Multiple Collection Estimation of Population Size

Lawrence D. Brown
Department of Statistics, University of Pennsylvania
Philip Ernst
Department of Statistics, Rice University
and
Larry Shepp
Department of Statistics, University of Pennsylvania*

November 30, 2016

Abstract

The classical capture-recapture model involves two independent surveys and hence (usually) two collectors ($m = 2$). Modern statistical applications suggest the desirability of considering models with more than two collectors ($m \geq 3$). We generalize classical results to this case and derive further results for all m. Building from Darroch (1958), the general problem is explicitly formulated, minimal sufficient statistics are described, and an equation for the maximum likelihood estimate is explicitly given. We then provide exact conditions for the existence and uniqueness of the MLE. We also derive a discrete version of the Cramer-Rao lower bound and prove the asymptotic efficiency of the MLE.

Asymptotics for this problem are studied in two separate domains. In the first, the number of collectors tends to infinity while the population size is fixed. In the other asymptotic domain m is fixed while the population size and the size of samples obtained by each collector is allowed to grow. Darroch (1958) studied asymptotics for this situation under specific conditions on the sample sizes. We extend his results so as to establish necessary and nearly sufficient conditions in this setting for asymptotic normality of the MLE. This enables practical asymptotic statistical inference as well as providing a simple, intuitive diagnostic for the validity of this inference.

Keywords: Capture-recapture; maximum likelihood estimator; asymptotic statistics

1 Introduction

Capture-recapture methods for estimation of population size have been studied for over a century. The classical model for capture-recapture involves two independent surveys, and hence (usually) two "collectors," as mentioned in the works of Good (1953), Darroch (1958), Darroch (1959), Mao and Lindsay (2007), Seber (1982), Darroch et al. (1993). Darroch et al. (1993) developed the triple-system estimation framework, adding an additional re-capture to the dual system, resulting in a 2^3 table of counts.

Since Darroch et al. (1993), there has been little literature rigorously analyzing an estimation framework with arbitrarily many samples or "re-captures." But such studies continue to be used in practical settings, and indeed it is now feasible to conduct studies with arbitrarily many collectors.

*Deceased April 23, 2013

Consider Amazon Mechanical Turk (MTurk), a website owned by Amazon that allows researchers to achieve a multiple collection framework (Chandler and Kapelner (2013)). Or, for a recent study involving a limited number of collectors, see Bales et al. (2015). Further remarks are made about this study in Section 2.5.

Our main approach is that of maximum likelihood estimation (MLE). For early use of this traditional approach see Craig (1953) and Darroch (1958). Craig (1953) considers a single collector who collects butterflies in order to estimate the population of butterflies. He discusses the MLE and moment estimators and derives expressions for the probability distribution of the frequency counts. The single collector question is a special case of our multiple collection framework in terms of a sequence of collectors, operating independently, who each collect only one butterfly. Throughout this work, we assume that a predetermined list for the size of each collector's sample is given at the start. In Section 2, minimal sufficient statistics for this model are described, the MLE is derived, and its existence is established. In Theorem 2 we prove our key result on the uniqueness of the MLE.

In Section 3, asymptotic properties of this MLE are studied in two separate asymptotic regimes. Section 3.1 concerns an asymptotic regime in which the number of collectors grows while the population size remains fixed. Asymptotic results for boundary optimization, strong consistency of the MLE, and convergence rates are presented. Results for this regime may be appropriate for studies involving large numbers of collectors such as those that might be conducted by MTurk. Section 3.2 concerns the more traditional asymptotic regime in which the number of collectors remains fixed while the population size and the collection sizes grow larger. Darroch (1958) studied such a regime when the collection sizes grow proportionally to the true sample size. He showed the MLE is then asymptotically normal, although his proof is logically incomplete because he assumed asymptotic normality of the total collection, rather than providing a proof. Our main result, Theorem 9 in this section, gives a complete proof under a mild condition for speed of growth of collection size relative to population size. Our condition is necessary and sufficient for asymptotic normality when the ratio of the largest collection size to the full population size is bounded below one. Along the way to this result we prove in Theorem 8 a central limit theorem for the total collection size under this condition. This result may be viewed as an extension of standard results on asymptotic normality of the entries in a 2^m contingency table under the usual independence assumption. Section 3.3 describes statistical inference based on the asymptotic distribution established. Darroch (1958) did not explicitly consider asymptotic efficiency of the MLE. This is the topic of our Section 4. In this section we derive a discrete version of the Cramer-Rao lower bound and prove the MLE to be asymptotically efficient.

2 Model formulation

We consider a class of statistical models generalizing the very useful, classical capture-recapture model. A simple paradigm that can help understand the basic general problem involves i-tunes listeners and distinct i-tunes songs. An i-tunes collection contains θ songs. θ is the unknown parameter. There are m listeners (collectors). Each collector's songs are played on an i-shuffle player. Each time the i-shuffle player plays a song it chooses with equal probability one song. Listeners record the names of the different songs they have heard, but not the number of times they have listened to songs nor the number of times they have heard each song. Let $n_1, \ldots, n_m$ be the number of distinct songs heard by each listener. Assume each $n_i \geq 1$. The goal is to estimate θ. The i-tunes setup is appealing because it guarantees the desired independence and equal probability of sampling among songs.

2.1 The "listening times" estimator

In this section and in Section 2.4, we describe two very differently motivated estimators of θ. Surprisingly, the two methods lead to the same estimator (see Remark 1). In the current section, we may think of the songs as being sampled with replacement at discrete times $t = 1, 2, \ldots$ However, each listener only records the total number of distinct songs he or she has heard (and not the time it took to collect them). Thus, the data recorded by each listener is the same as in the general model described in the previous paragraph. Let t_{ij}, $i = 1, ..., m; j = 1, ..., n_i$ denote the time at which listener i hears a song he or she has not heard before. Clearly, $t_{i,1} = 1$ and $t_{i,j+1} > t_{ij}$, $j = 1, ..., n_i - 1$. We emphasize that these times exist but that they are not part of the data gathered by the listeners. Let $\delta_{i,r} = t_{i,r} - t_{i,r-1}$ denote the additional time it takes listener i to hear his/her rth song, after hearing the $(r-1)th$ one. $\delta_r = \delta_{ir}$ is a geometric random variable with $p = 1 - \frac{r-1}{\theta}$ and

$$\mathbb{E}\left[\delta_r\right] = \frac{\theta}{\theta - r + 1}.$$

The expected time for a listener to collect all θ songs in the i-tunes collection is:

$$\sum_{r=1}^{r=\theta} \frac{\theta}{\theta - r + 1} = \theta \left(\frac{1}{\theta} + \frac{1}{\theta - 1} + ... + 1 \right) \approx \theta \, log(\theta). \tag{1}$$

Let $T(i)$ be the time it takes for collector i to collect n_i distinct songs. We compute the expected time $\mathbb{E}\left[T(i)\right]$ using the same logic as in (1).

$$\begin{aligned} \mathbb{E}\left[T(i)\right] = \theta \sum_{r=1}^{n_i - 1} \frac{1}{\theta - r + 1} &= \theta \left(\frac{1}{\theta} + \frac{1}{\theta - 1} + \frac{1}{\theta - 2} + \ldots + \frac{1}{\theta - n_i + 1} \right) \\ &\approx \quad \theta \left(log(\theta) - log(\theta - n_i) \right) = \theta log \left(\tfrac{\theta}{\theta - n_i} \right). \end{aligned} \tag{2}$$

In order to use the above approximation, the value of $\{n_i\}$ cannot equal θ.

Using equation (2), the total time spent by the m collectors to get $n_1, \ldots, n_m$ songs is calculated as:

$$\mathbb{E}\left[\sum_{i=1}^{m} T(i) \right] \approx \theta \sum_{i=1}^{m} log \left(\frac{\theta}{\theta - n_i} \right). \tag{3}$$

To simplify notation, let the random variable H be the total number of distinct songs heard by all listeners. When H is realized, it takes on the value h. Following in the spirit of Newman and Shepp (1960), we can also think of the collection of listeners as a single entity. We call that entity the "supercollector." The supercollector thus collects H different songs and listens for a total time of $T^* = T_1 + T_2 + \ldots + T_m$. Applying the approximation from (2) to T^* yields:

$$\mathbb{E}\left[T^*\right] \approx \theta log \left(\frac{\theta}{\theta - h} \right). \tag{4}$$

Since $T^* = T_1 + T_2 + \ldots + T_m$, we equate the right sides of the approximations in (3) and (5) to obtain:

$$\theta\left(log\left(\frac{\theta}{\theta-n_1}\times\frac{\theta}{\theta-n_2}\cdots\frac{\theta}{\theta-n_m}\right)\right)=\theta log\left(\frac{\theta}{\theta-h}\right), \tag{5}$$

which we rewrite as:

$$\theta\left(log\left(\frac{\theta}{\theta-n_1}\times\frac{\theta}{\theta-n_2}\cdots\frac{\theta}{\theta-n_m}\right)\right)-\theta log\left(\frac{\theta}{\theta-h}\right)=0. \tag{6}$$

Any solution of equation (6) defines the "listening times" estimator. For $m = 2$, after we cancel the $\theta' s$, exponentiate, and cancel the $\theta' s$ again, the listening times estimator becomes

$$\hat{\theta}=\frac{n_1 n_2}{n_1+n_2-h}. \tag{7}$$

This is the hypergeometric estimator found in the capture-recapture literature; for example, see Yip et al. (1995).

