An Exploration of Schnabel (1938)'s Capture-Recapture Estimators

Final Project for Probability Theory IV, taught by Dr. Ciprian Crainiceanu

Michelle Qin

May 15, 2024

Foreword: Statistical Opportunities in Capture-Recapture Studies

I leaped into this project inspired by the statistical computations in Z. E. Schnabel's 1938 paper "The Estimation of the Total Fish Population of a Lake." I was eager to use my computer to simulate the performance of the author's estimators, which were computed using the "calculating machines" of the time (White et al., 1982, p. 28) against numerically computed maximum likelihood estimators. A brief literature review revealed several ways in which my project could be extended further.

Scientists have simulated capture-recapture studies, also known as mark-recapture or tag-recapture studies, since the 1960s (White et al., 1982, p. 39). In this project, I simulated drawing a fixed number of mark-recapture samples from three differently sized populations of fish, directly inspired by three datasets cited in Schnabel (1938). For each population of fish, I drew samples using either of two possible sampling schemes: Binomial (i.e., fixed sample sizes, which for simplicity I set to be equal across all capture-recapture samples) or Poisson (i.e., each sample size is a Poisson random variable, whose rate I set to be equal across all capture-recapture samples). I assumed that both the samples and the fish within each sample are independent and identically distributed—an assumption that is likely untrue in real life. **A natural extension of my simulations would be to simulate correlated fish and/or correlated samples, to see how Schnabel's, others', and my estimators perform when they misspecify the data structure.**

A DuckDuckGo search of "capture-recapture methods" and "Schnabel method" (which refers to the second estimator in Schnabel's 1938 paper) revealed that ecologists have made **additional ecological estimators** (e.g., the Schumacher and Eschmeyer method, which uses a regression line rather than maximum likelihood), **real mark-recapture sample data**[1], and **tools for practioners and students** publicly available (Krebs, 2014, Chapter 2; Vierling, 2012). I used one such real-life dataset, from Gerking (1953), to compare my estimators against Schnabel's.

In public health, meanwhile, epidemiologists have developed methods deriving from capture-recapture to estimate specific human subpopulations using individuals' presence across multiple datasets ("lists"). Note that enrollment lists may not be as cleanly time-ordered as capture-recapture samples of wildlife. **Membership in different lists may be modeled in multiple ways to allow for dependence within or across patients.** Chao et al. (2001) delineate between log linear models (which include interaction terms between pairwise list membership, and can be made more flexible using Bayesian, random-effects, latent class, non-parametric, or semiparametric methods, for example in Wesson, 2018 and Wesson, 2023) versus "sample coverage" methods.

The key question for any project is, Who is the audience? Statistical, epidemiological, ecological, or public? Theoretical or practical? Accordingly, should we create a website, a "primer"[2] or guidebook, an app, or a paper?

[1] However, one advantage of simulations like mine is that the truth is known, and capture-recapture studies of the same underlying truth may be generated arbitrarily many times. Thus, my project can compute estimators' variance using such Monte Carlo simulations, rather than deriving from parametric assumptions. Another option for estimating variance is the jackknife (Burnham and Overton, 1978, cited in Chao et al., 2001).

[2] Actually, White et al. (1982) was written as a primer for students, to accompany Otis et al. (1978)'s more theoretical

Problem Statement: Estimating Total Population from Membership in Successive Samples

Imagine you are a fish surveyor in Illinois (D. H. Thompson, cited in Schnabel, 1938, p. 351), or a graduate student at Johns Hopkins estimating the rat population of Baltimore City (Calvin Zippin, cited in White et al., 1982, p. 28), or a state health department estimating the number of people living with HIV so that you can provide appropriate services for them (Wesson, 2024). You may have either the aggregated statistics or the individual-level data from $n \geq 1$ successive tag-and-recapture samples. For the purposes of this project, however, and following Schnabel (1938), I will assume that we have only aggregated data from each sample $i \in \{1, \ldots, n\}$, as schematized in Table 1 below.

Figure 1: Image of fish seining (Linda Lorning Nature Foundation).

In 1938, *The American Mathematical Monthly* published a paper by then-graduate-student Zoe Schnabel at the University of Wisconsin, entitled “The Estimation of the Total Fish Population of a Lake”. The paper presented four estimators developed for closed capture-recapture settings, i.e., where the population size N, which is the estimand, is assumed to be constant across samples.

	Tagged at time *i*	**Not tagged at time *i***	**Total at time *i***
In sample *i*	r_i	d_i	t_i
Not in sample *i*	$M_i - r_i$	$N - M_i - d_i$	$N - t_i$
Total at time *i*	M_i	$N - M_i$	N

Table 1: Data Structure for One Capture-Recapture Sample, in Schnabel (1938)’s notation.

Schnabel’s paper was the first major work extending Petersen (1896)’s and Lincoln (1930)’s 1-sample estimator, $\frac{t_1 M_1}{r_1}$, to $n \geq 1$ mark-and-recapture samples. The second estimator in Schnabel (1938), $\frac{\sum_{i=1}^{n} t_i M_i}{\sum_{i=1}^{n} r_i}$, which may be used when all $M_i << N$, is now called the “Schnabel method” or “Schnabel index” and has been the “backbone of population size estimation, assuming closure, for the past [more than] 40 years” (White et al., 1982, p. 28).[3]

In this project, I was interested in assessing the bias and variance of Schnabel (1938)‘s four estimators, as well as three estimators that I developed, in three simulated datasets, as well as the estimators’ agreement on one real-world capture-recapture dataset.

monograph. Incidentally, The postscript of White et al. (1982) is quite funny. It constructs a capture-recapture study from the four authors’ captures of typos while editing, recommends a specific model, and notes ways in which the samples are not identically distributed. For example, one author was assisted by a spouse, and each authors spent a different amount of time editing.

