.000 | 0.963 | 1.000 |
| 250 | 0.749 | 1.000 | 0.042 | 0.170 | 0.040 |
| 250 | 0.822 | 0.250 | 0.000 | 0.000 | 0.000 |
| 250 | 0.900 | 0.025 | 0.000 | 0.000 | 0.000 |
| 1000 | 0.250 | 316.406 | 1.000 | 1.000 | 1.000 |
| 1000 | 0.330 | 201.511 | 1.000 | 1.000 | 1.000 |
| 1000 | 0.500 | 62.500 | 1.000 | 1.000 | 1.000 |
| 1000 | 0.660 | 13.363 | 1.000 | 1.000 | 1.000 |
| 1000 | 0.749 | 4.000 | 1.000 | 1.000 | 1.000 |
| 1000 | 0.822 | 0.999 | 0.046 | 0.786 | 0.041 |
| 1000 | 0.900 | 0.100 | 0.000 | 0.000 | 0.000 |

Results are reported in Tables 1, 2, 3, 4. From the tables it can be seen that the proposed test is powerful and always guarantees the 5% nominal level everytime $Np_0(\theta) < 1$; and most importantly in most cases the nominal level is achieved almost exactly when $Np_0(\theta) \approx c = 1$. Since we are testing against a composite null hypothesis it is inevitable that the test level is well below the nominal level when

 

**Table 2** Simulation study. Model $M_h$. Proportion of rejections for our test ($R_T$), inversion of confidence interval ($R_{CI}$), and our test after model selection ($R_{MS}$) for different values of $N$ and $\mu$. For the test, $c = 1$. Results are averaged over $B = 1000$ replicates.

| $N$ | $\mu$ | $Np_0(\theta)$ | $R_T$ | $R_{CI}$ | $R_{MS}$ |
|---|---|---|---|---|---|
| 250 | -1.000 | 70.328 | 1.000 | 1.000 | 1.000 |
| 250 | 0.000 | 15.625 | 1.000 | 1.000 | 1.000 |
| 250 | 1.000 | 1.363 | 0.383 | 0.312 | 0.338 |
| 250 | 1.106 | 1.000 | 0.040 | 0.030 | 0.050 |
| 250 | 1.551 | 0.250 | 0.000 | 0.000 | 0.000 |
| 250 | 2.000 | 0.055 | 0.000 | 0.000 | 0.000 |
| 250 | 3.000 | 0.001 | 0.000 | 0.000 | 0.000 |
| 1000 | -1.000 | 281.313 | 1.000 | 1.000 | 1.000 |
| 1000 | 0.000 | 62.500 | 1.000 | 1.000 | 1.000 |
| 1000 | 1.000 | 5.451 | 1.000 | 1.000 | 1.000 |
| 1000 | 1.106 | 4.000 | 1.000 | 0.842 | 1.000 |
| 1000 | 1.551 | 1.000 | 0.039 | 0.000 | 0.043 |
| 1000 | 2.000 | 0.219 | 0.000 | 0.000 | 0.000 |
| 1000 | 3.000 | 0.006 | 0.000 | 0.000 | 0.000 |

**Table 3** Simulation study. Model $Poisson$. Proportion of rejections for our test ($R_T$) and inversion of confidence interval ($R_{CI}$) for different values of $N$ and $\lambda$. For the test, $c = 1$. Results are averaged over $B = 1000$ replicates.

| $N$ | $\lambda$ | $Np_0(\theta)$ | $R_T$ | $R_{CI}$ |
|---|---|---|---|---|
| 250 | 2.000 | 33.834 | 1.000 | 1.000 |
| 250 | 3.000 | 12.447 | 1.000 | 1.000 |
| 250 | 4.000 | 4.579 | 1.000 | 1.000 |
| 250 | 5.000 | 1.684 | 0.971 | 0.811 |
| 250 | 5.521 | 1.000 | 0.043 | 0.646 |
| 250 | 6.000 | 0.620 | 0.000 | 0.224 |
| 250 | 6.908 | 0.250 | 0.000 | 0.000 |
| 250 | 7.000 | 0.228 | 0.000 | 0.000 |
| 250 | 8.000 | 0.084 | 0.000 | 0.000 |
| 1000 | 2.000 | 135.335 | 1.000 | 1.000 |
| 1000 | 3.000 | 49.787 | 1.000 | 1.000 |
| 1000 | 4.000 | 18.316 | 1.000 | 1.000 |
| 1000 | 5.000 | 6.738 | 1.000 | 1.000 |
| 1000 | 5.521 | 4.000 | 1.000 | 1.000 |
| 1000 | 6.000 | 2.479 | 1.000 | 1.000 |
| 1000 | 6.908 | 1.000 | 0.049 | 0.986 |
| 1000 | 7.000 | 0.912 | 0.004 | 0.999 |
| 1000 | 8.000 | 0.335 | 0.000 | 0.000 |

