Estimation of the Population Total Utilizing Estimators of the Population Size

By

Ma'aly Salah Mohammad Ababneh

Supervisor

Prof. Mohammad Fraiwan Al-Saleh

Program:

Statistics

December 29, 2011

Estimation of the Population Total Utilizing Estimators of the Population Size

By

Ma'aly Salah Mohammad Ababneh

B.Sc. (Mathematics), Yarmouk University, 2007

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistics, Yarmouk University, Irbid, Jordan

Approved by:

Prof. Mohammad Fraiwan Al-Saleh ---------------------------------- **Chairman**

Professor of Statistics, Yarmouk University.

Dr. Omar M. Eidous -- **Member**

Associate Professor of Statistics, Yarmouk University.

Dr. Amer Ibrahim Al-Omari ------------------------------------- **(External) Member**

Assistant Professor of Statistics, Al al-Bayt University.

December 29, 2011

بسم الله الرحمن الرحيم

"اني ما رأيت أنه يكتب أحد كُتاباً في يومه إلا قال في غده:

لو غُيّـرَ هذا لكان أحسن، ولو زيد هذا لكان يُستحسَن، ولو قدّمَ هذا لكان أفضل، ولو تُركَ هذا لكان أجمل، وهذا من أعظم العبر، وهو دليل على استيلاء النقص على جملة البشر"

العماد الأصفهاني

ACKNOWLEDGEMENT

In the Name of Allah, the Most Gracious, the Most Merciful.

First of all, I am grateful to Allah for giving me the faith and strength to complete this study.

I would like to express my deep thanks to my supervisor, Prof. Mohammad Fraiwan Al-Saleh for his encouragement and patience. It was really very hard to get this job done without your knowledge, guidance, and support. It's a unique opportunity for me to have such experience and knowledge from you. I want to say, thank you very much Prof. Mohammad and it really was a good experience for me.

This thesis is dedicated to my father's soul who inspired me and taught me all good morals and values. May his soul rest in peace. It is also dedicated to my mother, who offered me unconditional love and support throughout all the challenges. I would also like to convey my heartfelt gratitude to my sisters and brothers especially Dr. Mustafa for their love, patience and support.

thanks to all friends and to all faculty members.

Table of Contents

	Page
List of Tables	vii
Abstract	viii
Abstract in Arabic	ix

Chapter One: Introduction and Literature Review

1.1.	Introduction	1
1.2.	Literature Review and Some Related Works	2
1.2.1	Estimation of the Parameters of Finite Population	2
1.2.2	Estimation of the Population Size (Capture Recapture Techniques)	3
1.2.3	Estimation of the Population Total When N is Unknown	6
1.3.	Thesis Organization	9

Chapter Two: Estimation of the Population Total When the Population Size is Unknown Using Capture-Recapture Direct Sampling

2.1.	Introduction	10
2.2.	Estimation of Population Size Using Direct Sampling	10
2.3.	Comparison Between $\hat{N}_S$ and $\hat{N}_C$	15
2.4.	Estimation of Population Total Using Direct Sampling	16
2.5.	Numerical Comparisons Between $\hat{\tau}_C$ and $\hat{\tau}_S$	21

Chapter Three: Estimation of Population Total Using Indirect Sampling

3.1. Introduction 32

3.2. Estimation of Population Size Using Indirect Sampling 32

3.3. Mathematical Comparison of $\hat{N}_I$ and $\hat{N}_I^*$ 36

3.4. Numerical Comparison of $\hat{N}_I$ and $\hat{N}_I^*$ 38

3.5. Estimation of Population Total Using Indirect Sampling 39

3.6. Numerical Comparison of $\hat{\tau}_I$ and $\hat{\tau}_I^*$ 43

Chapter Four: Conclusions and Suggestions for Further Research

4.1. Introduction 54

4.2. Concluding Remarks 54

4.3. Suggestions for Future Work 55

References 56

List of Tables **Page**

Table (2.1): Mean and Variance of $\hat{N}_C$ for selected values of n_1, n_2, N 24

Table (2.2): Mean and Variance of $\hat{N}_S$ for selected values of n_1, n_2, N 25

Table (2.3): Efficiency of $\hat{N}_S$ w.r.t. $\hat{N}_C$ 26

Table (2.4): Mean and Variance of $\hat{\tau}_C$ and $\hat{\tau}$ for selected values of n_1, n_2, N 27

Table (2.5): Mean and Variance of $\hat{\tau}_S$ and $\hat{\tau}$ for selected values of n_1, n_2, N 28

Table (2.6): Efficiency of $\hat{\tau}_C$ w.r.t. $\hat{\tau}$; $Eff(\hat{\tau}_C;\hat{\tau})$ 29

Table (2.7): Efficiency of $\hat{\tau}_S$ w.r.t. $\hat{\tau}$; $Eff(\hat{\tau}_S;\hat{\tau})$ 30

Table (2.8): Efficiency of $\hat{\tau}_S$ w.r.t. $\hat{\tau}_C$; $Eff(\hat{\tau}_S;\hat{\tau}_C)$ 31

Table (3.1): Expectation of n 46

Table (3.2): Mean and Variance of $\hat{N}_I$ for selected values of n_1, T, N 47

Table (3.3): Variance of $\hat{N}_I^*$ for selected values of n_1, T, N 48

Table (3. 4): Efficiency of $\hat{N}_I$ w.r.t. $\hat{N}_I^*$ 49

Table (3. 5): Efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}$; $Eff(\hat{\tau}_I;\hat{\tau})$ 50

Table (3. 6): Efficiency of $\hat{\tau}_I^*$ w.r.t. $\hat{\tau}_I$; $Eff(\hat{\tau}_I^*;\hat{\tau}_I)$ 51

Table (3. 7): Efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}_S$; $Eff(\hat{\tau}_I;\hat{\tau}_S)$ 52

Table (3. 8): Efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}_C$; $Eff(\hat{\tau}_I;\hat{\tau}_C)$ 53

Abstract

Ababneh, Ma'aly Salah. Estimation of the Population Total Utilizing Estimators of the Population Size. Master of Science Thesis, Department of Statistics, Yarmouk University, 2011. (Supervisor: Prof. Mohammad Fraiwan Al-Saleh).

Our main concern in this thesis is the estimation of the population total when the population size is unknown. Two estimators are suggested using direct capture recapture sampling and another two estimators using indirect capture recapture sampling are proposed. These estimators are compared via the biases and mean square errors. Also these estimators are compared with the estimator of the population total, when the population size is assumed to be known.

Key Words: Population Total; Population Size; Direct Capture-Recapture Sampling; Indirect Capture-Recapture Sampling; Petersen Estimator; Chapman Estimator; Hypergeometric Distribution; Negative Hypergeometric Distribution.

الملخص

عبابنة، معالي صلاح. تقدير مجموع المجتمع باستخدام تقديرات لحجم المجتمع. رسالة ماجستير في قسم الإحصاء بجامعة اليرموك، 2010. (المشرف: الأستاذ الدكتور محمد فريوان الصالح)

إن إهتمامنا في هذه الرسالة هو تقدير مجموع المجتمع عندما يكون حجم المجتمع غير معلوم. تم ايجاد تقديرين لمجموع المجتمع بإستخدام طريقة الصيد وإعادة الصيد المباشرة وتقديرين آخرين بإستخدام طريقة الصيد وإعادة الصيد الغير مباشرة. تمت مقارنة التقديرات التي تم الحصول عليها بالنسبة الى التحيز ومجموع مربعات الأخطاء. أيضا تمت مقارنة هذه التقديرات مع تقدير مجموع المجتمع عندما يكون حجم المجتمع معلوم.

الكلمات المفتاحية: مجموع المجتمع، حجم المجتمع، طريقة الصيد وإعادة الصيد المباشرة، طريقة الصيد وإعادة الصيد الغير مباشرة، تقدير بيترسون، تقدير شابمان، التوزيع فوق الهندسي، التوزيع فوق الهندسي السلبي.

CHAPTER ONE

Introduction and Literature Review

1.1. Introduction

Statistics is the science of making inference about a population using the information contained in a sample selected from that population. So, the aim is to make inference about the population parameters such as the mean, total, proportion, etc, based on a representative sample. The sample can be chosen by one of the techniques such as simple, stratified, systematic and cluster random sampling, etc. The choice of the technique depends on the objectives of the study, information available about the population of interest and the budget.

The information obtained from the chosen sample is used to estimate the population parameters. When sampling from finite population is used, the main population parameters are the mean (μ), variance (σ^2), total (τ) and proportion (p). Less common parameter is the population size (N). Estimation of the parameters and their variances depends on the value of the population size (N), which is usually known. If N is unknown, then we may need to estimate it first, so that we can estimate the population total and the variance of the other estimators. Our main concern in this work is the estimation of the population total when N is unknown. If N is known, then some estimators of N can be used as a guard against unsuitable estimates of τ.

1.2. Literature Review and Some Related Works

1.2.1 Estimation of the Parameters of Finite Population

Assume that we have a population of size N (known). Let $\{O_1,\ldots,O_N\}$ be the population measurements and $\{X_1,\ldots,X_n\}$ be a random sample of size n from it. If the sample is a simple random sample, then the estimators of the population mean, $\mu = \frac{\sum_{i=1}^{N} O_i}{N}$, and total, $\tau = \sum_{i=1}^{N} O_i$, respectively are:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \hat{\tau} = N\hat{\mu} = \frac{N}{n}\sum_{i=1}^{n} X_i .$$

It is known that $\hat{\mu}$ and $\hat{\tau}$ are unbiased estimators of μ and τ, respectively. The variance of $\hat{\mu}$ and $\hat{\tau}$ are

$$Var(\hat{\mu}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \text{ and } Var(\hat{\tau}) = N^2 Var(\hat{\mu}),$$

where σ^2 is the population variance, $\sigma^2 = \frac{\sum_{i=1}^{N}(O_i - \mu)^2}{N}$.