A simple calculation for $m = 3$ shows that the overlaps amongst collectors, $S_{\{1,2\}}, S_{\{1,3\}}, S_{\{2,3\}}$ in the notation of Section 2.2 below, are missing from our estimator! At first glance, this result may appear counterintuitive. It would seem that the overlap between collectors should provide vital information that the listening times estimator lacks. Sections 2.2 and 2.3 resolve this issue.

2.2 The likelihood function

The n_j, $j = 1, ..., m$ are the number of different songs heard by the respective listeners. Conceptually, these are constants and are neither parameters nor part of the sample space. Define the sample space as: $\{(S_B) : B \subset \{1, ..., m\}\}$, where each S_B counts the number of songs heard by the listeners in B and by no other listeners. For example, if $m = 3$, the vector (S_B) would consist of the following elements: $S_{\{1\}}, S_{\{2\}}, S_{\{3\}}, S_{\{1,2\}}, S_{\{1,3\}}, S_{\{2,3\}}, S_{\{1,2,3\}}$. Thus, the S_B partition the sample space. The parameter of interest, as before, is θ, the total number of songs in the i-tunes collection.

Recall that H is the total number of distinct songs heard by all listeners. Thus $H = \sum_B S_B$. Following from the initial model formulation, we consider the distribution of the vector (S_B), where $B \in \mathcal{P}(\{1, ..., m\}) \setminus \emptyset$.

Fixing $\mathbf{X} = (n_1, \ldots, n_m)$, the distribution of the vector (S_B) is:

$$\mathbb{P}\left((S_B) \mid \mathbf{X}=(n_1,\ldots,n_m)\right) = \frac{\begin{pmatrix}\theta\\ (S_B),\ \theta-\sum_B S_B\end{pmatrix}}{\prod_j \binom{\theta}{x_j}}. \tag{8}$$

In equation (8), the numerator counts the number of ways the distribution of songs gives rise to a particular vector (S_B). The denominator counts the number of ways to distribute songs giving rise to $n_1, n_2, ..., n_m$. Out of θ possible songs, the first collector hears n_1 different songs, the second collector hears n_2 different songs, and so on. Ignoring some constants, equation (8) simplifies to the following likelihood function:

$$L(\theta)=\frac{\theta!}{h!(\theta-h)!}\prod_{j=1}^{m}\frac{(\theta-n_j)!n_j!}{\theta!}. \tag{9}$$

2.3 A sufficient statistic

We now show that $H = h$ is a sufficient statistic.

Theorem 1. *$H = h$ is a sufficient statistic.*

Proof. For the multinomial coefficient in equation (8), note that:

$$\binom{\theta}{(S_B), \theta - \sum_B S_B} = \binom{\theta}{h}\binom{h}{(S_B)}. \tag{10}$$

Hence, equation (8) can be rewritten as:

$$\mathbb{P}_\theta\left((S_B)\right) = \frac{\binom{\theta}{h}}{\prod\limits_{j=1,\dots,m} \binom{\theta}{n_j}} \times \binom{h}{(S_B)}. \tag{11}$$

By the Neyman factorization theorem, H is a sufficient statistic. □

The exact distribution of H is:

$$\mathbb{P}_\theta\left(h\right) = \frac{\dbinom{\theta}{h}}{\prod\limits_{j=1,\dots,m} \dbinom{\theta}{n_j}} \times \sum_{(S_B) s.t. \sum_B S_B = h} \binom{h}{(S_B)}. \tag{12}$$

The computation of the last summation is needed to calculate the exact (non-asymptotic) distribution of proposed estimators. Such calculations are performed in Section 2.5.

2.4 The maximum likelihood estimator (MLE) and its uniqueness

We now derive the MLE for the multiple collection framework and show it exists and is unique except in boundary cases. Let $\mathbb{P}_\theta\left(h\right)$ denote the probability that h is observed given that θ is the true parameter. Consider the following ratio:

$$\frac{L(\theta)}{L(\theta-1)} = \frac{\prod\limits_{i=1}^{m}(\theta - n_i)}{\theta^{m-1}(\theta - h)} \triangleq \rho(\theta; h). \tag{13}$$

In solving for the MLE, we first find a continuous solution to $\rho(\theta; h) = 1$. Thus, $\theta^* \in (h, \infty)$ is a solution to the following equation:

$$\prod_{i=1}^{m}(\theta - n_i) - \theta^{m-1}(\theta - h) = 0. \tag{14}$$

Remark 1. *Rather surprisingly, the listening times estimator in (6) is the same as the MLE estimator in (14).This follows from simple algebra. Recall equation (5) , which states:*

$$\theta\left(log\left(\frac{\theta}{\theta - n_1} \times \frac{\theta}{\theta - n_2} \cdots \frac{\theta}{\theta - n_m}\right)\right) = \theta log\left(\frac{\theta}{\theta - h}\right) \tag{15}$$

After canceling the θ's and exponentiating, we can rewrite the above expression as:

$$\frac{\theta^m}{\prod_{i=1}^{m}(\theta - n_i)} = \frac{\theta}{\theta - h}.$$

Dividing by a factor of θ on both sides and subtracting the right hand side from the left hand side yields equation (14).

The MLE $\hat{\theta}$ is one of the two integers adjacent to θ^* (or both). Before proceeding, we note that we are not the first to derive this equation for the MLE. In fact, Darroch (1958) writes this same equation for the MLE (it is his equation (1) in Section 3.1), but he does not prove that the MLE has a unique solution in the feasible region $\theta > h$. We now prove the key result that equation (14) has a unique root in the domain $\hat{\theta} \geq h$.

Theorem 2. *Consider equation (14). The following three statements about the MLE then hold:*

(a) If $h < \sum_{i=1}^{m} n_i$, the equation has a unique root in [h, ∞) and the MLE $\hat{\theta}$ is one of the two integers adjacent to θ^* (If θ^* is an integer then $\hat{\theta}$=θ^*.)

(b) If $h = \sum_{i=1}^{m} n_i$, the equation has no root in [h, ∞) and the MLE does not exist.

(c) If $h = n_j$ for some j then $h = n_j$ is the unique root in [h, ∞) and $\hat{\theta} = h = n_j$.

Proof. We first look at the case in which $h > n_i$ for all $i = 1, ..., m$. We consider the rational function (13) which is $R(\theta) = \rho(\theta; h)$. Since $h > n_i \; \forall \, i$, neither $\theta = 0$ nor $\theta = h$ is a solution to equation (14). Hence, the solutions of (14) are identical to the solutions of $R(\theta) = 1$. We must show that this equation can have at most one root larger than h and that such a root exists if and only if $h < \sum_{i=1}^{m} n_i$.

First, to account for the fact that several collectors might collect the same number of songs, we group terms as follows:

$$\prod_{i=1}^{m} (\theta - n_i) = \prod_{j=1}^{p} (\theta - y_j)^{p_j},$$

where $y_j = n_i$ for exactly p_j different indices i. Hence $\sum_{j=1}^{p} p_j = m$ and p is the number of distinct values of n_i. We then rewrite $R(\theta)$ as:

$$R(\theta) = \frac{\prod_{j=1}^{p} (\theta - y_j)^{p_j}}{\theta^{m-1}(\theta - h)}.$$

This function has p roots in the range $\theta \in (0, h)$, namely $y_1, \ldots, y_p$. The function is also differentiable there. By Rolle's theorem, $R'(\theta)$ has at least one root in the interval strictly between each successive value of y_j. Hence $R'(\theta)$ has at least $p - 1$ roots in $(0, h)$. Explicitly calculating $R'(\theta)$ and simplifying, we arrive at:

$$R'(\theta) = \frac{\prod_{j=1}^{p} (\theta - y_j)^{p_j - 1} \left[\theta(\theta - h) \sum_{j=1}^{p} p_j \prod_{l \neq j} (\theta - y_l) - (m\theta - (m-1)h) \prod_{j=1}^{p} (\theta - y_j) \right]}{\theta^m (\theta - h)^2}. \tag{16}$$

Since the first term inside the brackets in the numerator is zero when $\theta = 0$ or $\theta = h$, but the second term is not, neither $\theta = 0$ nor $\theta = h$ is a root of the numerator. Thus, $R'(\theta)$ is a well-defined function except as $\theta = 0, h$. The roots of $R'(\theta)$ in $(0, h)$ are then found by setting the numerator equal to zero. The solutions are y_j for each index j such that $p_j > 1$ and those of equation (17) below:

$$\theta(\theta - h) \sum_{j=1}^{p} p_j \prod_{l \neq j} (\theta - y_l) - (m\theta - (m-1)h) \prod_{j=1}^{p} (\theta - y_j) = 0. \tag{17}$$

For sake of clarity, we note that the left hand side of equation (17) is the expression in brackets in the right hand side of equation (16).

