UNIVERSIDAD AUTÓNOMA AGRARIA ANTONIO NARRO
SUBDIRECCIÓN DE POSTGRADO

IDEAS BÁSICAS EN LA TEORÍA DEL MUESTREO: ESTIMADORES DE EXPANSIÓN Y DISEÑOS

Tesis

Que presenta DULCE MARÍA SÁNCHEZ GUILLERMO

como requisito parcial para obtener el Grado de

MAESTRA EN ESTADÍSTICA APLICADA

Saltillo, Coahuila   Octubre de 2020

IDEAS BÁSICAS EN LA TEORÍA DEL MUESTREO:
ESTIMADORES DE EXPANSIÓN Y DISEÑOS

Tesis

Elaborada por DULCE MARÍA SÁNCHEZ GUILLERMO como requisito parcial para obtener el Grado de MAESTRA EN ESTADÍSTICA APLICADA con la supervisión y aprobación del Comité de Asesoría

Dr. Rolando Cavazos Cadena
Asesor Principal

Dr. Mario Cantú Sifuentes
Asesor

M. C. Luis Rodríguez Gutiérrez
Asesor

Dr. Marcelino Cabrera De la Fuente
Subdirector de Postgrado
UAAAN

Saltillo, Coahuila

Octubre de 2020

Acknowledgement

Al Dr. Rolando Cavazos Cadena, mi asesor, por los conocimientos compartidos, la ayuda brindada durante este proyecto, por su paciencia, tiempo y comprensión.

Al Dr. Mario Cantú Sifuentes, por el apoyo y disposición de solucionar cualquier duda y compartir sus conocimientos.

A la Universidad Autónoma Agraria Antonio Narro, por la oportunidad de estudir en su programa de graduados.

Dedication

A mi esposo, Alberto Damián Flores Araujo, por todo su apoyo y motivación para iniciar y concluir este proyecto, y por estar siempre dispuesto a ayudarme.

A mi madre, Margarita Guillermo Lucio, por su apoyo incondicional.

COMPENDIO

IDEAS BÁSICAS EN LA TEORÍA DEL MUESTREO:

ESTIMADORES DE EXPANSIÓN Y DISEÑOS

Por

DULCE MARÍA SÁNCHEZ GUILLERMO

MAESTRÍA EN

ESTADÍSTICA APLICADA

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, Octubre de 2020

Dr. Rolando Cavazos Cadena –Asesor–

Palabras clave: Muestreo probabílistico, Estimadores de expansión, Estimadores insesgados, Estimadores de la varianza, Esquemas de selección secuencial, Diseño de Bernoulli, Diseño simple.

Este trabajo trata sobre ideas fundamentales en la *Teoría del Muestreo* y tiene tres objetivos básicos: (i) Estudiar la idea de diseño de muestreo probabilístico, (ii) Formular la noción de estimadores de expansión como los únicos estimadores lineales insesgados del total poblacional, y (iii) Estimar la varianza de los estimadores de expansión. Las conclusiones obtenidas del análisis de estos problemas se ilustran por medio de varios ejemplos analizados detalladamente. La organización de este trabajo es como sigue: En el Capítulo 1 se presenta una perspectiva general del material subsecuente, mientras que en el Capítulo 2 se introducen los conceptos de población, muestra y parámetro, y además se formula el problema básico de estimación en la teoría del muestreo. A continuación, se discuten estrategias (esquemas) generales para seleccionar una muestra, ilustrando las ideas por medio de dos esquemas, a saber, el simple y el de Bernoulli. En el Capítulo 3 se definen los *estimadores de expansión*, también conocidos como estimadores de Horvitz-Thompson, y se estudia la estimación de su varianza. Finalmente, el Capítulo 4 trata sobre el muestreo con reemplazo y el diseño de Bernoulli. Se muestra que bajo el diseño

de Bernoulli la varianza muestral es un estimador asintóticamente insesgado de la varianza poblacional, y se estudian los estimadres de Hurtwitz-Hensen, los cuales son los estimadores de expansión en la teoría de muestreo con reemplazo.

ABSTRACT

FUNDAMENTAL IDEAS IN SAMPLING THEORY:
EXPANSION ESTIMATORS AND SAMPLING DESIGNS

BY

DULCE MARÍA SÁNCHEZ GUILLERMO

MASTER IN

APPLIED STATISTICS

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, October, 2020

Dr. Rolando Cavazos Cadena –Advisor–

Key Words: Probability sampling , Expansion estimators, Unbiased estimators, Variance estimation, Sequential selection scheme, Bernoulli design, Simple design.

This work is about basic ideas in *Sampling Theory* and has three main objectives: (i) To study the concept of probability sampling design, (ii) To introduce the expansion estimators as the unique linear unbiased estimators of the population total, and (iii) To estimate the variance of expansion estimators. The conclusions obtained from the analysis of these problems are illustrated using carefully analyzed examples. The subsequent material is organized as follows: Chapter 1 presents a general perspective of this work, whereas in Chapter 2 the notions of population, sample and parameter are introduced, and the basic problem in the theory of sampling is formally stated. Next, general strategies (or schemes) to select a sample are briefly described, and they are illustrated using two important schemes, namely, the *simple and Bernoulli* strategies. In Chapter 3 the Horvitz-Thompson (expansion) estimators of the population total are introduced, and the estimation of their variances is studied. Finally, Chapter 4 is concerned with the Bernoulli design and sample *with* replacement. It is shown that, under the Bernoulli design, the bias of the sample variance as estimator of the population variance is asymptotically negligible, and the Hurwitz-Hansen estimators are studied.

Contents

Chapter 1

This Work in Perspective

1.1. Introduction

This work concerns with sampling theory, a branch of Classical Statistics which has to do with the following problem: To establish inferences about a finite population based on the knowledge of just a *part of it*, which is referred to as *the sample*. Usually, the inference takes the form of point estimation, and in this case it must be accompanied with a measure of the error of the estimate. In this chapter the main objectives of this work are stated and the organization of the subsequent material is briefly described.

With the objective of establishing conclusions about a whole population, nowadays results of sampling surveys on diverse topics are frequently reported in newspapers and magazines, as well as on Radio and TV shows. For instance, the winner of the 2020 edition of *La Academia* singer contest was Dalú, who obtained 24.29% of the phone votes, whereas Angie got 24.05% and was awarded the second position. Also, *El Financiero* reports every day the result of the *#AMLOTrackingPoll*, which is described by Roy Campos as a 'digital measure of the performance of public administration'; on March 25, 2020, it was reported that 50.2% of Mexican citizens approve president's work. On the other hand, on a quarterly basis, INEGI publishes the result of surveys on employment, particularly, the percentage of unemployed people in the country; on February 2020, the reported unemployment rate was 3.5% of the economically active population. What all of these figures say? They all are intend to reflect the 'behavior' of a whole population, but the reported quantities were obtained studying just a part (and, usually, a 'very small' part) of the population. So, the following is a most important question:

• How is it possible the establish conclusions about a whole population by studying just a (small) sample?

This question will be addressed later, after the discussion in the following section

1.2. Estimation Problem

To introduce the fundamental estimation problem of sampling theory, consider the following situation. In a small town with 1000 workers (the population), an analyst will use a subset of 10 workers (the sample) to estimate the total monthly income of all the workers in town, which is denoted by t (this is the unknown parameter). Let $y_1, y_2, \ldots, y_{10}$ be the monthly income of the ten workers in the sample, so that the total income for the sample is $(y_1+y_2+\cdots+y_{10})$; since the population has 100 times the workers in the sample, it is natural to estimate the total population monthly income t by

$$\hat{t} = 100(y_1+y_2+\cdots+y_{10}) = 1000\,\bar{y}_{10},$$

where $\bar{y}_{10} = (y_1+y_2+\cdots+y_{10})/10$ is the average income of the ten workers in the sample. Now suppose that the $\bar{y}_{10} = 4000$ has been observed, so that $\hat{t} = 4000000$ is the estimate of the population total t. What is the meaning of such a value of $\hat{t}$? *A first answer* is that $\hat{t}$ 'approximates' the unknown value t. Next, suppose that a different sample of 10 workers is chosen and that $\bar{y}_{10} = 3000$ was observed, and in this case $\hat{t} = 300000$ is the estimate of t; this figure is also an 'approximation' for t but, how far are these two numbers form t? At this point it seems clear that declaring that '$\hat{t}$ is an approximation for t' is not a very useful statement, *unless* a bound B for the *error* $|t-\hat{t}|$ is provided, so that $|t-\hat{t}| \le B$. Therefore, after analyzing the sample of 10 workers, a useful conclusion would be like the following one:

$$\triangleright \qquad \hat{t} \text{ approximates } t \text{ and } |\hat{t}-t| \le B \qquad \triangleleft \tag{1.2.1}$$

where, possibly, B depends on the sample data y_i, $i = 1, 2, \ldots, 10$. The main point in this discussion is that declaring that $\hat{t}$ approximates the unknown value t is not useful if no bound about the difference $|\hat{t}-t|$ is provided. Suppose now that the analyst has devised a procedure to associate, with each sample data $\mathbf{y} = (y_1, y_2, \ldots, y_{10})$, a bound $B(\mathbf{y})$ such that $|\hat{t}-t| \le B(\mathbf{y})$ or, more explicitly,

$$\hat{t}(\mathbf{y}) - B(\mathbf{y}) \le t \le \hat{t}(\mathbf{y}) + B(\mathbf{y}); \tag{1.2.2}$$

recall that $\hat{t} = \hat{t}(\mathbf{y})$ depends on the sample data vector $\mathbf{y}$. Now have a glance at this relation, and note that the extreme terms depend only on the 10 monthly incomes $y_1, \ldots, y_{10}$ of the 10 workers in the sample, whereas the middle quantity t, the total monthly income of all the workers in the population, is given by $t = Y_1 + Y_2 + \cdots + Y_{1000}$, where Y_i is the monthly income of the i-th worker in town. Thus, the above display can be explicitly written as

$$\begin{aligned}\hat{t}(y_1, y_2, \ldots, y_{10}) - B(y_1, y_2, \ldots, y_{10}) &\le Y_1 + Y_2 + \ldots + Y_{1000} \\ &\le \hat{t}(y_1, y_2, \ldots, y_{10}) + B(y_1, y_2, \ldots, y_{10}),\end{aligned} \tag{1.2.3}$$

where the sample vector $\mathbf{y}$ is given by

$$(y_1, \ldots, y_{|10}) = (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_{10}}),$$

and the the sample consists of workers $i_1, i_2, \ldots, i_{10}$. Hence, (1.2.3) is equivalent to

$$\begin{aligned}\hat{t}(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_{10}}) &- B(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_{10}}) \\ &\le Y_1 + Y_2 + \ldots + Y_{1000} \\ &\le \hat{t}(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_{10}}) + B(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_{10}}),\end{aligned} \tag{1.2.4}$$

A glance at this relation reveals that it can not be satisfied for every sample. In fact, the extreme terms in the above display depend only on the ten values $Y_{i_1}, \ldots, Y_{i_{10}}$, and if $j \neq i_1, \ldots, i_{10}$, then replacement of Y_j by $Y_j + h$ adds h to the middle term but leaves the extremes values invariable; hence, selecting h appropriately, the inequalities in (1.2.4) fail. Since (1.2.1)–(1.2.4) are all equivalent statements, it follows that it is not possible to design a procedure such that the goal (1.2.1) is *always* satisfied. The key point in this last sentence is the word *always*, and instead of looking for a method generating *a correct assertion every time* that a sample is analyzed, *the main goal* of the estimation problem in sampling theory is slightly less ambitious:

- To devise a method producing an estimate $\hat{t}$ for the parameter t, in such a way that

$$|\hat{t} - t| \le B \tag{1.2.5}$$

is true at least in a proportion γ of all the times in which the method is used.