Sometimes, the individuals of the population are divided into two groups, the first group has a specific characteristic and the second does not have that characteristic. If we are interested about the proportion of the first group we define the variable O_i to have the value 1 if the member is in the first group, and zero otherwise. If $\sum_{i=1}^{N} O_i = N_1$ then the proportion

$$p = \frac{N_1}{N} = \frac{\sum_{i=1}^{N} O_i}{N} = \mu .$$

Letting $\hat{p}$ be the proportion in the sample, then

$$\hat{p} = \frac{n_1}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \hat{\mu} ,$$

and

$$V\hat{a}r(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1} \frac{N-n}{N} ,$$

where n_1 is the number of individuals that have the characteristic in the sample and $\hat{q} = 1 - \hat{p}$.

1.2.2 Estimation of the Population Size (Capture Recapture Techniques)

As we can see from the previous section, N should be known to estimate τ or to estimate the variance of the estimators. In this section, we will outline techniques that are used to estimate N .

- **Direct Sampling**

The direct sampling method is known as Petersen's method in the ecological applications because of his work on it which goes back to 1894.

Assume that there is a closed population with equal chance of each member to be obtained. A random sample of size m is drawn, marked then released back into the population then, after waiting enough period of time until the marked sample mix with the remaining population, a second random sample of size n is drawn. Let T be the number of recaptured elements in the second sample. Then the Petersen estimator of the population size is:

$$\hat{N} = \frac{nm}{T}.$$

An approximate estimator of the variance of $\hat{N}$ is:

$$V\hat{a}r(\hat{N}) = \frac{mn(m-T)(n-T)}{T^3} \quad \text{(Sekar and Deming, 1949).}$$

Actually, $\hat{N}$ is the maximum likelihood estimator (MLE) of N. It's also the method of moment's estimator (MME).

A modified estimator of N was proposed by Chapman (1951):

$$\hat{N} = \frac{(m+1)(n+1)}{T+1} - 1,$$

with variance estimated by

$$V\hat{a}r(\hat{N}) = \frac{(m+1)(n+1)(m-T)(n-T)}{(T+1)^2(T+2)}.$$

This estimator has the advantage of being valid even when $T = 0$.

- **Inverse Sampling** (Indirect Sampling)

Inverse sampling is another method for estimating N. In this method, sampling continues until a fixed number of recaptured elements are obtained. So, a first random sample of size m is chosen, marked and released. Later, we select elements randomly from the population until k elements are being recaptured, then

$$\hat{N} = \frac{T^{*}m}{k}, k \leq m,$$

where T^{*} is the total number of elements selected in the second random sample to obtain k previously captured elements. The variance of $\hat{N}$ is estimated by

$$V\hat{a}r(\hat{N}) = \frac{m^{2}t^{*}(t^{*} - k)}{k^{2}(k+1)}$$

(Scheaffer et al., 1995).

A slight modification of this procedure was used by Ahmad et al. (2000) to estimate the population size N.

- **Application of Capture Recapture Technique**

Capture-Recapture technique is an old method used to estimate the size of fish and wildlife population. Later on, the method was used for estimating other population sizes. Some of the applications of this method are listed below.

- **Childhood Acute Leukemia Incidence**

Azevedo-Silva et al. (2009) analyzes the number of cases and incidence of childhood acute lymphoblastic leukemia (ALL) by using two source capture-recapture procedures in three different cities in Brazil. The estimated incidence was 5.76, 6.32 and 5.48 per 100,000 individuals and the estimated of childhood ALL was 15.5%, 35.4% and 29.2%, respectively.

- **Weevils in a Box of Wheat**

As Bishop et al. (1975) reports, Andrewartha (1961) describes an experiment in which about 2,000 weevils were placed in a box of wheat and allowed to disperse. Using the capture-recapture procedure, an estimate of the population size was 2739 weevils. (Thomas and Herzog, 2006).

- **Other Applications**

Estimating of birth and death rates in India was considered by SeKar and Deming (1949). Estimating the population size of Injecting Drug Users (IDU) was discussed by Luan et al. (2005) and estimating the number of people eligible for health service use is studied by Smith et al. (2002).

1.2.3 Estimation the Population Total When N is Unknown

In many situations, the ratio estimate is used to estimate τ for a population of size N (unknown) for the variable of interest. One way to overcome the difficulty of not knowing N is to use some auxiliary variable.

Let the population measurements to be $O_1,...,O_N$ and the corresponding values of an auxiliary variable be $V_1,...,V_N$ then the population elements can be expressed as $\{(O_1,V_1),...,(O_n,V_n)\}$. Assume that there is a fair degree of association between O and V. Let $\{(X_1,Y_1),...,(X_n,Y_n)\}$ be the elements of a simple random sample (SRS) from this population.

Now, using the relation

$$\frac{\tau_O}{\tau_V}=\frac{\mu_O}{\mu_V}$$

we can estimate τ_O by

$$\hat{\tau}_O=\frac{\overline{X}}{\overline{Y}}\tau_V.$$

Let

$$r=\frac{\overline{X}}{\overline{Y}}.$$

An estimate of the variance of $\hat{\tau}_O$ is given by:

$$V\hat{a}r(\hat{\tau}_O)=\tau_V^2\frac{N-n}{Nn}\frac{1}{\mu_V{}^2}S_r^2,$$

where

$$S_r^2=\frac{\sum_{i=1}^{n}(Y_i-rX_i)^2}{n-1}$$

(Scheaffer et al., 1995).

Ahmad et al. (2000) introduced another method to estimate the population total and the population size. They used sequential sampling with replacement until a fixed (k) members are repeated. If we continue sampling with replacement from a finite population of size N marking an element as we go along and stop when we have k previously marked units, then the probability of having t^* unmarked (or distinct) units is

$$p_k(t^*) = \frac{N!}{(N-t^*)!}\frac{H_k(t^*)}{N^{t^*+k}}, \; t^* = 0,1,2,\ldots, N,$$

where

$$H_1(t^*) = t^* \text{ and } H_k(t^*) = t^* \sum_{i=1}^{t^*} H_{k-1}(i).$$

Hence, an estimator of the population size was given as:

$$\hat{N}_k = \frac{H_{k+1}(\mathrm{T}^*)}{H_k(\mathrm{T}^*)}$$

and an estimator of the population total is:

$$\hat{\tau} = \frac{\hat{N}_k}{T^*}\sum_{i=1}^{T^*} y_i$$

Also, Ahmad (2000) suggested an approximate confidence interval (L,U) for the population total. Suppose (L_1,U_1) and (L_2,U_2) are $(1-\alpha)$ confidence intervals for N and $\bar{Y}$, respectively, then $L = L_1L_2$ and $U = U_1U_2$; (L,U) has confidence coefficient greater than $1-2\alpha$.

In their graduation project, Mohammad and Abdullah (2007) compared some Capture-Recapture techniques and cluster sampling for estimating the total number of times the word "الله" appears in the first 100 pages of the Holly Quran.

For more details about the estimation of population total and size, see also Gutierrez and Breidt (2009), Otieno et al. (2005), and Arnab (2004).

1.3. Thesis Organization

In this thesis, we are going to estimate the population total utilizing estimators of the population size. In Chapter (2), we consider the estimator of the population total when N is unknown using Direct Sampling. Two estimators are suggested for τ. One based on Chapman estimator of N and the other based on a suggested modified estimator of N. The efficiency, bias and variance of the two estimators are obtained.

In Chapter (3), τ is estimated using Indirect Sampling. The suggested estimators are compared with those in Chapter (2). Concluding remarks and suggested future works are outlined in Chapter (4).

CHAPTER TWO

Estimation of Population Total When the Population Size is Unknown Using Capture-Recapture Direct Sampling

2.1. Introduction

Capture-Recapture (CR) sampling goes back to the work of Petersen (1984) to estimate the size of fish population. We will use two CR methods for estimating the population total (τ) when its size N is unknown: "Direct sampling" and "Indirect Sampling". Direct sampling will be considered in this chapter. In Section (2.2), the suggested estimators will be introduced and their properties will be discussed. The estimators will be compared with the estimator obtained when N is known. Numerical comparisons are given in Section (2.3) and (2.5).

2.2. Estimation of Population Size Using Direct Sampling

Assume that we have a closed population with an equal chance of each member to be selected in a random sample. In Direct Sampling, a simple random sample of n_1 elements is drawn, marked and the value of the random variable of interest, Y, is noted for each element. Then, the n_1 items are released back into the population. After waiting enough period of time until the marked elements mix with the remaining population elements, a second SRS of n_2 elements is drawn. Let T be the number of recaptured elements in the second sample, i.e. T is the number of common elements in both samples. The values of the variable Y are noted for each of the $n_2 - T$ elements. From the first and second sample we obtain a net random sample of size n, where

$$n = n_1 + n_2 - T. \tag{2.1}$$

Note that n is a random variable (not fixed). T has a hypergeometric distribution with probability function:

$$f(t) = P(T = t) = \frac{\binom{n_1}{t}\binom{N - n_1}{n_2 - t}}{\binom{N}{n_2}}, t = 0,1,2,\ldots,\min(n_1, n_2).$$

Now,

$$E(n) = E(n_1 + n_2 - T)$$

$$= n_1 + n_2 - \frac{n_1 n_2}{N} \tag{2.2}$$

$$Var(n) = Var(n_1 + n_2 - T)$$

$$= \left(\frac{N - n_2}{N - 1}\right)\left(\frac{n_1 n_2}{N}\right)\left(1 - \frac{n_1}{N}\right). \tag{2.3}$$

Petersen estimator of the population size (N) is:

$$\hat{N}_P = \frac{n_1 n_2}{T}, T = 0,1,\ldots,\min(n_1, n_2). \tag{2.4}$$

Note that T may equal zero with positive probability:

$$P(T=0)=\frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}}=\frac{(N-n_1)!(N-n_2)!}{N!(N-n_1-n_2)!}>0$$

in this case, $\hat{N}_P$ is ∞.

