Exploring Capture Recapture Methods: From Historical Origins to Modern Applications

Estimating the size of a population is often of utmost importance, and capture-recapture methods are the standard approach to doing so. This paper explores the history of capture-recapture methodology and delves into their effectiveness in different scenarios. Different scenarios include varying sample sizes and sample counts, as well as whether samples are independent or not. Derivations of the formulas used to arrive at estimates are presented first, justifying their use. Then, an exploration into different scenarios and the associated resulting estimator is performed. Bootstrapping is also shown to be an effective way of estimating the variance of such estimators, and is then used to compare variances across the different scenarios. Current applications and work in the field are also briefly discussed.

Introduction

Accurately estimating the size of a population is a key goal in the field of statistics. One of the most common methods of doing this is using capture-recapture methodology. The general idea of capture-recapture is gathering multiple samples from the same population and using information about the *overlap* to estimate the population's size.

The first known use of capture-recapture methodology was by Pierre-Simon, Marquis de Laplace (Seber and Schofield 2019) when Laplace used the method to attempt to estimate the total population of France. While this seemed to be an incredibly daunting task when he performed it in 1786, the use of this methodology allowed him to get an accurate estimate without actually counting everyone in the population.

The first time this method was used for its most commonly known purpose was in 1917 when Dahl (Cren 1965) used capture-recapture to estimate the size of a trout population in Norway. Since then there have been hundreds of papers published applying this method to estimate animal population sizes. There has been a large variety of work in this field, with papers ranging from those using the classic methods to estimate salmon populations (Schwarz and Dempson 1994), to using spatial capture-recapture data to estimate bear populations (Sun et al. 2017).

However, a relatively recent development in the use of these methods is applying them to the field of epidemiology (Seber and Schofield 2019). It is often the case that epidemiologists are trying to estimate the number of people that have a certain disease. Obviously they can't just go around asking everyone in a certain area if they have the disease, for logistic and budget reasons. As such, that makes capture-recapture the perfect method to apply to this situation.

One of the most comprehensive papers detailing the application of this method in this setting is provided by Chao et al. (2001). They highlight some of the initial formulations of capture-recapture methods, as well as detailing some approaches that work even when complications are present, such as dependence exists between the samples. These methods require much more in terms of discussion of assumptions, and will not be discussed in this paper.

The goal of this paper is to give an expository review of some of the methods present in the paper by Chao and colleagues mentioned above, as well as undertaking a simulation study to evaluate the effectiveness of these methods in different situations.

In this paper there is a section devoted to the methods of capture-recapture in both a two list and a three list setting, as well as the assumptions needed. The methods section also contains a discussion on the use of bootstrapping to generate variances for these estimates. The results section will contain the results of a simulation study highlighting the performance of these methods under different coverage rates, different list counts, and when dependence exists between lists. A conclusion and a brief discussion of limitations and next step will be included at the end.

The variance of our estimator is also of utmost importance to know. Since a closed form variance can be hard to attain, bootstrap resampling can be used to estimate variances around these estimates (Chao 1987). The validity of this method and will be confirmed against a large sample approximation of the variance.

Methods

As stated before, capture-recapture methodology was first described in 1786, when Pierre-Simon, Marquis de Laplace (Seber and Schofield 2019) used the method to attempt to estimate the population of France. The first "capture" was a list based on the birth registries in all of France. The second "capture" consisted of looking at the birth registries in just a few parishes in France whose populations were more easily counted. By looking at the births in just these parishes he was able to estimate the total population of France. The idea here is that the proportion of the number of births in those parishes to their population sizes should be equivalent to the proportion of the number of births in France to the population size of France, which we are trying to determine. As an equation, this can be thought of as the following:

$$\frac{\text{Number of births in select parishes}}{\text{Population of select parishes}} \approx \frac{\text{Number of births in France}}{\text{Population of France}}$$

We know both values on the left hand side, and the numerator on the right, so using this information we can get an estimate of the total population of France by solving for that quantity. In this situation, we would have:

$$\text{Population of France} \approx \frac{\text{Population of select parishes} * \text{Number of births in France}}{\text{Number of births in select parishes}}$$

The same idea was used much later in 1917 when Dahl (Cren 1965) applied a similar methodology to estimate the size of trout populations. In this method, which is the more commonplace use nowadays, first one sample of size n_1 is taken and all the fish in it are marked in some way. Then, fish are caught a second time, and this time n_2 are observed and we note how many are already marked, calling this m_{12} since they appeared in both samples 1 and 2. We will denote our total population with N, thus, our estimate of N will be $\widehat{N}$. Using the same theory as before, we would hope that the proportion of fish we capture in our sample from the total population is the same as the proportion of marked fish we saw again in the second sample.

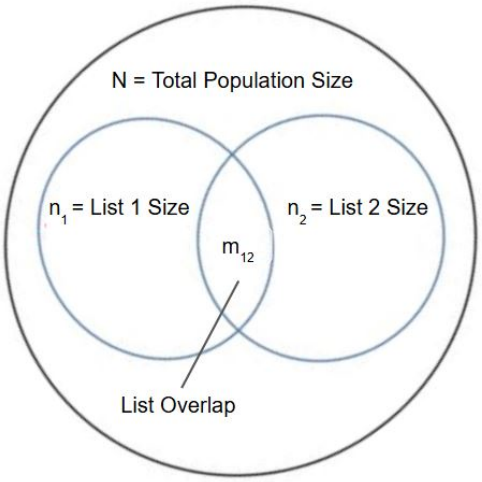


Figure 1: Venn Diagram Displaying Idea of Capture-Recapture.

This can be seen visually in Figure 1, and is present in equation form here:

$$\frac{n_1}{N} \approx \frac{m_{12}}{n_2}$$

Thus, we have our estimate of N as:

$$\widehat{N} = \frac{n_1 n_2}{m_{12}}$$

This estimate allows us to get an idea of the total size of an animal population using what can be relatively small samples compared to the entire population. We must note here there is an assumption that the probability of showing up in one sample is independent of appearing in the other sample (More on the importance of this assumption will come later). This estimate is often referred to as the "Lincoln Petersen" estimate, named for the ornithologist Frederick Charles Lincoln who described its creation, and marine biologist Carl Georg Johannes Petersen who pioneered its use.