In this last display $\hat{t}$ is a *statistic*, that is, a function of the data obtained after analyzing the sample, $|\hat{t} - t|$ is the *error* and B is the *bound* on the error, and $\gamma \in (0, 1)$ is the confidence level. Both γ and B are prescribed by the analyst, and the estimation problem consists in devising a procedure such that the inequality (1.2.5) holds at least in a fraction γ of all the times that the procedure is used.

1.3. Random Samples and Main Objectives

The key clue to achieve the goal (1.2.5) is to use randomization to select the subset of the population to be analyzed. Suppose that the sample $s = \{u_1, u_2, \ldots, u_n\}$ is selected via a *random* procedure and that, after analyzing each unit u_i, the corresponding relevant information y_i is determined, $i = 1, 2, \ldots, n$. Next, using the data vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ the estimate $\hat{t} = \hat{t}(\mathbf{y})$ is computed. Since $\mathbf{y}$ depends on s and $\hat{t}$ depends on $\mathbf{y}$, after all $\hat{t}$ is a function of s and then, it is a random variable. Therefore, given $\gamma \in (0, 1)$, there exists a constant C_γ (depending on the the distribution of $\hat{t}$), such that

$$P[|\hat{t} - t| \leq C_\gamma] \geq \gamma;$$

Thus, if after computing $\hat{t}$ it is declared that $|\hat{t} - t| \leq C_\gamma$, then this assertion will be correct in at least a fraction γ of all possible cases, satisfying (1.2.5) if

$$C_\gamma \leq B. \tag{1.3.1}$$

Selecting appropriately the procedure to choose the sample (including the number of selected elements), it is possible to satisfy the above inequality and achieve the goal (1.2.5). Frequently, C_γ has the form $c_\gamma/\sqrt{n}$, for a certain constant c_γ, and then the above relation will be satisfied if $c_\gamma/B \leq \sqrt{n}$, that is,

$$\frac{c_\gamma^2}{B^2} \leq n$$

For details see Lohr (2000). Thus, *the answer to the question posed at the end of Section 1 is*: Using a randomization procedure to choose the sample.

The importance of randomization in sampling theory provided the motivation for the present work, and the main objectives can be stated as follows:

(i) To analyze two fundamental methods to choose a random sample, namely, the draw sequential and listing selection procedures, and to illustrate their application using the simple and Bernoulli schemes;

(ii) To study the construction of the Horvitz-Thompson (expansion) estimators for the population total, and the conditions under which the corresponding variance of such estimators can be ubiasedly estimated.

(iii) To provide carefully analyzed examples on the topics under consideration, including the formulation of the Hansen Hurwitz estimators for the case of sampling with replacement.

1.4. The Origin of This Work

This work is a byproduct of the seminar entitled *Mathematical Statistics: Elements of Theory and Examples*, relaunched on July 2016 by the Graduate Program in Statistics at the Universidad Autónoma Agraria Antonio Narro. The basic aims of the project are:

(i) To be a framework were statistical problems can be freely and fruitfully discussed;

(ii) To promote the *understanding* of basic statistical and analytical tools through the analysis and detailed solution of exercises.

(iii) To develop the *writing skills* of the participants, generating an organized set of neatly solved examples, which can used by other members of the program, as well as by the statistical communities in other institutions and countries.

(iv) To develop the *communication skills* of the students and faculty through the regular participation in seminars, were the results of their activities are discussed with the members of the program.

The activities of the seminar are concerned with fundamental statistical theory at an intermediate (non-measure theoretical) level, as in the book *Mathematical Statistics* by Dudewicz and Mishra (1998). When necessary, other more advanced references that have been useful are Lehmann and Casella (1998), Borobkov (1999) and Shao (2002), whereas deeper probabilistic aspects have been studied in the classical text by Loève (1984). On the other hand, statistical analysis requires algebraic and analytical tools, and the basic references on these disciplines are Apostol (1980), Fulks (1980), Khuri (2002) and Royden (2003), which concern mathematical analysis, whereas the algebraic aspects are covered in Graybill (2000, 2001) and Harville (2008). Initially, the project was concerned with the theory of Point Estimation and Hypothesis Testing. During the last two years the seminar has been focused on Sampling Theory at the level of Lohr (2000), Tucker (1992), Hansen *et al.* (2002), and Sarndal *et al.* (1992); the examples presented in the following chapters were selected from the unsolved exercises in this last reference.

1.5. The Organization

The remainder of this work has been organized as follows: In Chapter 2 the notions of population, sample and parameter are introduced, and the basic problem in the theory of sampling is formally stated. Next, two general strategies (or schemes) to select a sample are briefly described, and they are illustrated by means of two important schemes, namely, the *simple and Bernoulli* strategies. Then, the concept of sampling (probability) design is formulated and an alternative implementation of the simple design is studied. The chapter concludes studying the ideas of inclusion probabilities and membership indicators.

Next, Chapter 3 introduces that the expansion estimator for the population total, it is shown that it is unbiased and the estimator for the corresponding variance is formulated. Also, an alternative (and 'appealing') expression for the variance and its estimators is provided for the case of a constant sample size.

Finally, in Chapter 4 the simple and Bernoulli sampling schemes are studied. It is shown that, conditionally on the observed sample size, the sample obtained from a Bernoulli scheme is a simple random sample. Also, it is proved that under the Bernoulli scheme the sample variance is a biased estimator of the population variance, although the relative bias converges to zero as the population size grows. Next, the Hurwitz-Hansen estimators are introduced as expansion estimators in the case of sampling with replacement. Finally, the exposition concludes with the derivation of basic properties of the multivariate hypergeometric distribution and the Bernoulli sampling design.

Chapter 2

Probability Samples

2.1. Introduction

The basic problem studied in the Theory of Sampling consists in formulating inferences about a whole population $\mathcal{U}$ using knowledge of just one part (a subset) of $\mathcal{U}$. In principle, the population is finite, the subset of the population which is analyzed to state the inferences is called the *sample* and, generally, it is required to accompany the stated conclusions about the population with an assessment of their precision or reliability. Such a requirement can be fulfilled if the analyzed sample is chosen via a random procedure, and this chapter introduces the basic ideas of 'probability sampling schemes'. The subsequent material has been organized as follows: In Section 2 the notions of population, sample and parameter are introduced, and the basic problem in the theory of sampling is formally stated. Next, in Section 3 two general strategies (or schemes) to select a sample are briefly described, and they are illustrated by means of two important schemes, namely, the *simple and Bernoulli* strategies. Then, the concept of sampling (probability) design is formulated in Section 4, and an alternative implementation of the simple design is presented in Section 5. Finally, the chapter concludes in Section 6, which concerns with two notions that will pay important roles in the study of estimation problems, namely, the ideas of inclusion probabilities and membership indicators.

2.2. Population and Random Samples

The environment of a sampling problem has an essential component, namely, *the population*, which is an abstract representation of a collection of objects (entities) that contain relevant information.

In these note the population is represented by a set

$$\mathcal{U} = \{U_1, U_2, U_3, \ldots, U_N\} \tag{2.2.1}$$

and the information conveyed by *the units* U_i is given by a function $\mathcal{Y}$ defined on $\mathcal{U}$ and taking values in $\mathbb{R}$ or $\mathbb{R}^k$; the function $\mathcal{Y}$ is frequently referred to as the *study variable.* The notation

$$\mathcal{Y}(U_i) = Y_i, \quad i = 1, 2, 3, \ldots, N \tag{2.2.2}$$

will be used for the value associated to U_i by the function $\mathcal{Y}$. For instance, if the units U_i are persons, Y_i might be the weight of the i-th person. It is assumed that N, the number of elements of the population, is known, but the function $\mathcal{Y}$ is *unknown.* Thus, the value Y_i associated to U_i can be determined only after analyzing the unit U_i. A *parameter* θ is a value that depends on the whole set of values $Y_1, Y_2, \ldots, Y_N$, that is,

$$\theta = f(Y_1, Y_2, Y_3, \ldots, Y_N) \tag{2.2.3}$$

for a certain function f. Common examples of parameters are the population *total*

$$t = Y_1 + Y_2 + Y_3 + \cdots + Y_N \equiv Y \tag{2.2.4}$$

and the population *average*

$$\bar{t} = \frac{Y_1 + Y_2 + Y_3 + \cdots + Y_N}{N} \equiv \bar{Y}. \qquad (2.2.5).$$

The main problem in sampling theory can be now stated as follows:

> *To estimate a population parameter based on the knowledge*
> of $Y_i = \mathcal{Y}(U_i)$ for U_i in a subset S of the population $\mathcal{U}$ (2.2.6)

The importance of this problem stems from the fact that, frequently, it is impossible, impractical or expensive to examine all of the units in the population to determine the whole set of values $Y_1, Y_2, \ldots, Y_N$ and then compute exactly the value of the parameter. However, it is possible that the available resources (time, budget) allow to examine some units $U_{i_1}, U_{i_2}, \ldots, U_{i_n}$ so that the corresponding $\mathcal{Y}$-values $Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n}$ can be determined, and the problem is to obtain 'a reasonable approximation' of the parameter value using only the information obtained from the analyzed units.

A subset of the population is called *a sample* and the problem stated above can be rephrased as follows:

> To estimate a population parameter based on the knowledge
> of the values Y_i corresponding to units U_i in a sample S. (2.2.7)

Since the parameter θ is unknown, every time that a 'reasonable approximation' $\hat{\theta}$ for θ is proposed, it is important to provide a measure of the 'estimation error' $|\hat{\theta} - \theta|$. In general such an assessment is possible if the sample used in the analysis was selected via a random mechanism. All of the samples considered below will be obtained from $\mathcal{U}$ using a procedure that involves randomness.

2.3. Sample Selection Schemes

Two general methods can be used used to select a sample via a random mechanism:

• A *Draw-Sequential Scheme* consists of a series of random experiments which lead to the selection of population elements, whose number depends on the result of the experiments. Each experiments that leads to select one of the (possible) units is called *a draw*, and draws are performed as many times as necessary until a certain stopping condition is fulfilled, for instance, when the desired number of elements has been selected.

• A *List-Sequential Scheme* consists in traveling down the list of units, performing random experiments each time that a new element is visited. As a result, the set of elements previously selected is modified, for instance, adding the current element to the selection, or removing some units already included. The process ends according to a sopping rule, so that it is possible that the process concludes before the N-th unit is reached.

Example 2.3.1. [A Draw-Sequential Scheme]. The *simple random sampling* scheme (without replacement), which is used to obtain a sample of size $n < N$, is as follows:

1. Select a member of the population using a random mechanism assigning probability $1/N$ to each one of the N elements of $\mathcal{U}$;

2. Remove from the population the unit selected in the previous draw and, with equal probability $1/(N-1)$, select from the remaining $N-1$ elements a new member of the population;

$\vdots$

n. Remove from the population the units selected in the $n-1$ draws already performed and, with equal probability $1/(N-n+1)$, select a new element from the remaining $N-n+1$ units.