We wish to know how many roots there are in Equation (17). From Rolle's theorem, we know that at least $p-1$ of these roots are less than h. Again, by Rolle's theorem, the roots lie strictly in the intervals between the y_j, $j = 1, \ldots, p$, and they are not equal to any of these values.

We observe that in equation (17) the coefficients of θ^{p+1} cancel, so this equation is a polynomial equation of degree at most p. The coefficient of θ^p is (after some simplification) $\sum_{i=1}^m n_i - h$. Hence equation (17) has at most p roots if $h < \sum_{i=1}^m n_i$ and at most $p-1$ roots if $h = \sum_{i=1}^m n_i$. In the latter case we have shown that all of these roots must be less than h. This proves case (b).

For case (a), note that the above discussion shows that $R'(\theta)$ has at most one root in the range $\theta > h$. In this case, however, we observe that the leading term is positive and hence that as $\theta \to \infty$ the derivative of (17) is positive. It is easy to see that the function $R(\theta)$ approaches 1 in the limit as $\theta \to \infty$. At $\theta = h$ the function $R(\theta)$ explodes to positive infinity. Since we can have at most one local maximum or local minimum for $\theta > h$ and the horizontal asymptote in (17) is 1, $R(\theta) = 1$ either has no solutions for $\theta > h$, in which case the derivative is negative in this region, or it has exactly one solution. In the latter case, the derivative will eventually be positive. This proves case (a). For case (c), we note that if $h = n_j$ then $\theta^* = n_j$ is a solution to equation (14). Any additional solutions must satisfy

$$\frac{\prod_{i \neq j}^m (\theta - n_j)}{\theta^{m-1}} = 1.$$

However, in the range that $\theta > n_i$ for all $i = 1, ..., m$ the numerator is strictly smaller than the denominator and hence there are no solutions. Finally, $h = n_j$ implies that $n_j = \max_i(n_i)$, so $\theta > h$ implies $\theta > n_i \ \forall i$. This proves case (c). □

2.5 Calculation of the MLE

Having proved uniqueness and existence of the MLE, we turn now to a simple calculation of the MLE. For a table of values from which Figure 1 is drawn, see Appendix A. The results were produced by computing the roots of equation (14) and searching for the unique root satisfying the conditions of Theorem 2. Figure 1 below, utilizing Table 1 as its input, displays a bar plot of the distribution of the MLE for $m = 3, n_1 = 100, n_2 = 75, n_3 = 50, \theta = 200$.

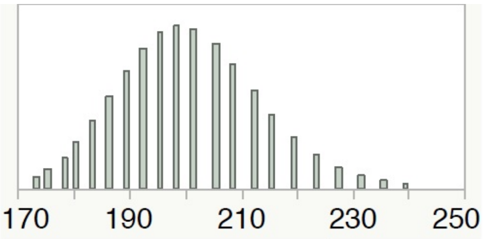

Figure 1: Bar plot of the distribution of the MLE for $m = 3, n_1 = 100, n_2 = 75, n_3 = 50, \theta = 200$. The x-axis is the value of the MLE and the y axis is $\mathbb{P}\left(\hat{\theta}\right)$.

The results shown in Figure 1 are encouraging. The distribution is unimodal. The maximum probability occurs at $\hat{\theta}$=199 (not every value of $\hat{\theta}$ is possible, and in particular $\hat{\theta} = 200$, the actual value of θ, is not a possible value). The values of h that lead to the values of $\hat{\theta}$ shown in the plot are $h = 144, 145, ..., 164$. Values above and below this range are possible but cumulatively amount to a total probability of less than .008. Apart from the fact that it is a discrete distribution, one may note that the plot has a roughly normal shape, though closer inspection shows it is slightly skewed to the right. The mean and variance of this distribution are (approximately) 200.3 and 151.7, respectively. The asymptotic values of these moments as calculated in Theorem 9 are 200 and 150, respectively. These asymptotic values agree quite well with the exact values corresponding to the plot.

Remark 2. *As mentioned in Section 1, Bales et al. (2015) recently reported a multiple collection study. Using the data provided in their manuscript, equation (14) gives a value for the MLE of 12,214. Interestingly, this value is bigger than any estimate reported there (the largest estimate reported by Bales et al. (2015) is 11,418). The standard deviation of our MLE is approximately 676.5, which is somewhat smaller than the standard deviation values reported for their various models. The article discusses possible correlations between the capture lists and produces analyses designed to accommodate possible correlations. Bales et al. (2015) implements a different type of multiple collector model that is closely related to the model we describe. In the Bales et al. (2015) model the total population size, θ, as well as the individual collection sizes are Poisson random variables. The estimator they implement could be viewed as resulting from an E-M estimation of the total in the missing cell corresponding to unobserved members of the population. Asymptotic normality of their estimator holds in their model as in ours, though the regularity conditions vary slightly and the asymptotic variances are different. Under the natural independence assumption, their estimator will be the same as ours, but the standard errors will be different. See also Van der Heijden et al. (2012) for further examination and elaboration of such a model. The formulation in Bales et al. (2015) has the advantage that it extends in a fairly natural way to situations that involve correlation among the individuals in different collections, and they consider such extensions. As Darroch (1958) remarks, our multiple hypergeometric framework could in principle also be extended to such settings. But as he writes, the extension seems complicated and neither he nor we pursue such an extension.*

3 Asymptotics

In this section, we consider asymptotic regimes for the multiple collection setting. Throughout our discussion, we assume that a predetermined list $\{n_i : i = 1, 2, \ldots\}$ of each collector's collection number is given at the start. We further assume that all $n_i \geq 1$. In this first setting, which has never been previously discussed in the literature, we keep the population parameter θ constant and analyze how the MLE estimator behaves as we let the number of collectors $m \to \infty$. In the second setting, which was first, albeit incompletely, discussed by Darroch (1958), we let both the population $\theta \to \infty$ and the amounts collected $n_i \to \infty$, $i = 1, \ldots, m$, while keeping the number of collectors m fixed.

3.1 Asymptotic scenario I

Consider the scenario in which the population θ remains constant, while the number of collectors $m \to \infty$. The i^{th} collector has gathered n_i distinct songs and the total number of songs collected by all collectors combined is h. The goal is to estimate θ, the total number of songs in circulation.

Lemma 3 (Boundary Optimization)**.** *Let $B > \max\{n_i\} + 1$ satisfy*

$$\sum_{i=1}^{m} \log(B - n_i) - (m-1)\log B = 0. \tag{18}$$

Then $h \leq B - 1$ implies $\hat{\theta} = h$. Further, define $M_0(\theta)$ as

$$M_0(\theta) = \left(1 - \frac{\log\theta}{\log(\theta+1)}\right)^{-1}. \tag{19}$$

Then $m > M_0(\theta)$ implies $\hat{\theta} = h$.

Proof. Expression (13) yields:

$$\log\frac{L(\theta)}{L(\theta-1)} = \sum \log(\theta - n_i) - \log(\theta - h) - (m-1)\log\theta. \tag{20}$$

Let $h < B - 1$ and $\theta_0 = h + 1$. Then $\log(\theta_0 - h) = 0$. Furthermore, the right hand side of (20) is negative since $\theta_0 < B$ and $\sum\log(\theta - n_i) - (m-1)\log\theta$ is increasing in θ and takes the value 0 when $\theta = B$. It follows that:

$$\log\frac{L(\theta_0)}{L(\theta_0-1)} < 0.$$

The proof of Theorem (2) guarantees that there will be one root, θ^*, of (14) in $[h, \infty)$ and that $\hat{\theta} = \lfloor\theta^*\rfloor$ and that $\theta^* < \theta_0$. This yields that $\hat{\theta} < h + 1$ and this implies that $\hat{\theta} = h$ since $\hat{\theta} \geq h$. For the second assertion of the lemma, note that:

$$\sum_{i=1}^{m}\log(h - n_i) - (m-1)\log h \leq \sum_{i=1}^{m}\log(h-1) - (m-1)\log h.$$

When $m > M_0(\theta)$ then

$$\sum_{i=1}^{m}\log(\theta - 1) - (m-1)\log\theta < 0.$$

Hence in this case $B > \theta + 1$. This implies that $h \leq \theta < B - 1$ and hence yields $\hat{\theta} = h$ by the first assertion of the lemma. □

3.1.1 Consistency

We proceed to show that the maximum likelihood estimator is consistent.

