International Journal of Statistics and Probability; Vol. 12, No. 4; July 2023
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

A Close Look at the Estimation of the Population Size

Mohammad Fraiwan Al-Saleh[1] & Bayan Abdel-Wahab[1]

[1] Department of Statistics-Yarmouk University, Jordan

Correspondence: Mohammad Fraiwan Al-Saleh, Department of Statistics, Yarmouk University, Jordan

Received: February 12, 2023 Accepted: July 17, 2023 Online Published: July 20, 2023

doi:10.5539/ijsp.v12n4p19 URL: https://doi.org/10.5539/ijsp.v12n4p19

Abstract

The main purpose of this paper is to take a close look at the main methods of estimation the population size. In particular, the main concentration is on Capture-Recapture technique (direct sampling and the inverse (indirect) sampling), highlighting some of the hidden properties of these techniques, that are rarely mentioned in classroom and can be of interest to statistics teachers. Another Capture-Recapture estimator is introduced. The content of the paper may add significant contribution to educational statistics.

Keywords: population size, capture-recapture technique, direct-indirect sampling, Petersen estimator, chapman estimator

1. Introduction

Statistics is the science of collecting information from a suitable portion of the population of interest(sample), summarizing and analyzing the collected information, making inference about a population and measuring the accuracy of the inference. The accuracy of the inference depends mainly on the sample size and the method of selecting the sample (sampling technique). The sample taken from a population of interest contains information that can be used to estimate the population parameters such as the mean (μ), proportion (p), total (τ) and variance (σ^2), etc. Simple, Stratified, Systematic and Cluster random sampling are the main sampling techniques of selecting an appropriate sample.

In some real life applications, the population size, N, is also unknown and in this case, it is regarded as an additional population parameter; for example, the size of fish population in a sea (not the Red Sea), the size of camels in a desert, the number of beggars in a city. Estimation some of the population parameters such as τ may require the estimation of the population size first.

Capture-Recapture method is the most popular procedure used to estimate N. The main contribution of this manuscript is the highlighting of some hidden properties of this technique that are rarely mentioned in classrooms. We use(●) for each highlighted property. This technique and some of its hidden properties are discussed in section 2. A new estimator of N is introduced in section 3. Concluding remarks are given in section 4.

2. Capture-Recapture Technique

If the study population consists of N elements, where N, population size, is unknown, then Capture-Recapture techniques can be used to estimate N. This method was introduced by Lincoln-Petersen and the estimator of N using this method, called “Lincoln-Petersen Estimator”, was first used by Laplace in 1786 to estimate the size of the French population. The estimator is based on the fact that the proportion of marked elements in the second sample is an estimator of proportion of the marked elements in the population before the second sample is taken. Bohninget et al. (2004) used Capture–Recapture approach to estimate the number of drug users in Bangkok. For more details, see Schaeffer et al. (1995). Plante et al. (1998) provided an estimator of population size using this technique based on Stratified random sampling. Estimation the size of the population of drug users in Washington was considered by Ruiz et al. (2016). Koski et al. (2013) used this method to estimate the size of the Whale population in West Greenland Bowhead**.** Burnham and Overyon (1978) considered the estimation of the size of a closed population when capture probabilities vary among animals. Razzak and Luby (1997) used the Capture-Recapture method to estimate deaths and injuries of traffic accidents in Karachi, Pakistan. Al-Saleh & Ababneh (2022) considered the use of Capture-Recapture method in estimating the population total parameter τ, taking advantage of known size. For more on this technique see Pedro et al. (2020) and Stephen (1996).

There are two forms of Capture-Recapture technique:

(i) Capture-Recapture Technique-Direct Sampling:

This method can be performed using the following steps:

i. Select a random sample of n_1 elements from the population of interest, mark(tag) them, then release them back to the population.

ii. After waiting a suitable period of time, so that the marked elements spread out into the population (especially in the case of mobile population), select a second random sample of size n_2 and count the number of marked elements in the sample(the number of elements that are selected twice).