After these steps, a (random) sequence

$$\tilde{S} = (U_{i_1}, U_{i_2}, \ldots, U_{i_n}) \tag{2.3.1}$$

is obtained, where U_{i_k} is the unit selected in the k-th draw. This is a vector of distinct units taking values in the space

$$\tilde{\mathcal{S}}_n := \{\tilde{s} = (u_1, u_2, \ldots, u_n) \,|\, u_1, u_2, \ldots, u_n \text{ are different elements of } \mathcal{U}\}. \tag{2.3.2}$$

The elements of $\tilde{\mathcal{S}}_n$ are the *ordered samples* without replacement of size n and are also referred to as the *permutations* of size n of the population $\mathcal{U}$. From the above description it follows that

$$P[\tilde{S} = \tilde{s}] = \frac{1}{(N)_n} = \frac{1}{N(N-1)\cdots(N-n+1)}, \quad \tilde{s} \in \tilde{S}_n,$$

that is, all of the ordered samples (permutations) of size n have the same probability of selection. Finally, a set S is immediately determined form $\tilde{S}$ forgetting the order in which the units were selected:

$$S = \{U_{i_1}, U_{i_2}, \ldots, U_{i_n}\}.$$

This set is a member of the family

$$\mathcal{S}_n = \{s | \, s \text{ is a subset of size } n \text{ of } \mathcal{U}\}.$$

which consists of all subsets (samples) of size n of $\mathcal{U}$. Since the elements of a set of size n can be arranged into a sequence in $n!$ forms, it follows that

$$P[S = s] = \frac{n!}{(N)_n} = \frac{1}{\binom{N}{n}}, \quad s \in \mathcal{S}, \tag{2.3.3}$$

so that all of the samples of size n have the same probability of selection. □

Example 2.3.2. [A List-Sequential Scheme]. Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ be independent random variables with $\mathcal{U}(0,1)$ distribution, *i.e.*, the uniform distribution in $(0,1)$. Given a number $\pi \in [0,1]$, the *Bernoulli (sequential) sampling scheme* is as follows:

$$\begin{array}{l}\text{For each } i = 1, 2, \ldots, N, \text{ include the unit} \\ U_i \text{ in the sample if and only if } \varepsilon_i < \pi.\end{array} \tag{2.3.4}$$

Denote by $\mathcal{S}$ the family of all subsets of $\mathcal{U}$ and let S be the random sample (subset) obtained by using the above Bernoulli scheme, so that

$$\begin{aligned} P[U_i \in S] &= P[\varepsilon_i < \pi] \\ &= \pi \\ &= 1 - P[\varepsilon_i \geq \pi] = 1 - P[U_i \notin S], \quad S \in \mathcal{S}, \quad i = 1, 2, 3, \ldots, N, \end{aligned}$$

so that for $i \neq j$ the events $[U_i \in S]$ and $[U_j \in S]$ are independent with probability π. Hence, the corresponding indicator functions $I[U_i \in S]$ and $I[U_j \in S]$ are independent with common distribution *Bernoulli*(π). It follows that

$$P[S = s] = \pi^{n_s}(1-\pi)^{N-n_s}, \quad s \in \mathcal{S}, \tag{2.3.5}$$

where n_s is the number of elements of the (sample) subset s. □

2.4. Sampling Designs

As already noted, in this work all the sample under consideration will be obtained via a random procedure, which determines a probability distribution on the space of possible samples.

Definition 2.4.1. Let $\mathcal{S}$ be the space of all samples (subsets) of the population $\mathcal{U}$. A *sampling design* is a probability function $p: \mathcal{S} \to [0,1]$ such that

$$p(s) = \text{probability of selecting the sample } s, \quad s \in \mathcal{S};$$

a sampling design is also referred to as a *sampling plan*.

Note that a sampling design $p(\cdot)$ satisfies two conditions: $p(s) \geq 0$ for every $s \in \mathcal{S}$, and $\sum_{s\in\mathcal{S}} p(s) = 1$.

Example 2.4.1. (i) Let n be a positive integer less than N. The simple random sampling design without replacement is

$$p(s) = \frac{1}{\binom{N}{n}}, \quad s \in \mathcal{S}_n, \quad p(s) = 0, \quad s \in \mathcal{S} \setminus \mathcal{S}_n, \tag{2.4.1}$$

where $\mathcal{S}_n$ is the class of all samples with n elements, and $\mathcal{S}$ is the family of all subsets of the population $\mathcal{U}$. This design will be denoted by *SI* where the sample size n is understood from the context. Note that under (2.4.1) all of the samples outisde $\mathcal{S}_n$ have probability zero of being observed.

(ii) The Bernoulli design corresponding to a number $\pi \in [0,1]$ is defined by

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s} \tag{2.4.2}$$

where n_s is the number of elements of s. Note that every sample has positive probability of being selected under (2.4.2), which will be denoted by *BE*, where the value of π will be clear from the context. □

Remark 2.4.1. In principle, any sampling design can be implemented as follows:

1. Determine the class $\mathcal{S}^* = \{s \in \mathcal{S} \,|\, p(s) > 0\}$, and label its elements as $s_1, s_2, \ldots, s_M$, where M = number of elements of $\mathcal{S}^*$.

2. Generate a random variable ε with distribution $\mathcal{U}(0,1)$ and define the random sample S by

$$S = s_k \quad \text{if } \sum_{r<k} p(s_r) \leq \varepsilon < \sum_{r\leq k} p(s_r) = \sum_{r<k} p(s_r) + p(s_k).$$

It follows that $P[S = s_k] = p(s_k)$ for every $k = 1, 2, \ldots, M$, that is,

$$P[S = s] = p(s),$$

when $s \in \mathcal{S}^*$, and then for every $s \in \mathcal{S}$, since both sides of the above display are null if $s \in \mathcal{S} \setminus \mathcal{S}^*$. Although the above procedure looks simple and direct it has a serious drawback, namely, usually the number M of samples with positive probability is really huge, even for 'moderate'population sizes, and then it is impossible to store the set $\mathcal{S}^*$ in a computer's memory. For instance, consider the design *BE* in Example 2.4.1(ii). In that case all of the samples of $\mathcal{S}$ have positive probability, and then $M = 2^N$. If $N = 2000$ then $M = 2^{2000} \sim 2.14 \times 10^{602}$; this figure is too large to implement the above procedure in a computer. Now consider the *SI* design and suppose that the population has $N = 5000$ elements, and that a sample of size $n = 250$ is going to be selected. In this case the number M of samples with positive probability is $M = \binom{5000}{250} \sim 10^{429}$, again an astronomical number. These comments highlight the importance of a good sampling scheme, allowing to implement a given design in practice. □

2.5. Implementing the SI Scheme

Usually, the schemes in Examples 2.3.1 and 2.3.2 are employed to implement the *SI* and *BE* designs, respectively. An alternative implementation of the *SI* design is discussed in the example below.

Example 2.5.1. The following draw sequential scheme implements the *SI* design:

Successively select with equal probability $1/N$ a unit from of the population $\mathcal{U}$ and do not remove the element chosen. Perform independent repetitions of the draw until n different units have been selected. □

As it is shown in the following proposition, this scheme implements the *SI* design.

Proposition 2.5.1. If S is the set consisting of the n different elements obtained from the scheme described in Example 2.5.1, then

$$P[S = s] = \frac{1}{\binom{N}{n}}, \quad s \text{ is a subset of } \mathcal{U} \text{ with } n \text{ elements.}$$

Proof. Let $\tilde{s} = (U_{k_1}, U_{k_2}, \ldots, U_{k_n}) \in \tilde{\mathcal{S}}_n$ be an arbitrary ordered sample with n elements and, for nonnegative integers $r_1, r_2, \ldots, r_{n-1}$, consider the event

$$A(U_{k_1}, r_1, U_{k_2}, r_2, \ldots, U_{k_{n-1}}, r_{n-1}, U_{k_n})$$

determined by the following conditions, whose probability is indicated in parenthesis:

1 (a). Unit U_{k_1} is selected in the first draw; $(1/N)$:

1 (b). In the next r_1 draws unit U_{k_1} is chosen. $((1/N)^{r_1})$

2 (a). The next draw (number r_1+2) yields unit U_{k_2}; $(1/N)$

2 (b). In the next r_2 draws either units U_{k_1} or U_{k_2} are chosen. $(2/N)^{r_2}$

3 (a). Unit U_{k_3} is selected in the next draw (number r_1+r_2+3); $(1/N)$

3 (b). In the next r_3 draws either units U_{k_1}, U_{k_2} or U_{k_3} are chosen. $(3/N)^{r_3}$.

$\vdots$

$(n-1)$(a). Draw number $r_1+r_2+\cdots+r_{n-2}+n-1$ yields unit $U_{k_{n-1}}$; $(1/N)$

$(n-1)$(b). In each one of the the next r_{n-1} draws one of U_{k_1}, $U_{k_2}, \ldots U_{k_{n-1}}$ is chosen. ($((n-1)/N)^{r_{n-1}}$).

n(a). Draw number $r_1+r_2+\cdots+r_{n-1}+n$ yields unit U_{k_n}; $(1/N)$

Since successive draws are independent, it follows that

$$P[A(U_{k_1}, r_1, U_{k_2}, r_2, \ldots, U_{k_{n-1}}, r_{n-1}, U_{k_n})] = (1/N)^n \prod_{k=1}^{n-1} (k/N)^{r_k} \tag{2.5.1}$$

Now, let $\tilde{S}$ be the random ordered sample of size n whose components are the different units that are selected using the above scheme *and preserving the order of selection*, so that

$$[\tilde{S} = \tilde{s}] = \bigcup_{r_1, r_2, \ldots, r_{n-1} \geq 0} A(U_{k_1}, r_1, U_{k_2}, r_2, \ldots, U_{k_{n-1}}, r_{n-1}, U_{k_n}).$$

Since the different events in this union are disjoint, combining these two last displays it follows that

$$P[\tilde{S} = \tilde{s}] = (1/N)^n \sum_{r_1, r_2, \ldots, r_{n-1} \geq 0} \prod_{k=1}^{n-1} (k/N)^{r_k}.$$

Next, observe $\sum_{r_k \geq 0} (k/N)^{r_k} = \sum_{r_k=0}^{\infty} (k/N)^{r_k} = 1/(1-k/N) = N/(N-k)$, and then

$$\sum_{r_1, r_2, \ldots, r_{n-1} \geq 0} \prod_{k=1}^{n-1} (k/N)^{r_k} = \prod_{k=1}^{n-1} \sum_{r_k=0}^{\infty} (k/N)^{r_k} = \prod_{k=1}^{n-1} \frac{N}{N-k}$$

a relation that combined with the above formula for $P[\tilde{S} = \tilde{s}]$ leads to

$$P[\tilde{S} = \tilde{s}] = (1/N)^n \prod_{k=1}^{n-1} \frac{N}{N-k} = \frac{1}{N} \prod_{k=1}^{n-1} \frac{1}{N-k} = \frac{1}{(N)_n}$$

Finally, since the $n!$ posssible orderings of $\tilde{s}$ generate the same (unordered) sample $s = \{U_{k_1}, U_{k_2}, \ldots, U_{k_n}\}$, it follows that $P[S = s] = n!/(N)_n = 1/\binom{N}{n}$. $\square$

After a sample s has been chosen, the problem is to use the data obtained from s to establish inferences about a certain parameter θ, and an estimator $\hat{\theta}$ is required at this step.

Definition 2.5.1. *The sampling strategy* is the pair that consists of the *sampling design* and the *estimator(s)* used to estimate the parameter(s) of interest.

Example 2.5.2. Consider the problem of estimating the population mean $\bar{Y}$ in (2.2.5). An example of a sampling strategy is $(SI, \hat{\theta}))$, where the design SI is based on a sample of size n, and

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

is the sample mean of the values $y_i = \mathcal{Y}(U_{k_i})$ corresponding to the units U_{k_i} in the sample, $i = 1, 2, \ldots, n$. Other example of a sampling strategy is $(SI, \tilde{\theta})$, where

$$\tilde{\theta} = \frac{\max\{y_1, y_2, \ldots, y_n\} + \min\{y_1, y_2, \ldots, y_n\}}{2}. \qquad \square$$

2.6. Inclusion Probabilities and Membership Indicators

Suppose that a sampling plan $p(\cdot)$ is used to select a sample S from the population $\mathcal{U}$. In this case S is a random object whose distribution is given by

$$P[S = s] = p(s), \quad s \in \mathcal{S}.$$

Definition 2.6.1. (i) For each $k = 1, 2, 3, \ldots, N$, the membership indicator I_k of the unit $U_k \in \mathcal{U}$ is defined by

$$I_k \equiv I_k(S) = 1 \quad \text{if } U_k \in S, \qquad I_k \equiv I_k(S) = 0 \quad \text{if } U_k \notin S,$$

whereas

$$\pi_k{:}= P[I_k(S) = 1] = P[k \in S]$$

is the (first order) inclusion probability of the unit U_k.