To overcome this difficulty, an alternative estimator of N was proposed by Chapman (1951) as:

$$\hat{N}_C=\frac{(n_1+1)(n_2+1)}{T+1}-1, \tag{2.5}$$

with

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

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

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

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

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

$$= N - \frac{(n_2+1)\binom{N-n_1}{n_2+1}}{\binom{N}{n_2}}. \qquad (2.6)$$

Clearly, $\hat{N}_C$ is negatively biased.

Also,

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

$$= \sum_{t=0}^{\min(n_1,n_2)}\left(\left(\hat{N}_C - E(\hat{N}_C)\right)^2 \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}\right) \qquad (2.7)$$

Using the estimators (2.4) and (2.5), we suggest the following estimator of N as:

$$\hat{N}_S = \begin{cases}\hat{N}_P, T \neq 0\\ \hat{N}_C, T = 0\end{cases}$$

$$= \hat{N}_P I(T \neq 0) + \hat{N}_C I(T = 0). \qquad (2.8)$$

Now,

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

$$= n_1 n_2 \sum_{t=1}^{\min(n_1,n_2)} \left(\left(\frac{1}{t}\right) \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}} \right) + ((n_1+1)(n_2+1)-1) \frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}}. \quad (2.9)$$

$$Var(\hat{N}_S) = E(\hat{N}_S - E(\hat{N}_S))^2.$$

But,

$$\left(\hat{N}_S - E(\hat{N}_S)\right)^2 = \left(\left(\hat{N}_P I(T \neq 0) - E(\hat{N}_S)I(T \neq 0)\right) + \left(\hat{N}_C I(T=0) - E(\hat{N}_S)I(T=0)\right)\right)^2$$

$$= \left(\hat{N}_p \; I(T \neq 0) - E(\hat{N}_S)I(T \neq 0)\right)^2 + \left(\hat{N}_C \; I(T=0) - E(\hat{N}_S)I(T=0)\right)^2$$

$$= \left(\hat{N}_p - E(\hat{N}_S)\right)^2 I(T \neq 0) + \left(\hat{N}_C - E(\hat{N}_S)\right)^2 I(T=0).$$

Thus,

$$Var(\hat{N}_S) = \sum_{t=1}^{\min(n_1,n_2)} \left(\frac{n_1 n_2}{t} - E(\hat{N}_S)\right)^2 P(T=t) + ((n_1+1)(n_2+1)-1-E(\hat{N}_S))^2 P(T=0)$$

$$= \sum_{t=1}^{\min(n_1,n_2)} \left(\frac{n_1 n_2}{t} - E(\hat{N}_S)\right)^2 \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}} + ((n_1+1)(n_2+1)-1-E(\hat{N}_S))^2 \frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}} \quad (2.10)$$

The above results are summarized in the following lemma:

Lemma (2.1)

The expected value and variance of the suggested estimator, $\hat{N}_S$, of the population size N are given by

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

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

2.3. Comparison Between $\hat{N}_S$ and $\hat{N}_C$

The efficiency of $\hat{N}_S$ w.r.t. $\hat{N}_C$ is:

$$Eff(\hat{N}_S, \hat{N}_C) = \frac{MSE(\hat{N}_C)}{MSE(\hat{N}_S)}.$$

The expected value, variance, bias and MSE of $\hat{N}_S$ and $\hat{N}_C$ are computed for different values of n_1, n_2, N. Tables (2.1) and (2.2) contain the results. The efficiency of $\hat{N}_S$ w.r.t. $\hat{N}_C$ is given in Table (2.3).

Based on these tables, we have the following observations:

1. The bias of $\hat{N}_C$ is negative and decreases (in absolute value) when the expected sample size increases, for example for $E(n) = 51.9, 57.6, 61.4, 67.1$ the $bias(\hat{N}_C)$ values are $-855.52, -624.02, -505.09, -367.2$, respectively.

2. For $N = 1000$, the bias of $\hat{N}_S$ decreases (in absolute value) when the expected sample size increases until some point, then it increases (in absolute value).
3. Similar comments can be said about the mean square error of each estimator.
4. $\hat{N}_S$ is more efficient than $\hat{N}_C$ for small to moderate expected sample size but less efficient for large expected sample size, for example for $E(N) = 72.8$ the efficiency of $\hat{N}_S$ with respect to $\hat{N}_C$ is 1.712622.

2.4. Estimation of Population Total Using Direct Sampling

Given n, let $Y_1, Y_2, ..., Y_n$ be the values of the variable Y for the sample elements. The suggested estimators of the population total (τ) are:

$$\hat{\tau}_C = \hat{N}_C \overline{Y}, \qquad (2.11)$$

and

$$\hat{\tau}_S = \hat{N}_S \overline{Y}, \qquad (2.12)$$

where

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

We conjecture here that given n, $Y_1, Y_2, ..., Y_n$ is a SRS from the population. We have not been able to prove this conjecture yet.

Now,

$$E(\hat{\tau}_S) = E(E(\hat{\tau}_S \mid n))$$

$$= E(E(\hat{N}_S \overline{Y} \mid n))$$

$$= E(\hat{N}_S E(\overline{Y} \mid n))$$

$$= E(\hat{N}_S)\mu$$

$$= \mu\left(n_1 n_2 \sum_{t=1}^{\min(n_1,n_2)} \left(\left(\frac{1}{t}\right) \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}} \right) + ((n_1+1)(n_2+1)-1) \frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}} \right) \quad (2.13)$$

Similarly, for $\hat{\tau}_C$ we have

$$E(\hat{\tau}_C) = E(\hat{N}_C)\mu$$

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

$$= \tau - \frac{\mu(n_2+1)\binom{N-n_1}{n_2+1}}{\binom{N}{n_2}}, \quad (2.14)$$

where $\mu = \frac{\sum_{i=1}^{N} O_i}{N}$ and O_i are the population measurements. Also,

$$Var(\hat{\tau}_S) = E(Var(\hat{\tau}_S \mid n)) + Var(E(\hat{\tau}_S \mid n))$$

$$= E(Var(\hat{N}_S \overline{Y} \mid n)) + Var(E(\hat{N}_S \overline{Y} \mid n))$$

$$= E\left(\hat{N}_S^2 Var(\overline{Y} \mid n)\right) + Var(\hat{N}_S \mu)$$

but

$$Var(\overline{Y} \mid n) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

where σ^2 is the population variance; $\sigma^2 = \frac{\sum_{i=1}^{N}(O_i - \mu)^2}{N}$. Therefore,

$$Var(\hat{\tau}_S) = E\left(\hat{N}_S^2 \frac{\sigma^2}{n}\frac{N-n}{N-1}\right) + \mu^2 Var(\hat{N}_S)$$

$$= \frac{\sigma^2}{N-1}E\left(\hat{N}_S^2\left(\frac{N}{n}-1\right)\right) + \mu^2 Var(\hat{N}_S)$$

$$= \frac{\sigma^2}{N-1}E\left(\hat{N}_S^2\frac{N}{n} - \hat{N}_S^2\right) + \mu^2 Var(\hat{N}_S).$$

Now,

$$E\left(\hat{N}_S^2 \frac{N}{n}\right) = NE\left(\frac{\hat{N}_S^2}{n}\right)$$

$$= Nn_1^2 n_2^2 \sum_{t=1}^{\min(n_1,n_2)}\left(\frac{1}{t^2(n_1+n_2-t)}\frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}\right) + N\frac{((n_1+1)(n_2+1)-1)^2}{n_1+n_2}\frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}}.$$

Thus,

$$Var(\hat{\tau}_S)=\frac{\sigma^2}{N-1}\left(Nn_1^2n_2^2\sum_{t=1}^{\min(n_1,n_2)}\left(\frac{1}{t^2(n_1+n_2-t)}\frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}\right)+N\frac{((n_1+1)(n_2+1)-1)^2}{n_1+n_2}\frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}}-\left(Var(\hat{N}_S)+\left(E(\hat{N}_S)\right)^2\right)\right)+\mu^2Var(\hat{N}_S). \tag{2.15}$$

Similarly for $\hat{\tau}_C$, we have

$$Var(\hat{\tau}_C)=\frac{\sigma^2}{N-1}\left(N\sum_{t=0}^{\min(n_1,n_2)}\left(\frac{\hat{N}^2}{n}\frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}\right)-\left(Var(\hat{N}_C)+\left(E(\hat{N}_C)\right)^2\right)\right)+\mu^2Var(\hat{N}_C). \tag{2.16}$$

The above results are given in the following lemma:

Lemma (2.2)

The expected value and the variance of the estimators of the population total are given by

- $E(\hat{\tau}_S)=\mu E(\hat{N}_S)$
- $E(\hat{\tau}_C)=\mu E(\hat{N}_C)$
- $$Var(\hat{\tau}_S)=\frac{\sigma^2}{N-1}\left(Nn_1^2n_2^2\sum_{t=1}^{\min(n_1,n_2)}\left(\frac{1}{t^2(n_1+n_2-t)}\frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}\right)+N\frac{((n_1+1)(n_2+1)-1)^2}{n_1+n_2}\frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}}-\left(Var(\hat{N}_S)+\left(E(\hat{N}_S)\right)^2\right)\right)+\mu^2Var(\hat{N}_S)$$