[3] It is strange that Z. E. Schnabel does not have a Wikipedia page. Filling that gap could be an interesting project as well.

Modeling Assumptions in Schnabel (1938) and This Project

Before discussing any estimators of N, their derivations, or their guarantees, we should first discuss what probability space[4] [5] our data $(r_1, d_1, t_1, M_1), \dots, (r_n, d_n, t_n, M_n)$, as well as our estimand N and our estimators $\hat{N}_1, \dots, \hat{N}_7$, might live on. The following are some key modeling questions that stood out to me.

Key Questions: Randomness, Variability, Dependence, Sampling Distribution

1. What is random? What varies within or across capture-recapture samples? Which random variables are conditionally independent?

N	(i) Shall we treat the total population, which is never observed, as a constant or a random variable? (ii) Should we assume N may or may not change between the n capture-recapture samples?
n	(iii) Is the number of capture-recapture samples a pre-determined constant or a random variable (e.g., if surveyors collect fish until a certain number of tags are given out or a certain amount of time has passed)?
t_i	(iv) Similarly, does each capture-recapture sample have a fixed or random size?
M_i	(v) Is the total number of tags in the lake at any point in time a (non-random) function of $t_0, \dots, t_{i-1}$? For example, $M_i = t_0 + \dots + t_{i-1}$ if all captured fish are tagged.[6] (vi) Is M_i known to the researcher? This could be accomplished by digital/remote sensing or by assuming that no previously tagged fish lose their tags or leave the population.
r_i	(vii) Clearly random. Is the number of recaptures the sum of independent and/or identically distributed indicators for each fish?
d_i	(viii) Also random, unless t_i and r_i are conditioned on. Could be independent from r_i marginally.
ϕ_i?	(ix) Should we include other sources of randomness, i.e., measurable variables or inferred overdispersion or zero-inflation parameters that influence (r_i, d_i, t_i, M_i) and/or (n, N)?
(r_i', d_i', t_i')	(x) Are samples $i \neq i'$ independent conditional on N, M_i, M_i'?

2. What is the sampling distribution of the data $(r_1, d_1, t_1, M_1), \dots, (r_n, d_n, t_n, M_n)$? What happens, e.g., to the true and estimated bias and variance of an estimator $\hat{N}$, if we misspecify it?

Schnabel (1938)'s Assumptions: Constant Population; Not Modeling Possibly Variable Sample Sizes as Random; Independence Between Fish and Samples; Identical Sampling Probability Among Fish; No Unexplained Variance Parameters

1. Schnabel assumes that N is constant. Since samples are stated to usually be taken about every 24 hours (for a total of 79 or 39 samples, respectively, in Datasets I and II in the paper), this assumption seems plausible.

   Note: In my Bayesian estimator, I let N be a random variable, but I keep the assumption that a single N is shared over all n capture-recapture samples.

[4] We can think of the probability space of one set of n capture-recapture samples as $(\Omega, \mathcal{F}, P)$, where the observed data $(r_1(\omega), d_1(\omega), t_1(\omega), M_1(\omega)), \dots, (r_n(\omega), d_n(\omega), t_n(\omega), M_n(\omega))$—and possibly $n(\omega)$ as well if n is not pre-fixed by the surveyor but rather depends on nature—are random variables, i.e., functions of one realization ω from the sample space Ω. (In my simulations, I generated `N_SIMS = 30` such $\omega \in \Omega$.) Then, the probability measure $P(\{\omega\}) = P((r_1, d_1, t_1, M_1), \dots, (r_n, d_n, t_n, M_n), (n))$ can have multiple forms, which we can put parametric assumptions on. Notice that the estimand N can be modeled either as a constant that influences P (frequentist), or as $N(\omega)$, a random variable and another input in P (Bayesian).

[5] WLOG, let $\mathcal{F}$ be the sigma-algebra generated by the data, i.e., the smallest sigma-algebra containing the data. Then, $\mathcal{F}$ is countable, and $\mathcal{F}$ is finite if $N \in \mathbb{N}$ is fixed or if $N(\omega)$ is bound by some practical limit of the possible numbers of fish in an Earthen lake. Meanwhile, $P(\{\omega\}) = 0$ for any $\{\omega\} \in \mathcal{F}$ such that $r_i(\omega) + d_i(\omega) \neq t_i(\omega)$ or $t_i(\omega) > M_i(\omega)$ for any $i \in \{1, \dots, n\}$.

[6] I am using t_0 to denote the size of the initial sample of fish that are captured from the lake, tagged, and thrown back in to be recaptured in the first tag-recapture sample.

2. Even though she does not fix any formulas, structures, or equalities for the (n, t_i, M_i), Schnabel's models do not afford them the status of being random variables—perhaps depending on latent variables, distributional structures, etc.—in (n, t_i, M_i), instead treating them as observed data in the likelihood function.

3. Independence and Identical Sampling Probability Among Fish in One Sample: $r_i|t_i, M_i, N \sim \text{Bin}(t_i, p_i)$, where $p_i = \frac{M_i}{N}$.

4. Independence Between Samples: $P(r_i, r_i'|t_i, t_i', M_i, M_i', N) = P(r_i|t_i, M_i, N)P(r_i'|t_i', M_i', N)$ for $i \neq i'$.