Brief Example

To make the method a little more concrete, we will consider a small working example. For this example, imagine we are trying to estimate the size of a fish population in a pond.

To perform this procedure, we first take a sample of 100 fish and mark them in some way. Then, after letting the fish disperse through the pond, we come back and take another sample

of 100 fish. Now, we count how many marked fish appear in the second sample. In this example, suppose it was 20.

Now, we use these values to calculate $\widehat{N}$:

$$\widehat{N} = \frac{n_1 n_2}{m_{12}} = \frac{100 * 100}{20} = 500$$

Thus, we see in this situation we would estimate that the number of fish in the pond is 500. It should be noted that in most situations we won't have a perfect estimate of our population like we do here, simply because the same amount won't always overlap, but generally we will have an unbiased estimate. This will be shown in greater detail later.

Epidemiological setting

While the idea of capture-recapture was used extensively in the estimation of animal populations, this isn't the only scenario in which it works. The extension to the epidemiological setting follows naturally, with a slight adjustment to our procedure for collecting the samples. Instead of thinking about collecting subsequent samples where we are marking the captured, we instead consider the existence of two or more lists, hopefully containing overlapping information. These lists are simply lists of people known to have some disease, perhaps by a local hospital, a school's health center, a disease registry, or an insurance company. When used in reference to human populations the term "multiple-record system," is used, and our captures become ascertainments, which indicate someone appears on a given list.

Chao et al. (2001) noted three main differences between the animal population and the human population estimation. Firstly, we normally have a larger amount of lists for animals, whereas epidemiological situations often only have two to four lists. Second, when we sample from an animal population, the samples have a chronological ordering, whereas all the lists could be collected at the same time for epidemiological data. Lastly, in animal studies the method of obtaining samples is often the same, whereas the lists may be created in different ways, leading to some level of bias between the lists.

Assumptions

In all settings, we must assume independence for all our samples and lists. Without this, we run into the issue of getting inaccurate estimates based on our lists. Ideally, the rate of recapture in the second sample is equal to the proportion of individuals captured in the initial sample, giving us the formula from above $\frac{n_1}{N} \approx \frac{m_{12}}{n_2}$.

If the probability of showing up in one list is *positively* correlated with showing up in the other, then our estimate will be an underestimate. We can see that if there exists a positive correlation between appearing on the two lists, the chance of showing up on the second list

will be greater than the proportion of individuals from the population on the first list. In a formula, $\frac{n_1}{N} < \frac{m_{12}}{n_2}$, and rearranging to get our estimate alone, we would have $N > \frac{n_1 n_2}{m_{12}}$. Since we are still estimating using the right hand side, we would have an underestimate of the true population size in this case.

If the probability of showing up in one list is *negatively* correlated with showing up in the other, then our estimate will be an overestimate. In the case where there exists a *negative* correlation between appearing on the two lists, we have that the chance of showing up on the second list will be less than the proportion of individuals on the first list. Thus, we have $\frac{n_1}{N} > \frac{m_{12}}{n_2}$ which simplifies to $N < \frac{n_1 n_2}{m_{12}}$. Since our estimate is based on the right hand side, we would have an overestimate of the true population size in this case.

In that vein, we also assume *homogeneity* within our lists. This means the probability that person 1 appears in the list is the same probability that person 2 appears in the list. This doesn't necessarily mean that appearing in each list has the same probability, just that each individual has the same chance of showing up in any given list.

Lastly, we assume that the population we are estimating the size of is *closed*. All this means is that no individuals are entering or leaving our population between our list making. In practice even if the population isn't closed we can assume the population size is roughly constant by assuming that the rates of entering and leaving are roughly equal.

Further explorations of the issues that occur without independence are present in the results section.

Three list case

The case of working with just two lists is relatively straightforward, and unfortunately not always much use. Since we only have two lists, we are limited to looking at the overlap between just the two. It's no surprise that with more data we get more accurate estimates of the true value, so the natural next step is looking at the case of three lists. We will go through the development of a three list estimator, as described in Chao et al. (2001).

Unfortunately, our old method of estimation does not commute immediately, so we must define some new notation. We first define Z_{00}, Z_{10}, Z_{01}, and Z_{11} as the counts of people that appear in neither list, the first list, the second list, and both, respectively. Also, we define the probability person 1 through N appears in list 1 as $p_{11}, p_{21}, ..., p_{N1}$. Similarly, we denote the probability person 1 through N appears in list j as $p_{1j}, p_{2j}, ..., p_{Nj}$. Last, we define X_{ij} as follows: if person i is in list j we say $X_{ij} = 1$, whereas if they do not appear in list j we say $X_{ij} = 0$. Thus, the indicator function $I[X_{ij} > 0]$ tells us whether or not person i appeared on list j. Now that we have notation, we will discuss sample coverage in the two list case.

We want to define a measure of coverage between our lists. We would like a proportion of the number of people in list 2 and list 1 over the number of people in list 2. Thus, we define the sample coverage of list 1 with respect to list 2 as:

$$C^*_{II}(L_1) = \frac{\sum_i p_{i2} I[X_{i1} > 0]}{\sum_i p_{i2}}$$

The numerator is the sum of the probabilities of each person in list 1 appearing in list 2, while the denominator is the number of people in list 2. Taking the expectation will thus give us the ratio we desire. Doing so allows us to get a better idea of what our estimator should be:

$$E[C^*_{II}(L_1)] = E\left[\frac{\sum_i p_{i2} I[X_{i1} > 0]}{\sum_i p_{i2}}\right] = \frac{E[\sum_i p_{i2} I[X_{i1} > 0]]}{E[\sum_i p_{i2}]}$$