- $$Var(\hat{\tau}_C)=\frac{\sigma^2}{N-1}\left(N\sum_{t=0}^{\min(n_1,n_2)}\left(\frac{\hat{N}^2}{n}\frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}\right)-\left(Var(\hat{N}_C)+\left(E(\hat{N}_C)\right)^2\right)\right)+\mu^2 Var(\hat{N}_C)$$

Now, if N was known, then τ can be estimated based on a simple random sample (SRS) of size $n=n_1+n_2-T$ by

$$\hat{\tau}=N\overline{Y}, \tag{2.17}$$

with

$$E(\hat{\tau})=E(E(\hat{\tau}\mid n))=N\mu=\tau.$$

Thus, $\hat{\tau}$ is an unbiased estimator of τ.

$$Var(\hat{\tau})=Var(N\overline{Y})$$

$$=E(Var(N\overline{Y}\mid n))+Var(E(N\overline{Y}\mid n))$$

$$=E\left(N^2\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)+Var(N\mu)$$

$$=\frac{N^2\sigma^2}{N-1}\left(NE\left(\frac{1}{n}\right)-1\right)$$

$$=\frac{N^2\sigma^2}{N-1}\left(N\sum_{t=0}^{\min(n_1,n_2)}\left(\frac{1}{n_1+n_2-t}\right)\frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}-1\right) \tag{2.18}$$

$$=K\sigma^2,$$

where

$$K = \frac{N^2}{N-1}\left(N \sum_{t=0}^{\min(n_1,n_2)} \left(\frac{1}{n_1+n_2-t} \right) \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}} - 1 \right).$$

The mean and variance of $\hat{\tau}_C$ and $\hat{\tau}_S$ are given in Tables (2.4) and (2.5), respectively. Also the variance of $\hat{\tau}$ is given in these tables.

2.5. Numerical Comparisons Between $\hat{\tau}_C$ and $\hat{\tau}_S$

The efficiency of $\hat{\tau}_S$ with respect to $\hat{\tau}$ is

$$Eff\ (\hat{\tau}_S;\hat{\tau}) = \frac{MSE\ (\hat{\tau})}{MSE\ (\hat{\tau}_S)},$$

where

$$MSE\ (\hat{\tau}) = Var\ (\hat{\tau}),$$

and

$$MSE\ (\hat{\tau}_S) = \left(bias\ (\hat{N}_S)\right)^2 \mu^2 + Var\ (\hat{\tau}_S).$$

Now, let

$$Var(\hat{\tau}_S) = L\sigma^2 + Var(\hat{N}_S)\mu^2,$$

where

$$L=\frac{1}{N-1}\left(\begin{array}{l}Nn_1^2n_2^2\sum_{t=1}^{\min(n_1,n_2)}\left(\frac{1}{t^2(n_1+n_2-t)}P(T=t)\right)+N\frac{((n_1+1)(n_2+1)-1)^2}{n_1+n_2}P(T=0)\\-\left(Var(\hat{N}_S)+\left(E(\hat{N}_S)\right)^2\right)\end{array}\right)$$

then

$$Eff(\hat{\tau}_S;\hat{\tau})=\frac{Var(\hat{\tau})}{MSE(\hat{N}_S)\mu^2+L\sigma^2}$$

$$=\frac{K\sigma^2}{MSE(\hat{N}_S)\mu^2+L\sigma^2}.$$

The efficiency can be rewritten in terms of the coefficient of variation (CV) given by $CV=\frac{\sigma}{\mu}$, (assume $\mu\neq 0$) as

$$Eff\ (\hat{\tau}_S;\hat{\tau})=\frac{\frac{K}{MSE\ (\hat{N}_S)}(CV\ (y))^2}{1+\frac{L}{MSE\ (\hat{N}_S)}(CV\ (y))^2} \qquad (2.19)$$

The efficiency of $\hat{\tau}_S$ and $\hat{\tau}_C$ w.r.t. $\hat{\tau}$ is given in Tables (2.6) and (2.7) respectively. Also, the efficiency of $\hat{\tau}_S$ w.r.t. $\hat{\tau}_C$ is given in Table (2.8).

Based on these tables, we have the following observations:

1. $\hat{\tau}_C$ is negatively biased; the bias is decreasing in $E(n)$ (in absolute value) to zero.

2. $\hat{\tau}_S$ is negatively biased for small $E(n)$ when $N = 1000$ and the absolute bias is decreasing for $E(n)$ less than 73. It starts increasing and becomes positively biased for $E(n)$ greater than 73. For $N = 5000$, $\hat{\tau}_S$ is positively biased. The bias tends to decrease as $E(n)$ gets large.

3. $\hat{\tau}_C$ has less absolute bias than $\hat{\tau}_S$.

4. $\hat{\tau}_S$ is more efficient than $\hat{\tau}_C$ for small expected sample size.

5. $\hat{\tau}_S$ and $\hat{\tau}_C$ are more efficient than $\hat{\tau}$ when $E(n)$ is small and for large CV.

Table (2.1): Mean and Variance of $\hat{N}_C$ for selected values of n_1, n_2, N

N	n_1	n_2	$E(n)$	$E(\hat{N}_C)$	$Var(\hat{N}_C)$	$bias(\hat{N}_C)$	$MSE(\hat{N}_C)$
1000	50	2	51.9	144.48	525.41	-855.52	732439.8804
1000	50	8	57.6	375.98	13513	-624.02	402913.9604
1000	50	12	61.4	494.91	33889	-505.09	289004.9081
1000	50	18	67.1	632.8	78218	-367.2	213053.84
1000	50	24	72.8	733.58	130720	-266.42	201699.6164
1000	50	25	73.75	747.5	139630	-252.5	203386.25
1000	50	32	80.4	826.85	199030	-173.15	229010.9225
1000	50	45	92.75	914.72	279350	-85.28	286622.6784
1000	50	70	116.5	978.76	310650	-21.24	311101.1376
5000	150	80	227.6	4591.0	7324600	-409	7491881
5000	150	90	237.3	4700.7	7813900	-299.3	7903480.49
5000	150	100	247.0	4781.0	8087000	-219	8134961
5000	150	106	252.82	4818.5	8161000	-181.5	8193942.25
5000	150	107	253.79	4824.1	8167400	-175.9	8198340.81
5000	150	120	266.4	4883.1	8122100	-116.9	8135765.61
5000	150	145	290.65	4946.8	7563400	-53.2	7566230.24
5000	150	170	314.9	4975.9	6725100	-24.1	6725680.81
5000	150	200	344.0	4990.7	5689100	-9.3	5689186.49

Table (2.2): Mean and Variance of $\hat{N}_S$ for selected values of n_1, n_2, N

N	n_1	n_2	$E(n)$	$E(\hat{N}_S)$	$Var(\hat{N}_S)$	$Bias(\hat{N}_S)$	$MSE(\hat{N}_S)$
1000	50	2	51.9	146.8	255.66	-853.2	728205.9
1000	50	8	57.6	426.73	3955.1	-573.27	332593.5929
1000	50	12	61.4	596.16	14034	-403.84	177120.7456
1000	50	18	67.1	815.82	50440	-184.18	84362.2724
1000	50	24	72.8	991.49	117700	-8.51	117772.4201
1000	50	25	73.75	1016.6	131820	16.6	132095.56
1000	50	32	80.4	1161.4	249150	161.4	275199.96
1000	50	45	92.75	1311.8	501350	311.8	598569.24
1000	50	70	116.5	1351.8	794820	351.8	918583.24
5000	150	80	227.6	6673.4	13946000	1673.4	16746267.56
5000	150	90	237.3	6804.8	16511000	1804.8	19768303.04
5000	150	100	247.0	6863.2	18584000	1863.2	22055514.24
5000	150	106	252.82	6871.5	19560000	1871.5	23062512.25
5000	150	107	253.79	6871.2	19703000	1871.2	23204389.44
5000	150	120	266.4	6835.1	21039000	1835.1	24406592.01
5000	150	145	290.65	6660.4	21239000	1660.4	23995928.16
5000	150	170	314.9	6440.8	19341000	1440.8	21416904.64
5000	150	200	344.0	6192.9	15941000	1192.9	17364010.41

Table (2.3): Efficiency of $\hat{N}_S$ w.r.t. $\hat{N}_C$

N	$E(n)$	$MSE(\hat{N}_S)$	$MSE(\hat{N}_C)$	$Eff(\hat{N}_S;\hat{N}_C)$
1000	51.9	728205.9	732439.8804	1.005814
1000	57.6	332593.5929	402913.9604	1.21143
1000	61.4	177120.7456	289004.9081	1.631683
1000	67.1	84362.2724	213053.84	2.525464
1000	72.8	117772.4201	201699.6164	1.712622
1000	73.75	132095.56	203386.25	1.53969
1000	80.4	275199.96	229010.9225	0.832162
1000	92.75	598569.24	286622.6784	0.478846
1000	116.5	918583.24	311101.1376	0.338675
5000	227.6	16746267.56	7491881	0.447376
5000	237.3	19768303.04	7903480.49	0.399806
5000	247.0	22055514.24	8134961	0.36884
5000	252.82	23062512.25	8193942.25	0.355293
5000	253.79	23204389.44	8198340.81	0.35331
5000	266.4	24406592.01	8135765.61	0.333343
5000	290.65	23995928.16	7566230.24	0.315313
5000	314.9	21416904.64	6725680.81	0.314036
5000	344.0	17364010.41	5689186.49	0.327642