5. Similarly to bullet point #2, there are also no unaccounted-for variance parameters in the probability model of $r_i|t_i, M_i, N$.

From the above assumptions, Schnabel arrives at the product-of-Binomials likelihood (discussed in the next section).

Additional Assumptions in My Simulations: Sample Sizes May or May Not Be Fixed; All Fish are Tagged

1. My simulations, to the detriment of mimicking the real world more flexibly and challenging Schnabel's and my estimators more, did not break any of the above assumptions.

   However, my simulations had the added task of generating the t_i, affording me the opportunity to do so in multiple ways.

2. I chose (i) $t_0 = t_1 = \cdots = t_n =$ [pre-fixed] constant size, which I denoted the "Binomial sampling scheme", and (ii) $t_i \sim$ Pois([pre-fixed] constant rate), which I denoted the "Poisson sampling scheme".

3. I made the additional assumption that all captured fish were tagged and no tags fell off: $M_i = t_0 + \cdots + t_{i-1}$.

Derivations of Likelihood from Assumptions

Finally, we have arrived at the likelihood calculations!

Schnabel (1938)'s first three estimators come from the same likelihood; they are different approximations of the same Binomial MLE. Schnabel (1938)'s fourth estimator is the minimizer of the Pearson χ^2 goodness-of-fit statistic, where the expected count of the ith sample's recaptures (r_i) is still $\frac{t_i M_i}{N}$.

Schnabel's Binomial Likelihood

$$P(r_1, \ldots, r_n|t_1, \ldots, t_n, M_1, \ldots, M_n, N) = \prod_{i=1}^{n} \binom{t_i}{r_i} \left(\frac{M_i}{N}\right)^{r_i} \left(\frac{N - M_i}{N}\right)^{d_i} \qquad (r_i \sim \text{Bin}(t_i, \frac{M_i}{N}))$$

$$\implies \ell(N; r_1, \ldots, r_n, t_1, \ldots, t_n, M_1, \ldots, M_n) \propto \sum_{i=1}^{n} d_i \log(N - M_i) - (r_i + d_i) \log N$$

$$\implies \frac{d\ell}{dN} = \sum_{i=1}^{n} \frac{d_i}{N - M_i} - \frac{r_i + d_i}{N} = \sum_{i=1}^{n} \frac{d_i N - (r_i + d_i)(N - M_i)}{N(N - M_i)} = \sum_{i=1}^{n} \frac{-r_i(N - M_i) + d_i M_i}{N(N - M_i)} \propto \sum_{i=1}^{n} -r_i + \frac{d_i M_i}{N - M_i}.$$

My Hypergeometric Likelihood

$$P(r_1, \ldots, r_n | t_1, \ldots, t_n, M_1, \ldots, M_n, N) = \prod_{i=1}^{n} \frac{\binom{M_i}{r_i}\binom{N-M_i}{d_i}}{\binom{N}{t_i}} \qquad (r_i \sim \text{HGeom}(M_i, N - M_i, t_i))$$

Notice that $r_i | t_i, M_i, N \sim \text{HGeom}(M_i, N - M_i, t_i)$ has the same mean as $r_i | t_i, M_i, N \sim \text{Bin}(t_i, \frac{M_i}{N})$ but smaller variance because its likelihood puts more structure on the data, explaining more of the variance.

My Poisson-Gamma Bayesian Mixture Likelihood

Let p_s denote the sampling probability of any fish. We can use r_i and M_i to infer about p_s, which we can use to infer about N from t_i:

$$\begin{aligned}
N | \lambda &\sim \text{Pois}(\lambda) \\
t_i | \lambda, p_s &\overset{i.i.d.}{\sim} \text{Pois}(\lambda p_s) && \text{(Poisson thinning property)} \\
\lambda p_s &\sim \text{Gamma}(s_0, r_0) && \text{(conj. prior on } \lambda p_s) \\
\implies \lambda p_s | t_1, \ldots, t_n &\sim \text{Gamma}(s_0 + \sum_{i=1}^{n} t_i, r_0 + n) && \text{(posterior of } \lambda p_s) \\
\implies \lambda | p_s, t_1, \ldots, t_n = \frac{\lambda p_s}{p_s}\Big| p_s, t_1, \ldots, t_n &\sim \text{Gamma}(s_0 + \sum_{i=1}^{n} t_i, p_s(r_0 + n)) && \textbf{(conditl. post. of } \boldsymbol{\lambda}\text{; Gamma scaling)}
\end{aligned}$$

$$\begin{aligned}
r_i | M_i, p_s &\overset{\perp\!\!\!\perp}{\sim} \text{Bin}(M_i, p_s) \\
p_s &\sim \text{Beta}(a_0, b_0) && \text{(conj. prior on } p_s) \\
\implies p_s | r_1, \ldots, r_n, M_1, \ldots, M_n &\sim \text{Beta}(a_0 + \sum_{i=1}^{n} r_i, b_0 + \sum_{i=1}^{n} M_i - r_i) && \textbf{(posterior of } \mathbf{p_s}\textbf{)}.
\end{aligned}$$

Estimators Examined in This Project

Estimator 1 from Schnabel (1938)—Binomial MLE, Geometric Series Approximation

$$\text{Since } 1+\frac{M_i}{N}+\cdots+\left(\frac{M_i}{N}\right)^{\infty}=\frac{1}{1-\frac{M_i}{N}}=\frac{N}{N-M_i} \text{ and } \frac{d\ell}{dN}\propto\sum_{i=1}^{n}-r_i+\frac{d_iM_i}{N-M_i},$$

$$\hat{N}_{1,k}=\boxed{\text{positive root of } \sum_{i=1}^{n}-r_i+\frac{d_iM_i}{N}\left(1+\frac{M_i}{N}+\cdots+\left(\frac{M_i}{N}\right)^{k}\right)=0}.$$