- The purpose of the first step is to create a binary population (a population of two types: tagged and untagged, with a known number of tagged elements). *Sometimes the population of interest consists of two types and thus, the first step is deleted*. For example, the total number of residents in a population is usually known, but the total number of nonresidents (refugees) is unknown. Thus, to estimate the size of the population N , the first step is not needed because the population consists of two types.

Let p be the proportion of marked elements in the population; $p = n_1 / N$. If T is the number of marked elements in the second sample then p can be estimated by $\hat{p} = T / n_2$. Equating the proportion of marked elements in the population to the proportion of marked elements in the second sample, we get an estimator of N , $\hat{N}_d$ given by(method of moments estimator):

$$\hat{N}_d = \frac{n_1 n_2}{T}.$$

- This can simply be put in a more interesting way: if two subsets A, B are taken from a population, the size of the population can be estimated by:

$$\hat{N}_d = \frac{\#elements\ in\ A \times \#elements\ in\ B}{\#elements\ in\ A \cap B},$$

Which can be rewritten as:

$$\hat{N}_d = N \frac{\left(\#elements\ in\ A / N\right) \times \left(\#elements\ in\ B / N\right)}{\left(\#elements\ in\ A \cap B\right) / N} = N \frac{P(A)P(B)}{P(A \cap B)}.$$

If A, B are independent then $\hat{N}_d = N$. Thus, *the closer A, B to be independent, the closer is the estimator $\hat{N}_d$ to N.*

Note that T is a discrete Hypergeometric r.v. with probability mass function:

$$P(T = t) = p_T(t) = \frac{\binom{n_1}{t}\binom{N - n_1}{n_2 - t}}{\binom{N}{n_2}}, \ \max(0, n_1 + n_2 - N) \le t \le \min(n_1, n_2).$$

$$E(T) = \frac{n_1 n_2}{N} \ \ \& \ \ Var(T) = n_2 \frac{n_1}{N}\left(1 - \frac{n_1}{N}\right)\left(\frac{N - n_2}{N - 1}\right).$$

- The domain of T can be obtained using the trivial conditions on the two combinations, $\binom{n_1}{t} \& \binom{N - n_1}{n_2 - t}$:

$$0 \le t \le n_1 \ \& \ n_2 - t \ge 0 \rightarrow t \le \min(n_1, n_2);$$

$$n_2 - t \le N - n_1 \ \& t \ge 0 \rightarrow t \ge n_1 + n_2 - N \ \& t \ge 0 \rightarrow t \ge \max(0, n_1 + n_2 - N).$$

Thus, $Min(\hat{N}_d) = \dfrac{n_1 n_2}{\min(n_1, n_2)} = \max(n_1, n_2)$.

For large N, the possible values of T are $0,1,2,...,\min(n_1,n_2)$.

If $T = 0$ then $\hat{N} = \infty$. T might be zero with positive probability:

$$P(T=0) = \binom{N-n_1}{n_2} / \binom{N}{n_2}.$$

For example , if $N = 500, n_1 = 50, n_2 = 50$, then $P(T=0) \approx 0.004$.

- Thus, **<u>the expected value of this estimator is infinite</u>**. This estimator is used in practice and an approximate mean and variance of it are given in almost every book in sampling, but it is not mentioned that we are using $\hat{N}_d$ under the condition that $T \neq 0$. Thus, the mean and variance are actually *the conditional mean and variance*; $E\left(\hat{N}_d \mid T \neq 0\right)$ and $Var\left(\hat{N}_d \mid T \neq 0\right)$.