(ii) For $U_j, U_k \in \mathcal{U}$ the corresponding second order inclusion probability is

$$\pi_{jk} = P[U_j \in S,\ U_k \in S] = P[I_k(S) = 1, I_j(S) = 1].$$

Note that I_k is a random variable with Bernoulli distribution of parameter π_k, so that

$$E[I_k] = \pi_k, \quad \text{and} \quad \mathrm{Var}\,[I_k] = \pi_k(1-\pi_k). \tag{2.6.1}$$

The second order inclusion probability π_{jk} is

$$\pi_{jk} = P[I_j = 1, I_k = 1] = E[I_j I_k] \tag{2.6.2}$$

and with this notation

$$\mathrm{Cov}\,(I_j, I_k) = \pi_{jk} - \pi_j\pi_k =: \Delta_{jk}; \tag{2.6.3}$$

note that $\pi_{kk} = \pi_k = E[I_k]$ and $\Delta_{kk} = \mathrm{Var}\,[I_k]$ for every k. The *sample size* n_S is defined by

$$n_S = \sum_{\mathcal{U}} I_k \equiv \sum_{\mathcal{U}} I_k(S), \tag{2.6.4}$$

where $\sum_{\mathcal{U}}$ is used as an abbreviated form of $\sum_{k\in\mathcal{U}}$. Observe that

$$\begin{aligned} E[n_S] &= \sum_{\mathcal{U}} E[I_k] = \sum_{\mathcal{U}} \pi_k \\ \mathrm{Var}\,[n_S] &= \sum_{\mathcal{U}} \mathrm{Var}\,[I_k] + \sum_{j,k\in\mathcal{U},\ j\neq k} \mathrm{Cov}\,(I_j, I_k) \\ &= \sum_{\mathcal{U}} \Delta_{kk} + \sum_{j,k\in\mathcal{U},\ j\neq k} \Delta_{jk} \end{aligned} \tag{2.6.5}$$

Example 2.6.1. (i) Consider the *SI* sampling design in Example 2.4.1(i). In this case

$$\pi_k = P_{SI}[U_k \in S] = \sum_{s\in\mathcal{S}_n : U_k\in s} p(s) = \sum_{s\in\mathcal{S}_n : U_k\in s} \frac{1}{\binom{N}{n}},$$

and then, since there are $\binom{N-1}{n-1}$ samples of size n that contain U_k, it follows that

$$\pi_k = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

Next, observe that

$$\pi_{jk} = P_{SI}[U_j \in S, U_k \in S] = \sum_{s\in\mathcal{S}_n : U_j\in S,\, U_k\in s} p(s) = \sum_{s\in\mathcal{S}_n : U_j\in S,\, U_k\in s} \frac{1}{\binom{N}{n}},$$

and using that there are $\binom{N-2}{n-2}$ samples of size n that include U_j and U_k, it follows that

$$\pi_{jk} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}, \quad j \neq k,$$

and then

$$\Delta_{jk} = \pi_{jk} - \pi_j\pi_k = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n}{N}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}, \quad j \neq k.$$

Of course, in the present context the sample size n_S is the constant n, so that $E[n_S] = n$ and $\mathrm{Var}\,[n_S] = 0$. To verify these equalities have a glance at (2.6.1) and note that

$$E[n_S] = \sum_{\mathcal{U}} \pi_k = \sum_{\mathcal{U}} \frac{n}{N} = n$$

where the fact that $\mathcal{U}$ has N elements was used to set the last equality. On the other hand,

$$\begin{aligned}
\mathrm{Var}\,[n_S] &= \mathrm{Var}\left[\sum_{\mathcal{U}} I_k\right] \\
&= \sum_{\mathcal{U}} \mathrm{Var}\,[I_k] + \sum_{j,k\in\mathcal{U}\,:\,j\neq k} \mathrm{Cov}\,(I_j, I_k) \\
&= \sum_{k} \pi_k(1-\pi_k) + \sum_{j,k\in\mathcal{U}\,:\,j\neq k} \Delta_{jk} \\
&= \sum_{\mathcal{U}} \frac{n}{N}\left(1-\frac{n}{N}\right) - \sum_{j,k\in\mathcal{U}\,:\,j\neq k} \frac{n}{N}\left(1-\frac{n}{N}\right)\frac{1}{N-1}.
\end{aligned}$$

Since $\mathcal{U}$ has N elements, and there are $N(N-1)$ pairs (j,k) with different components in $\mathcal{U}$ it follows that

$$\mathrm{Var}\,[n_S] = N\frac{n}{N}\left(1-\frac{n}{N}\right) - N(N-1)\frac{n}{N}\left(1-\frac{n}{N}\right)\frac{1}{N-1} = 0.$$

(ii) For the *BE* sampling design in Example 2.4.1 the membership indicators are independent with common *Bernoulli*(π) distribution. Thus,

$$\pi_k = E_{BE}[I_k] = \pi, \quad \Delta_{jk} = \mathrm{Cov}_{BE}[I_j, I_k] = 0, \quad j \neq k,$$

and n_S has the *Binomial*(N,π) distribution, a property that leads to $E[n_S] = N\pi$ and $\mathrm{Var}\,[n_S] = N\pi(1-\pi)$. □

Proposition 2.6.1. Assume that the sample size n_S is *constant and equal to* n under certain sampling design. In that context, the following relation hold;

(a) $\sum_{\mathcal{U}} \pi_k = n$;

(b) For each fixed j, $\sum_{k\in\mathcal{U}} \pi_{kj} = n\pi_j$;

(c) $\sum_{k\neq j} \pi_{kj} = n(n-1)$.

Proof. Recall that

$$n_S = \sum_{\mathcal{U}} I_k$$

so that $E[n_S] = \sum_{\mathcal{U}} E[I_k] = \sum_{\mathcal{U}} \pi_k$, and part (a) follows since n_S is equal to n with probability 1. Now, multiply both sides of the above display by I_j and, recalling that n_S is constant and equal to n, take expectation in both sides of the resulting equality to obtain

$$n\pi_j = E[n_S I_j] = \sum_{k\in\mathcal{U}} E[I_k I_j] = \sum_{k\in\mathcal{U}} \pi_{kj}$$

establishing part (b). Using this conclusion it follows that

$$\sum_{j,k\in\mathcal{U}} \pi_{kj} = n \sum_{j\in\mathcal{U}} \pi_j = n^2,$$

where part (a) was used to set the second equality. Next, recalling that $\pi_{k\,k} = \pi_k$, observe now that $\sum_{j,k\in\mathcal{U}} \pi_{kj} = \sum_{j,k\in\mathcal{U},\, j\neq k} \pi_{kj} + \sum_{k\in\mathcal{U}} \pi_k = \sum_{j,k\in\mathcal{U},\, j\neq k} \pi_{kj} + n$. These relations and the above display together lead to $n^2 = \sum_{j,k\in\mathcal{U},\, j\neq k} \pi_{kj} + n$, and part (c) follows. □

To conclude this section note that the whole vector of membership functions and the sample S determine each other. In fact, given $s \in \mathcal{S}$, define define

$$i^*(s) = (i_1^*(s), i_2^*(s), i_3^*(s), \ldots, i_N^*(s)),$$

where $i_k^*(s) = 1$ if $U_k \in s$ and $i_k^*(s) = 0$ if $U_k \notin s$. In this case,

$$(I_1(S), I_2(S), \ldots I_N(S)) = i^*(s) \iff S = s. \tag{2.6.6}$$

Chapter 3

Horvitz-Thompson Estimators

3.1. Introduction

This chapter analyzes the problem of estimating a population parameter on the basis of a sample obtained via a probability selection scheme. The main objective is to introduce the expansion estimator for the population total, to show that it is unbiased, and to introduce the estimator for the corresponding variance. The presentation has been organized as follows: In Section 2 the expansion estimator is defined, and it is shown that it is unique and unbiased in the class of linear estimators. Next, in Section 3, the 'measurability condition' is introduced, and it is shown that its is sufficient to estimate the variance of the expansion estimator via an unbiased statistic. Also, an alternative formula for the estimation of the variance is established for the case of designs with constant sample size. In Section 4 the previous ideas are illustrated via a detailed example with constant sample size, whereas Section 5 studies a Bernoulli design, which has a non-constant sample size. Finally, the exposition concludes in Section 6, using a population with three elements to provide a global illustration of the main ideas introduced in the chapter.

3.2. The Expansion Estimators

Before going any further, its is convenient to introduce some notation: A random sample obtained via a given sampling design is denoted by $s = \{U_{k_1}, U_{k_2}, \ldots, U_{k_n}\}$ and $y_i = \mathcal{Y}(U_{k_i})$ stands for the value of the study variable at unit U_{k_i} in the sample. The lower case y_i indicates that the value was obtained from a unit in the random sample s under consideration. On the other hand, $\sum_s y_i$ is the summation of the values y_i over all indices k_i such that $U_{k_i} \in s$. Thus, if $s = \{U_{k_1}, U_{k_2}, \ldots, U_{k_n}\}$, then $\sum_s y_i = y_1 + y_2 + \cdots + y_n = Y_{k_1} + Y_{k_2} + \ldots + Y_{k_n}$. Note that, by Definition 2.6.1(i),

$$\sum_s y_i = \sum_{\mathcal{U}} I_k Y_k.$$

Consider the problem of estimating the population total

$$t = \sum_{\mathcal{U}} Y_k = Y_1 + Y_2 + Y_4 + \cdots + Y_N.$$

When a sample $s = \{U_{k_1}, U_{k_2} \ldots, U_{k_n}\}$ is available only the values $Y_{k_1}, Y_{k_2}, \ldots, Y_{k_n}$ are known and t can not be calculated exactly. In that case, estimations of t must be constructed. The following linear estimators will be studied:

$$\hat{t} = \sum_{\mathcal{U}} c_k I_k Y_k,$$

where the c_k's are constants. Note that $E[\hat{t}] = \sum_{\mathcal{U}} c_k E[I_k] Y_k = \sum_{\mathcal{U}} c_k \pi_k Y_k$, and then $E[\hat{t}] = t$ if and only if $c_k = 1/\pi_k$ for every k, that is, there is only one choice of the coefficients c_k so that $\hat{t}$ is an unbiased estimator of t.

Definition 3.2.1. (i) A sampling design $p(\cdot)$ is a *probability design* if

$$\pi_k > 0, \quad k = 1, 2, \ldots, N.$$

(ii) For a probability design, the π-expanded estimator of t is

$$\hat{t} := \sum_s \frac{y_i}{\pi_{k_i}} = \sum_s \check{y}_i = \sum_{\mathcal{U}} I_k \check{Y}_k \tag{3.2.1}$$

where

$$\check{y}_i := \frac{y_i}{\pi_{k_i}} = \frac{Y_{k_i}}{\pi_{k_i}}, \quad i = 1, 2, \ldots, n \tag{3.2.2}$$

is the expanded i-th sample value, and

$$\check{Y}_k = \frac{Y_k}{\pi_k}$$

is the k-th expanded population value.

The statistic $\hat{t}$ in (3.2.1) is also known as the Horvitz-Thompson estimator.

3.3. Mean and Variance

By construction, the expansion estimator in Definition 3.2.1 is unbiased, so that $E[\hat{t}] = t$. In the following proposition this result is stated for future reference, and the variance of $\hat{t}$ is computed.

Proposition 3.3.1. (i) The estimator $\hat{t}$ in (3.2.1) satisfies

$$E[\hat{t}] = t \quad \text{and} \quad \mathrm{Var}\left[\hat{t}\right] = \sum_{j,k\in\mathcal{U}} \check{y}_j \Delta_{jk} \check{y}_k.$$

see (2.6.3).