Bound 1b from Schnabel (1938), derived from geometric series remainder

Note: Bound 1b bounds the error of Estimator 1 from the Binomial MLE, not necessarily from the true N.

$$\hat{N}_{\text{Bound }1b}=\boxed{\text{positive root of } (\sum_{i=1}^{n}r_i)N^k-(\sum_{i=1}^{n}d_iM_i)N^{k-1}-\cdots-\sum_{i=1}^{n}d_iM_i^k-\frac{M_n}{\hat{N}_{1,k}-M_n}\sum_{i=1}^{n}d_iM_i^k=0}.$$

Estimator 2 from Schnabel (1938), a.k.a. "Schnabel Method" in Modern Capture-Recapture—Binomial MLE, Approximated when $M_i << N$

$$\hat{N}_2=\boxed{\frac{\sum_{i=1}^{n}t_iM_i}{\sum_{i=1}^{n}r_i}}.$$

Estimator 3 from Schnabel (1938)—Binomial MLE, Approximated when $M_1 = \cdots = M_n$

$$\hat{N}_3=\boxed{\frac{M_1\sum_{i=1}^{n}t_i}{\sum_{i=1}^{n}r_i}}.$$

Estimator 4 from Schnabel (1938)—Pearson χ^2 GOF Minimizer, Appropriate when $r_i >> 0$

Schnabel (1938)'s fourth estimator is the minimizer of the Pearson χ^2 goodness-of-fit statistic, where the expected count of the ith sample's recaptures (r_i) is still $\frac{t_iM_i}{N}$.

$$\hat{N}_4=\boxed{\sqrt{\frac{\sum_{i=1}^{n}t_iM_i}{\sum_{i=1}^{n}\frac{r_i^2}{t_iM_i}}}}.$$

My Estimator 5—Hypergeometric MLE, found by Inspection

I found the MLE by inspection in my first pass, though an iterative algorithm like [stochastic] gradient descent, BFGS, or Newton-Raphson (probably using the Gamma function generalization of the factorial function, for the Hypergeometric likelihood) likely would be more robust or faster, as well as checking if the likelihood for hypergeometric is convex, if I were to rewrite this code.

My Estimator 6—Binomial MLE, found by Inspection

Similarly to the comment for Estimator 5, this estimator would have been better estimated by Newton-Raphson.

Estimator 6 should be very similar to Estimator 1 with a sufficient number of terms.

My Estimator 7 (Posterior Mean from Poisson-Gamma Bayesian Mixture)

Let $a_0 = b_0 = s_0 = r_0 = 1$.

Then, using the distributions derived in the previous section,

$$\hat{N}_7 = \mathbb{E}[N|t, r, M] = \mathbb{E}[\mathbb{E}[N|\lambda, t, r, M]|t, r, M] = \mathbb{E}[\lambda|t, r, M] = \mathbb{E}[\mathbb{E}[\lambda|p_s, t, r, M]|t, r, M]$$

$$= \mathbb{E}\left[\frac{s_0 + \sum_{i=1}^n t_i}{p_s(r_0 + n)}\middle| t, r, M\right] = \boxed{\frac{s_0 + \sum_{i=1}^n t_i}{r_0 + n}\mathbb{E}\left[\frac{1}{p_s}\middle| t, r, M\right]}$$

where $\mathbb{E}\left[\frac{1}{p_s}\middle| t, r, M\right]$ is estimated by 1,000 Monte Carlo samples from the posterior $p_s|r_1, \ldots, r_n, M_1, \ldots, M_n \sim \text{Beta}(a_0 + \sum_{i=1}^n r_i, b_0 + \sum_{i=1}^n M_i - r_i)$.

Datasets Used in This Project: Three Simulated, One Real-World

Simulated Datasets Inspired by Schnabel (1938)

I used `N_SIMS = 30` Monte Carlo simulations of three scenarios (large, medium, and small lake, respectively) using two implementations each (Binomial and Poisson sampling), as described in the Modeling Assumptions section.

Scenario I

I set 1,300,542 as the true number of fish because Schnabel (1938) computed 1,300,543 as a strict upper bound for Estimator 1, which I trusted over the paper's other estimators, which were largely approximations to Estimator 1.

Scenario II

Similarly to Scenario I, I set 15,344 as the true number of fish because Schnabel (1938) computed 13,345 as a strict upper bound for Estimator 1.

Scenario III

I set 1,872 as the true number of beans because Schnabel (1938) reported that as the truth in her experiment.

Preview of One Simulation, from Scenario (Dataset) III

Below is an excerpt from one of my simulations inspired by Dataset III, the bean dropping experiment that Schnabel performed in her paper, where there were 1,872 beans in a jar and the author drew 14 tag-recapture samples. In the implementation below (which is duly denoted as the "Binomial sampling scheme" in the boxplots to follow), I simulated drawing 14 samples of exactly (i.e., fixed) size 150.