$Np_0(\theta) < c$. The empirical level of the test after model selection (Tables 1 and 2) is still below 5%, indicating that in these scenarios adjustments might not be strictly necessary.

Finally, inversion of the confidence interval clearly exceeds the nominal level, with the exception of $M_h$ model where it is conservative.

 

**Table 4** Simulation study. Model $MixPoisson$. Proportion of rejections for our test ($R_T$) and inversion of confidence interval ($R_{CI}$) for different values of $N$ and $\xi$. For the test, $c = 1$. Results are averaged over $B = 1000$ replicates.

| $N$ | $\xi$ | $Np_0(\theta)$ | $R_T$ | $R_{CI}$ |
|---|---|---|---|---|
| 250 | 1.000 | 18.780 | 1.000 | 1.000 |
| 250 | 1.842 | 1.000 | 0.040 | 0.714 |
| 250 | 2.000 | 0.447 | 0.000 | 0.002 |
| 250 | 2.104 | 0.249 | 0.000 | 0.000 |
| 250 | 3.000 | 0.000 | 0.000 | 0.000 |
| 250 | 4.000 | 0.000 | 0.000 | 0.000 |
| 250 | 5.000 | 0.000 | 0.002 | 0.000 |
| 1000 | 1.000 | 75.119 | 1.000 | 1.000 |
| 1000 | 1.842 | 4.000 | 1.000 | 1.000 |
| 1000 | 2.000 | 1.787 | 1.000 | 0.987 |
| 1000 | 2.104 | 0.997 | 0.050 | 0.997 |
| 1000 | 3.000 | 0.001 | 0.000 | 0.000 |
| 1000 | 4.000 | 0.000 | 0.000 | 0.000 |
| 1000 | 5.000 | 0.000 | 0.000 | 0.000 |

## 4 Real Data Examples

We revisit in this section two benchmark data sets and one original data example in epidemiology. For the benchmark data sets completion of sampling has been previously discussed in some extension in the literature. For the epidemiology data example, ascertainment bias was one of the main scientific questions. In all cases to the best of our knowledge the problem of assessing how many subjects were not sampled has been tackled so far by comparing a population size estimate, and its confidence interval, with the observed sample size.

### 4.1 House Mouse Data

The house mouse data involves a live trapping study on feral House mice (*Mus musculus*) in the December of 1962 in Ballana Creek (California, USA). There are $S = 10$ sampling occasions with trapping repeated twice a day for five days, for a total observed of $n = 173$ animals.

In Otis et al. (1978) different models are compared, and the general conclusion is that all mice have been sampled at least once. Notably, for models $M_0$ (defined in (3)), $M_t$ (defined in (5)) and $M_h$ (defined in (4)) the confidence interval for $\hat{N}$ includes $n$. Additionally, only two new mice are caught after the seventh trapping occasion. In Otis et al. (1978) this is argued as additional evidence that the entire population might have been observed.

Surprisingly enough, with model $M_0$ and $c = 1$ the resulting likelihood ratio test statistic is 33.21; and the null hypothesis is rejected. Also with $c = 2$ the test statistic is 5.09, well above the 5% significance threshold of 2.705. Only with $c \geq 3$ the LRT is not rejected. Very similar results are observed also with model $M_h$. Our formal test, contradicting previous results, indicates therefore that some mice might have not been sampled during the experiment.