Theorem 4. *Assume θ is fixed and that $\{n_i\}$ is a fixed sequence with each $n_i \leq \theta$. Let θ be the true number of distinct songs, and let $\hat{\theta}_m$ be the MLE estimator with m collectors. Then*

$$\hat{\theta}_m \overset{a.s.}{\to} \theta \quad \textit{as} \quad m \to \infty.$$

Proof. By the conclusion of Lemma 3, there exists some N depending only on θ and n_i such that if $m \geq N$ then $\hat{\theta}_m = h$. We first note that

$$\mathbb{P}\left(\text{collectors do not get song } j\right) = \prod_{i=1}^{m}\left(1 - \frac{n_i}{\theta}\right).$$

Let $M(m) = \sum_{i=1}^{m} n_i$. Observe that on a particular try the probability of not picking the j^{th} song is $(1 - 1/\theta)$. Arithmetic manipulation shows for each i that

$$\left(1-\frac{n_i}{\theta}\right) \le \left(1-\frac{1}{\theta}\right)^{n_i}.$$

Thus,

$$\prod_{i=1}^{m}\left(1-\frac{n_i}{\theta}\right) \le \left(1-\frac{1}{\theta}\right)^{M(m)}.$$

It follows that

$$\mathbb{P}\left(\text{there exists a song that is not collected}\right) \le \theta(1-1/\theta)^{M(m)} \to 0. \tag{21}$$

Equivalently,

$$\mathbb{P}\left(H<\theta\right) \le \theta(1-1/\theta)^{M(m)} \to 0, \tag{22}$$

where the limit is taken as the number of collectors $m \to \infty$. The conclusion of Lemma 3 combined with the inequality in (22) yield the conclusion of the theorem. □

3.1.2 Convergence rates

We would like to bound $\mathbb{P}\left(|H-\theta| \ge 1\right)$ as $M=\sum_{i=1}^{m} n_i$ becomes large (or, equivalently, as the number of collectors becomes large).

Theorem 5 (Convergence Rate)**.** *Assume θ is fixed. As $M=\sum_{i=1}^{m} n_i$ becomes large,*

$$\mathbb{P}\left(H=\theta\right) \ge e^{\frac{-\theta^2+\theta}{M+1}}. \tag{23}$$

Proof. First, note that

$$\mathbb{P}\left(|H-\theta| \ge 1\right)$$

is bounded above by the probability that at least one song is not collected in M trials (the total number of songs collected is higher because each collector throws out duplicates). Thus:

$$\begin{aligned}\mathbb{P}\left(|H-\theta| \ge 1\right) &\le \mathbb{P}\left(\text{at least one song is not chosen if picking } M \text{ songs with replacement}\right)\\ &= 1-\mathbb{P}\left(\text{all songs chosen}\right).\end{aligned}$$

To estimate the probability that all songs are chosen, we consider "distributing" M draws to θ songs in a model in which there is sampling with replacement. We look at the ratio of the number of ways to distribute M picks so that each song gets at least one pick. To calculate this with replacement, we use the "stars and bars" method of counting that is germane to elementary combinatorics. Thus, sampling with replacement, we obtain:

$$P \triangleq \mathbb{P}\left(\text{all songs chosen}\right) = \frac{\binom{M-1}{\theta-1}}{\binom{M+\theta-1}{\theta-1}} = \frac{(M-1)\cdots(M-\theta+1)}{(M+\theta-1)\cdots(M+1)}. \tag{24}$$

Since we are trying to find an upper bound to $1-P$, we want a lower bound for P. We observe that P is a product of θ terms of the type in which the numerator is θ less than the denominator. Hence the smallest term in the product is the one with the smallest denominator and

$$P \ge \left(\frac{M-\theta+1}{M+1}\right)^{\theta-1} = \left(1-\frac{\theta}{M+1}\right)^{\theta-1}.$$

$\log(1+x) < x$ for $x < 0$. Hence,

$$\log P \geq (\theta - 1) \log \left(1 - \frac{\theta}{M+1}\right) \leq (\theta - 1) \left(\frac{-\theta}{M+1}\right). \tag{25}$$

In particular,

$$P \geq e^{\frac{-\theta^2+\theta}{M+1}}.$$

Thus for sufficiently large M,

$$\mathbb{P}\left(H = \theta\right) \geq e^{\frac{-\theta^2+\theta}{M+1}}.$$

□

Corollary 6. *If $m > M_0(\theta)$ then we find that the number of collectors is indeed large enough in the settings of both Lemma 3 and Theorem 4. Thus,*

$$\mathbb{P}\left(\hat{\theta} = \theta\right) \geq e^{\frac{-\theta^2+\theta}{M+1}},$$

where M is as in Theorem 4, i.e. $M = \sum_{i=1}^{m} n_i$.

Proof. From Lemma 3, we have that $m > M_0(\theta)$ implies that $\hat{\theta} = h$, i.e. we are in a boundary optimization regime. Hence the following sets are equal: $\{H = \theta\} = \{\hat{\theta} = \theta\}$. Thus the lower bound on the probability is also the same. □

3.2 Asymptotic scenario II

3.2.1 The Distribution of H

In asymptotic scenario II, the population θ is large and the number of collectors remains bounded. In order to construct a suitable asymptotic stochastic formulation, we consider a triangular array of collection problems, indexed by $k = 1, ...$ The sample sizes in these problems are $\theta^{[k]}$ and the collection sizes are $n_i^{[k]}$, $i = 1, \ldots, m$. The number of collectors, m, is fixed throughout. Darroch (1958) established results for such a setting under the additional assumption that the ratios $c_i^{[k]} = n_i^{[k]}/\theta^{[k]}$ satisfy $c_i^{[k]} \to \gamma_i$, with $0 < \gamma_i < 1$. However, in many applications the values of c_i may be near 0. Darroch's results do not apply in such a situation. Indeed, as we will show, in such a case the limiting Gaussian distribution has a different order of variance from that in Darroch's setting. In the sequel, for convenience, we will throughout suppress the superscript $[k]$.

As a practical example, at its crudest level the dual system estimate in the 2000 U.S. census can be considered as a situation having two enumerations—the original census is the first enumeration. If one takes the original March 2010 figures (before additional adjustments were made) the census appeared to contain about 98.8% of the total population. (Hence $c_i \approx .988$.) The follow-up (recapture) survey consisted of roughly .3% of the population (Hence $c_2 \approx .003$.) The total population was of course quite large, with $\theta > 280,000,000$ so one might hope that any asymptotic theory would yield good approximations for the distribution of the MLE.

However, the actual analysis was carried out within substrata having much smaller total populations, so one cannot count on theory involving total sample sizes in the range of millions. Our main point in discussing this type of application is to emphasize that theory is needed that applies to values of c_i very near 0. At the same time, we do not in any way wish to suggest that the theory in the present paper would completely suffice for realistic applications of this sort. There are many important special features and practical subtleties of any dual system application, especially of the coverage evaluation program of the recent U.S. censuses. See Citro et al. (2003) (in particular, Chapter 6) for detailed discussion of many of these in connection with the 1990 and 2000 censuses. For additional discussion of dual system estimation see Brown et al. (1999), Brown and Zhao (2008), Darroch et al. (1993), Rao (2003) (especially, Chapter 3.4), and Zaslavsky (1993).

The first theorem of this section proves that H is asymptotically normal. For this purpose, consider a $2{\times}2 \times ... \times 2 = 2^m$ way contingency table under standard assumptions of independence (e.g., see Bishop et al. (1986)). Let θ be the total number of observations in the table and let $X_{\mathbf{i}}, \mathbf{i} \in \prod_1^m \{0,1\}$ denote the entries in the cells of the table. Let n_i denote the marginal totals. Note that the $X_{\mathbf{i}}$ correspond to the numbers $S_B, B \subset \{1, ...m\}$ (recall that the S_B were first defined in Section 2.1) with $\mathbf{i} = \mathbf{I}_B$. Let $X_{\mathbf{0}} = S_\emptyset = \theta - H$. We are interested in $H = \theta - X_{\mathbf{0}}$, namely the number of songs collected. The exact formula for the mean was given in Darroch (1958) as:

$$\mathbb{E}\left[H\right] = \theta - \frac{\prod_i (\theta - n_i)}{\theta^{m-1}}, \tag{26}$$

Based on expressions in Darroch (1958), the variance is easily derived as:

$$\begin{aligned}\sigma_H^2 &= \theta(\theta-1)\prod_{j=1}^{m}\left(1-\frac{n_j}{\theta}\right)\left(1-\frac{n_j}{\theta-1}\right) + \theta\prod_{j=1}^{m}\left(1-\frac{n_j}{\theta}\right) - \theta^2\prod_{j=1}^{m}\left(1-\frac{n_j}{\theta}\right)^2 \\ &= \theta\left(\prod_{j=1}^{m}\left(1-\frac{n_j}{\theta}\right) - \prod_{j=1}^{m}\left(1-\frac{n_j}{\theta}\right)^2\right) - \prod_{j=1}^{m}\left(1-\frac{n_j}{\theta}\right)n_j. \end{aligned} \tag{27}$$

Without loss of generality, we assume that

$$n_1 \geq ... \geq n_m. \tag{28}$$

Further, assume that

$$\text{limsup}\frac{n_1}{\theta} < 1. \tag{29}$$

Under these assumptions, it is straightforward to see that the expression in (27) has exact order $\frac{n_1 n_2}{\theta}$.