Table 2: Example of a simulation where Fish #5 is captured and tagged in the fourth sample and Fish #1 is captured and tagged in the ninth sample. Rows of this table represent individual fishes, which are assumed to remain tagged after being captured. Columns of the table represent successive samples. Cell entries denote each fish's tagged/untagged status at the time of sampling.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	0	0	0	0	0	0	0	1	1	1	1	1	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	1	1	1	1	1	1	1	1	1	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3: Summary statistics of one simulation of capture-recapture.

sample_id	sample_size	n_tags_in_pond	n_tags_in_sample	n_untagged_in_sample
1	150	150	12	138
2	150	288	23	127
3	150	415	27	123
4	150	538	44	106
5	150	644	54	96
6	150	740	69	81
7	150	821	65	85
8	150	906	66	84
9	150	990	77	73
10	150	1063	91	59
11	150	1122	85	65
12	150	1187	91	59
13	150	1246	112	38
14	150	1284	103	47

Real-World Dataset: 14 samples from an Indiana stream by Gerking (1953)

The Gerking (1953) dataset is available in `fishmethods` R package.

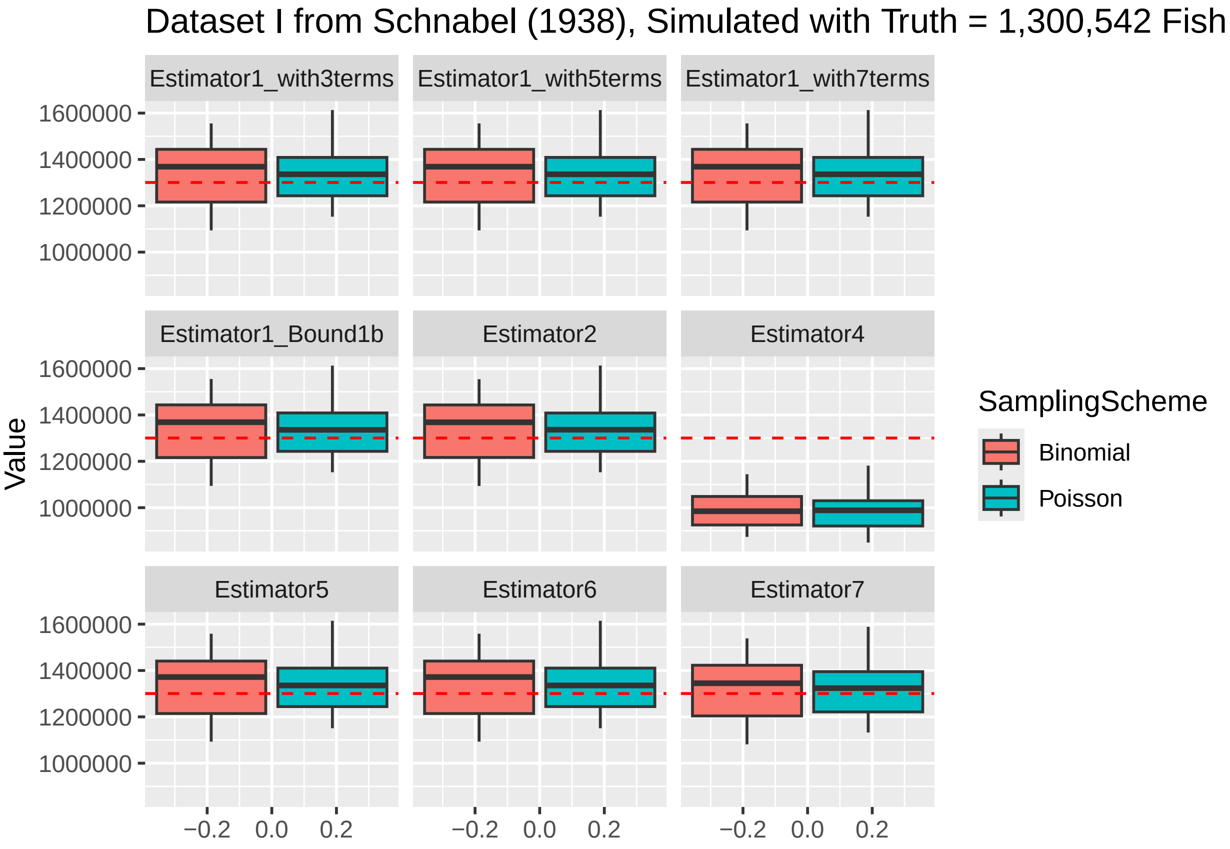
	sample_size	n_tags_in_sample	n_new_tags	n_tags_in_pond	n_untagged_in_sample
1	27	0	27	10	27
2	17	0	17	37	17
3	7	0	7	54	7
4	1	0	1	61	1
5	5	0	5	62	5
6	6	2	4	67	4
7	15	1	14	71	14
8	9	5	4	85	4
9	18	5	13	89	13
10	16	4	10	102	12
11	5	2	3	112	3
12	7	2	4	115	5
13	19	3	0	119	16