### 4.2 Meadow Voles Data

A second benchmark data set involves a population of meadow voles (*Microtus pennsylvanicus*) in Laurel, Maryland. Here $S = 5$, $n = 56$ adult males. Data come from Nichols et al. (1984). For these data and model $M_0$ (defined in (3)), $\hat{N} = n = 56$, with a 95% profile confidence interval $(56 - 58)$. The probability

 

to be captured at least once is estimated as being as high as 99%. We now proceed with our test: under model $M_0$ the likelihood ratio test statistic is equal to 0.0001 with $c = 1$; similarly with model $M_h$ (defined in (4)) we have $T(1) = 0.00000596$. The null hypothesis is never rejected. There is no evidence therefore of the existence of meadow voles that have not been sampled, in agreement with previous results in the literature.

### 4.3 Amyothrophic Lateral Sclerosis Data

Coffman et al. (2005) collects five data sets, each with $S = 4$ lists of cases of Amyotrophic Lateral Sclerosis (ALS) in the U.S. military population. The population includes Gulf war (1990-1991) veterans, who are known to be at higher risk of ALS, probably due to exposure to certain risk factors. The data has been revisited also in Alunni Fegatelli and Farcomeni (2016). The main issue is ascertainment bias: are there militaries with the disease that have not been diagnosed?

There five data sets to be considered are: (i) overall ($n = 107$), (ii) deployed to Iraq according to the Defense Manpower Data Center (DMDC) ($n = 40$), (iii) not deployed to Iraq according to the DMDC ($n = 67$), (iv) deployed to Iraq according to self-reporting ($n = 49$) and (v) not deployed to Iraq according to self-reporting ($n = 58$). Surprisingly enough counts in (ii) and (iv) (and consequently (iii) and (v)) are slightly different due to errors in the DMDC database and/or in self-reporting.

For each of the five data sets the sources were: the veterans affairs database, the department of defense database, the database of the ALS patients association, and calls to a toll-free telephone number that individuals could call if they believed were eligible for the study. The toll-free number was publicized both in military and non-military media. It is clear that the probability of identification of individuals is differential among lists, so we use model $M_{th}$ (defined in (6)).

In Table 5 we report the test statistic for $c = 1, \ldots, 6$ and each of the five data sets. In the last column we also report the population size estimate $\hat{N}$ together with 95% confidence interval in parentheses.

**Table 5** ALS data. Test statistics for $c = 1, \ldots, 6$ and each of the available databases (DMDC-D: deployed according to DMDC, DMDC-ND: not deployed according to DMDC, Self-D: deployed according to self-reporting, Self-ND: not deployed according to self-reporting). Results are based on model $M_{th}$. The critical value at $\alpha = 0.05$ is 2.705. In the last column we report the population size estimate $\hat{N}$ together with 95% confidence interval in parentheses.

| | $c=1$ | $c=2$ | $c=3$ | $c=4$ | $c=5$ | $c=6$ | $\hat{N}$ |
|---|---|---|---|---|---|---|---|
| Overall | 35.17 | 13.97 | 5.51 | 1.70 | 0.20 | 0.00 | 113 (111-122) |
| DMDC-D | 1.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 42 (40-45) |
| DMDC-ND | 22.02 | 7.21 | 2.02 | 0.21 | 0.00 | 0.00 | 72 (69-80) |
| Self-D | 5.42 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 52 (50-55) |
| Self-ND | 7.15 | 0.79 | 0.00 | 0.00 | 0.00 | 0.00 | 60 (59-66) |

It can be seen that at the stringent threshold $c = 1$ only the DMDC database about veterans of the Gulf War does not show evidence of uncompleted sampling. This is reasonable as officially deployed veterans were accurately screened due to known higher risk of developing the disease. For all other databases we reject the null hypothesis, at least for $c = 1$ and $c = 2$. Indeed, it can be expected that military that were not deployed to Iraq were not as accurately screened; and some deployed veterans were probably missing from the DMDC database and were therefore not contacted for diagnostic assessment. In summary, there is some evidence of ascertainment bias for ALS in US military, with Gulf War veterans more likely to be included in the databases.

 

## 5 Conclusions

We have proposed, to the best of our knowledge, the first formal test for assessing that sampling has been completed in population size estimation experiments. The test can be seen as a goodness-of-fit test, where failure to reject the null hypothesis is the desired outcome in many cases. The null hypothesis is that the expected number of unobserved individuals ($Np_0(\theta)$) is below a user-specified positive threshold. The resulting test is non-standard, but its critical constant can be approximated and tabulated quite simply. We have seen in simulation that the resulting null distribution is very well approximated and the nominal level is closely achieved. We have argued and shown in simulation that the common pratice of checking whether $n$ falls in the confidence interval for $\hat{N}$ is not appropriate as it often exceeds the expected nominal level.