The proof of normality of H in Theorem 8 involves a chain of hypergeometric random variables. It requires an extension of the classical hypergeometric limit theorem proven in Lemma 7 below. Let us consider a hypergeometric random variable. Fix a rational number $0 < p < 1$ and $q = 1 - p$. Let the sample space contain θp red balls and θq black balls for a total of θ balls. We sample n times from this space, at random and without replacement. Let the number of red balls drawn be denoted as k. Theorems 1 and 2 of Pinsky (2003) prove asymptotic normality for this hypergeometric random variable when p is fixed and either $n/\theta \to t \in (0,1)$ or $n^3/\theta^2 \to 0$. For the following Lemma, the parameters p and n depend on θ as $\theta \to \infty$ The Lemma generalizes Pinsky's result so as to provide a necessary and sufficient condition for normality. The proof begins along the lines of Pinsky's argument, but then includes additional steps (we need only to state and prove sufficiency, as Pinsky establishes the conditions of Lemma 7 to be necessary for asymptotic normality.) See Nicholson (1956) for further information relevant to the standard hypergeometric setting in which p and n/θ are fixed constants in (0,1).

For convenience, let $x_k := \frac{k-np}{\sqrt{npq}}$. We now prove Lemma 7.

Lemma 7. *k is asymptotically normal with mean np and variance $npq(1-t)$ when $p = p(\theta)$ satisfies $\theta p \to \infty$ and $\theta q \to \infty$ and where $np \to \infty$.*

Proof. It suffices to assume that either:

$$p = p(\theta) \to p_0 \in (0,1) \tag{30}$$

or

$$p \to 0 \tag{31}$$

and

$$n \leq \theta p \text{ with } np \to \infty \tag{32}$$

and either

$$\theta p \to t \in (0,1) \tag{33}$$

or

$$\theta p \to 0. \tag{34}$$

(Note that the hypergeometric problem is invariant to relabeling of rows and columns, so there is no loss of generality in assumng $n \leq \theta p$.) Pinsky (2003) shows via a sequence of algebraic simplifications that the probability function for k can be written as

$$\mathbb{P}\left(k; p, n, \theta\right) = B(k; n, p)R(k; p, n, \theta). \tag{35}$$

(we have added the argument "p" to Pinsky's notation, since p now depends on θ). Here, B denotes the usual binomial probability function and R is given in equation (2) of Pinsky (2003) as:

$$R = \frac{\prod_{j=1}^{k-1}\left(1 - \frac{j}{\theta p}\right)\prod_{j=1}^{n-k-1}\left(1 - \frac{j}{\theta q}\right)}{\prod_{j=1}^{n-1}\left(1 - \frac{j}{\theta}\right)}. \tag{36}$$

We are interested in $x = O(1)$, which implies $k = np + O(\sqrt{np})$. Under condition (30), the DeMoivre-Laplace CLT establishes that

$$B(k; n, p) \sim \frac{1}{\sqrt{npq}}\phi(x), \tag{37}$$

where $\phi(x)$ is the Gaussian probability density function. In addition, $\liminf(q) > 0$ because of the first part of (32) along with (30) and (31). The limiting expression in (37) also holds under (31) by the binomial to Poisson limit theorem together with the Poisson CLT. Under condition (33), Pinsky shows that for $x = O(1)$,

$$R \sim \frac{1}{\sqrt{1-t}}\exp\left(\frac{-tx^2}{2(1-t)}\right). \tag{38}$$

It now remains to check that $R \sim 1$ under (34). First, we take the logarithm of both sides of (36). The first term on the right side becomes

$$\log \prod_{j=1}^{k-1} \left(1 - \frac{j}{\theta p}\right) = \sum_{j=1}^{k-1} \log \left(1 - \frac{j}{\theta p}\right) \sim \theta p \int_0^{k/\theta p} \log(1-v) dv, \tag{39}$$

since the summation is the Riemann sum for the integral that follows. Note that $k = np + O(\sqrt{np})$. Under (32) and (34), $n \leq \theta p = o(\theta)$, allowing the value of k in the limit in the sum and integral in equation (39) to be satisfactorality approximated as np. Making a substitution of $\delta = x - np$, we proceed to evaluate the integral as:

$$\begin{aligned} & \log \prod_{j=1}^{k-1} \left(1 - \frac{j}{\theta p}\right) \sim \theta p \left[w \text{log} w - w\right] \Big|_{1-k/\theta p}^{1} \\ = \quad & -\theta p \left(-1 - (1 - (np+\delta)/\theta p) \log\left(1 - (np+\delta)/\theta p\right) + 1 - (np+\delta)/\theta p\right) \\ = \quad & \theta p \left((1 - n/\theta)\text{log}(1 - n/\theta) - n/\theta - x/\theta p\right) + o(1). \end{aligned}$$

The second and third terms of the logged version of equation (36) can be evaluated similarly. Combining these results together yields:

$$\log R = \left((1-n/\theta)\text{log}(1-n/\theta) - n/\theta\right)\left(\theta p + \theta q - \theta + \delta - \delta\right) + o(1) = o(1).$$

Thus, under (34), $R \sim 1$, our desired result. □

We are now ready to prove Theorem 8.

Theorem 8. *Assuming the conditions in (28) and (29) hold, and, further, that*

$$\frac{n_1 n_2}{\theta} \to \infty, \tag{40}$$

then H is asymptotically normal.

Proof. The proof is by induction on m. The case $m = 1$ is included in the statement of the theorem since the condition in (40) does not apply when $m = 1$. When $m = 1$, $X_{\mathbf{0}} = \theta - n_1$, a constant. Consider the collectors arriving in sequential order. The first collector ($j = 1$) collects n_1 songs and fails to collect $\theta - n_1$ songs. The second collection can be viewed as a standard hypergeometric random variable, with the second collector choosing n_2 songs without replacement from θ possible songs with $\theta - n_1$ not seen by observer $j = 1$ and n_1 seen by observer $j = 1$. Let X_{01} be the number of songs collected by the first collector.

$$\mathbb{E}\left[X_{01}\right] = n_2 \left(1 - \frac{n_1}{\theta}\right)$$

and variance

$$var(X_{01}) = \frac{n_1}{\theta}\left(1 - \frac{n_1}{\theta}\right) n_2 \left(1 - \frac{n_2 - 1}{\theta - 1}\right).$$

In addition,

$$H = n_1 + X_{01}.$$

Hence, for $m = 2$,

$$\mathbb{E}\left[H\right] = n_1 + n_2 - \frac{n_1 n_2}{\theta}. \tag{41}$$

Since n_1 is a constant,

$$var(H) = var(X_{01}) = \frac{n_1}{\theta} n_1 \left(1 - \frac{n_1}{\theta}\right) n_2 \left(1 - \frac{n_2 - 1}{\theta - 1}\right). \tag{42}$$

By assumptions (29) and (40),

$$var(H) \to \infty \tag{43}$$

and

$$var(H) = O\left(\mathbb{E}\left[H\right]\right). \tag{44}$$

Using straightforward algebra, the ratio of the variance of H to its expectation is

$$\frac{var(H)}{\mathbb{E}\left[H\right]} \to \frac{1}{\frac{n_1+n_2}{n_1 n_2/\theta} - 1}.$$

Since $2n_2 \leq n_1 + n_2$, the denominator is bounded below by $2\theta/n_1 - 1 \geq 1$.