(ii) Assume that the following 'measurability condition' holds:

$$\pi_{jk} \neq 0, \quad j,k = 1,2,\ldots,N. \tag{3.3.1}$$

In this case an unbiased estimator of $\mathrm{Var}\left[\hat{t}\right]$ is

$$\hat{V}(\hat{t}) = \sum_{j,k\in S} \check{y}_j \check{\Delta}_{jk} \check{y}_k, \quad \text{where } \check{\Delta}_{jk} := \frac{\Delta_{jk}}{\pi_{kj}}. \tag{3.3.2}$$

(iii) If the sample size is constant, then

$$\mathrm{Var}\left[\hat{t}\right] = -\frac{1}{2} \sum_{j,k\in\mathcal{U}} \Delta_{jk} (\check{y}_j - \check{y}_k)^2$$

and, under the measurabilty condition (3.3.1), an unbiased estimator of $\mathrm{Var}\left[\hat{t}\right]$ is given by

$$\hat{V}(\hat{t}) = -\frac{1}{2} \sum_{j,k\in S} \check{\Delta}_{jk} (\check{y}_j - \check{y}_k)^2. \tag{3.3.3}$$

Proof. (i) Observe that

$$E[\hat{t}] = E\left[\sum_{\mathcal{U}} I_k \check{y}_k\right] = \sum_{\mathcal{U}} \pi_k \check{y}_k = \sum_{\mathcal{U}} y_k = t,$$

where (3.2.2) was used to set the last equality. To conclude note that

$$\mathrm{Var}\left[\hat{t}\right] = \mathrm{Var}\left[\sum_{\mathcal{U}} I_k \check{y}_k\right] = \sum_{j,k\in\mathcal{U}} \check{y}_j \mathrm{Cov}\left(I_j, I_k\right) \check{y}_k = \sum_{j,k\in\mathcal{U}} \check{y}_j \Delta_{jk} \check{y}_k.$$

(ii) To begin with, observe that

$$\hat{V}(\hat{t}) = \sum_{j,k\in S} \check{y}_j \check{\Delta}_{jk} \check{y}_k = \sum_{j,k\in\mathcal{U}} I_j I_k \check{y}_j \check{\Delta}_{jk} \check{y}_k,$$

and then

$$E\left[\hat{V}(\hat{t})\right] = E\left[\sum_{j,k\in\mathcal{U}} I_j I_k \check{y}_j \check{\Delta}_{jk} \check{y}_k\right] = \sum_{j,k\in\mathcal{U}} E[I_j I_k] \check{y}_j \check{\Delta}_{jk} \check{y}_k = \sum_{j,k\in\mathcal{U}} \pi_{jk} \check{y}_j \check{\Delta}_{jk} \check{y}_k,$$

and via (3.3.2) it follows that $E[\hat{V}(\hat{t})] = \sum_{j,k\in\mathcal{U}} \check{y}_j \Delta_{jk} \check{y}_k = \mathrm{Var}\left[\hat{t}\right]$.

(iii) Suppose that n_S is constant, say n, so that $\sum_{k\in\mathcal{U}} I_k = n$. It follows that

$$0 = \text{Cov}\,(n, I_j) = \text{Cov}\left(\sum_{k\in\mathcal{U}} I_k, I_j\right) = \sum_{k\in\mathcal{U}} \Delta_{k,j}.$$

Then, multiplying by $\check{y}_j^2$, the above equality yields that $\sum_{k\in\mathcal{U}} \Delta_{k,j}\check{y}_j^2 = 0$, and then

$$\sum_{j,k\in\mathcal{U}} \Delta_{k,j}\check{y}_j^2 = 0.$$

Similarly,

$$\sum_{j,k\in\mathcal{U}} \Delta_{k,j}\check{y}_k^2 = 0.$$

Therefore,

$$\begin{aligned}
\text{Var}\left[\hat{t}\right] &= \sum_{j,k\in\mathcal{U}} \check{y}_j\Delta_{jk}\check{y}_k \\
&= \sum_{j,k\in\mathcal{U}} \check{y}_j\Delta_{jk}\check{y}_k - \frac{1}{2}\sum_{j,k\in\mathcal{U}} \Delta_{jk}\check{y}_k^2 - \frac{1}{2}\sum_{j,k\in\mathcal{U}} \Delta_{jk}\check{y}_j^2 \\
&= -\frac{1}{2}\sum_{j,k\in\mathcal{U}} \Delta_{jk}\left[-2\check{y}_j\check{y}_k + \check{y}_k^2 + \check{y}_j^2\right] \\
&= -\frac{1}{2}\sum_{j,k\in\mathcal{U}} \Delta_{jk}(\check{y}_j - \check{y}_k)^2
\end{aligned}$$

To conclude, note that

$$\hat{V}(\hat{t}) = -\frac{1}{2}\sum_{j,k\in S} \check{\Delta}_{jk}(\check{y}_j - \check{y}_k)^2 = -\frac{1}{2}\sum_{j,k\in\mathcal{U}} I_k I_j\check{\Delta}_{jk}(\check{y}_j - \check{y}_k)^2$$

and then

$$E[\hat{V}(\hat{t})] = -\frac{1}{2}\sum_{j,k\in\mathcal{U}} E[I_k I_j]\check{\Delta}_{jk}(\check{y}_j - \check{y}_k)^2 = -\frac{1}{2}\sum_{j,k\in\mathcal{U}} \pi_{jk}\check{\Delta}_{jk}(\check{y}_j - \check{y}_k)^2;$$

via (3.3.2), it follows that $E[\hat{V}(\hat{t})] = -\dfrac{1}{2}\displaystyle\sum_{j,k\in\mathcal{U}} \Delta_{jk}(\check{y}_j - \check{y}_k)^2 = \text{Var}\left[\hat{t}\right].$ □

3.4. An Example with Constant Sample Size

This section illustrates the idea of inclusion probability as well as the results in Proposition 2.6.1 concerning designs with constant sample size. Also, the problem of determining the sample inclusion probability in an *SI* design is studied.

Example 3.4.1. In planning an office network study, the following draw sequential sampling scheme was proposed for selecting a random sample of two nonadjacent office hours intervals $[9, 10), [10, 11), \ldots, [15, 16), [16, 17)$ (labeled 1–8).

1. Draw the first interval with equal probability from the eight intervals.

2. Draw, without replacement, the second interval from the intervals that are nonadjacent to the first one selected.

(a) Determine the first and second order inclusion probabilities;

(b) Is the sampling design induced by the proposed selection scheme measurable?

(c) Determine the covariance of the sample membership indicators

(d) Verify that the fixed sample relations are satisfied in this case. □

Solution. The application of the sampling scheme produces an ordered sample $\tilde{s} = (\tilde{u}_1, \tilde{u}_2)$, and $s = \{x \,|\, x = \tilde{u}_1 \text{ or } x = \tilde{u}_2\}$ is the corresponding (unordered) sample that is finally obtained. Note that the intervals $\tilde{u}_1 = 1$ and $\tilde{u}_1 = 8$ (*i.e.*, [9, 10) and [16, 17)) have just one adjacent interval ([10, 11) for $\tilde{u}_1 = 1$ and [15, 16) for $\tilde{u}_1 = 8$), whereas if $\tilde{u}_1 = x \in \{2,3,4,5,6,7\}$ then x has five nonadjacent intervals. Since the second unit is selected without replacement from intervals which are nonadjacent to $\tilde{u}_1$, it follows that

$$\begin{array}{rcll}
P[\tilde{u}_2 = x|\tilde{u}_1 = 1] & = & 1/6, & x \in \{3,4,5,6,7,8\} \\
P[\tilde{u}_2 = x|\tilde{u}_1 = 2] & = & 1/5, & x \in \{4,5,6,7,8\} \\
P[\tilde{u}_2 = x|\tilde{u}_1 = 3] & = & 1/5, & x \in \{1,5,6,7,8\} \\
P[\tilde{u}_2 = x|\tilde{u}_1 = 4] & = & 1/5, & x \in \{1,2,6,7,8\} \\
P[\tilde{u}_2 = x|\tilde{u}_1 = 5] & = & 1/5, & x \in \{1,2,3,7,8\} \\
P[\tilde{u}_2 = x|\tilde{u}_1 = 6] & = & 1/5, & x \in \{1,2,3,4,8\} \\
P[\tilde{u}_2 = x|\tilde{u}_1 = 7] & = & 1/5, & x \in \{1,2,3,4,5\} \\
P[\tilde{u}_2 = x|\tilde{u}_1 = 8] & = & 1/6, & x \in \{1,2,3,4,5,6\}
\end{array}$$

Note that after each equality the set of units that are nonadjacent to $\tilde{u}_1$ is explicitly indicated. Recalling that $P[\tilde{u}_1 = y] = 1/8$ for every $y \in \{1,2,\ldots,8\}$ it follows from the multiplication rule that

$$\begin{array}{rcll}
P[(\tilde{u}_1,\tilde{u}_2) = (1,x)] & = & 1/48, & x \in \{3,4,5,6,7,8\} \\
P[(\tilde{u}_1,\tilde{u}_2) = (2,x)] & = & 1/40, & x \in \{4,5,6,7,8\} \\
P[(\tilde{u}_1,\tilde{u}_2) = (3,x)] & = & 1/40, & x \in \{1,5,6,7,8\} \\
P[(\tilde{u}_1,\tilde{u}_2) = (4,x)] & = & 1/40, & x \in \{1,2,6,7,8\} \\
P[(\tilde{u}_1,\tilde{u}_2) = (5,x)] & = & 1/40, & x \in \{1,2,3,7,8\} \\
P[(\tilde{u}_1,\tilde{u}_2) = (6,x)] & = & 1/40, & x \in \{1,2,3,4,8\} \\
P[(\tilde{u}_1,\tilde{u}_2) = (7,x)] & = & 1/40, & x \in \{1,2,3,4,5\} \\
P[(\tilde{u}_1,\tilde{u}_2) = (8,x)] & = & 1/48, & x \in \{1,2,3,4,5,6\}
\end{array}$$

The same information is presented in the following matrix where the i, j entry gives the probability $P[\tilde{u}_1 = i, \tilde{u}_2 = j]$:

$$P[\tilde{S} = (i,j)]$$

$$\begin{array}{cccccccc}
0 & 0 & 1/48 & 1/48 & 1/48 & 1/48 & 1/48 & 1/48 \\
0 & 0 & 0 & 1/40 & 1/40 & 1/40 & 1/40 & 1/40 \\
1/40 & 0 & 0 & 0 & 1/40 & 1/40 & 1/40 & 1/40 \\
1/40 & 1/40 & 0 & 0 & 0 & 1/40 & 1/40 & 1/40 \\
1/40 & 1/40 & 1/40 & 0 & 0 & 0 & 1/40 & 1/40 \\
1/40 & 1/40 & 1/40 & 1/40 & 0 & 0 & 0 & 1/40 \\
1/40 & 1/40 & 1/40 & 1/40 & 1/40 & 0 & 0 & 0 \\
1/48 & 1/48 & 1/48 & 1/48 & 1/48 & 1/48 & 0 & 0
\end{array}$$

This table can be used to determine the probability of selecting any unordered sample s as follows:

$$\text{If } s = \{i, j\},$$
$$p(s) = P[S = s] = P[\tilde{S} = (i, j)] + P[\tilde{S} = (j, i)].$$

The following upper triangular matrix gives the distribution of S; for each pair (i, j) with $1 \leq i < j < 8$ and $i \leq 6$, $P[S = \{i, j\}]$ is given in the (i, j) entry.

$$P[S = \{s_1, s_2\}], \quad s_1 < s_2$$
$$\begin{array}{cccccccc}
0 & 0 & 11/240 & 11/240 & 11/240 & 11/240 & 11/240 & 10/240 \\
0 & 0 & 0 & 12/240 & 12/240 & 12/240 & 12/240 & 11/240 \\
0 & 0 & 0 & 0 & 12/240 & 12/240 & 12/240 & 11/240 \\
0 & 0 & 0 & 0 & 0 & 12/240 & 12/240 & 11/240 \\
0 & 0 & 0 & 0 & 0 & 0 & 12/240 & 11/240 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 11/240
\end{array} \tag{3.4.1}$$

In decimal notation this matrix is

$$\begin{array}{cccccccc}
0 & 0 & 0.046 & 0.046 & 0.046 & 0.046 & 0.046 & 0.042 \\
0 & 0 & 0 & 0.05 & 0.05 & 0.05 & 0.05 & 0.046 \\
0 & 0 & 0 & 0 & 0.05 & 0.05 & 0.05 & 0.046 \\
0 & 0 & 0 & 0 & 0 & 0.05 & 0.05 & 0.046 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.05 & 0.046 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.046
\end{array} \tag{3.4.2}$$

For instance, $p(\{2, 5\}) = 12/240 = 0.05$, and $p(\{4, 8\}) = 11/240 = 0.046$.