Table (2.4): Mean and Variance of $\hat{\tau}_C$ and $\hat{\tau}$ for selected values of n_1, n_2, N

N	n_1	n_2	$E(n)$	$E(\hat{\tau}_C)$	$Var(\hat{\tau}_C)$	$Var(\hat{\tau})$
1000	50	2	51.9	144.48μ	$390.71\sigma^2+525.41\mu^2$	$18287\sigma^2$
1000	50	8	57.6	375.98μ	$2523.2\sigma^2+13513\mu^2$	$16379\sigma^2$
1000	50	12	61.4	494.91μ	$4235.6\sigma^2+33889\mu^2$	$15304\sigma^2$
1000	50	18	67.1	632.8μ	$6596.5\sigma^2+78218\mu^2$	$13920\sigma^2$
1000	50	24	72.8	733.58μ	$8427.7\sigma^2+130720\mu^2$	$12752\sigma^2$
1000	50	25	73.75	747.5μ	$8675.1\sigma^2+139630\mu^2$	$12575\sigma^2$
1000	50	32	80.4	826.85μ	$9969.3\sigma^2+199030\mu^2$	$11452\sigma^2$
1000	50	45	92.75	914.72μ	$10759\sigma^2+279350\mu^2$	$9794.1\sigma^2$
1000	50	70	116.5	978.76μ	$9474.8\sigma^2+310650\mu^2$	$7593.3\sigma^2$
5000	150	80	227.6	4591.0μ	$118360\sigma^2+7324600\mu^2$	$104870\sigma^2$
5000	150	90	237.3	4700.7μ	$119280\sigma^2+7813900\mu^2$	$100380\sigma^2$
5000	150	100	247.0	4781.0μ	$118290\sigma^2+8087000\mu^2$	$96239\sigma^2$
5000	150	106	252.82	4818.5μ	$117040\sigma^2+8161000\mu^2$	$93908\sigma^2$
5000	150	107	253.79	4824.1μ	$116790\sigma^2+8167400\mu^2$	$93530\sigma^2$
5000	150	120	266.4	4883.1μ	$112820\sigma^2+8122100\mu^2$	$88866\sigma^2$
5000	150	145	290.65	4946.8μ	$103100\sigma^2+7563400\mu^2$	$81034\sigma^2$
5000	150	170	314.9	4975.9μ	$93072\sigma^2+6725100\mu^2$	$74409\sigma^2$
5000	150	200	344.0	4990.7μ	$82318\sigma^2+5689100\mu^2$	$67691\sigma^2$

Table (2.5): Mean and Variance of $\hat{\tau}_S$ and $\hat{\tau}$ for selected values of n_1, n_2, N

N	n_1	n_2	$E(n)$	$E(\hat{\tau}_S)$	$Var(\hat{\tau}_S)$	$Var(\hat{\tau})$
1000	50	2	51.9	146.8μ	$398.33\sigma^2 + 255.66\mu^2$	$18287\sigma^2$
1000	50	8	57.6	426.73μ	$3039.7\sigma^2 + 3955.1\mu^2$	$16379\sigma^2$
1000	50	12	61.4	596.16μ	$5633\sigma^2 + 14034\mu^2$	$15304\sigma^2$
1000	50	18	67.1	815.82μ	$9909.5\sigma^2 + 50440\mu^2$	$13920\sigma^2$
1000	50	24	72.8	991.49μ	$13928\sigma^2 + 117700\mu^2$	$12752\sigma^2$
1000	50	25	73.75	1016.6μ	$14535\sigma^2 + 131820\mu^2$	$12575\sigma^2$
1000	50	32	80.4	1161.4μ	$18113\sigma^2 + 249150\mu^2$	$11452\sigma^2$
1000	50	45	92.75	1311.8μ	$21467\sigma^2 + 501350\mu^2$	$9794.1\sigma^2$
1000	50	70	116.5	1351.8μ	$19563\sigma^2 + 794820\mu^2$	$7593.3\sigma^2$
5000	150	80	227.6	6673.4μ	$243900\sigma^2 + 13946000\mu^2$	$104870\sigma^2$
5000	150	90	237.3	6804.8μ	$250620\sigma^2 + 16511000\mu^2$	$100380\sigma^2$
5000	150	100	247.0	6863.2μ	$251150\sigma^2 + 18584000\mu^2$	$96239\sigma^2$
5000	150	106	252.82	6871.5μ	$249070\sigma^2 + 19560000\mu^2$	$93908\sigma^2$
5000	150	107	253.79	6871.2μ	$248580\sigma^2 + 19703000\mu^2$	$93530\sigma^2$
5000	150	120	266.4	6835.1μ	$239040\sigma^2 + 21039000\mu^2$	$88866\sigma^2$
5000	150	145	290.65	6660.4μ	$210920\sigma^2 + 21239000\mu^2$	$81034\sigma^2$
5000	150	170	314.9	6440.8μ	$179540\sigma^2 + 19341000\mu^2$	$74409\sigma^2$
5000	150	200	344.0	6192.9μ	$145820\sigma^2 + 15941000\mu^2$	$67691\sigma^2$

Table (2.6): Efficiency of $\hat{\tau}_C$ w.r.t. $\hat{\tau}$, $Eff(\hat{\tau}_C;\hat{\tau})$

N	$E(n)$	$CV=0.5$	$CV=1$	$CV=5$	$CV=10$
1000	51.9	0.006241	0.024954	0.615966	2.370284
1000	57.6	0.010147	0.040398	0.878713	2.499718
1000	61.4	0.01319	0.052189	0.968865	2.147734
1000	67.1	0.016208	0.063373	0.920717	1.595043
1000	72.8	0.015642	0.060687	0.773051	1.220907
1000	73.75	0.015294	0.059299	0.748042	1.17425
1000	80.4	0.012367	0.04792	0.598649	0.93414
1000	92.75	0.008463	0.032934	0.440701	0.718821
1000	116.5	0.006056	0.023686	0.346428	0.603322
5000	227.6	0.003486	0.01378	0.250864	0.542584
5000	237.3	0.003163	0.012512	0.230536	0.506165
5000	247.0	0.002947	0.011661	0.216907	0.482064
5000	252.82	0.002855	0.011299	0.211125	0.471948
5000	253.79	0.002842	0.011248	0.21031	0.470536
5000	266.4	0.002721	0.010773	0.202774	0.457653
5000	290.65	0.002668	0.010566	0.199714	0.453306
5000	314.9	0.002756	0.010912	0.205493	0.464102
5000	344.0	0.002964	0.011728	0.218439	0.486251

Table (2.7): Efficiency of $\hat{\tau}_S$ **w.r.t.** $\hat{\tau}$, $Eff(\hat{\tau}_S;\hat{\tau})$

N	$E(n)$	$CV=0.5$	$CV=1$	$CV=5$	$CV=10$
1000	51.9	0.006277	0.025099	0.619341	2.380999
1000	57.6	0.012284	0.0488	1.002176	2.573034
1000	61.4	0.021431	0.083741	1.20335	2.066933
1000	67.1	0.040074	0.147658	1.047878	1.294508
1000	72.8	0.026292	0.096826	0.684161	0.844183
1000	73.75	0.023162	0.08576	0.634498	0.793077
1000	80.4	0.010235	0.039044	0.393256	0.548862
1000	92.75	0.004054	0.015796	0.215683	0.356763
1000	116.5	0.002056	0.008094	0.134857	0.264126
5000	227.6	0.00156	0.006172	0.114769	0.254933
5000	237.3	0.001265	0.005014	0.096394	0.223911
5000	247.0	0.001088	0.004314	0.084914	0.204024
5000	252.82	0.001015	0.004028	0.080156	0.195766
5000	253.79	0.001005	0.003988	0.079481	0.194601
5000	266.4	0.000908	0.003606	0.073122	0.183947
5000	290.65	0.000842	0.003348	0.069215	0.179724
5000	314.9	0.000867	0.003445	0.071808	0.188995
5000	344.0	0.000973	0.003866	0.080548	0.211892

Table (2.8): Efficiency of $\hat{\tau}_S$ w.r.t. $\hat{\tau}_C$, $Eff(\hat{\tau}_S;\hat{\tau}_C)$

N	$E(n)$	$CV = 0.5$	$CV = 1$	$CV = 5$	$CV = 10$
1000	51.9	1.005811	1.005801	1.005478	1.004521
1000	57.6	1.210561	1.207977	1.140504	1.02933
1000	61.4	1.624743	1.604566	1.24202	0.962378
1000	67.1	2.472407	2.329969	1.138111	0.811582
1000	72.8	1.680817	1.595495	0.885014	0.69144
1000	73.75	1.514448	1.446229	0.848211	0.675391
1000	80.4	0.827601	0.814762	0.656905	0.587559
1000	92.75	0.479045	0.47962	0.489408	0.496317
1000	116.5	0.339446	0.341712	0.389279	0.437785
5000	227.6	0.447514	0.44792	0.457494	0.46985
5000	237.3	0.400046	0.400759	0.418129	0.442368
5000	247.0	0.36913	0.36999	0.391477	0.42323
5000	252.82	0.355601	0.356517	0.379659	0.414804
5000	253.79	0.353621	0.354545	0.377924	0.413574
5000	266.4	0.333682	0.334688	0.36061	0.401936
5000	290.65	0.315694	0.316825	0.34657	0.396475
5000	314.9	0.314463	0.315735	0.349444	0.407227
5000	344.0	0.328139	0.329615	0.368744	0.435766

CHAPTER THREE

Estimation of Population Total Using Indirect Sampling

3.1. Introduction

Indirect sampling is another Capture Recapture method for estimating N. It will be considered in this chapter. In Section (3.2), the suggested estimators will be introduced and their properties will be discussed. The estimators will be compared with the estimators obtained in Chapter (2). Mathematical and numerical comparisons are given in Sections (3.3) and (3.4), respectively.