We now continue with the induction hypothesis on m that H is normal. Assume that for some $m \geq 2$ we have that the result (H is asymptotically normal) holds for all $m_I < m$, or simply that the theorem holds for $m - 1$ collectors $m_I = m - 1$. We can then view the new collector n_m as having a collection whose distribution is hypergeometric. There are h_I songs already seen by the first $m - 1$ collectors, and θ songs total. Let $X_{0(m-1)}$ be the number of songs collected by the mth collector that were not already collected by the previous collectors. Using the standard formulas for the hypergeometric distribution, we have:

$$\mathbb{E}\left[X_{0(m-1)}\right] = n_m \left(\frac{\theta - h_I}{\theta}\right),$$

and

$$var(X_{0(m-1)}) = n_m \left(\frac{\theta - h_I}{\theta}\right)\left(\frac{h_I}{\theta}\right)\left(\frac{\theta - n_m}{\theta - 1}\right).$$

$X_{0(m-1)}$ is a hypergeometric variable with parameters $\theta - h_I$ and n_m. However, $\theta - h_I = \mathbb{E}\left[X_{0(m-1)}\right]$ $+o(\mathbb{E}\left[X_{0(m-1)}\right])$. Hence if $(n_1 n_m)/\theta$ does not converge to 0, then $X_{0(m-1)}$ converges to normality by Lemma 7, with mean and variance asymptotically constant and independent of h_I. On the other hand, if $(n_1 n_m)/\theta \to 0$, then $X_{0(m-1)}$ is negligible relative to h_I. This completes the proof.

Remark 3. *The growth condition in (40) is necessary as well as sufficient. If (40) fails, then* $\liminf \mathbb{P}\left(H = k\right) > 0$ *for each finite k, and thus normality of H cannot hold. Then neither H nor the MLE can be asymptotically normal.*

□

Theorem 9. *Assume (28), (29), and (40). Then the MLE $\hat{\theta}$ is asymptotically normal with mean θ and variance:*

$$\mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \sim \left[\frac{\theta^{m-1}}{\prod_{i=1}^{m}(\theta - n_i)} + \frac{m-1}{\theta} - \sum_{i=1}^{m} \frac{1}{\theta - n_i}\right]^{-1} \tag{45}$$

Proof. We first show that

$$\hat{\theta}(\mathbb{E}_{\theta}\left[H\right]) = \theta. \tag{46}$$

From (26),

$$\mathbb{E}_{\theta}\left[H\right] = \theta\left(1 - \prod \frac{\theta - n_j}{\theta}\right).$$

We proceed to substitute the right-hand side of the above equation into equation (14), the defining equation for the MLE. Simple algebra shows that (46) holds.

We now turn our attention to the asymptotic variance. Under the assumptions of the current theorem,

$$\frac{\sigma_H}{\mu_H} \to 0, \tag{47}$$

where μ_H and σ_H denote the mean and variance of H as in Theorem (8). In the context of the theorem, $\mu_H \cong (n_1 n_2/\theta)$, where "$\cong$" denotes "is the exact order of." In addition, $\sigma_H^2 \cong (n_1 n_2/\theta)$. Hence,

$$\frac{\sigma_H}{\mu_H} \cong \left(\frac{n_1 n_2}{\theta}\right)^{-1} \to 0.$$

A consequence of (47) is that

$$H = \mu_H(1 + o_p(1)) \tag{48}$$

From standard calculus arguments, it follows that:

$$\hat{\theta} = \theta(1 + o_p(1)), \tag{49}$$

and for $\tilde{h}$ between μ_H and H,

$$\left[\frac{d\hat{\theta}}{dh}\right]_{h=\tilde{h}} = \left[\frac{d\hat{\theta}}{dh}\right]_{\mu_H} (1 + o_p(1)), \tag{50}$$

where the "little o" term holds uniformly in the choice of $\tilde{h}$ between μ_H and H. The above being established, we note that the defining equation for the MLE (14) can be equivalently stated as:

$$\sum \log(\theta - n_i) = (m-1)\log(\theta) + \log(\theta - h).$$

Implicit differentiation yields:

$$\left[\frac{d\hat{\theta}}{dh}\right]_{h=\tilde{h}} = \frac{1}{\hat{\theta}(\tilde{h}) - \tilde{h}} \left(\frac{1}{\hat{\theta}(\tilde{h}) - \tilde{h}} + \frac{m-1}{\hat{\theta}(\tilde{h})} - \sum \frac{1}{\hat{\theta}(\tilde{h}) - n_j}\right)^{-1}. \tag{51}$$

We can now employ a standard delta-method argument. Using the above equation, as well as the results in (48), (49), and (50), an exact Taylor expansion yields that for some $\tilde{h} = \tilde{h}(H)$ between μ_H and H,

$$\hat{\theta}(H) - \theta = \left[\frac{d\hat{\theta}}{dh}\right]_{\tilde{h}} (H - \mu_H) = \left[\frac{d\hat{\theta}}{dh}\right]_{\mu_H} (H - \mu_H)(1 + o_p(1)).$$

The theorem now follows from Theorem (8). Expression (51) evaluated at μ_H along with $\theta = \hat{\theta}(\mu_H)$ from (46) yields the asymptotic variance formula for $\hat{\theta}$.

□

Remark 4. *(45) is the same variance formula as in Darroch (1958), but he only established it under the special condition $n_i/\theta \to c_i \neq 0$.*

Remark 5. *In the special case considered by Darroch, the right-hand side (RHS) of (45) is $\asymp \theta$. In our more general setting,*

$$RHS(45) \asymp \frac{\theta^3}{max_{i,j}\{n_i n_j\}}.$$

The above documents a non-standard rate of convergence of the MLE. For example, suppose $n_i \sim k_i\theta^{a_i}, k_i > 0, a_1 \geq a_2 \geq ... \geq a_m$. Assume $1 < a_1 + a_2 < 2$. Then

$$Var(\hat{\theta}) \cong \theta^{3-(a_1+a_2)} > \theta.$$

The rate in this expression is worse than in the case discussed by Darroch, but the multiple collector estimator is still useful since $\hat{\theta}/sd(\hat{\theta}) \to 0$.

3.3 Inference for MLE

Theorem 9 enables reliable asymptotic inference in the form of confidence intervals for θ. Under the conditions of the theorem $\hat{\theta}$ is asymptotically normal and asymptotically unbiased in the sense that

$$\mathbb{E}\left[\hat{\theta} - \theta\right] = o(\mathrm{sd}(\hat{\theta})).$$

Its variance is given by (27). This variance can be consistently estimated in the obvious fashion, since $\hat{\theta}/\theta \xrightarrow{p} 1$. Hence under (40) and (29) asymptotically correct $1 - \alpha$ confidence intervals are given by $\theta \pm \mathbf{z}_{1-\frac{\alpha}{2}}\hat{\sigma}$, where

$$\hat{\sigma^2} = \hat{\theta}\left[\frac{1}{\prod_{i=1}^m \left(1 - \frac{n_i}{\hat{\theta}}\right) + m - 1 - \sum_{i=1}^m \frac{1}{1-\frac{n_i}{\hat{\theta}}}}\right]^{-1}. \tag{52}$$

Validity of the expression (45) is proven in Theorem 9 under the conditions (29) and (40).These two conditions can be empirically validated to within statistical error.

With respect to (40), note that (27) shows that:

$$\mathbb{E}\left[\sum_{i=1}^m n_i - H\right] = \sum_{i=1}^m n_i - \theta\left(1 - \prod_{i=1}^m \left(1 - \frac{n_i}{\theta}\right)\right). \tag{53}$$

where the above is the exact order of $\frac{n_1 n_2}{\theta}$. Conversely,

$$\sum_{i=1}^m n_i - h = O_P\left(\frac{n_1 n_2}{\theta}\right),$$

as can be seen from (53) and the formula for the variance in (27). Thus a reasonably large value of $\sum_{i=1}^m n_i - h$ is a reliable diagnostic for a large value of $\frac{n_1 n_2}{\theta}$, as required by (40). The remaining condition (29) requires n_i/θ to be bounded away from 1. Examination of the formulas for the variance of H and of $\hat{\theta}$ reveals that $\hat{\theta}/\theta \xrightarrow{p} 1$ whenever (40) holds and $\theta \to \infty$. Hence, $n_1/\hat{\theta}$ not too close to 1 is a reliable diagnostic for (29). Calculations like those in Table 1 in the Appendix show that (52) has coverage satisfactorily close to the nominal value in benign setting in which $\sum_{i=1}^m n_i - h$ is not small and $n_i/\hat{\theta}$ is not close to 1.

4 Asymptotic efficiency of the MLE

4.1 A variant of the Cramer-Rao lower bound

A standard bound for evaluating efficiency of the MLE is the Cramer-Rao lower bound; see Le Cam (1986). However, our problem of interest involves a discrete parameter space and integer valued data. Thus, we seek to prove a discrete analog of the Cramer-Rao lower bound to be used for any asymptotic regime. The key insight in doing so is to use first differences instead of derivatives, the latter of which was done in Papathanasiou (1993). Our restriction on h is, as before, that $\max_i\{n_i\} \leq h \leq \sum_{i=1}^{m} n_i$. Without loss of generality, we can formulate the problem in terms of the following restriction: $\max_i\{n_i\} \leq h \leq \theta$.