If $T = 0$, then we may use Chapman (1951) estimator:

$$\hat{N}_C = \frac{(n_1+1)(n_2+1)}{T+1} - 1.$$

Now,

$$E\left(\hat{N}_C\right) = \sum_{t=0}^{\min(n_1,n_2)} \left(\frac{(n_1+1)\,(n_2+1)}{t+1} - 1\right) \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}} = N - \frac{(n_2+1)\binom{N-n_1}{n_2+1}}{\binom{N}{n_2}},$$

$$Var\left(\hat{N}_C\right) = \sum\nolimits_{t=0}^{\min(n_1,n_2)} \left(\frac{(n_1+1)\ (n_2+1)}{t+1} - 1 - E\left(\hat{N}_C\right)\right)^2 \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}.$$

For more details, see Al-Saleh and Ababneh (2022), Lohe (1999).

Clearly, $\hat{N}_C$ is symmetric in n_1 and n_2 i.e. $\hat{N}_C(n_1,n_2) = \hat{N}_C(n_2,n_1)$. Also, its mean and variance are symmetric in n_1 and n_2. $E\left(\hat{N}_C\right) \leq N$, so $\hat{N}_C$ is negatively biased.

- An estimate of $Var\left(\hat{N}_C\right)$ is $\hat{V}\left(\hat{N}_C\right) = \frac{(n_1+1)\,(n_2+1)(n_1-t)(n_2-t)}{(t+1)^2(t+2)}$, (Seber 1970). A strange thing about this estimate is that <u>its unit is not the square of the unit of the measurement</u>; if the unit of measurement is cm then the mean is in cm but the variance should be in cm^2. Here, if we are estimating the total number of fish in a sea, then if $n_1 = 54$, $n_2 = 30$, $t = 10$ then based on the above estimates " N is estimated by 154 fish with variance about 252 fish not $fish^2$.

Al-Saleh and Ababneh(2022) introduced the following estimator of N based on direct sampling method:

$$\hat{N}_S = \begin{cases} \hat{N}_d & if\ \ T \neq 0 \\ \hat{N}_C & if\, T = 0 \end{cases} = \widehat{N}_c I_{(T=0)} + \widehat{N}_d I_{(T\neq 0)}.$$

It was concluded that $\hat{N}_S$ is more efficient than Chapman estimator for small to moderate sample size, but less efficient for large sample size.

(ii) Capture-Recapture Technique-Inverse Sampling

This is another method to estimate the population size N. This technique can be performed using the following steps:

1. Select a random sample of size n_1, mark all the elements in the sample, then release the marked elements to the population.

2. Wait a suitable period of time for the marked units to spread out into the population (in the case of mobile population), then select elements continuously and independently from the population until t marked elements are obtained (t is specified in advance). t is a suitable integer ;$1 \leq t \leq n_1$.

Let N_2 be the required sample size to obtain t marked elements(recaptured one). Note that n_1 and t are fixed in advance, but N_2 is a random variable that has a negative hypergeometric distribution with prob. function given by: (See Al-Saleh and Ababneh, 2022):

$$p_{N_2}(n_2) = \frac{\binom{n_1}{t-1}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2-1}} \frac{n_1 - t + 1}{N - n_2 + 1},\ \ n_2 = t, t+1, t+2, ..., N - n_1 + t.\ ^{\cdot}$$

$$E(N_2) = \frac{t(N+1)}{n_1+1}, \text{ (Balakrishnan et al. 2003)}$$

$$Var(N_2) = \frac{t(N+1)(N-n_1)(n_1+1-t)}{(n_1+1)^2(n_1+2)}, \text{ (Khan 1994).}$$

Using the method of moments, the population size is estimated by (see Schaeffer et al. (1995)):

$$\hat{N}_I = \frac{n_1 N_2}{t};\ \ 1 \leq t \leq n_1.$$

$$E(\hat{N}_I) = = \frac{n_1}{t} E(N_2) = \frac{n_1(N+1)}{(n_1+1)},\ Var(\widehat{N}_I) = \ \frac{n_1^{\ 2}}{t} \frac{(N+1)(N-n_1)(n_1+1-t)}{(n_1+1)^2(n_1+2)}.$$

Note that $E(\hat{N}_I)$ is free of t and increasing in n_1 increases, but the variance is decreasing t. $\hat{N}_I$ is negatively biased; its bias is: $Bias(\widehat{N}_I) = E(\widehat{N}_I) - N = (n_1 - N)/(n_1 + 1)$. Clearly, the bias is decreasing in n_1 but free of t.