3.2. Estimation of Population Size Using Indirect Sampling

In this method, sampling continues until a fixed number (T) of recaptured elements are obtained. So, a first random sample of size n_1 is chosen, marked and released. Later, we select elements randomly from the population until T elements are being recaptured. Let n_2 be the total number of elements selected in the second random sample to obtain T previously captured elements, then from the first and the second sample we obtain a random sample of size n where

$$n = n_1 + n_2 - T. \tag{3.1}$$

Here, n_2 is a random variable (not fixed). It has a negative hypergeometric distribution with probability function given by:

$$f(n_2) = \frac{\binom{n_2 - 1}{T - 1}\binom{N - n_2}{n_1 - T}}{\binom{N}{n_1}}, \quad n_2 = T, T+1, \ldots, N - n_1 + T. \tag{3.2}$$

The expected value of n_2 is

$$E(n_2) = \frac{T(N+1)}{n_1 + 1} \qquad \text{(N. Balakrishnan 2003)}$$

and the variance of n_2 is given by

$$Var(n_2) = \frac{T(N+1)(N - n_1)(n_1 + 1 - T)}{(n_1 + 1)^2 (n_1 + 2)} \qquad \text{(Khan, 1994).}$$

Now,

$$E(n) = E(n_1 + n_2 - T)$$

$$= n_1 - T + \sum_{n_2 = T}^{N - n_1 + T} n_2 \frac{\binom{n_2 - 1}{T - 1}\binom{N - n_2}{n_1 - T}}{\binom{N}{n_1}}$$

$$= n_1 - T + \frac{T}{\binom{N}{n_1}} \sum_{n_2 = T}^{N - n_1 + T} \binom{n_2}{T}\binom{N - n_2}{n_1 - T}$$

$$= n_1 - T + \frac{T(N+1)}{n_1 + 1}$$

$$= n_1 + T\left(\frac{N-n_1}{n_1+1}\right) \quad (3.3)$$

Also,

$$Var(n) = Var(n_2)$$

$$= \frac{T(N+1)(N-n_1)(n_1+1-T)}{(n_1+1)^2(n_1+2)} \quad (3.4)$$

The expected value of the sample size is computed for different values of n_1, T, N. Table (3.1) contains the results. It can be seen that $E(n)$ is increasing in T for fixed n_1 and decreasing in n_1 for fixed T.

The estimator of the population size N is

$$\hat{N}_I = \frac{n_1 n_2}{T}, \qquad 1 \le T \le n_1, T \text{ is integer.} \quad (3.5)$$

Now,

$$E(\hat{N}_I) = E\left(\frac{n_1 n_2}{T}\right)$$

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

$$= \frac{n_1(N+1)}{n_1+1} \quad (3.6)$$

The bias of $\hat{N}_I$ is

$$Bias(\hat{N}_I) = \frac{n_1 - N}{n_1 + 1}. \tag{3.7}$$

Clearly, $E(\hat{N}_I)$ does not depend on T and increasing in n_1, $\hat{N}_I$ is negatively biased. Now,

$$Var(\hat{N}_I) = Var\left(\frac{n_1 n_2}{T}\right)$$

$$= \frac{n_1^2}{T^2} Var(n_2)$$

$$= \frac{n_1^2}{T} \frac{(N+1)(N-n_1)(n_1+1-T)}{(n_1+1)^2(n_1+2)}. \tag{3.8}$$

Hence,

$$MSE(\hat{N}_I) = \frac{1}{(n_1+1)^2}\left(\frac{n_1^2}{T}\frac{(N+1)(N-n_1)(n_1+1-T)}{(n_1+2)} + (n_1 - N)^2\right). \tag{3.9}$$

It can be seen from (3.6) that this estimator of N can be corrected to be unbiased estimator of N as follows

$$E(\hat{N}_I) = \frac{n_1(N+1)}{n_1+1},$$

which gives

$$E(\hat{N}_I \frac{n_1+1}{n_1}) = N+1.$$

© Arabic Digital Library - Yarmouk University

So, the estimator

$$\hat{N}_I\left(\frac{n_1+1}{n_1}\right)-1=\frac{n_2(n_1+1)}{T}-1$$

is an unbiased estimator of N.

Let

$$\hat{N}_I^* = \frac{n_2(n_1+1)}{T}-1, \tag{3.10}$$

then

$$E(\hat{N}_I^*) = N.$$

Therefore, the mean square error of $\hat{N}_I^*$ is

$$MSE(\hat{N}_I^*) = Var(\hat{N}_I^*) = \frac{(n_1+1)^2}{T^2}Var(n_2)$$

$$= \frac{(N+1)(N-n_1)(n_1+1-T)}{T(n_1+2)}. \tag{3.11}$$

3.3. Mathematical Comparison of $\hat{N}_I$ and $\hat{N}_I^*$

Because $Var(\hat{N}_I) = \frac{n_1^2}{T}\frac{(N+1)(N-n_1)(n_1+1-T)}{(n_1+1)^2(n_1+2)}$, then the variance of $\hat{N}_I$ and $\hat{N}_I^*$ are related as follows:

$$Var(\hat{N}_I) = \left(\frac{n_1}{n_1+1}\right)^2 Var(\hat{N}_I^*),$$

which indicates that for any values of N, n_1 and T,

$$Var(\hat{N}_I) < Var(\hat{N}_I^*)$$

This case is demonstrated clearly for different numerical values of N, n_1 and T that considered in the next section. However, the $MSE(\hat{N}_I)$ is not necessary less than $MSE(\hat{N}_I^*)$ as illustrated below:

$MSE(\hat{N}_I) < MSE(\hat{N}_I^*)$ if $\frac{MSE(\hat{N}_I)}{MSE(\hat{N}_I^*)} < 1$, which is equivalent to,

$$\frac{\left(\frac{n_1}{n_1+1}\right)^2 MSE(\hat{N}_I^*) + \left(\frac{n_1 - N}{n_1+1}\right)^2}{MSE(\hat{N}_I^*)} < 1$$

or

$$\left(\frac{n_1}{n_1+1}\right)^2 + \frac{T(N-n_1)(n_1+2)}{(n_1+1)^2(N+1)(n_1+1-T)} < 1 \qquad (3.12)$$

or

$$\frac{T}{(n_1+1-T)} \frac{(N-n_1)(n_1+2)}{(n_1+1)^2(N+1)} < \frac{2n_1+1}{(n_1+1)^2}$$

or

$$T < \frac{(2n_1+1)(n_1+1)(N+1)}{(n_1+2)(N-n_1)+(2n_1+1)(N+1)} \quad (3.13)$$

In other words, if the number of common elements (recaptured elements), T is fixed to be greater than the right hand of (3.13) then the proposed estimator $\hat{N}_I^*$ is more efficient than $\hat{N}_I$. Otherwise, $\hat{N}_I$ is better than $\hat{N}_I^*$ for estimating N.

3.4. Numerical Comparison Between $\hat{N}_I$ and $\hat{N}_I^*$

The efficiency of $\hat{N}_I$ w.r.t. $\hat{N}_I^*$ is:

$$Eff(\hat{N}_I^*, \hat{N}_I) = \frac{MSE(\hat{N}_I)}{MSE(\hat{N}_I^*)}.$$

The expected value, variance, bias and MSE of $\hat{N}_I$ and the variance of $\hat{N}_I^*$ are computed for different values of n_1, T, N are given in Tables (3.2) and (3.3). The efficiency of $\hat{N}_I$ w.r.t. $\hat{N}_I^*$ is given in Table (3.4).

Based on these tables, we have the following observations:

1. The bias of $\hat{N}_I$ is negative and the absolute bias decreases when the sample size increases.

2. From Table (3.2), it is obvious that $E(\hat{N}_I)$ does not depend on T while the variance of $\hat{N}_I$ decreases when the sample size increases.