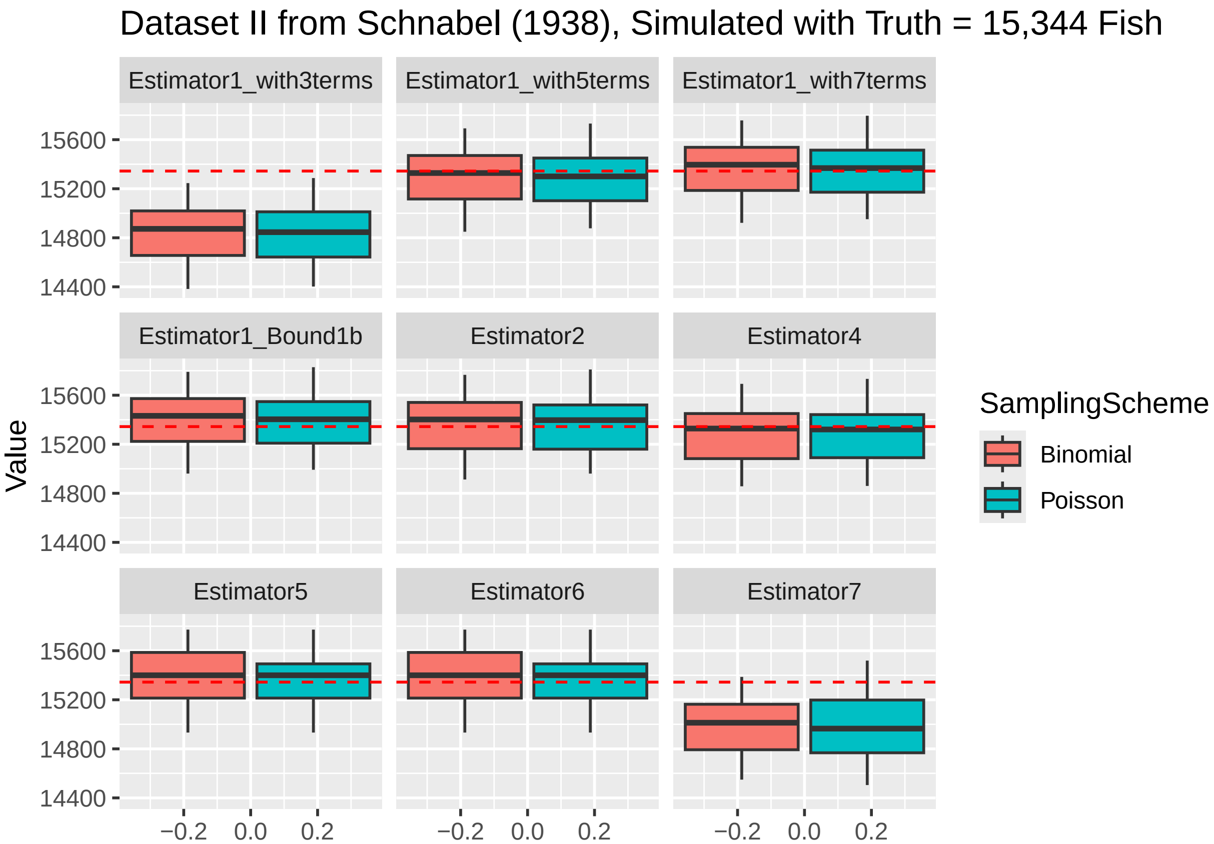
Results: Performance of Estimators on Simulated and Real Capture-Recapture Datasets

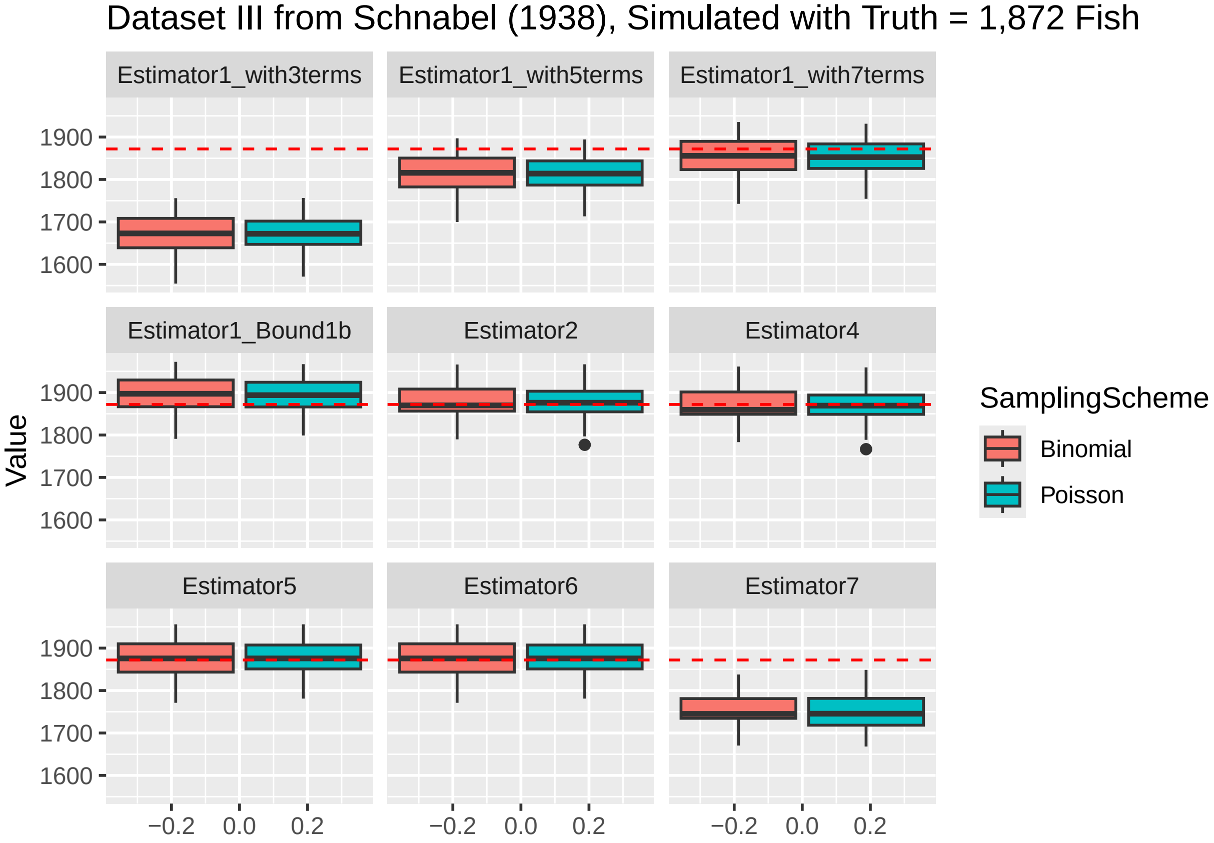
Simulated Datasets Inspired by Schnabel (1938)