- The same comment about the unit of the variance mentioned above is applied here.
- The net sample size is $n = n_1 + N_2$; it is a r.v. with

$$E(n) = n_1 + \frac{t(N+1)}{n_1+1}.$$

- If n_1 is a simple random sample, what can we say about n or $E(n)$?

3. Suggested Estimator

The following technique can be used when getting infinite $\hat{N}_d$. The technique is a combination of the two Capture-Recapture methods (direct and indirect). The following is a description of suggested technique:

1. **1.** Based on direct sampling: a random sample of n_1 elements are selected, tagged and released back to the population. After waiting enough time for the tagged elements to spread out into a population, a second random sample of size n_2 elements is chosen and the number of tagged elements T (recaptured ones) is determined.

If $T \neq 0$, then Petersen estimator can be used to estimate $N : \hat{N}_d = n_1 n_2 / T$.

2. **2.** If $T = 0$, then continue sampling until obtaining $T = 1$.

Let N_3 be the required number of chosen elements to obtain the first tagged element. In this case a suitable estimator of N is

$$\hat{N}^* = \frac{n_1\left(n_2 + N_3\right)}{1}.$$

Note that N_3 is a r.v. that has negative hypergeometric distribution:

$$P(N_3 = n_3) = p(n_3) = \frac{\binom{n_1}{0}\binom{N - n_2 - n_1}{n_3 - 1}}{\binom{N - n_2}{n_3 - 1}} \frac{n_1}{N - n_2 - n_3 + 1},$$

$$n_3 = 1,2,3,\ldots, N - n_1 - n_2 + 1.\ E(N_3) = \frac{(N - n_2 + 1)}{n_1 + 1};\ Var(N_3) = \frac{(N - n_2 + 1)(N - n_2 - n_1)n_1}{(n_1 + n_2 + 1)^2 (n_1 + n_2 + 2)}.$$

The net sample size is $n = n_1 + n_2 + N_3$;it is a random variable with;

$$E(n) = n_1 + n_2 + \frac{(N - n_2 + 1)}{n_1 + 1}.$$

The suggested estimator is:

$$\widehat{N}_A = \begin{cases} \widehat{N}_d & if \quad T \neq 0 \\ \widehat{N}^* & if \quad T = 0 \end{cases} = \widehat{N}_d \ I_{(T\neq 0)} + \widehat{N}^* \ I_{(T=0)}.$$

Now,

$$E\left(\hat{N}_A\right) = E\left(E\left(\hat{N}_A \mid \mathrm{T}\right)\right),$$

$$E\left(\hat{N}_A \mid T\right) = \frac{1}{T} E\left(n_1 n_2\right) I_{(T\neq 0)} + E\left(n_1\left(n_2 + n_3\right)\right) I_{(T=0)}$$

$$= \frac{1}{T}\left(n_1 n_2\right) I_{(T\neq 0)} + n_1\left(n_2 + \frac{(N - n_2 + 1)}{n_1 + 1}\right) I_{(T=0)}.$$

Thus,

$$E\left(E\left(\hat{N}_A \mid \mathrm{T}\right)\right) = E\left(\frac{1}{T}\left(n_1 n_2\right) I_{(T\neq 0)} + n_1\left(n_2 + \frac{(N - n_2 + 1)}{n_1 + 1}\right) I_{(T=0)}\right)$$

$$= n_1 n_2\, E\left(\frac{1}{T} I_{(T\neq 0)}\right) + n_1\left(n_2 + \frac{(N - n_2 + 1)}{n_1 + 1}\right) E\left(I_{(T=0)}\right).$$