(a) Recall that $I_k = I[u_k \in S]$ is the membership indicator of the k-th unit, and that the sampling design is given in (3.4.1) or (3.4.2). Observe now that in the present case $\pi_k = P[k \in S] = \sum_{j>k} P[S = \{k, j\}] + \sum_{j<k} P[S = \{j, k\}]$ is the summation of the k-th row and column of the matrix (3.4.1). For instance,

$$\begin{aligned}
\pi_4 &= \sum_{j>4} P[S = \{4, j\}] + \sum_{j<4} P[S = \{j, 4\}] \\
&= (12/240 + 12/240 + 11/240) + (11/240 + 12/240) = 58/240
\end{aligned}$$

and

$$\begin{aligned}
\pi_7 &= \sum_{j>7} P[S = \{7, j\}] + \sum_{j<7} P[S = \{j, 7\}] \\
&= (0) + (11/240 + 12/240 + 12/240 + 12/240 + 12/240) = 59/240
\end{aligned}$$

The vector π of first order inclusion probabilities is given below:

$$\begin{array}{cccc} \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ 65/240 & 59/240 & 58/240 & 58/240 \\ \pi_5 & \pi_6 & \pi_7 & \pi_8 \\ 58/240 & 58/240 & 59/240 & 65/240 \end{array} \tag{3.4.3}$$

On the other hand, the matrix $[\pi_{j,k}]$ is given by

$$\pi_{j,k} = P[u_j \in S, u_k \in S] = P[I_j = 1, I_k = 1] = P[S = \{j,k\}],$$

where $\pi_{k,k} = P[k \in S] = \pi_k$. Thus, the matrix $[\pi_{j,k}]$, can be immediately determined combining (3.4.1) or (3.4.2) with (3.4.3):

$$[\pi_{j,k}] = \frac{1}{240}\begin{bmatrix} 65 & 0 & 11 & 11 & 11 & 11 & 11 & 10 \\ 0 & 59 & 0 & 12 & 12 & 12 & 12 & 11 \\ 11 & 0 & 58 & 0 & 12 & 12 & 12 & 11 \\ 11 & 12 & 0 & 58 & 0 & 12 & 12 & 11 \\ 11 & 12 & 12 & 0 & 58 & 0 & 12 & 11 \\ 11 & 12 & 12 & 12 & 0 & 58 & 0 & 11 \\ 11 & 12 & 12 & 12 & 12 & 0 & 59 & 0 \\ 10 & 11 & 11 & 11 & 11 & 11 & 0 & 65 \end{bmatrix} \tag{3.4.4}$$

(b) The sampling design p is measurable if

$$\pi_{j,k} = P[U_j \in S, U_k \in S] > 0$$

for each pair of different units U_j and U_k. In the present case $\pi_{3,4} = 0$ and then the design p is not measurable.

(c) The covariance matrix $\lambda = [\mathrm{Cov}\,(I_k, I_j)] = [\pi_{j,k} - \pi_j \pi_k]$ is given by

$$\lambda = \begin{bmatrix} 0.197 & -0.067 & -0.02 & -0.02 & -0.02 & -0.02 & -0.021 & -0.032 \\ -0.067 & 0.185 & -0.059 & -0.009 & -0.009 & -0.009 & -0.01 & -0.021 \\ -0.02 & -0.059 & 0.183 & -0.058 & -0.008 & -0.008 & -0.009 & -0.02 \\ -0.02 & -0.009 & -0.058 & 0.183 & -0.058 & -0.008 & -0.009 & -0.02 \\ -0.02 & -0.009 & -0.008 & -0.058 & 0.183 & -0.058 & -0.009 & -0.02 \\ -0.02 & -0.009 & -0.008 & -0.008 & -0.058 & 0.183 & -0.059 & -0.02 \\ -0.021 & -0.01 & -0.009 & -0.009 & -0.009 & -0.059 & 0.185 & -0.067 \\ -0.032 & -0.021 & -0.02 & -0.02 & -0.02 & -0.02 & -0.067 & 0.197 \end{bmatrix} \tag{3.4.5}$$

(d) The fixed sample relations are:

(i) $\sum_k \pi_k = n$. In the present case (3.4.3) yields that

$$\sum_{k=1}^{8} \pi_k = \frac{65 + 59 + 58 + 58 + 58 + 58 + 59 + 65}{240} = \frac{480}{240} = 2.$$

(ii) $\sum_{k,j \in U, j \neq k} \pi_{k,j} = n(n-1)$.

The matrix $[\pi_{j,k}]$ in (3.4.4) was computed and saved in R under the name `PIMat`. The summation of interest was obtained with following R-code:

```
AUX <- PIMat; diag(AUX) <- 0; sum(AUX)
```

The result is 2; since $n * (n-1) = 2 * (2-1)$ it follows that the equality (ii) holds. Note that `AUX` has zeros along the main diagonal, and that the command `sum(AUX)` returns the sum of all elements of the matrix `AUX`.

(iii) The third relation is

$$\sum_{j:j \in U,\, j \neq k} \pi_{k,j} = (n-1)\pi_k.$$

Since $n = 2$, the right-hand side equals the k-th component of the vector `diag(PIMat)`. The left-hand side is the k-th component of `apply(AUX, 1, sum)`. Thus, to verify the equality in the present context, it is sufficient to issue the following R-command:

```
round( apply(AUX, 1, sum) - diag(PIMat), 5 )
```

and to check that a vector of zeros is produced. The output is the null vector of size 8, verifying the third equality; note that, because of unavoidable rounding errors, the use of the `round` function is necessary. □

Example 3.4.2. A sample s of n individuals is drawn by the *SI* design from a frame that contains N individuals. The households corresponding to the selected individuals are identified. Compute the inclusion probability of a household composed by M individuals, where $M < n$. Obtain approximate expressions for the inclusion probability for $M = 1, 2, 3$, supposing that both N and n are large with $n/N = f_N \to f > 0$. □

Solution. A household of M inhabitants is included for analysis if and only if one of the M individuals is selected in the *SI* sample s. Thus, the probability of inclusion of a household of size M is

$$\alpha_M = 1 - \frac{\binom{N-M}{n}}{\binom{N}{n}}$$

Observe that

$$\begin{aligned}\frac{\binom{N-M}{n}}{\binom{N}{n}} &= \frac{(N-M)_n}{(N)_n} \\ &= \frac{(N-n)(N-n-1)\cdots(N-M-n+1)}{N(N-1)\cdots(N-M+1)} \\ &= \frac{(1-f_N)(1-f_N-1/N)\cdots(1-f_N-(M-1)/N)}{1(1-1/N)\cdots(1-(M-1)/N)}\end{aligned}$$

Thus, if M is fixed, then as n and N go to ∞ in such a way that $n/N = f_N \to f$ it follows that

$$\frac{\binom{N-M}{n}}{\binom{N}{n}} \to (1-f)^M,$$

and then $\alpha_M \to 1-(1-f)^M$. □

3.5. Inclusion Probabilities in Bernoulli Designs

This section contains two examples about the inclusion probabilities in Bernoulli sample designs, where the underlying population is subdivided in clusters.

Example 3.5.1. Consider a population $\mathcal{U}$ with three subpopulations $\mathcal{U}_1, \mathcal{U}_2$ and $\mathcal{U}_3$ of sizes $N_1 = 600$, $N_2 = 300$ and $N_3 = 100$, so that $\mathcal{U}$ is of size $N = 1000$. For each k in $\mathcal{U}$, the inclusion in the sample s is determined by a Bernoulli experiment that gives the element k the probability π_k of being selected. The experiments are independent.

(a) Let $\pi_k = 0.1$ for $k \in \mathcal{U}_1$, $\pi_k = 0.2$ for $k \in \mathcal{U}_2$, and $\pi_k = 0.8$ for $k \in \mathcal{U}_3$. Find the expected value and variance of n_S under this design.

(b) Suppose that π_k is constant for every $k \in \mathcal{U}$. Determine this constant so that the expected value of the sample size agrees with the expected value obtained in the previous part (a). Next, determine the variance of the sample size and compare it with the variance in case (a). □

Solution. (a) Let $n_S^{(i)}$ be the number of elements in the sample that belong to the subpopulation $\mathcal{U}_i$, so that

(i) $n_S^{(1)}, n_S^{(2)}, n_S^{(3)}$ are independent;

(ii) $n_S^{(1)} \sim Ber(\pi_1, 600) = Ber(0.1, 600)$, $n_S^{(2)} \sim Ber(\pi_2, 300) = Ber(0.2, 300)$ and $n_S^{(3)} \sim Ber(\pi_3, 100) = Ber(0.8, 100)$;

(iii) $n_S = n_S^{(1)} + n_S^{(2)} + n_S^{(3)}$.

It follows that

$$\begin{aligned} E[n_S] &= E[n_S^{(1)}] + E[n_S^{(2)}] + E[n_S^{(3)}] \\ &= 600(0.1) + 300(0.2) + 100(0.8) = 60 + 60 + 80 = 200 \end{aligned}$$

and

$$\begin{aligned} \text{Var}\,[n_S] &= \text{Var}\left[n_S^{(1)}\right] + \text{Var}\left[n_S^{(2)}\right] + \text{Var}\left[n_S^{(3)}\right] \\ &= 600(0.1)(0.9) + 300(0.2)(0.8) + 100(0.8)(0.2) = 54 + 48 + 16 = 118. \end{aligned}$$

(b) Let $\pi_k = \pi$ for every $k \in \mathcal{U}$. In this case, the number of elements n_S in the sample S is a random variable with distribution $Ber(\pi, 1000)$ so that $E[n_S] = 1000\pi$, and $\text{Var}\,[n_S] = 1000\pi(1-\pi)$. Thus, in order to have that the expected value 1000π coincides with the one in part (a) the equality $1000\pi = 200$ must be satisfied, so that $\pi = 0.20$. In this case the variance is $\text{Var}\,[n_S] = 1000(0.2)(0.8) = 160$, which is larger than the one in part (a). □

Example 3.5.2. A Population of $1,600$ individuals is divided into 800 clusters (households) with the number of clusters of size a is N_a for $a = 1, 2, 3, 4$ as indicated below:

a :	1	2	3	4
N_a :	250	350	150	50

A sample of individuals is selected as follows: 300 clusters are drawn from the 800 by the SI design and all individuals in the selected clusters constitute the sample. Determine $E[n_S]$ and $\text{Var}\,[n_S]$ □

The argument below relies on the formulas for the expectation and variance of a random vector with multidimensional hypergeometric distribution, which are established at the end of the Chapter 4.