By the linearity of expectations and the fact that $E[p_{i2}] = E[X_{i2}]$ we have:

$$\frac{E[\sum_i p_{i2} I[X_{i1} > 0]]}{E[\sum_i p_{i2}]} = \frac{\sum_i E[X_{i2} I[X_{i1} > 0]]}{\sum_i E[X_{i2}]} = \frac{Z_{11}}{n_2}$$

Taking the average over this value for both lists we can get an estimate of our sample coverage. Thus, our estimate of the sample coverage is:

$$\widehat{C} = \frac{1}{2}\left(\frac{Z_{11}}{n_1} + \frac{Z_{11}}{n_2}\right) = 1 - \frac{1}{2}\left(\frac{Z_{10}}{n_1} + \frac{Z_{01}}{n_2}\right)$$

Given independence in both ways, within and between the lists, we can reduce C to:

$$\widehat{C} = \frac{1}{2}\left(\frac{Z_{11}}{n_1} + \frac{Z_{11}}{n_2}\right) = \frac{1}{2}\left(\frac{n_1}{N} + \frac{n_2}{N}\right) = \frac{\frac{n_1+n_2}{2}}{N}$$

We will call this top quantity $D = \frac{n_1+n_2}{2}$. Notice that this is equivalent to $D = \frac{1}{2}(Z_{10} + Z_{11} + Z_{01} + Z_{11})$ Thus, our estimate of N is:

$$\widehat{N} = \frac{D}{\widehat{C}}$$

Substituting D and $\widehat{C}$ into this, we get that in the two list case our estimate is:

$$\widehat{N} = \frac{\frac{n_1+n_2}{2}}{\frac{1}{2}\left(\frac{Z_{11}}{n_1} + \frac{Z_{11}}{n_2}\right)} = \frac{1}{Z_{11}} \frac{n_1 + n_2}{\frac{n_2+n_1}{n_1 n_2}} = \frac{n_1 n_2}{Z_{11}}$$

This can be recognized as the same estimate we had earlier, but setting up the two list case in this way makes the intuition of the three list case much more clear.

First, we would like an analog of the sample coverage in the three list case. Similar to what we did earlier, we define the sample coverage of list 3 with respect to lists 1 and 2 as:

$$C^*_{III}(L_1 \cup L_2) = \frac{\sum_i p_{i3} I[X_{i1} + X_{i2} > 0]}{\sum_i p_{i3}}$$

Once again, we can take the expectation and get:

$$E[C^*_{III}(L_1 \cup L_2)] = E\left[\frac{\sum_i p_{i3} I[X_{i1} + X_{i2} > 0]}{\sum_i p_{i3}}\right] = \frac{E[\sum_i p_{i3} I[X_{i1} + X_{i2} > 0]]}{E[\sum_i p_{i3}]}$$

Once again, by the linearity of expectations and the fact that $E[p_{i3}] = E[X_{i3}]$ we have:

$$\frac{E[\sum_i p_{i3} I[X_{i1} + X_{i2} > 0]]}{E[\sum_i p_{i3}]} = \frac{\sum_i E[x_{i3}] I[X_{i1} + X_{i2} > 0]]}{\sum_i E[x_{i3}]} = \frac{Z_{011} + Z_{101} + Z_{111}}{n_3}$$

Notice that since $n_3 = Z_{001} + Z_{011} + Z_{101} + Z_{111}$, we can rewrite the above as:

$$\frac{Z_{011} + Z_{101} + Z_{111}}{n_3} = \frac{n_3 - Z_{001}}{n_3} = 1 - \frac{Z_{001}}{n_3}$$

We will once again take an average over these values for all three lists to get an estimate of our sample coverage. Thus, our estimate of the sample coverage is:

$$\widehat{C} = 1 - \frac{1}{3}\left(\frac{Z_{100}}{n_1} + \frac{Z_{010}}{n_2} + \frac{Z_{001}}{n_3}\right)$$

We define D in a similar way as before, although it looks more complicated:

$$D = \frac{1}{3}\left(\sum_i I[X_{i2} + X_{i3} > 0] + \sum_i I[X_{i1} + X_{i3} > 0] + \sum_i I[X_{i1} + X_{i2} > 0]\right)$$

This can be thought of as the average number of people observed in each pair of two of the lists. Naturally, we want a formula for quickly finding D. First, we define M as the total distinct people we observe over all the lists. Notice we have $M - Z_{100} = Z_{010} + Z_{001} + Z_{110} + Z_{101} + Z_{011} + Z_{111}$. The same is true for $M - Z_{010}$ and $M - Z_{001}$. Now, our formula for finding D is:

$$D = \frac{1}{3}\left((M - Z_{100}) + (M - Z_{010}) + (M - Z_{001})\right) = M - \frac{1}{3}(Z_{100} + Z_{010} + Z_{001})$$

We combine these two values into our estimate $\widehat{N}$ to get:

$$\widehat{N} = \frac{D}{\widehat{C}}$$

When we have independence, this will yield an unbiased estimate of the total population size. Simulations confirming this and the effectiveness of the bootstrap in this situation will follow.

Example

To make this a bit more tangible we will consider a small example. Suppose there is some infection going around a college campus, and lists of infected students are kept by three different groups. Imagine these lists each contain 100 students, although each student is not necessarily on only one list.

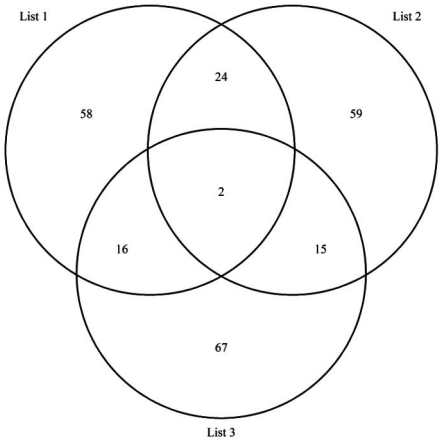


Figure 2: Venn Diagram of Small Example.