Let $f(h;\theta)$ denote the probability of obtaining h given that the true population total is θ. Consider the "score function" defined as follows:

$$U(h,\theta) \triangleq \begin{cases} \frac{f(h;\theta)-f(h;\theta-1)}{f(h;\theta)} & \max_i\{n_i\} \leq h < \theta \\ 1 & h = \theta \\ 0 & \text{otherwise.} \end{cases} \tag{54}$$

In this score function, $f(h;\theta) = 0$ unless $\max_i\{x_i\} \leq h \leq \theta$. Let $\Delta_x f(x,y) \triangleq f(x+1,y) - f(x,y)$ denote the first difference in the respective parameter/function input. The subscript below Δ indicates which of the two input variables should be incremented.

Theorem 10. *Let $E_\theta(\cdot)$ and Var_θ denote respectively expectation and variance when the true population is θ. Then we have a discrete analog of CRLB as follows:*

$$Var_\theta(\hat{\theta}) \geq \frac{(\Delta_\theta E_{\theta-1}(\hat{\theta}))^2}{E_{\theta-1}(-\Delta_\theta U(h,\theta-1))}. \tag{55}$$

Proof. The proof of Theorem 10 follows from ideas in Papathanasiou (1993) and the standard proof of the Cramer-Rao lower bound for continuous distributions; see section 8.2 of Young and Smith (2005). Note that for any two random variables X and Y it follows from Cauchy-Schwarz that

$$Cov(X,Y)^2 \leq Var(X)Var(Y). \tag{56}$$

Let $X = \hat{\theta}$ and $Y = U$. First, we manipulate the covariance term to show that

$$Cov(U(h,\theta),\hat{\theta}) = \Delta_\theta E_{\theta-1}(\hat{\theta}). \tag{57}$$

Let Q denote $\max_i\{n_i\}$. The proof of equation (57) is as follows:

$$\begin{aligned} E_\theta[U(h,\theta)\hat{\theta}_h] &= \hat{\theta}_\theta f(\theta;\theta) + \sum_{Q\leq h<\theta} \frac{f(h;\theta)-f(h;\theta-1)}{f(h;\theta)}\hat{\theta}_h f(h;\theta) \\ &= \hat{\theta}_\theta f(\theta;\theta) + \sum_{Q\leq h<\theta} (f(h;\theta)-f(h;\theta-1))\hat{\theta}_h \\ &= \sum_{Q\leq h\leq\theta} \hat{\theta}_h f(h;\theta) - \sum_{Q\leq h\leq\theta-1} \hat{\theta}_h f(h;\theta-1) \\ &= E_\theta(\hat{\theta}_h) - E_{\theta-1}(\hat{\theta}_h). \end{aligned}$$

We now show that

$$Var_\theta(U(h,\theta)) = E_{\theta-1}[-\Delta_\theta U(h,\theta-1)]. \tag{58}$$

The right hand side of (58) is an analog of the Fisher information. In order to find a suitable expression for the variance, we must first show that the expectation of the score function is zero.

$$\begin{aligned}
E_\theta[U(h,\theta)] &= f(\theta;\theta) + \sum_{Q\le h\le\theta-1} \frac{f(h;\theta)-f(h;\theta-1)}{f(h;\theta)} f(h;\theta) \\
&= f(\theta;\theta) + \sum_{Q\le h\le\theta-1} f(h;\theta)-f(h;\theta-1) \\
&= E_\theta(1) - E_{\theta-1}(1) \\
&= 0.
\end{aligned}$$

We now expand $Var_\theta(U(h,\theta))$:

$$\begin{aligned}
Var_\theta(U(h,\theta)) &= E_\theta[U(h,\theta)^2] \\
&= f(\theta;\theta) + \sum_{Q\le h<\theta} \left(\frac{f(h;\theta)-f(h;\theta-1)}{f(h;\theta)}\right)^2 f(h;\theta) \\
&= f(\theta;\theta) + \sum_{Q\le h\le\theta-1} \frac{f(h;\theta)^2 - 2f(h;\theta)f(h;\theta-1) + f(h;\theta-1)^2}{f(h;\theta)} \\
&= E_\theta(1) - 2E_{\theta-1}(1) + E_{\theta-1}\left[\frac{f(h;\theta-1)}{f(h;\theta)}\right] \\
&= E_{\theta-1}\left[\frac{f(h;\theta-1)}{f(h;\theta)}\right] - 1.
\end{aligned}$$

We finish the proof by showing

$$\begin{aligned}
E_{\theta-1}[-\Delta_\theta U(h,\theta-1)] &= -E_{\theta-1}[U(h,\theta)-U(h,\theta-1)] \\
&= -\sum_{Q\le h\le\theta-2} \frac{f(h;\theta)-f(h;\theta-1)}{f(h;\theta)} f(h;\theta-1) \\
&\quad -\left(\frac{f(\theta-1;\theta)-f(\theta-1;\theta-1)}{f(\theta-1;\theta)}\right) f(\theta-1;\theta-1) \\
&\quad +\sum_{Q\le h\le\theta-2} \frac{f(h;\theta-1)-f(h;\theta-2)}{f(h;\theta-1)} f(h;\theta-1) + f(\theta-1;\theta-1) \\
&= -\sum_{Q\le h\le\theta-1} f(h;\theta-2) + \sum_{Q\le h\le\theta-1} \frac{f(h;\theta-1)}{f(h;\theta)} f(h;\theta-1) \\
&= E_{\theta-1}\left[\frac{f(h;\theta-1)}{f(h;\theta)}\right] - 1. \qquad (59)
\end{aligned}$$

Combining equations (56), (57), and (58), the proof is complete. □

4.1.1 Asymptotic efficiency of the MLE

Theorem 11. *Under the discrete CRLB bound in (55), the MLE is asymptotically efficient.*

Proof. We first work with the numerator of the right hand side of (55). Recall the formula for bias β found in Darroch (1958) can be simplified. When simplified, it is clear that the bias is asymptotically negligible, and thus

$$\Delta_\theta E_{\theta-1}(\hat{\theta}) = 1 + \beta_\theta - \beta_{\theta-1} = 1.$$

We now consider the denominator of the right hand side of (55). We first employ a key identify in the proof of (55), appearing in equation (59).

$$E_{\theta-1}[-\Delta_\theta U(h,\theta-1)] = E_{\theta-1}\left[\frac{f(h;\theta-1)}{f(h;\theta)}\right] - 1,$$

where $f(h;\theta)$ is the probability of observing a supercollector with h songs when the true population size is θ. Let u_w be the number of individuals in sample j only in which $j \in w$ (note that this is the same as S_B in our notation from equation (11), but we use this new notation in the context of this proof for sake of simplicity of arguments). Equation (11) calculates the probability of obtaining a specific sequence $\{u_w\}$ which leads to a supercollector of size h to be:

$$P_\theta\left[\{u_w\}|\{n_i\}\right] = \frac{\theta!}{(\theta-h)!\prod_w u_w!}\prod_{i=1}^m \binom{\theta}{n_i}^{-1}.$$

To obtain $f(h;\theta)$, we must sum over all possible ways $\{u_w\}$ which lead to a supercollector of size h. In comparing $f(h;\theta)$ and $f(h;\theta-1)$ we must assume that $h < \theta-1$ so that the sequence to be summed over $\{u_w\}$ is the same. The expression for $f(h;\theta)$ is:

$$f(h;\theta) = \frac{\theta!}{(\theta-h)!}\prod_{i=1}^m \binom{\theta}{n_i}^{-1} \sum_{\sum u_w = h} \prod_w u_w!. \tag{60}$$

and the expression for $f(h;\theta-1)$ is:

$$f(h;\theta-1) = \frac{(\theta-1)!}{(\theta-1-h)!}\prod_{i=1}^m \binom{\theta-1}{n_i}^{-1} \sum_{\sum u_w = h} \prod_w u_w!. \tag{61}$$

Combining equations (60) and (61), we obtain:

$$E_{\theta-1}\left[\frac{f(h;\theta-1)}{f(h;\theta)}\right] = E_{\theta-1}\left[\frac{\theta-h}{\theta}\prod_{i=1}^m \frac{\theta}{\theta-n_i}\right] = \frac{\theta - E_{\theta-1}(h)}{\theta}\prod_{i=1}^m \frac{\theta}{\theta-n_i}.$$

In order to utilize the above equalities more explicitly, we note the following equalities which appear in Darroch (1958):

$$\theta - 1 - E_{\theta-1}(h) = E_{\theta-1}[\theta-1-h] = \frac{\prod_{i=1}^m(\theta-1-n_i)}{(\theta-1)^{m-1}}.$$