3. For $N = 1000$ the variance of N_I^* is decreases when the sample size increases.

3.5. Estimation of Population Total Using Indirect Sampling

Given T, $Y_1, Y_2, ..., Y_n$ are the values of variable for the sample element. Two estimators of the population total, τ, are:

$$\hat{\tau}_I = \hat{N}_I \overline{Y},$$

and

$$\hat{\tau}_I^* = \hat{N}_I^* \overline{Y},$$

where, $\overline{Y} = \frac{\sum_{j=1}^{n} Y_j}{n}$. Now,

$$E(\hat{\tau}_I) = E(E(\hat{\tau}_I | n_2))$$

$$= E\left(E\left(\hat{N}_I \overline{Y} | n_2\right)\right)$$

$$= E\left(\hat{N}_I E\left(\overline{Y} | n_2\right)\right)$$

$$= \mu E(\hat{N}_I)$$

$$= \frac{n_1(N+1)}{n_1+1}\mu . \qquad (3.14)$$

Also,

$$E(\hat{\tau}_I^*) = N\mu = \tau . \qquad (3.15)$$

The variance of $\hat{\tau}_I$ and $\hat{\tau}_I^*$ can be derived as follow

$$Var(\hat{\tau}_I) = E(Var(\hat{\tau}_I | n_2)) + Var(E(\hat{\tau}_I | n_2))$$

$$= E(Var(\hat{N}_I \overline{Y} | n_2)) + Var(\mu \hat{N}_I)$$

$$= E(\hat{N}_I^2 Var(\overline{Y} | n_2)) + \mu^2 Var(\hat{N}_I)$$

$$= E(\hat{N}_I^2 Var(\overline{Y} \mid n_2)) + \mu^2 Var(\hat{N}_I)$$

$$= E\left(\frac{n_1^2 n_2^2}{T^2} \frac{\sigma^2}{n_1+n_2-T} \frac{N-(n_1+n_2-T)}{N-1}\right) + \mu^2 Var(\hat{N}_I)$$

$$= \frac{n_1^2\sigma^2}{T^2(N-1)} E\left(\frac{n_2^2(N-(n_1+n_2-T))}{n_1+n_2-T}\right) + \mu^2 Var(\hat{N}_I)$$

$$= \frac{n_1^2\sigma^2}{T(N-1)} \sum_{n_2=T}^{N-n_1+T} \left(\frac{n_2^2(N-(n_1+n_2-T))}{n_1+n_2-T} \frac{\binom{n_2-1}{T-1}\binom{N-n_2}{n_1-T}}{\binom{N}{n_1}}\right) + \mu^2 Var(\hat{N}_I)$$

$$= B\sigma^2 + \mu^2 Var(\hat{N}_I), \tag{3.16}$$

where

$$B = \frac{n_1^2}{T(N-1)} \sum_{n_2=T}^{N-n_1+T} \left(\frac{n_2^2(N-(n_1+n_2-T))}{n_1+n_2-T} \frac{\binom{n_2-1}{T-1}\binom{N-n_2}{n_1-T}}{\binom{N}{n_1}} \right). \tag{3.17}$$

and $Var(\hat{N}_I)$ is given by Formula (3.8).

Now,

$$Var(\hat{\tau}_I^*) = E(Var(\hat{\tau}_I^* | n_2)) + Var(E(\hat{\tau}_I^* | n_2))$$

$$= E(Var(\hat{N}_I^* \bar{Y} | n_2)) + Var(E(\hat{N}_I^* \bar{Y} | n_2))$$

$$= E(\hat{N}_I^{*2} Var(\bar{Y} | n_2)) + \mu^2 Var(\hat{N}_I^*)$$

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

$$+ \frac{(N+1)(N-n_1)(n_1+1-T)}{T(n_1+2)} \mu^2$$

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

$$+ \frac{(N+1)(N-n_1)(n_1+1-T)}{T(n_1+2)} \mu^2 \tag{3.18}$$

$$= Z\sigma^2 + J\mu^2,$$

where

$$Z = \frac{1}{T^2(N-1)} E((n_2(n_1+1)-T)^2 (\frac{N}{n_1+n_2-T} - 1))$$

and

$$J = \frac{(N+1)(N-n_1)(n_1+1-T)}{T(n_1+2)}.$$

Note that, the previous derivations depend on the fact that n_2 has a negative hypergeometric distribution.

Now, if N was known, then τ can be estimated based on a simple random sample of size $n = n_1 + n_2 - t$ by

$$\hat{\tau} = N\overline{Y},$$

with

$$Var(\hat{\tau}) = Var(N\overline{Y})$$

$$= E(Var(N\overline{Y} \mid n)) + Var(E(N\overline{Y} \mid n))$$

$$= E\left(N^2 \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right) + Var(N\mu)$$

$$= \frac{N^2\sigma^2}{N-1} E\left(\frac{N}{n} - 1\right)$$

$$= \frac{N^2\sigma^2}{N-1} E\left(\frac{N}{n}\right) - \frac{N^2\sigma^2}{N-1}$$

$$=\frac{N^3\sigma^2}{N-1}E\left(\frac{1}{n}\right)-\frac{N^2\sigma^2}{N-1}$$

$$=\frac{N^3\sigma^2}{N-1}\sum_{n_2=T}^{N-n_1+T}\left(\frac{1}{n_1+n_2-t}\frac{\binom{n_2-1}{T-1}\binom{N-n_2}{n_1-T}}{\binom{N}{n_1}}\right)-\frac{N^2\sigma^2}{N-1}$$

$$=H\sigma^2,$$

where

$$H=\frac{N^2}{N-1}\left(N\sum_{n_2=T}^{N-n_1+T}\left(\frac{1}{n_1+n_2-t}\frac{\binom{n_2-1}{T-1}\binom{N-n_2}{n_1-T}}{\binom{N}{n_1}}\right)-1\right).$$

3.6. Numerical Comparison of $\hat{\tau}_I$ and $\hat{\tau}_I^*$

The efficiency of $\hat{\tau}_I$ with respect to $\hat{\tau}$ (obtained for a sample size equal the expected sample size) is:

$$Eff(\hat{\tau}_I;\hat{\tau})=\frac{MSE(\hat{\tau})}{MSE(\hat{\tau}_I)},$$

$$MSE(\hat{\tau})=Var(\hat{\tau}),$$

$$MSE(\hat{\tau}_I)=(bias(\hat{\tau}_I))^2+Var(\hat{\tau}_I)$$

$$=(bias(\hat{N}_I))\mu^2+B\sigma^2+Var(\hat{N}_I)\mu^2$$

$$= B\sigma^2 + MSE(\hat{N}_I)\mu^2$$

$$Eff(\hat{\tau}_I;\hat{\tau}) = \frac{H\sigma^2}{B\sigma^2 + MSE(\hat{N}_I)\mu^2}.$$

The efficiency can be rewritten in terms of the coefficient of variation (CV), given by $CV = \frac{\sigma}{\mu}$, (assume $\mu \neq 0$), as:

$$Eff(\hat{\tau}_I;\hat{\tau}) = \frac{\frac{H}{MSE(\hat{N}_I)}(CV(y))^2}{1 + \frac{B}{MSE(\hat{N}_I)}(CV(y))^2}.$$

The efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}$ is given in Table (3.5).

Similarly, we have

$$Eff(\hat{\tau}_I^{*};\hat{\tau}_I) = \frac{B(CV(y))^2 + MSE(\hat{N}_I)}{Z(CV(y))^2 + J}.$$

The efficiency of $\hat{\tau}_I^{*}$ w.r.t. $\hat{\tau}_I$ is given in Table (3. 6).

Based on the tables, we have the following observations:

1. From Table (3.5), the efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}$ increases when the coefficient of variation (CV) increases.

2. For small sample size, the efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}_S$ is greater than or close to one, also the efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}_C$ is greater than one for small sample size and small value of the coefficient of variation.

3. For large sample size, the efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}_S$ is less than one.

4. For large value for the coefficient of variation, the efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}_C$ is less than one, and it decreases when the sample size increase.

5. $\hat{\tau}_I^*$ is more efficient than $\hat{\tau}_I$ for large value of the expected sample size and coefficient of variation.

Table (3.1): Expectation of n

N	T	n_1	$E(n)$
1000	1	50	68.627
1000	3	50	105.88
1000	4	50	124.51
1000	5	100	144.55
1000	10	100	189.11
1000	15	100	233.66
1000	25	250	324.7
1000	63	250	438.25
1000	75	250	474.1
5000	1	100	148.51
5000	2	100	197.03
5000	8	400	491.77
5000	32	400	767.08
5000	63	400	812.97
5000	50	500	949.1
5000	65	500	1083.8

Table (3.2): Mean and Variance of $\hat{N}_I$ for selected values of n_1, T, N

N	T	n_1	$E(n)$	$E(\hat{N}_I)$	$Var(\hat{N}_I)$	$bias(\hat{N}_I)$	$MSE(\hat{N}_I)$
1000	1	50	68.627	981.37	878870	-18.63	879217.1
1000	3	50	105.88	981.37	281240	-18.63	281587.1
1000	4	50	124.51	981.37	206530	-18.63	206877.1
1000	5	100	144.55	991.09	166240	-8.91	166319.4
1000	10	100	189.11	991.09	78791	-8.91	78870.39
1000	15	100	233.66	991.09	49641	-8.91	49720.39
1000	25	250	324.7	997.01	26717	-2.99	26725.94
1000	63	250	438.25	997.01	8819.5	-2.99	8828.44
1000	75	250	474.1	997.01	6935.5	-2.99	6944.44
5000	1	100	148.51	4951.5	23551000	-48.5	23553352
5000	2	100	197.03	4951. 5	11658000	-48.5	11660352
5000	8	400	491.77	4988. 5	2797200	-11.5	2797332
5000	32	400	767.08	4988. 5	656590	-11.5	656722.3
5000	63	400	812.97	4988. 5	305489.3	-11.5	305621.5
5000	50	500	949.1	4991.0	402750	-9	402831
5000	65	500	1083.8	4991.0	299500	-9	299581

Table(3.3): Variance of $\hat{N}_l^*$ for selected values of n_1, T, N

N	T	n_1	$E(n)$	$Var(\hat{N}_l^*)$
1000	1	50	68.627	914375
1000	3	50	105.88	292600
1000	4	50	124.51	214878.1
1000	5	100	144.55	169581.2
1000	10	100	189.11	80374.41
1000	15	100	233.66	50638.82
1000	25	250	324.7	26931.67
1000	63	250	438.25	8890.212
1000	75	250	474.1	6991.111
5000	1	100	148.51	24024412
5000	2	100	197.03	11892084
5000	8	400	491.77	2811196
5000	32	400	767.08	659880.1
5000	63	400	812.97	307018.7
5000	50	500	949.1	404363.7
5000	65	500	1083.8	300703.7