First, we must calculate the amount of singletons. That is, the number of students detected in one of the lists, but not appearing in either of the other two. From Figure 2, we can see that for this example, the lists end up containing 58, 59, and 67 singletons. These are the values of z_{100}, z_{010}, and z_{001}, respectively.

Thus, our value of $\widehat{C}$, the sample coverage, is:

$$\widehat{C} = 1 - \frac{1}{3}\left(\frac{58}{100} + \frac{59}{100} + \frac{67}{100}\right) = 1 - \frac{61.3}{100} = 0.387$$

Now, to get D we first need to get M, the total number of observed people across all lists. In this example our value of M ended up being 241. So we can calculate D as follows:

$$D = 241 - \frac{1}{3}(58 + 59 + 67) = 241 - 61.3 = 179.7$$

Thus, we can get our estimate of $\widehat{N}$ as:

$$\widehat{N} = \frac{D}{\widehat{C}} = \frac{179.7}{.387} = 464.34 \approx 464$$

So our final estimate in this example is that the total population is of size 464. This data was generated by sampling from population with size 500. Thus, our estimate is a bit of an underestimate, but overall close to the true value.

Variance of the Estimate

Recall from earlier our estimate of the population size in the two list case, $\widehat{N} = \frac{n_1 n_2}{m_{12}}$. While this estimator may be unbiased, the next most important thing we must know about it is the variability behind its estimates. In the two list case we can do so as follows:

First, notice that since m_{12} is the number of observations present in list 2 that were already present in list 1, we can model its distribution as a hypergeometric random variable. This makes intuitive sense, as m_{12} is generated by counting the number of people already present in list 1 (successes) obtained in our sample from N (the population). Thus, we can write m_{12} $\sim H(n_2, n_1, N)$, where H represents a hypergeometric distribution with a sample size of n_2, n_1 successes in the whole population, and a population size of N that we are sampling from. The variance of m_{12} is thus given to us by the variance of a hypergeometric random variable:

$$V(m_{12}) = \left(\frac{N - n_2}{N - 1}\right)(n_2)\left(\frac{n_1}{N}\right)\left(\frac{N - n_1}{N}\right)$$

We will use the fact that $\widehat{N} = \frac{n_1 n_2}{m_{12}}$ and the delta method (Liu 2012) to find this variance. We are able to apply the following formula to find the variance, where $f(m_{12})$ is a function of m_{12}:

$$\text{Var}(f(m_{12})) \approx (f'(m_{12}))^2 \,\text{Var}(m_{12})$$

Notably, $\widehat{N}$ is a function of m_{12}, with $\widehat{N} = f(m_{12}) = \frac{n_1 n_2}{m_{12}}$. Although n_1 and n_2 have variability associated with them, it is small relative to the variance of m_{12}. Hence, they will be treated as constants for the calculation of the variance with no effect on the final result. Thus, taking a derivative yields $f'(m_{12}) = -\frac{n_1 n_2}{(m_{12})^2}$. So, we have:

$$\text{Var}(\widehat{N}) \approx \left(-\frac{n_1 n_2}{(m_{12})^2}\right)^2 \left(\frac{\widehat{N}-n_2}{\widehat{N}-1}\right)(n_2)\left(\frac{n_1}{\widehat{N}}\right)\left(\frac{\widehat{N}-n_1}{\widehat{N}}\right)$$

Noting that $\left(-\frac{n_1 n_2}{(m_{12})^2}\right)^2 = \widehat{N}^2\left(\frac{1}{m_{12}}\right)^2$, we can perform some algebra to get:

$$\text{Var}(\widehat{N}) \approx \left(\frac{1}{m_{12}}\right)^2 \left(\frac{\widehat{N}-n_2}{\widehat{N}-1}\right)(n_2)(n_1)\left(\widehat{N}-n_1\right)$$

Now, observe that $\widehat{N}-n_2 = \frac{n_1 n_2}{m_{12}} - \frac{n_2 m_{12}}{m_{12}} = \frac{n_2}{m_{12}}(n_1 - m_{12})$ and by similar calculations, $\widehat{N}-n_1 = \frac{n_1}{m_{12}}(n_2 - m_{12})$. So now we have:

$$\text{Var}(\widehat{N}) \approx \left(\frac{1}{m_{12}}\right)^2 \left(\frac{n_2}{m_{12}}(n_1 - m_{12})\right)\left(\frac{1}{\widehat{N}-1}\right)(n_2)(n_1)\left(\frac{n_1}{m_{12}}(n_2 - m_{12})\right)$$

Rearranging this gives us:

$$\text{Var}(\widehat{N}) \approx \left(\frac{1}{m_{12}}\right)^4 \left(\frac{n_1 n_2}{\widehat{N}-1}\right)(n_2(n_1 - m_{12}))(n_1(n_2 - m_{12}))$$

Noticing that $\frac{1}{m_{12}}\left(\frac{n_1 n_2}{\widehat{N}-1}\right) = \frac{n_1 n_2}{n_1 n_2 - m_{12}}$. Asymptotically this term approaches 1, so our ultimate approximation of the variance is thus:

$$\text{Var}(\widehat{N}) \approx \left(\frac{1}{m_{12}}\right)^3 (n_2(n_1 - m_{12}))(n_1(n_2 - m_{12}))$$

However, this variance is only able to be computed in the two list case. The three list case doesn't have a nice closed form like this does. Thus, we would like to confirm that bootstrapping yields appropriate variance estimates in the two list case using this formula and the bootstrap. Once we know the bootstrap works for two lists, we can assume it commutes and will work for three lists as well. Thus, we will be able to find variances in the three list case, where it otherwise may not be possible to do so.

Results

The demonstration of application of this method will use only simulated data, as real world data won't provide us with a known population size we can compare to our estimates.

Generally, our simulated data is created by assuming some population size N. Then, we sample from this population to create our lists, where each person has probability p of appearing in a given list.

Table 1: Sample of List Data Format

ID	List 1	List 2	List 3
1	1	1	0
2	1	0	1
3	1	0	0
4	1	1	1
...	...	...	...