Solution. The sample of $n = 300$ households is selected form the population $\mathcal{U}$, which is the union of four subpopulations $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3, \mathcal{U}_4$ of sizes $N_1 = 250, N_2 = 350, N_3 = 150, N_4 = 50$, respectively. If X_i is the number of units in the sample that belong to $\mathcal{U}_i$, it follows that

$$\mathbf{X} = (X_1, X_2, X_3, X_4) \sim \mathcal{H}_4(300, 800; 250, 350, 150, 50)$$

and then

$$E[\mathbf{X}] = n\mathbf{p} \quad \text{and} \quad \text{Var}\,[\mathbf{X}] = n(1 - \tilde{f})\,[diag(\mathbf{p}) - \mathbf{p}'\mathbf{p}]$$

where

$$\mathbf{p} = (N_1/N, N_2/N, N_3/N, N_4/N) = (0.3125,\ 0.4375,\ 0.1875,\ 0.0625)$$

and

$$\tilde{f} = \frac{n-1}{N-1} = \frac{299}{799}.$$

Consequently

$$E[\mathbf{X}] = n\mathbf{p} = (93.75, 131.25, 56.25, 18.75).$$

and

$$\mathrm{Var}\,[\mathbf{X}] = \begin{bmatrix} 40.3336 & -25.6668 & -11.0001 & -3.6667 \\ -25.6668 & 46.2003 & -15.4001 & -5.1334 \\ -11.0001 & -15.4001 & 28.6002 & -2.2 \\ -3.6667 & -5.1334 & -2.2 & 11.0001 \end{bmatrix}$$

The number of individuals in the selected clusters is

$$n_S = X_1 + 2X_2 + 3X_3 + 4X_4 = (1,2,3,4) \cdot \mathbf{X}$$

and then

$$E[n_S] = (1,2,3,4) \cdot E[\mathbf{X}] = (1,2,3,4) \cdot (93.75, 131.25, 56.25, 18.75) = 600,$$

and

$$\mathrm{Var}\,[n_S] = (1,2,3,4)\mathbf{V} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = 140.801$$

completing the argument. □

3.6. An Example with Variable Sample Size

In this section two simple examples are used to illustrate the main ideas introduced in this chapter.

Example 3.6.1. Consider a population of size $N = 3$, say $\mathcal{U} = \{1, 2, 3\}$ and let the sampling design $p(\cdot)$ be determined as follows:

s :	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$p(s)$:	0.4	0.3	0.2	0.1

(a) Compute the vector $\pi = (\pi_k)$ and the matrix $[\pi_{j,k}]$.

(b) Find $E[n_S]$ by direct calculation using the table above;

(c) Find $E[n_S]$ by using the formula in terms of the inclusion probabilities π_k. □

Solution. Recall the $\pi_k = P[k \in S]$ is the probability of inclusion of the unit k in the selected sample S, whereas $\pi_{j,k} = P[j \in S, k \in S]$ is the probability of having that both units j and k belong to S.

(a) The inclusion probabilities are given by

$$\begin{array}{rccc} k: & 1 & 2 & 3 \\ \pi_k = P[k \in S]: & 0.8 & 0.7 & 0.6 \end{array}$$

that is,

$$(\pi_1, \pi_2, \pi_3) = (0.8, 0.7, 0.6)$$

For instance

$$\begin{aligned} \pi_2 &= P[2 \in S] \\ &= P[S = \{1,2\}] + P[S = \{2,3\}] + P[S = \{1,2,3\}] = 0.4 + 0.2 + 0.1 = 0.7. \end{aligned}$$

On the other hand,

$$[\pi_{j,k}] = \begin{bmatrix} 0.8 & 0.5 & 0.4 \\ 0.5 & 0.7 & 0.3 \\ 0.4 & 0.3 & 0.6 \end{bmatrix}$$

As an example, $\pi_{1,3} = P[1 \in S, 3 \in S] = P[S = \{1,3\}] + P[S = \{1,2,3\}] = 0.3 + 0.1 = 0.4$.

(b) From the definition of the sampling design, n_S attains two values, namely 2 and 3. Note that $P[n_S = 2] = P[S = \{1,2\}] + P[S = \{1,3\}] + P[S = \{2,3\}] = 0.9$, and $P[n_S = 3] = P[S = \{1,2,3,\}] = 0.1$. Consequently,

$$E[n_S] = 2P[n_S = 2] + 3P[n_S = 3] = 2 \cdot 0.9 + 3 \cdot 0.1 = 2.1.$$

(c) Note that $E[n_S] = \pi_1 + \pi_2 + \pi_3 = 0.8 + 0.7 + 0.6 = 2.1$. □

Example 3.6.2. In the context of Exercise 3.6.1, let the values of the the study variables be

$$y_1 = 16, \quad y_2 = 21, \quad y_3 = 18,$$

so that the total is

$$t = 55.$$

(a) Compute the expectation and variance of the π-estimator $\hat{t}_\pi$.

(b) Compute the variance of $\hat{t}_\pi$ using the general formula in terms of the covariances $\Delta_{j,k}$.

(c) Compute the coefficient of variation of the π estimator.

(d) Compute the estimator of the variance $\hat{V}(\hat{t}_\pi)$ using the π expansion formula.

(e) Find the expectation of $\hat{V}(\hat{t}_\pi)$ using the definition of expected value. □

Solution. The expanded values of y_i, namely, $\check{y}_i = y_i/\pi_i$ are given by

$$\check{y}' = (\check{y}_1, \check{y}_2, \check{y}_3) = (20, 30, 30).$$

(a) Observe now that $\hat{t}_\pi(\{1,2\}) = \check{y}_1 + \check{y}_2 = 50$ and $\hat{t}_\pi(\{1,2,3\}) = \check{y}_1 + \check{y}_2 + \check{y}_3 = 80$. Proceeding similarly, the following table is obtained:

$s:$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$p(s):$	0.4	0.3	0.2	0.1
$\hat{t}_\pi:$	50	50	60	80
$\hat{t}_\pi - 55:$	-5	-5	5	25

It follows that $E[\hat{t}_\pi] = 50\cdot 0.7 + 60\cdot 0.2 + 80\cdot 0.1 = 35 + 12 + 8 = 55$, verifying that $\hat{t}_\pi$ is an unbiased estimator, and

$$V(\hat{t}_\pi) = (5)^2 \cdot 0.9 + 25^2 \cdot 0.1 = 85.$$

(b) First observe that the second order probabilities $\pi_{i,j}$ are given by

$$\begin{aligned}
\pi_{2,1} = \pi_{1,2} =& P[\{1,2\} \subset S] = P[S = \{1,2\}] + P[S = \{1,2,3\}] = 0.5\\
\pi_{3,1} = \pi_{1,3} =& P[\{1,3\} \subset S] = P[S = \{1,3\}] + P[S = \{1,2,3\}] = 0.4\\
\pi_{3,2} = \pi_{2,3} =& P[\{2,3\} \subset S] = P[S = \{2,3\}] + P[S = \{1,2,3\}] = 0.3
\end{aligned}$$

and

$$\begin{aligned}
\pi_{1,1} = \pi_1 = P[1 \in S]&\\
&= P[S = \{1,2\}] + P[S = \{1,3\}] + P[S = \{1,2,3\}]\\
&= 0.8
\end{aligned}$$

whereas $\pi_{2\,2}$ and $\pi_{3,3}$ are computed similarly. The matrix $[\pi_{i\,j}]$ was introduced in the R environment under the name `pimat` and then the matrix

$$\Delta = [\Delta_{j,k}] = [\pi_{jk} - \pi_j \pi_k] = \mathrm{Cov}\,(I_j, I_k)$$

was computed using the following R code:

```
Delta <- pimat - crossprod (rbind(diag ( pimat) ) )
```

and the following result was obtained:

$$\Delta = \begin{bmatrix} 0.16 & -0.06 & -0.08 \\ -0.06 & 0.21 & -0.12 \\ -0.08 & -0.12 & 0.24 \end{bmatrix}$$

In terms of the covariance matrix Δ, the variance of $\hat{t}_\pi$ is given by

$$\begin{aligned}
V[\hat{t}_\pi] &= (\check{y}_1, \check{y}_2, \check{y}_3)\Delta \begin{bmatrix} \check{y}_1 \\ \check{y}_2 \\ \check{y}_3 \end{bmatrix}\\
&= (20, 30, 30) \begin{bmatrix} 0.16 & -0.06 & -0.08 \\ -0.06 & 0.21 & -0.12 \\ -0.08 & -0.12 & 0.24 \end{bmatrix} \begin{bmatrix} 20 \\ 30 \\ 30 \end{bmatrix} = (20, 30, 30) \begin{bmatrix} -1 \\ 1.5 \\ 2 \end{bmatrix} = 85.
\end{aligned}$$

(c) $CV(\hat{t}_\pi) = (V[\hat{t}_\pi])^{1/2}/E[\hat{t}_\pi] = 85^{1/2}/55 = 0.1676281$

(d) The estimator of the variance $\hat{V}(\hat{t}_\pi)$ is given by

$$\hat{V}(\hat{t}_\pi) = \sum_S \check{y}_j \check{\Delta}_{j,k} \check{y}_k$$

where

$$\check{\Delta} = [\Delta_{j,k}/\pi_{j,k}] = \begin{bmatrix} 0.2 & -0.12 & -0.2 \\ -0.12 & 0.3 & -0.4 \\ -0.2 & -0.4 & 0.4 \end{bmatrix}$$

Using these two last displays, for each possible sample s, the estimate $\hat{V}(\hat{t}_\pi)$ can be immediately computed. For instance, if $s = \{1, 2\}$, then

$$\hat{V}(\hat{t}_\pi)(\{1,2\}) = (20, 30, 0)\check{\Delta} \begin{bmatrix} 20 \\ 30 \\ 0 \end{bmatrix} = 206$$

The entries in the third line of the following table are computed similarly.

s :	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$p(s)$:	0.4	0.3	0.2	0.1
$\hat{V}(\hat{t}_\pi)(s)$:	206	200,	−90	−394

It is interesting to observe that $\hat{V}[\hat{t}_\pi]$ attains negative values at some samples.

(d) Note that

$$E[\hat{V}(\hat{t}_\pi)] = 206 \cdot 0.4 + 200 \cdot 0.3 - 90 \cdot 0.2 - 394 \cdot 0.1 = 85,$$

confirming that $\hat{V}[\hat{t}_\pi]$ is an unbiased estimator of $V[\hat{t}_\pi]$. □

Chapter 4

Simple and Bernoulli Schemes

4.1. Introduction

The simple and Bernoulli sampling schemes have been previously studied, and in this chapter they will be analyzed more deeply. To begin with, in Section 2 it is shown that, conditionally on the observed sample size, the sample obtained from a Bernoulli scheme is a simple random sample, and it is shown that, under the *SI* design, the sample variance as an unbiased estimator of the population variance. Next, in Section 3 it is proved that under the Bernoulli scheme the sample variance is a biased estimator, but that the relative bias converges to zero as the population size grows, and the section concludes analyzing the covariance between two sample means obtained from disjoint simple random samples. Then, in Section 4 sampling with replacement is considered, the estimation of the population total is analyzed via the the Hurwitz-Hansen expansion estimator, and the results are illustrated in Section 5 for the problem of estimating the income per household; an interesting feature of the of the analysis is that the sampling units are not the population elements (the individuals), but small clusters (the households). Finally, Sections 6 and 7 contain a formal statement and proofs of basic properties of the multivariate hypergeometric distribution and the Bernoulli sampling design.

4.2. Relation Between Simple and Bernoulli Samples

The main objective of this section is to show that, conditionally on the observed sample size, the sample obtained from a Bernoulli scheme is a simple random sample. The analysis is used to provide, under the *SI* design, a short proof of the unbiasedness of the sample variance as an estimator of the population variance.

Example 4.2.1. Let S be a sample realized from the *BE* design with $\pi_k = \pi$ for every k and, as usual, let n_S denote the (random) sample size of S. Show that, conditionally on $n_S = n$, the probability of any sample s of size n is $1/\binom{N}{n}$, the same probability as in the *SI* design. $\square$

Solution. Under Bernoulli sampling, $n_S \sim B(N, \pi)$, where N is the population size, so that

$$P[n_S = n] = \binom{N}{n} \pi^n (1-\pi)^{N-n}.$$

Now, let s be an arbitrary sample (subset of the population $\mathcal{U}$) with n elements, and note that under Bernoulli sampling

$$P[S = s] = \pi^n (1-\pi)^{N-n}$$

Thus,

$$P[S = s | n_S = n] = \frac{P[S = s, n_S = n]}{P[n_S = n]} = \frac{P[S = s]}{n_S = n]} = \frac{\pi^n (1-\pi)^{N-n}}{\binom{N}{n} \pi^n (1-\pi)^{N-n}}$$

and then

$$P[S = s | n_S = n] = \frac{1}{\binom{N}{n}}.$$

Thus, conditionally on the event $n_S = n$, all samples of size n have the same probability $1/\binom{N}{n}$, as in the *SI* design. $\square$

Example 4.2.2. The objective of this exercise is to show that, in the *SI* design, the equality

$$E_{SI}[S^2_{ys}] = S^2_{yU} \tag{4.2.1}$$

holds, so that the expected value of the (corrected) sample variance equals the (corrected) population variance. Note that, for every set $A \subset \mathcal{U}$,

$$S^2_{yA} = \frac{1}{n_A - 1} \sum_{i \in A} (y_i - \bar{y}_A)^2, \text{ where } \bar{y}_A = \sum_{i \in A} y_i / n_A$$

and n_A is the number of elements of A.