In the figures below, the red dashed line indicates the true number of fish in my simulations. Observe:

- Estimates from Poisson sampling tended to be slightly smaller in value, closer to the truth in the large-lake scenario (Dataset I), and smaller variance than estimates from Binomial sampling.
- Unsurprisingly, Estimator 1 performed better when the number of terms, k, is larger. $k = 5$ terms were usually enough for Estimator 1 to match the performance of Estimator 6, at least in these settings that I tested.
- As Schnabel (1938) predicted, Estimator 4 did not perform well when recapture counts r_i are small. However, when conditions for Estimators 2, 3, and 4 were met, they performed comparably to the other estimators, which surprised me.
- Estimator 5 had usually-identical point estimates but smaller variance than Estimator 6, as I predicted in the Likelihood section of this writeup.
- My Bayesian estimator (Estimator 7) did not perform well (biased downward and poor coverage) in the medium- and small-lake scenarios.




Dataset II from Schnabel (1938), Simulated with Truth = 15,344 Fish
Estimator1_with3terms
Estimator1_with5terms
Estimator1_with7terms
Estimator1_Bound1b
Estimator2
Estimator4
Estimator5
Estimator6
Estimator7
Value
15600
15200
14800
14400
−0.2
0.0
0.2
SamplingScheme
Binomial
Poisson


Dataset III from Schnabel (1938), Simulated with Truth = 1,872 Fish
Estimator1_with3terms
Estimator1_with5terms
Estimator1_with7terms
Estimator1_Bound1b
Estimator2
Estimator4
Estimator5
Estimator6
Estimator7
Value
1900
1800
1700
1600
−0.2
0.0
0.2
SamplingScheme
Binomial
Poisson

Gerking (1953) Dataset

Most estimators produced similar results, except for Schnabel's Estimator 4; the number of recaptures (`n_tags_in_sample` in the table below) is small in this dataset.

Estimator1_with3terms	Estimator1_with5terms	Estimator1_with7terms	Estimator1_Bound1b
438.5	442.1	442.3	442.4

Estimator2	Estimator4	Estimator5	Estimator6	Estimator7
447.5	355.3	444.7	443	447.8

As a sanity check, I compared my implementation Estimator 2 with the existing `schnabel()` function in the `fishmethods` package. It matched!

	N	invSE	LCI	UCI	CI_Distribution
Schnabel	447.5000	0.0004561	315.8824	716.0000	Poisson
Schumacher-Eschmeyer	422.9712	0.0004465	299.6589	718.7388	t

Conclusions

Discussion of Results

- Given that estimates from Poisson sampling tended to be slightly smaller in value, closer to the truth in the large-lake scenario (Dataset I), and smaller variance than estimates from Binomial sampling, should Poisson sampling be recommended over binomial sampling, i.e., surveyors should collect as many fish as enter a seine in a preset amount of time instead of collecting a preset number of fish? Perhaps Poisson sampling contains more information than Binomial sampling.

- Most estimators produced similar results in the smallest dataset I tested (Gerking, 1953), except for Schnabel's Estimator 4, which was not designed for small-recapture datasets anyway. Perhaps capture-recapture estimators differ from each other more when the population of fish is larger (that makes sense intuitively) and not so much when the population is small (though perhaps there is a limit to their performance on populations smaller than Gerking's).

- As Schnabel (1938) noted, certain conditions are necessary for Estimators 2, 3, and 4 to be used. When those conditions were met, though, these estimators performed as well as Estimator 6 (Hypergeometric MLE). Schnabel (1938) seems timeless indeed! I imagine that ecologists have favored her second estimator over the others because it is easier to compute than Estimator 1 but more generally applicable than Estimators 3 and 4.

- I am not sure why my Bayesian estimator (Estimator 7) did not perform well (biased downward and poor coverage) in the medium- and small-lake scenarios. I must have specified the model poorly somewhere.

Limitations of Results

- `N_SIMS`, the number of Monte Carlo replicates of each simulated dataset, was only 30, so the results in this writeup may not generalize.
- My simulated scenarios were closely aligned with the assumptions in both Schnabel's and my likelihood models, so they likely overestimated the performance of our estimators on real-world data.
  - It would have been beneficial to simulate from correlated data, for example, clusters of fish who get captured together (maybe because they travel together or hang out in the same part of the lake).
  - It would have been beneficial to simulate from heterogeneous (i.e., not all identically distributed) data, for example having a mix of small/fast fish who never get captured and large/slow fish who get captured at higher rates than other fish.
  - It would have been beneficial to include a latent variable or two to add some unmodeled noise.
- I could have used more real-world data, as available on several ecological departments' websites such as Vierling (2012).
- Inspection (Estimators 5 and 6) will be an intractable strategy on big data; better to use an iterative algorithm like Newton-Raphon, [stochastic] gradient descent, or a quasi-Newton method like BFGS.

Further Directions

- Quantifying the bias and variance of Schnabel's estimators as a function of n, N, and k for Estimator 1.

- Including overdispersion and/or zero-inflation parameters in my estimators.