Table 1 displays what the data look like. A 1 in a cell indicates that person is present on that list, while a 0 indicates they are not present. For example, the individual with ID 1 is present on lists 1 and 2, but not 3. Notice, there are some individuals not present on any lists, such as individual 8.

Bootstrapping for Variance

Our first endeavor into simulation will be to show that the bootstrap variance is a valid substitute for the large sample approximation of the variance, as this will allow us to use it in our comparisons of different scenarios. This is a necessary step, as the large sample approximation will only apply in the two sample case, and most situations we are interested in involve at least three lists. This is because of the fact that the three sample case does not have a simple formula for direct computation of the variance, although we would still like an idea of the variability of our estimates.

First, we confirm that the bootstrap will yield variances similar to those we obtain through the use of the large sample approximation. For this simulation, in each iteration we will generate two lists of size 100 by sampling from our population of size 500. Then, for each iteration we will directly calculate the variance using the formula for the large sample approximation, as well as creating 100 bootstrap resamples.

Our method for bootstrapping is as follows. For each individual captured on at least one list, define their "coverage history" as the lists they are present on. For example, if person 1 was on lists 1 and 2, their coverage history would be (1,1). To bootstrap, we consider all individuals we have on at least one list, then resample from these capture histories with replacement.

Our new sample size should be equal to the total number of people we observed, as with any bootstrap procedure. After doing this, we can compute our estimate $\widehat{N}$ for each bootstrap sample to get an idea of the variability around our estimate. This method is taken from method 1 of Norris and Pollock (1996).

To compute the variance we use the classic definition of sample variance applied to the bootstrap estimates. Our formula for this is:

$$\hat{Var}(\hat{N})_{boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{N}_b - \bar{\hat{N}})^2$$

Where B is the number of bootstrap iterations and $\bar{\hat{N}} = \frac{1}{B}\sum_{b=1}^{B}\hat{N}_b$. This calculation is taken from page 24 of Laura Gruber's master thesis (Gruber 2023).

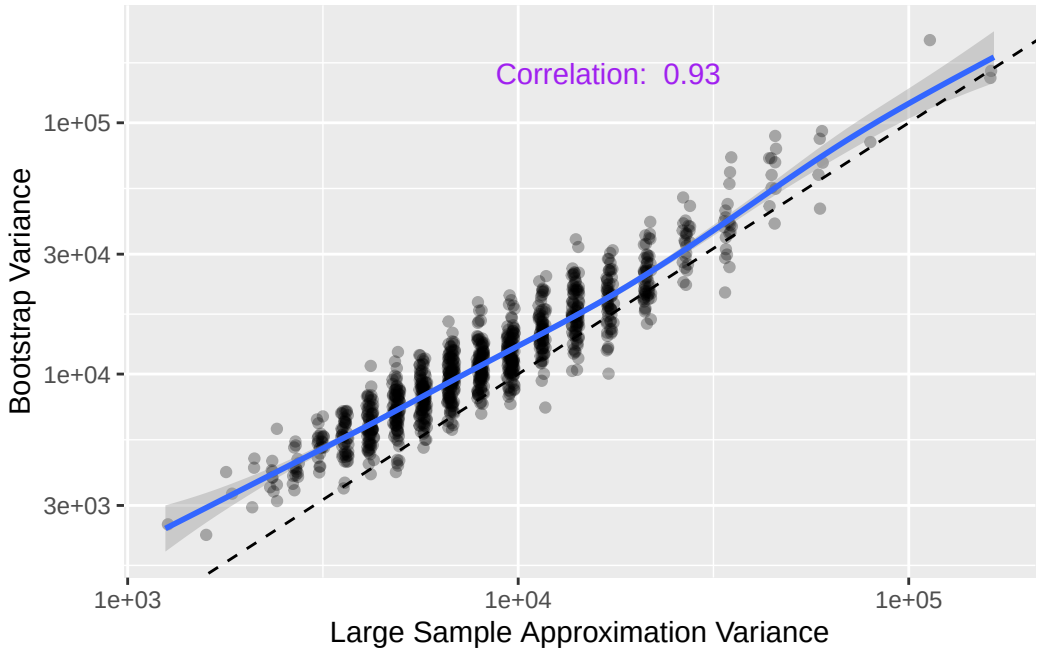


Figure 3: Scatterplot of Large Sample Approximation of Variance vs Bootstrap Estimate of Variance with Line y=x and Smoothing Spline. We observe that the bootstrap estimate tends to be slightly larger than the LSA, though the correlation between the two estimates is quite strong (0.93, 95% CI ranges from 0.92 to 0.94).

From Figure 3 we can see that our bootstrap estimate of variance is almost always overestimating the large sample approximation, as the points generally lie above the line $y = x$. Also, our bootstrap estimate is generally worse as our large sample approximation increases, which is of moderate concern, but since those variances appear to be outliers, we can safely use

the bootstrap as an estimate of the variance when we have at least 100 bootstrap resamples. Additionally, since the estimates are close enough, it seems reasonable that we can compare two bootstrap estimates of the variance to each other.