Table (3. 4): Efficiency of $\hat{N}_I$ w.r.t. $\hat{N}_I^*$

N	$E(n)$	$Eff(\hat{N}_I, \hat{N}_I^*)$
1000	68.627	0.96155
1000	105.88	0.962362
1000	124.51	0.962765
1000	144.55	0.980766
1000	189.11	0.981287
1000	233.66	0.981863
1000	324.7	0.992361
1000	438.25	0.993052
1000	474.1	0.993324
5000	148.51	0.980392
5000	197.03	0.980514
5000	491.77	0.995068
5000	767.08	0.995215
5000	812.97	0.995449
5000	949.1	0.99621
5000	1083.8	0.996266

Table (3. 5): Efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}$; $Eff(\hat{\tau}_I;\hat{\tau})$

N	$E(n)$	$CV = 0.5$	$CV = 1$	$CV = 5$	$CV = 10$
1000	68.627	0.004084	0.016098	0.275107	0.553307
1000	105.88	0.008043	0.030201	0.254925	0.332162
1000	124.51	0.009028	0.032967	0.217609	0.263777
1000	144.55	0.0087	0.030742	0.162453	0.187566
1000	189.11	0.012318	0.036371	0.096949	0.102273
1000	233.66	0.013651	0.034478	0.067367	0.069437
1000	324.7	0.013146	0.026527	0.039341	0.039944
1000	438.25	0.011182	0.014516	0.016046	0.016099
1000	474.1	0.01013	0.012492	0.013499	0.013533
5000	148.51	0.001876	0.007452	0.1538	0.398104
5000	197.03	0.002887	0.011358	0.186294	0.359178
5000	491.77	0.003979	0.014462	0.092053	0.110597
5000	767.08	0.007933	0.018096	0.030669	0.031349
5000	812.97	0.008032	0.0173	0.027427	0.027938
5000	949.1	0.007608	0.014631	0.019868	0.020093
5000	1083.8	0.008022	0.012369	0.015381	0.015499

Table (3. 6): Efficiency of $\hat{\tau}_I^*$ w.r.t. $\hat{\tau}_I$; $Eff(\hat{\tau}_I^*;\hat{\tau}_I)$

N	$E(n)$	$CV=0.5$	$CV=1$	$CV=5$	$CV=10$
1000	68.627	0.961554074	0.961567	0.961839	0.962132
1000	105.88	0.976538358	1.017843	1.782482	2.402905
1000	124.51	0.986293669	1.054636	2.265405	3.177549
1000	144.55	1.016554076	1.120112	2.862745	4.070053
1000	189.11	1.098159758	1.430944	6.043467	8.431787
1000	233.66	1.199760755	1.813903	9.464754	12.88522
1000	324.7	1.447352181	2.713834	16.74673	22.13178
1000	438.25	3.130990361	8.738736	49.20767	58.61586
1000	474.1	3.790414517	11.03604	59.69973	70.17584
5000	148.51	0.98039283	0.980394	0.980422	0.98047
5000	197.03	0.983243013	0.99134	1.194488	1.497824
5000	491.77	1.025178903	1.11397	3.103942	5.416284
5000	767.08	1.315245288	2.23671	16.78553	25.90941
5000	812.97	1.711857968	3.694578	24.88031	32.22224
5000	949.1	1.632535803	3.445735	28.77917	42.04627
5000	1083.8	1.939294326	4.608143	39.26858	55.66183

Table(3.7): Efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}_S$, $Eff(\hat{\tau}_I;\hat{\tau}_S)$

N	$E(n)(Direct\ Sampling)$	$E(n)(Indirect\ Sampling)$	$CV = 0.5$	$CV = 1$	$CV = 5$	$CV = 10$
1000	145	144.55	4.463795	3.994293	1.188815	0.653906
1000	190	189.11	2.120387	1.601488	0.294603	0.179763
1000	235	233.66	1.320195	0.863234	0.141618	0.096202
1000	325	324.7	0.869154	0.459933	0.068055	0.049615
1000	437.5	438.25	0.348173	0.124108	0.021223	0.017657
1000	475	474.1	0.280008	0.095774	0.017212	0.014552
5000	149	148.51	0.090072	0.09364	0.187262	0.343551
5000	198	197.03	1.014306	1.012957	0.985086	0.957543
5000	492	491.77	3.283976	3.001701	0.912219	0.412856
5000	768	767.08	0.936362	0.548585	0.068849	0.042877
5000	814	812.97	0.870743	0.482657	0.058613	0.037213
5000	950	949.1	0.694676	0.328184	0.037717	0.025251
5000	1085	1083.8	0.566316	0.237751	0.027019	0.018768

Table (3. 8): Efficiency of $\hat{\tau}_I$ w.r.t. $\hat{\tau}_C$, $Eff(\hat{\tau}_I;\hat{\tau}_C)$

N	$E(n)(Direct\ Sampling)$	$E(n)(Indirect\ Sampling)$	$CV=0.5$	$CV=1$	$CV=5$	$CV=10$
1000	145	144.55	1.428131	12.29764	0.458121	0.299687
1000	190	189.11	1.126041	3.109297	0.189701	0.130806
1000	235	233.66	1.267287	2.616909	0.130124	0.086305
1000	325	324.7	0.732619	0.63094	0.061542	0.046095
1000	437.5	438.25	0.328808	0.13547	0.020646	0.017285
1000	475	474.1	0.267574	0.101676	0.016854	0.014317
5000	149	148.51	0.226501	5.843564	0.256812	0.305555
5000	198	197.03	0.572615	14.98489	0.536627	0.502703
5000	492	491.77	1.325156	13.62571	0.410536	0.217967
5000	768	767.08	0.805546	0.879507	0.061971	0.039709
5000	814	812.97	0.763334	0.728728	0.053538	0.034835
5000	950	949.1	0.634184	0.430249	0.035533	0.024173
5000	1085	1083.8	0.529153	0.288019	0.025888	0.018188

CHAPTER FOUR

Conclusions and Suggestions for Further Research

4.1 Introduction

In this chapter, we summarize our findings throughout this thesis, in addition; we give some research ideas for further studies in the same field of estimating the population total.

4.2 Concluding Remarks

Four different estimators for the population total are discussed in this thesis and from the results obtained we can conclude the following:

- For large expected sample size, when we use Direct Sampling we found that the estimator of the population total $\hat{\tau}_C$ based on Chapman estimator $\hat{N}_C$ is better than the estimator $\hat{\tau}_S$ based on the suggested modified estimator $\hat{N}_S$ for the population size *N*. On the other hand if the expected sample size is small then $\hat{\tau}_S$ is more efficient than $\hat{\tau}_C$.

- For small expected sample size and coefficient of variation, we found that it is better to use Indirect Sampling to estimate τ than using Direct Sampling.

- The bias in $\hat{N}_I$ can be corrected to obtain an unbiased estimator of N, $\hat{N}_I^*$, and also an unbiased estimator of τ.

4.3 Suggestions for Future Work

- If N is known, then an estimator of N (pretending it is unknown) can be used as a guard against unsuitable or insufficient sample. So, we may suggest estimator of τ conditioning on $\hat{N}$ to be between $N-\varepsilon$ and $N+\varepsilon$ for some ε.

- Estimation of τ based on other sampling techniques when N is unknown can also be considered next.

References

- **Ahmad, M., Alalouf, S. and Chaubey, Y. (2000).** Estimation of the population total when the population size is unknown. Statistics and Probability Letters **49**, 211-216.

- **Andrewatha, H. (1961).** Introduction to the study of animals population. University of Chicago Press.

- **Arnab, R. (2004).** Optimum estimation of a finite population total in PPS sampling with replacement for multi-character surveys. Journal of the Indian Society of Agricultural Statistics **58,** 231-243

- **Azevedo-Silva, F., Reis, R., Santos,M., Luiz, R. and Pombo-de-oliveira, M. (2009).** Evaluation of childhood acute leukemia incidence and underreporting in Brazil by capture-recapture methodology. Cancer Epidemiology **33**, 403-405.

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

- **Bishop, Y., Fienberg, S. and Holland, P. (1975).** Discrete Multivariate Analysis. MIT press, Cambridge.

- **Chapman, D. (1951).** Some properties of the Hyper-geometric distribution with application to zoological censuses. University of California Publication in Statistics.

- **Gutierrez, H. and Breidt, J. (2009).** Estimation of the population total using the generalized difference estimator and Wilcoxon ranks. Revista Colombiana de Estadistica **32**, 123-143.

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

- **Luan, R., Liang, B., Yuan, P., Fan, L., Huang, Y., Zeng, G., Wang, L., and Wang, S. (2005).** A study on the Capture-Recapture method for estimating the population size of injecting drug users in Southwest China. Journal of Health Science **51**, 405-409.

- **Mohammad, M. and Abdullah, M. (2007).** Undergraduate graduation project. Qatar University.

- **Otieno, R., Mwita, P., Obudho, G.** and **Wafula, C. (2005).** Model-based estimation of finite population total in stratified sampling. East African Journal of Statistics **1**, 23-40.

- **Scheaffer, Mendenhall and Ott, (1995).** Elementary survey sampling. Fifth edition, Duxbury Press.

- **Sekar, C. and Deming, W. (1949).** On a method of estimation birth and death rates and the extent of registration. Journal of the American Statistical Association **44**, 101-115. (See

Thomson Steven, K. (1992). Sampling/Wiley series in probability and mathematical statistics. Applied Probability and Statistics, Wiley, New York.

- **Smit, F., Reinking, D. and Reijerse, M. (2002).** Estimating the number of people eligible for health use, Evaluation and Program Planning **25,** 101-105.

- **Thomas, N. Herzog. (2006).** www.soa.org/library/research/actuarial-research-clearing-house/2006/january/arch06v40n1-i.pdf.