As could be seen in the simulation study, if $Np_0(\theta) << c$ the test is still valid, but conservative. The researcher might also switch the focus from testing to estimation of the number of unobserved individuals. In this case, one can proceed with a preliminary estimation of $Np_0(\theta)$ and then set $c = \hat{N}p_0(\hat{\theta})$. The resulting test might have to be adjusted for post-selection-inference. Derivation of a conditional selective approach for this scenario is left for further work.

We have discussed a very general framework for discrete-time experiments and closed populations. Multiple systems estimation with repeated entries in multiple lists, and some particular cases of continuous-time experiments are special cases. The latter include situations in which time homogeneity assumptions can be made, so that sufficient statistics are the number of times a subject has been observed. Further work is needed for more general cases in continuous time, and open population experiments. A further open problem is the presence of individual covariates, as we do not have information for subjects that were never observed. The common strategy would be to perform inference based on the conditional likelihood, while the proposed test is based on ratio of the full likelihoods. An extension of the test when individual covariates are used is therefore also left for further work.

## Supplementary Materials

Code referenced in Section 1 is available on the publisher's website along with the electronic version of this paper.

## Acknowledgments

The author is grateful to an Associate Editor and three referees for constructive comments.

**Conflict of Interest** *The author declares no conflict of interest*

## References

Alunni Fegatelli, D. and Farcomeni, A. (2016). On the design of closed recapture experiments. *Biometrical Journal* **58**, 1273–1294.

Böhning, D. (2008). A simple variance formula for population size estimators by conditioning. *Statical Methodology* **5**, 410–423.

Bohning, D.; van der Heijden, P. G. M. and Bunge, J. (editors) (2018). *Capture-recapture methods for the social and medical sciences*. Chapman and Hall/CRC.

Chao, A. (1981). On estimating the probability of discovering a new species. *Annals of Statistics* **9**, 1339–1342.

Coffman, C. J.; Horner, R. D.; Grambow, S. C. and Lindquist, J. (2005). Estimating the occurrence of Amyotrophic Lateral Sclerosis among Gulf War (1990-1991) veterans using capture-recapture methods. *Neuroepidemiology* **24**, 141–150.

Esty, W. W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *Annals of Statistics* **11**, 905–911.

 

Farcomeni, A. (2008). A Review of Modern Multiple Hypothesis Testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* **17**, 347–388.

Farcomeni, A. and Tardella, L. (2012). Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electronic Journal of Statistics* **6**, 2602–2626.

Favaro, S.; Lijoi, A. and Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics* **68**, 1188–1196.

Kuchibhotla, A. K.; Kolassa, J. E. and Kuffner, T. A. (2022). Post-Selection Inference. *Annual Review of Statistics and Its Application* **9**, 505–527.

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *Journal of the American Statistical Association* **99**, 1108–1118.

McCrea, R. S. and Morgan, B. J. T. (2014). *Analysis of Capture-recapture Data*. CRC Press, Boca Raton.

Nichols, J. D.; Pollock, K. H. and Hines, J. E. (1984). The use of a robust capture-recapture design in small mammal population studies: A field example with Microtus Pennsylvanicus. *Acta Theriologica* **29**, 357–365.

Otis, D. L.; Burnham, K. P.; White, G. C. and Anderson, D. R. (1978). *Statistical inference from capture data on closed animal populations*, volume 62. Wildlife Monographs.

Powell, M. J. D. (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation. S. Gomez and J.-P. Hennart (editors), *Advances in Optimization and Numerical Analysis*, pages 51–67. Kluwer Academic Publishers, Dordrecht.

Sanathanan, L. (1972). Estimating the Size of a Multinomial Population. *The Annals of Mathematical Statistics* **43**(1), 142 – 152.

Silvapulle, M. and Sen, P. (2004). *Constrained Statistical Inference*. Wiley.

Silverman, B. W. (2020). Multiple Systems Analysis for the quantification of Modern Slavery: Classical and Bayesian approaches (with Discussion). *Journal of the Royal Statical Society (Series A)* **183**, 691–736.

 