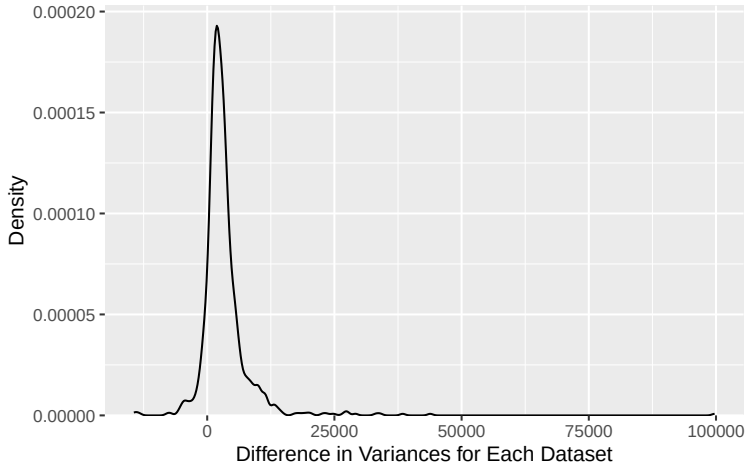
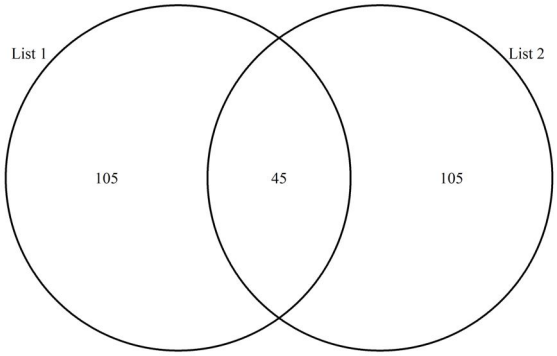


Figure 4: Density Plot of Difference Between Large Sample Approximation and Bootstrap Variance. We see that the bootstrap overestimates slightly in most situations, with some rare large outliers.

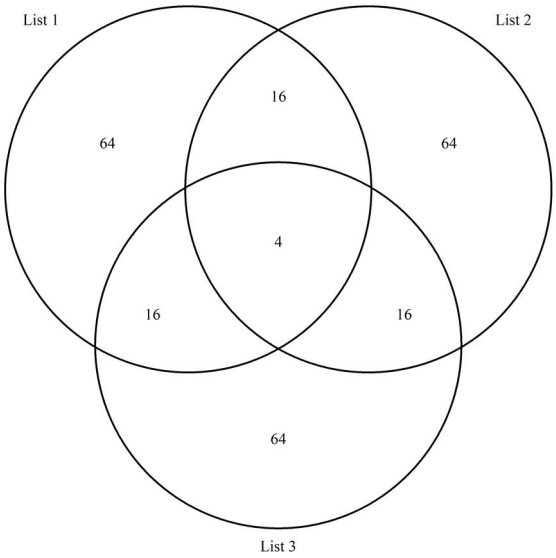
Another way to visualize these differences is through Figure 4. We can see that the distribution of differences between the two methods is extremely right skewed and centered slightly above 0. Thus, as noticed previously, the bootstrap estimates are often overestimating the large sample approximation of the variance. Additionally, the right tail contains very few values, with only 9 differences being greater than 25000 in magnitude.

Two Versus Three Lists

It is generally accepted that the more data we are able to obtain the better our estimates will be, however, there is often a difficulty in obtaining more data. In the case of capture-recapture and list ascertainment, there is an inherent push-pull relationship between our capture probabilities and the number of lists we have. This is owing to the fact that research often has logistic or budgetary constraints that must be balanced. With that in mind, we sought to find evidence in favor of using more lists, even if necessarily we will have fewer captures on each list.



(a) Two Lists (p = 0.3, N = 500)



(b) Three Lists (p = 0.2, N = 500)

Figure 5: Expected Overlap with Varying List Counts. Note the Total Population is of Size 500

Table 2: Table of Summary Statistics of Estimators for Different List Counts and Different Capture Probabilities p (Simulations = 1000)

	Three Lists (p = 0.2)	Two Lists (p = 0.2)	Two Lists (p = 0.3)
Median	500.00	500.00	500.00
Mean	504.95	518.18	506.22
Standard Deviation	50.50	107.96	56.98

From Table 2 we can see that the case with three lists performs the best, in terms of both reducing bias and variability around our estimate. The median of all our estimates is the same, the correct value of 500. However, when comparing the use of 3 lists with just 2 lists, both with capture probability of 0.2, we see that the additional list reduces the bias by about 75%. Additionally, the third list roughly halves the standard deviation, a huge improvement in terms of the variability of our estimate.

Based on the expected number of people in the overlap of the lists, we get that the two list case with capture probability $p = 0.3$ should capture 255 distinct people on average, while the three list case with capture probability $p = 0.2$ will capture 256. This can be seen by summing all the values present in Figure 5.

Although we only saw one more person in terms of distinct people observed, we can see from Table 2 that we can obtain better estimates solely by using three lists, even with lower capture

probability. Thus, making the more fair comparison by increasing the capture probability in the two list case, we obtain a similar number of people across the lists as in the three list case, yet it is still slightly more biased. Additionally, the standard deviation of our estimate is reduced by about 7 people in the three list case. Note that while these aren't exact estimates and rely on the bootstrap, we showed previously that the bootstrap yields similar variances to the large sample approximation.