Combining the above equations, we obtain:

$$E_{\theta-1}\left[\frac{f(h;\theta-1)}{f(h;\theta)}\right] - 1 = \left(\frac{\theta}{\theta-1}\right)^{m-1}\prod_{i=1}^m \frac{\theta-1-n_i}{\theta-n_i} + \theta^{m-1}\prod_{i=1}^m \frac{1}{\theta-n_1} - 1. \tag{62}$$

We now simplify equation (62) by making two preliminary calculations:

$$\left(\frac{\theta}{\theta-1}\right)^{m-1} = \left(1+\frac{1}{\theta-1}\right)^{m-1} = 1 + \frac{m-1}{\theta-1} + O(1/\theta^2) \tag{63}$$

and

$$\prod_{i=1}^{m} \frac{\theta - 1 - n_i}{\theta - n_i} = \prod_{i=1}^{m} \left(1 - \frac{1}{\theta - n_i}\right) = 1 - \sum_{i=1}^{m} \frac{1}{\theta - n_i} + O(1/\theta^2). \tag{64}$$

By multiplying equations (63) and (64) together and combining with equation(62), we obtain:

$$E_{\theta-1}\left(-\Delta_\theta U(h, \theta - 1)\right) = \theta^{m-1} \prod_{i=1}^{m} \frac{1}{\theta - n_i} - \sum_{i=1}^{m} \frac{1}{\theta - n_i} + \frac{m-1}{\theta - 1} + O(1/\theta^2).$$

Theorem 9 proves the asymptotic variance of $\hat{\theta}$ to be:

$$Var_\theta(\hat{\theta}) = \left[\frac{\theta^{m-1}}{\prod_{i=1}^{m} \theta - n_i} - \sum_i \frac{1}{\theta - n_i} + \frac{m-1}{\theta}\right]^{-1}.$$

Following the discrete CRLB bound in (55), we need now consider the following:

$$\left[\frac{\theta^{m-1}}{\prod_{i=1}^{m} \theta - n_i} - \sum_{i=1}^{m} \frac{1}{\theta - n_i} + \frac{m-1}{\theta}\right]^{-1} \geq \left[\frac{\theta^{m-1}}{\prod_{i=1}^{m} \theta - n_i} - \sum_{i=1}^{m} \frac{1}{\theta - n_i} + \frac{m-1}{\theta - 1}\right]^{-1}. \tag{65}$$

The two sides of the above expression differ only in the last term and the difference between these two terms is $O(1/\theta^2)$, and thus we immediately expect the MLE to be asymptotically efficient. We now prove that it is so. We calculate the difference between the left and the right hand sides of (65), and simply this difference as:

$$\frac{\frac{m-1}{\theta(\theta-1)}}{O\left(\frac{1}{\theta^2}\right)} = \frac{O\left(\frac{1}{\theta^2}\right)}{O\left(\frac{1}{\theta^2}\right)} = O(1).$$

Thus, the MLE is asymptotically efficient. □

5 Summary

We have considered broad aspects of a standard version of the multiple-collector capture-recapture problem. In this version, the sizes of the collections are viewed as fixed beforehand or as ancillary to the total population size, which is the parameter of interest. This is referred to as "Model B" in the seminal paper Darroch (1958) and is the topic of sections 3 and 4 of that paper. In common with the treatment there our focus has been on the Maximum Likelihood Estimator. Darroch describes the maximum likelihood equations for this problem and claims without proof that these have a unique solution. Our Theorems 1 and 2 describe the sufficient statistic and prove that the likelihood equation has a unique solution except for special boundary cases or very special cases when it can have two solutions that are adjacent integers. These results were claimed in Darroch (1958) without detailed proof. We also provide a new motivation for the MLE a the "listening times" estimator in Section 2.1.

Asymptotic properties of the MLE are then studied in two different variants of the formulation. The first variant was not treated by Darroch. In this variant the number of collectors converges to infinity while the total population size remains constant. Our Theorems 4 and 5 provide asymptotic convergence results in this setting. The second variant of this problem involves a fixed number of collectors and collection sizes that grow along with the population size and yields asymptotic normality of the MLE. Darroch also treats this variant but requires the collection sizes to grow proportional to the population size. As briefly argued at the start of Section 3.2, such an assumption is not suited to many capture-recapture applications. The general situation treated in Section 3.2 requires a new asymptotic normality result for 2^m contingency tables (Theorem 8), and as noted in Remark 5 yields asymptotic normality at different rates from those in Darroch's case. Section 4 establishes asymptotic efficiency for the MLE in this setting by developing and applying an appropriate Cramer-Rao type of inequality for this discrete sample space problem.

The problems we have treated here involve at least two collectors. Analogous capture tag and recapture situations involving a single collector are also of practical interest. See Sun et al. (2014) for recent examples. The theory in Section 2 applies to establish uniqueness of the MLE. But the asymptotic distribution of the MLE in the single collector setting is a separate problem and will be considered in a separate manuscript currently under preparation.

Appendix A

Table 1 below contains the data from which Figure 1 is drawn.

h	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164
MLE	174	176	179	181	184	187	190	193	196	199	202	206	209	213	216	220	224	228	232	236	240
$\mathbb{P}(h \mid \theta = 200)$	.006175	.011083	.018581	.029104	.04206	.058274	.074506	.089037	.099454	.10383	.101307	.092368	.078686	.062614	.046531	.032283	.020904	.012628	.007113	.003735	.001827

Table 1: Calculations of MLE for $m = 3$, $n_1 = 100$, $n_2 = 75$, $x_3 = 50$, $\mathbb{P}(h \mid \theta = 200)$.

References

Bales, K., Hesketh, O., and Silverman, B. (2015). *Modern slavery in the UK: how many victims? Significance* **12**, 16–21.

Bishop, Y., Fienberg, S., and Holland, P. (1986). *Discrete Multivariate Analysis: Theory and Applications.* Springer, New York.

Brown, L. D., Eaton, M., Freedman, D., Klein, S., Olshen, R., Wachter, K., Wells, M., and Ylvisaker, D.(1999). Statistical controversies in Census 2000. *Jurimetrics* **39**, 347–375.

Brown, L. D. and Zhao, Z. Alternative formulas for synthetic dual system estimation in the 2000 census. *Probability and Statistics: Essays in Honor of David A. Freedman; IMS Collections v.2.* Institute of Mathematical Statistics.

Chandler, D. and Kapelner, A. (2013). Breaking monotony with meaning: motivation in crowdsourcing Markets. *Journal of Economic Behavior and Organization* **90**, 123–133.

Citro, C. F., Cork, D. L., and Norwood, J. L.(2003). *The 2000 Census—Counting Under Adversity.* National Academy Press.

Craig, C. C. (1953). On the utilization of marked specimens in estimating populations of flying insects. *Biometrika* **40**, 170–176.

Darroch, J. N. (1958). The multiple-recapture Census I. Estimation of a closed population. *Biometrika* **45**, 343–359.

Darroch, J. N. (1959). The Multiple-recapture census II. Estimation where there is immigration or death. *Biometrika* **46**, 336–351.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88**, 1137–1148.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* B **40**, 237–264.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer.

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

Newman, D. and Shepp, L. (1960). The double dixie cup problem. *American Mathematical Monthly* **67**, 58–61.

Nicholson, W. L.(1956). On the normal approximation to the hypergeometric distribution. *The Annals of Mathematical Statistics* **27**, 471–483.

Papathanasiou, V. (1993). Some characteristic properties of the Fisher information matrix via Cacoullos-type Inequalities. *Journal of Multivariate Analysis*, 44:256–265.

Pinsky, M. A. (2003) The normal approximation to the hypergeometric distribution. *Note on webpage.*

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley.

Samuel, E. (1968). Sequential maximum likelihood estimation of the size of a population. *The Annals of Mathematical Statistics* **39**, 1058–1068.

Seber, G. (1982). *The Estimation of Animal Abundance and Related Parameters.* Macmillan.

Sun, C. C., Fuller, A. K., and Royle, J. A. (2014). Trap configuration and spacing influences parameter estimates in spatial capture-recapture models. *PLoS ONE* **9**(2): e88025.

Van der Heijden, P., Whittaker, J., Cruyff, M., Bakker, B., and Van der Vliet, R. (2012). People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics* **6**(3): 831–852.

Yip, P., Bruno, G., Seber, G., Buckland, S., Cormack, R., Unwin, N., Chang, Y., Fienberg, S., Junker, B., Laporte, R., Libman, I., and McCarty, D. (1995). Capture-recapture and multiple-record systems estimation. *American Journal of Epidemiology* **142(10)**, 1047–1058.

Young, G. and Smith, R. (2005). *Essentials of Statistical Inference.* Cambridge University Press, New York.

Zaslavsky, A. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association* **88**, 1092–1105.