Thus,

$$E\left(\hat{N}_A\right) = n_1 n_2 \sum_{t=1}^{\min(n_1,n_2)} \frac{1}{t} \frac{\binom{n_1}{t}\binom{N - n_1}{n_2 - t}}{\binom{N}{n_2}} + n_1\left(n_2 + \frac{(N - n_2 + 1)}{n_1 + 1}\right) \frac{\binom{N - n_1}{n_2}}{\binom{N}{n_2}}$$

$$Var\left(\hat{N}_A\right) = E\left(\hat{N}_A\right)^2 - \left(E\left(\hat{N}_A\right)\right)^2.$$

Now,

$$E\left(\hat{N}_A\right)^2 = E\left(\left(\hat{N}_d\ I_{(T\neq 0)} + \hat{N}^*\ I_{(T=0)}\right)\right)^2 \quad = E\left(\left(\hat{N}_d\ I_{(T\neq 0)}\right)^2 + \left(\hat{N}^*\ I_{(T=0)}\right)^2\right)$$

$$= E\left(E\left(\left(\hat{N}^2\quad {}_d I_{(T\neq 0)} |\, T\right)\right)\right) + E\left(E\ \left(\left(\hat{N}^*\quad {}^2 |\, T\right)\right)\right).$$

Similarly, we can find $E\left(E\left(\hat{N}_d{}^2 \,|\, T\right)\right)$ and $E\left(E\left(\hat{N}^{*2} \,|\, T\right)\right)$ as the expected value of $\hat{N}_A$.

$$E\left(\hat{N}^{*2}\right) = E\left(E\left(\left(n_1 n_2 + n_1 n_3\right)^2 I_{(T=0)} \,|\, T\right)\right) = E\left(\left(n_1^2 n_2^2 + n_1^2\, E\left(n_3^2\right) + 2n_1^2 n_2\, E\left(n_3\right)\right) I_{(T=0)}\right)$$

$$= \left(n_1^2 n_2^2 + n_1^2\left(\frac{(N-n_2+1)(N-n_2-n_1)n_1}{(n_1+n_2+1)^2(n_1+n_2+2)} + \left(\frac{(N-n_2+1)}{n_1+1}\right)^2\right) + 2n_1^2 n_2\left(\frac{(N-n_2+1)}{n_1+1}\right)\right)\frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}}$$

Therefore, the final formula of the variance is:

$$Var\left(\hat{N}_A\right) = \left(n_1 n_2\right)^2 \sum_{t=1}^{\min(n_1,n_2)} \frac{1}{t^2} \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}$$

$$+\left(n_1^2 n_2^2 + n_1^2\left(\frac{(N-n_2+1)(N-n_2-n_1)n_1}{(n_1+n_2+1)^2(n_1+n_2+2)} + \left(\frac{(N-n_2+1)}{n_1+1}\right)^2\right) + 2n_1^2 n_2\left(\frac{(N-n_2+1)}{n_1+1}\right)\right)\frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}} - \left(\ E\left(\hat{N}_A\right)\right)^2$$

To see the performance of $\hat{N}_A$, the efficiency of $\hat{N}_A$ w.r.t $\hat{N}_C$,

$Eff\left(\hat{N}_A; \hat{N}_C\right) = \frac{MSE(\hat{N}_C)}{MSE(\hat{N}_A)}$, is computed for different values of N and $E(n)$. The values are given in Table 1. It can be seen that $\hat{N}_C$ is more efficient than $\hat{N}_A$. Thus, $\hat{N}_C$ is a better choice to use when $\hat{N}_d$ can't be used.

4. Conclusions

In this paper, we have closely looked at the two methods of Capture-Recapture technique, the direct sampling and inverse sampling, to estimate the population size, we introduced a new estimator that combines the two methods of Capture-Recapture Technique. It turned out that Chapman (1951) estimator is a better choice. Some hidden properties are highlighted, some of these properties are rarely mentioned in classrooms. Thus, the content of the paper can be a good contribution to statistical education.