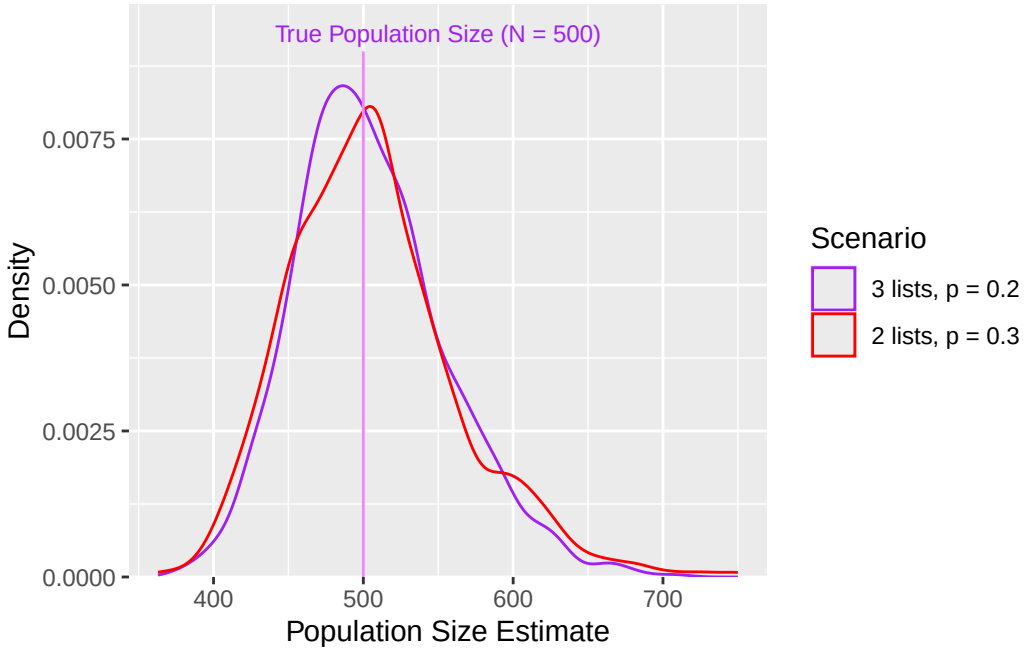


Figure 6: Density Plot of Population Estimates with Different List Counts. We see the two scenarios are very similar, although the 2 list case has heavier tails.

We draw the same conclusions from Figure 6 that we did previously. The two list case leads to a slightly more positively biased estimator on average, hence the peak slightly to the right of the true population size of 500. Additionally, the tail is a bit thicker on both ends for the two list case compared to the three lists. Besides these slight differences, the distributions look largely similar.

Given this information, it is preferable to use more lists when possible, even if the size of each list must be reduced to account for it.

Different capture rates

Another interesting perspective is to see how much better our estimates get as our capture probabilities increase. For this case we still suppose a population of size 500, but instead of

always capturing 100 people per list, we capture 50, 100, or 200, depending on the scenario. Thus, our capture rates vary from 10% to 20% to 40%. Note that in this case we have three lists again for all three capture probabilities.

Figure 7: Density Plot of Population Estimates with Different Capture Probabilities. We observe that our estimates become much less variable as the capture probability increases.

We can see from Figure 7 that as the capture probability increases the variance around our estimate decreases drastically. Additionally, while all the distributions are generally unbiased, the slight underestimate of the peak is also corrected by the increase in capture probability. This isn't a surprising result, as we get more people on our lists our estimates get better, but just how much the variance decreases as our probability increases is of note.

Table 3: Table of Summary Statistics of Estimators for Different Capture Probabilities

Capture Probability	10%	20%	40%
Median	510.71	500.00	500.78
Mean	527.08	504.95	501.17
Standard Deviation	137.17	50.50	18.19

Table 3 supplies summary statistics for the three different capture probability scenarios. We can see that all 3 methods have a positive bias, although the size of the bias decreases as our

capture probability increases. Similar to what we saw above, the standard deviation of our estimate decreases by roughly 60% as we double our capture probability.

With all that said, whenever possible, a higher capture probability should be strived for. This isn't always attainable, but any increase in capture probability will reduce the variance, so even a slight increase is beneficial.

Dependence

Last, we illustrate a very brief example where our estimates fail if there exists dependence between the lists. In this example, we consider three scenarios:

1. The probability of appearing on list 1 is completely independent of the probability of appearing on list 2. In both cases the probability of appearing on a given list is 0.2.
2. The probability of appearing on list 2 is somewhat dependent on appearing on list 1. In this case the probability of appearing in list 1 is 0.2, whereas the probability of appearing in list 2 is either 0.5 or 0.2 if someone is in list 1 or not, respectively. This set up induces a correlation of 0.43 between the two lists.
3. The probability of appearing on list 2 is highly dependent on appearing on list 1. In this case the probability of appearing in list 1 is 0.2, whereas the probability of appearing in list 2 is either 0.7 or 0.2 if someone is in list 1 or not, respectively. This set up induces a correlation of 0.69 between the two lists.

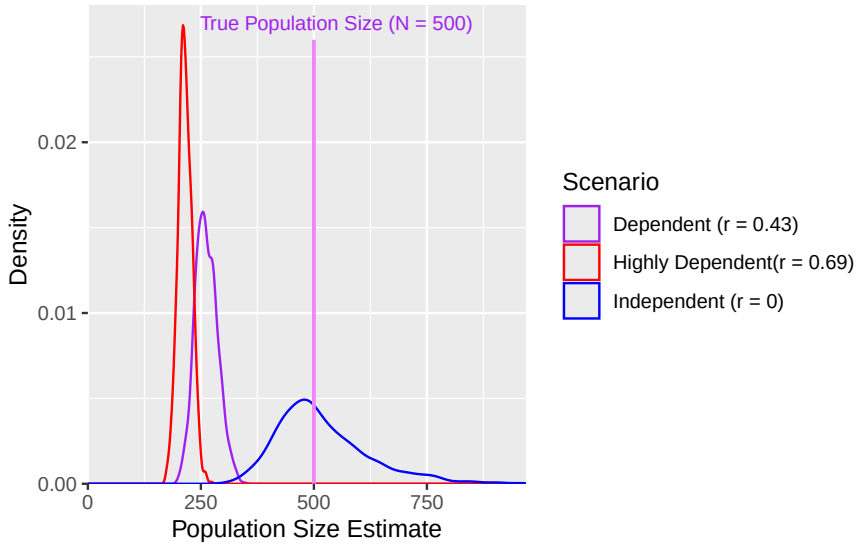


Figure 8: Density Plot of Population Estimates in Different Scenarios. We observe that as our lists become more dependent our estimate becomes more and more biased.

As we can see from Figure 8, our estimates get worse and worse as the correlation between the lists goes up. In the case of independence between the two lists our estimate is roughly centered around the true value of 500. However, with a correlation of 0.43 between the two lists our estimate appears centered around 500, and as the correlation increases to 0.69 it dips even lower. The exact values are displayed as follows:

Table 4: Table of Summary Statistics of Estimators in Different Scenarios

	Independent (r = 0)	Dependent (r = 0.43)	Highly Dependent (r = 0.69)
Median	498.64	259.77	213.87
Mean	518.49	262.02	214.87
SD	101.95	24.30	15.22

From Table 4 we can see that in the independent case our classic estimator produced a roughly unbiased result, whereas the dependent scenarios produce wildly biased results. Similar to the underestimate when the correlation is positive, there will be an overestimate when the correlation is negative.