- Generating confidence, prediction, and/or credible intervals, along with variance estimates, from my and Schnabel's estimators. Some ecological texts use parametric assumptions to do so; I could perhaps use bootstrap/jackknife or M-estimation theory.

  This is a finite-sample problem: N (the estimand) literally cannot go to infinity, and n may not be able to go to infinity either if we retain our closed population assumption (though maybe we don't have to; and maybe we could anyway and assume that the population is in equilibrium forever). However, maybe asymptotic approximations and assumptions may be valid enough if N, n, and/or M_i or $N - M_i$ are large.

- Alternative estimators:

  - Rao-Blackwell? May be harder without a closed-form MLE. Also, my estimators are already using sufficient statistics (r_i, d_i, t_i, M_i).
  - Regression estimator like Schumacher and Eschmeyer.
  - My Estimator 8 (in `code/ratio-estimator-functions.R`, though not explored in this writeup):

  $$\frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n \frac{r_i}{M_i}} \left(\overset{\text{SLLN, Slutsky}}{\rightarrow} \frac{E[t_i]}{E[\frac{r_i}{M_i}]} \text{ if } n \to \infty \right).$$

- Using individual-level data $I(\text{individual } j \text{ is in sample } i)$ rather than aggregated sample data (r_i, d_i, t_i, M_i).

- As alluded to in the Foreword, it would be interesting to explore epidemiological applications, e.g., non-time-ordered and non-experimental datasets, and methods, including log-linear or sample coverage models to allow for dependence between samples

- Open populations, which would require defining a higher-dimension $(\Omega, \mathcal{F}, P)$.

References

David L. Otis, Kenneth P. Burnham, Gary C. White, and David R. Anderson (1978). "Statistical inference from capture data on closed animal populations." *Wildlife Monographs*, Volume 62, pp. 1–135.

Anne Chao, P. T. Tsay, Sheng-Hsiang Lin, Wen-Yi Shau, and Day-Yu Chao (2001). "The applications of capture-recapture models to epidemiological data." *Statistics in Medicine*, Volume 20, pp. 3123–3157. DOI: 10.1002/sim.996.

Shelby D. Gerking (1953). "Evidence for the Concepts of Home Range and Territory in Stream Fishes." *Ecology*, Volume 34, Issue 2, pp. 347-365.

Charles J. Krebs (1989 ed.). "Ecological Methodology." *Harper & Row Publishers.*

Charles J. Krebs (2014 ed., in prep.). "Ecological Methodology." Accessed at *https://www.zoology.ubc.ca/~krebs/books.html.*

Frederick C. Lincoln, (1930). "Calculating waterfowl abundance on the basis of banding returns." *U.S. Department of Agriculture Circular*, No. 118, pp. 1-4.

Intern, Linda Lorning Nature Foundation (2022). "Seining for Long Pond Creatures." *https://llnf.org/blog/2022/8/25/seining-for-long-pond-creatures.*

Gary A. Nelson (2023). "`fishmethods`: Fishery Science Methods and Models." R package version 1.12-1. *https://CRAN.R-project.org/package=fishmethods.*

David L. Otis, Kenneth P. Burnham, Gary C. White and David R. Anderson (1978). "Statistical Inference from Capture Data on Closed Animal Populations." *Wildlife Monographs*, No. 62. Accessed at *https://www.jstor.org/stable/pdf/3830650.pdf*.

Carl Georg Johannes Petersen (1896). "The Yearly Immigration of Young Plaice into the Limfjord from the German Sea." *Report of the Danish Biological Station to the Ministry of Fisheries*, No. 6, pp. 1-48.

Zoe E. Schnabel (1938). "The Estimation of the Total Fish Population of a Lake." *American Mathematical Monthly*, Volume 45, Issue 6, pp. 348–352.

F. X. Schumacher and R. W. Eschmeyer (1943). "The estimation of fish populations in lakes or ponds." *J. Tenn. Acad. Sci.*, Volume 18, pp. 228-49.

Kerri Vierling (2012). "WLF 315 Wildlife Ecology I Lab: Population Estimation Concepts." Accessed at *https://www.webpages.uidaho.edu/wlf314/labs/Population_Estimation_Lab_Handout_2012.pdf*.

Paul Wesson, Ali Mirzazadeh, and Willi McFarland (2018). "A Bayesian approach to synthesize estimates of the size of hidden populations: the Anchored Multiplier." *Int J Epidemiol.*, Volume 47, Issue 5, pp. 1636–1644. DOI: 10.1093/ije/dyy132.

Paul Wesson, Manjari Das, Mia Chen, Ling Hsu, Willi McFarland, Edward Kennedy, and Nicholas P. Jewell (2023). "Evaluating a Targeted Minimum Loss-Based Estimator for Capture-Recapture Analysis: An Application to HIV Surveillance in San Francisco, California." *American Journal of Epidemiology*, Volume 193, Issue 4, pp. 673–683. *https://doi.org/10.1093/aje/kwad231*.

Paul Wesson (2024). "Inclusive Epidemiology: Methodologies for Accessing and Understanding Socially Marginalized Populations in Epidemiologic Research." Lecture on March 8, 2024 at Johns Hopkins Bloomberg School of Public Health.

Gary C. White, David R. Anderson, Kenneth P. Burnham, and David L. Otis (1982). "Capture-recapture and removal methods for sampling closed populations." *Los Alamos National Laboratory* Report LA-8787-NERP. Accessed at *https://sites.warnercnr.colostate.edu/gwhite/white-et-al-1982/.*

Appendix

Code is available in the `code/` folder in this project, also posted on GitHub: *https://github.com/m-qin/schnabel_capture_recapture_estimators*.