Table (1). $MSE\left(\hat{N}_A\right)$, $MSE\left(\hat{N}_C\right)$ & $Eff\left(\hat{N}_A, \hat{N}_C\right)$

N	$E(n)$	$MSE\left(\hat{N}_A\right)$	$MSE\left(\hat{N}_C\right)$	$Eff\left(\hat{N}_A, \hat{N}_C\right)$
1000	179	4. 6910×10⁵	1. 3663×10⁵	0.275
	189	3. 3105×10⁵	1. 1689×10⁵	0.353
	199	2. 4451×10⁵	1. 0097×10⁵	0.412
5000	205	3. 0526×10⁷	6. 9123×10⁶	0.226
	218	2. 2406×10⁷	6. 4354×10⁶	0.287

References

Al-Saleh, M. F., & Ababneh, M. (2022). Estimation of the Population Total Utilizing Estimators of the Population Size. *American Review of Mathematics and Statistics, 10*, 17-33.

Balakrishnan, N., Charalambides, C., & Papadatos, N. (2003). Bounds on expectation of order statistics from a finite population. *Journal of Statistical Planning and Inference, 113*, 569-588.

https://doi.org/10.1016/S0378-3758(01)00321-4

Bohning, D., Suppawattanabodee, B., Kusolvisitkul, W., & Viwatwongkasem, C. (2004). Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in one list. *European Journal of Epidemiology, 19,* 1075-1083. https://doi.org/10.1007/s10654-004-3006-8

Burnham, K., & Overyon, W. (1978). The estimation the size of a closed population when capture probabilities vary among animals. *Biometrika, 65,* 625-633. https://doi.org/10.1093/biomet/65.3.625

Chapman, D. (1951). *Some properties of the hypergeometric distribution with application to zoological censuses.* University of California publication in statistics.

Khan, R. (1994). A note on the generating function of a negative hypergeometric distribution. *Sankhya, 56,* 309-313.

Koski, W., da-Silva, C., Zeh, J., & Reeves, R. (2013). Evaluation of the potential to use capture-recapture analyses of photographs to estimate the size of the eastern Canada-West Greenland Bowhead Whale (Balaena mysticetus) population**.** *Canadian Wildlife Biology and Management, 2,* 23-35.

Laplace, P. S. (1786). “Sur les naissances, les mariages et les morts,” in *Histoire de L’Acad’Emie Royale Des Sciences*. [Google Scholar]

Lohr, S. (1999). *Sampling Design and Analysis.* Cole publishing company.

Pedro, L., Inês, S., Rui, S., William, H., Keith, G., Julie, C., Amândio, R., & Antonio, F. (2020). A Review of Capture-recapture Methods and Its Possibilities in Ophthalmology and Vision Sciences. *Ophthalmic Epidemiology, 27*(4). https://doi.org/10.1080/09286586.2020.1749286

Plante, N., Rivest, L., & Trembley, G. (1998). Stratified capture-recapture estimation of the size of a closed population. *Biometrics, 54*, 47-60. https://doi.org/10.2307/2533994

Razzak, J. A., & Luby, S. P. (1997). Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture-recapture method. *International Journal of Epidemiology, 27,* 866-870. https://doi.org/10.1093/ije/27.5.866

Ruiz, M. S., O'Rourke, A., & Allen, S. T. (2016). Using capture-recapture methods to estimate the population of people who inject drugs in Washington, DC. *AIDS Behav., 20,* 363-368. https://doi.org/10.1007/s10461-015-1085-z

Schaefer, R., Mendenhall, W., & Ott, R. (1995). Elementary survey sampling. Fifth edi.

Seber, G. (1970). The effect of trap response on tag-recapture estimates. *Biometrics, 26,* 13-22. https://doi.org/10.2307/2529040

Stephen, C. (1996). Capture-recapture methods in epidemiological studies, 17, 262–266. [Crossref], [PubMed], [Web of Science ®], [Google Scholar]. https://doi.org/10.1017/S019594170000388X

Copyrights