Chao proposes some solutions to the issue of dependence beyond the scope of this paper. The most commonly used correction for the dependence issue is to use something known as a "log linear model," which allows us to correct for this dependence and the associated bias in our estimates.

Summary of Results

The results generally confirmed intuition that was already present. Higher capture rates and more lists lead to better estimates, as we may have more data to make those estimates with. Additionally, dependence between the lists will lead to biased estimates, which can be corrected for with log-linear models.

Discussion

Limitations

While the methods in this paper exist as a simple, straightforward way to estimate a population size, their simplicity comes at the cost of lacking a degree of robustness.

First, it should be duly noted that the methods outlined in this will paper will only produce unbiased estimates in the case of independence between lists. If this assumption is violated, dependence can lead to over or under estimates. While there are ways to fix this, they are

beyond the scope of this paper and those interested should look to Chao et al. (2001) for a more comprehensive dive into the methodology.

Further, we assumed a closed population with our models, which is close enough to being true in most cases, as discussed in the Assumptions section. However, if we have reason to believe this is not true, different methods may have to be used.

Lastly, these methods only work if the sample coverage is adequate. If the lists capture a very small percent of the population, we could have issues with overlap counts. Thus, a low coverage approach may need to be taken, which can also be seen in Chao et al. (2001).

Application

Capture-recapture is most commonly used in animal populations, but has also been used in other fields. As discussed, these methods are popular in epidemiology for estimating the size of a diseased population (Chao et al. 2001). Additionally, one application that capture-recapture has been gaining traction in is that of estimating the number of people that are victims of modern slavery. By using lists provided by various monitoring organizations, such as local authorities or the National Crime Agency in the UK, the total number of victims was able to be estimated as somewhere in the range 10000-13000 (Silverman 2020). These estimates can play a pivotal role in policy making, as the Modern Slavery Act 2015 was heavily influenced by this work. With that in mind, it is of great importance to make sure these estimates are as accurate as possible, but also reflect the variability inherent in their estimation.

Conclusion

Ultimately, we saw that capture-recapture methods fair well in terms of bias even in situations with low coverage and a lower amount of lists. However, whenever possible more lists and more captures should be sought after, as long as they are of sufficient quality.

Additionally, we saw substantive bias in our estimates when we adjusted the lists to have just a moderate amount of dependence ($r = 0.43$). While there are methods to correct for this dependence, the estimators outlined in this paper are much simpler and more interpretable. Future work could delve into the dependent case more thoroughly, as there is most likely some amounts of dependence that are largely ignorable.

Lastly, capture-recapture is applicable to several fields, and more uses are being pioneered every year. One recent paper applying these methods has sought to estimate dementia prevalence in New Zealand (Ma'u et al. 2024). Another unique application used capture-recapture to estimate road traffic mortality in Zambia (Mwale et al. 2023).

References

Chao, Anne. 1987. "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." *Biometrics* 43 (4): 783–91. http://www.jstor.org/stable/2531532.

Chao, Anne, P. K. Tsay, Sheng-Hsiang Lin, Wen-Yi Shau, and Day-Yu Chao. 2001. "The Applications of Capture-Recapture Models to Epidemiological Data." *Statistics in Medicine* 20 (20): 3123–57. https://doi.org/https://doi.org/10.1002/sim.996.

Cren, E. David Le. 1965. "A Note on the History of Mark-Recapture Population Estimates." *Journal of Animal Ecology* 34: 453. https://api.semanticscholar.org/CorpusID:87284381.

Gruber, Laura. 2023. "Capture-Recapture Methods: A Comparison of Different Estimators and Confidence Intervals," December. https://epub.jku.at/obvulihs/download/pdf/9466367.

Liu, Xian. 2012. "Appendix a: The Delta Method." In *Survival Analysis*, 405–6. John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781118307656.app1.

Ma'u, E., S. Cullum, N. Mukadam, D. Davis, C. Rivera-Rodriguez, and G. Cheung. 2024. "Estimating the Prevalence of Dementia in New Zealand Using Capture-Recapture Analysis on Routinely Collected Health Data." *International Journal of Geriatric Psychiatry* 39 (8): e6131. https://doi.org/https://doi.org/10.1002/gps.6131.

Mwale, Moses, Kelvin Mwangilwa, Ernest Kakoma, and Kacem Iaych. 2023. "Estimation of the Completeness of Road Traffic Mortality Data in Zambia Using a Three Source Capture Recapture Method." *Accident Analysis & Prevention* 186: 107048. https://doi.org/https://doi.org/10.1016/j.aap.2023.107048.

Norris, James L., and Kenneth H. Pollock. 1996. "Including Model Uncertainty in Estimating Variances in Multiple Capture Studies." *Environmental and Ecological Statistics* 3 (3): 235–44. https://doi.org/10.1007/BF00453012.

Schwarz, Carl James, and J. Brian Dempson. 1994. "Mark-Recapture Estimation of a Salmon Smolt Population." *Biometrics* 50 (1): 98–108. http://www.jstor.org/stable/2533200.

Seber, George A. F., and Matthew R. Schofield. 2019. "A Brief History of Capture–Recapture." In *Capture-Recapture: Parameter Estimation for Open Animal Populations*, 1–11. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-18187-1_1.

Silverman, Bernard W. 2020. "Multiple-Systems Analysis for the Quantification of Modern Slavery: Classical and Bayesian Approaches." *Journal of the Royal Statistical Society Series A: Statistics in Society* 183 (3): 691–736. https://doi.org/10.1111/rssa.12505.

Sun, Catherine C., Angela K. Fuller, Matthew P. Hare, and Jeremy E. Hurst. 2017. "Evaluating Population Expansion of Black Bears Using Spatial Capture-Recapture." *The Journal of Wildlife Management* 81 (5): 814–23. https://doi.org/https://doi.org/10.1002/jwmg.21248.