(a) Establish (4.2.1) using that

$$E\left[\sum_{k \in s} y_k^2\right] = E\left[\sum_{k \in U} I_k y_k^2\right] = \frac{n}{N} \sum_U y_k^2 \tag{4.2.2}$$

and

$$E\left[\sum_{j \neq k, j,k \in s} y_j y_k\right] = E\left[\sum_{j \neq k, j,k \in U} I_j I_k y_j y_k\right] = \frac{n(n-1)}{N(N-1)} \sum_{j \neq k, j,k \in U} y_j y_k \tag{4.2.3}$$

(b) Prove (4.2.1) using that

$$\sum_s (y_j - y_k)^2 = 2n(n-1)S_{ys}^2 \tag{4.2.4}$$

with a similar relation for S_{yU}^2 $\square$

Solution. (a) Observe that under the *SI* design $n_s = n$ for every possible sample, and then

$$\begin{aligned}(n-1)S_{ys}^2 &= \sum_{i\in s}(y_i - \bar{y}_s)^2 = \sum_{i\in s} y_i^2 - n\bar{y}_s^2 \\ &= \sum_{i\in s} y_i^2 - \frac{1}{n}\left(\sum_{k\in s} y_k\right)^2 = \sum_{i\in s} y_i^2 - \frac{1}{n}\left(\sum_{k\in s} y_k^2 + \sum_{j\neq k, j,k\in s} y_j y_k\right) \\ &= \frac{n-1}{n}\sum_{i\in s} y_i^2 - \frac{1}{n}\sum_{j\neq k, j,k\in s} y_j y_k\end{aligned}$$

Combining this relation with (4.2.2) and (4.2.3) it follows that

$$\begin{aligned}(n-1)E[S_{ys}^2] &= \frac{n-1}{n}E\left[\sum_{i\in s} y_i^2\right] - \frac{1}{n}E\left[\sum_{j\neq k, j,k\in s} y_j y_k\right] \\ &= \frac{n-1}{n}\frac{n}{N}\sum_{i\in U} y_i^2 - \frac{1}{n}\frac{n(n-1)}{N(N-1)}\sum_{j\neq k, j,k\in U} y_j y_k \\ &= \frac{n-1}{N}\sum_{i\in U} y_i^2 - \frac{n-1}{N(N-1)}\sum_{j\neq k, j,k\in U} y_j y_k \\ &= \frac{n-1}{N}\left[\sum_{i\in U} y_i^2 - \frac{1}{N-1}\sum_{j\neq k, j,k\in U} y_j y_k\right]\end{aligned}$$

To continue, observe that

$$\sum_{j\neq k, j,k\in U} y_j y_k = \left(\sum_{k\in U} y_k\right)^2 - \sum_{k\in U} y_k^2 = N^2\bar{y}_U^2 - \sum_{k\in U} y_k^2,$$

an equality that together with the previous display yields that

$$\begin{aligned}(n-1)E[S_{ys}^2] &= \frac{n-1}{N}\left[\sum_{i\in U} y_i^2 - \frac{1}{N-1}\left(N^2\bar{y}_U^2 - \sum_{k\in U} y_k^2\right)\right] \\ &= \frac{n-1}{N}\left[\frac{N}{N-1}\sum_{i\in U} y_i^2 - \frac{N^2}{N-1}\bar{y}_U^2\right] \\ &= \frac{n-1}{N-1}\left[\sum_{i\in U} y_i^2 - N\bar{y}_U^2\right] = (n-1)S_{yU}^2\end{aligned}$$

and (4.2.1) follows.

(b) First, it will be verified that $2n(n-1)S_{ys}^2 = \sum_{i,j\in s}(y_i - y_j)^2$. Note that

$$\begin{aligned}
\sum_{i,j\in s}(y_i - y_j)^2 &= \sum_{i,j\in s}(y_i - \overline{y}_s - (y_j - \overline{y}_s))^2 \\
&= \sum_{i,j\in s}[(y_i - \overline{y}_s)^2 + (y_j - \overline{y}_s))^2 - 2(y_i - \overline{y}_s)(y_j - \overline{y}_s)] \\
&= \sum_{i,j\in s}(y_i - \overline{y}_s)^2 + \sum_{i,j\in s}(y_j - \overline{y}_s))^2 - 2\sum_{i,j\in s}(y_i - \overline{y}_s)(y_j - \overline{y}_s) \\
&= n\sum_{i\in s}(y_i - \overline{y}_s)^2 + n\sum_{j\in s}(y_j - \overline{y}_s))^2 - 2\sum_{i\in s}(y_i - \overline{y}_s)\sum_{j\in s}(y_j - \overline{y}_s) \\
&= 2n\sum_{i\in s}(y_i - \overline{y}_s)^2;
\end{aligned}$$

since $S_{ys}^2 = (n-1)^{-1}\sum_{i,j\in s}(y_i - y_j)^2$, it follows that

$$\sum_{i,j\in s}(y_i - y_j)^2 = 2n\sum_{i\in s}(y_i - \overline{y}_s)^2 = 2n(n-1)S_{y\,s}^2,$$

establishing the desired equality. Next, observe that

$$\sum_s (y_j - y_k)^2 = \sum_U I_j I_k (y_j - y_k)^2,$$

so that

$$\begin{aligned}
2n(n-1)E[S_{ys}^2] &= E\left[\sum_U I_j I_k (y_j - y_k)^2\right] \\
&= \frac{n(n-1)}{N(N-1)}\sum_U (y_j - y_k)^2 \\
&= \frac{n(n-1)}{N(N-1)} 2N(N-1)S_{yU}^2 = 2n(n-1)S_{yU}^2,
\end{aligned}$$

where equality (4.2.4) for S_{yU}^2 was used in the last step. □

4.3. Relative Bias Under BE Design

In this section it is shown that, under the Bernoulli scheme, the sample variance is a biased estimator, and that the relative bias converges to zero as the population size grows. Next, the covariance between two sample means obtained from disjoint simple random samples will be obtained.

Example 4.3.1. Let s be a sample drawn by the *BE* design with $\pi_k = \pi$ for all k. Set

$$S_{ys}^2 = \frac{\sum_s (y_k - \bar{y}_s)^2}{(n_s - 1)} \text{ if } n_s \geq 2, \quad S_{ys}^2 = 0, \text{ if } n_s \leq 1.$$

Show that, as an estimator of S_{yU}^2, the relative bias of S_{ys}^2, namely

$$\frac{E[S_{ys}^2] - S_{yU}^2}{S_{yU}^2},$$

is given by

$$\frac{E[S_{ys}^2] - S_{yU}^2}{S_{yU}^2} = -P[n_S \le 1] = -[(1-\pi)^N + N\pi(1-\pi)^{N-1}]. \qquad \square$$

Solution. Recall that, under the *BE* design, given $n_S = k$ the conditional distribution of the sample s is the same as if s were selected via the *SI* design; see Example 4.2.1. Since S_{ys}^2 is an unbiased estimator of S_{yU}^2 when the sample size is larger than 1, it follows that

$$E[S_{ys}^2 | n_S = k] = S_{yU}^2, \quad k \ge 2.$$

On the other hand, $S_{ys}^2 = 0$ when $n_s \le 1$, and combining this fact with the above display it follows that $E[S_{ys}^2] = S_{yU}^2 P[n_S \ge 2]$, so that

$$\frac{E[S_{ys}^2] - S_{yU}^2}{S_{yU}^2} = P[n_S \ge 2] - 1 = -P[n_S \le 1].$$

Too conclude recall that $n_S \sim B(N, \pi)$ under the *BE* design, and then

$$P[n_S \le 1] = (1-\pi)^N + N(1-\pi)^{N-1}\pi,$$

completing the argument. $\square$

Now, the covariance between to disjoint simple random samples will be obtained.

Example 4.3.2. Let s_A be an *SI* sample, and let s_B be an *SI* sample from $\mathcal{U} \setminus s_A$. Denote by $\hat{y}_A$ and $\hat{y}_B$ the sample means corresponding to s_A and s_B, respectively. Determine the covariance and the correlation between $\hat{y}_A$ and $\hat{y}_B$. $\square$

Solution. Given s_A, s_B is an *SI* sample from $\mathcal{U} \setminus s_A$, so that

$$\begin{aligned} E[\hat{y}_B | s_A] &= \frac{1}{N - n_A} \sum_{\mathcal{U} \setminus s_A} y_k \\ &= \frac{1}{N - n_A} \left[\sum_{\mathcal{U}} y_k - \sum_{s_A} y_k \right], \\ &= \frac{N\bar{Y} - n_A \hat{y}_{s_A}}{N - n_A} \end{aligned} \tag{4.3.1}$$

and then, since $\hat{y}_A$ is a function of s_A,

$$E[\hat{y}_A \hat{y}_B | s_A] = \frac{N\hat{y}_A \bar{Y} - n_A \hat{y}_A^2}{N - n_A};$$

recalling that $E[\hat{y}_A] = \bar{Y}$ (since s_A is an *SI* sample) it follows that

$$E[\hat{y}_A \hat{y}_B] = \frac{N\bar{Y}^2 - n_A E[\hat{y}_A^2]}{N - n_A}. \tag{4.3.2}$$

On the other hand, since $\hat{y}_A$ is the sample mean of an *SI* sample of size n_A,

$$E[\hat{y}_A] = \overline{Y}$$

whereas, via (4.3.1),

$$\begin{aligned}
E[\hat{y}_B] &= E\left[E[\hat{y}_B|s_A]\right] \\
&= E\left[\frac{N\bar{Y} - n_A\hat{y}_{s_A}}{N - n_A}\right] \\
&= \frac{N\bar{Y} - n_A E\left[\hat{y}_{s_A}\right]}{N - n_A} \\
&= \frac{N\bar{Y} - n_A\bar{Y}}{N - n_A}
\end{aligned}$$

and then, $E[\hat{y}_B] = \bar{Y}$. Hence, (4.3.2) leads to

$$\begin{aligned}
\mathrm{Cov}\left(\hat{y}_A, \hat{y}_B\right) &= E[\hat{y}_A\hat{y}_B] - E[\hat{y}_A]E[\hat{y}_B] \\
&= \frac{N\bar{Y}^2 - n_A E[\hat{y}_A^2]}{N - n_A} - \bar{Y}^2 \\
&= \frac{n_A\bar{Y}^2 - n_A E[\hat{y}_A^2]}{N - n_A} \\
&= -\frac{n_A}{N - n_A}\mathrm{Var}\left[\hat{y}_A\right] \\
&= -\frac{n_A}{N - n_A}\frac{1}{n_A}\frac{N - n_A}{N}S_{yU}^2
\end{aligned}$$

and then

$$\mathrm{Cov}\left(\hat{y}_A, \hat{y}_B\right) = -\frac{1}{N}S_{yU}^2.$$

Now observe that that the formula for the variance from an *SI* sample yields that

$$\begin{aligned}
\sqrt{\mathrm{Var}\left[\hat{y}_A\right]\mathrm{Var}\left[\hat{y}_B\right]} &= \sqrt{\frac{1}{n_A}\frac{N - n_A}{N}S_{yU}^2\frac{1}{n_B}\frac{N - n_B}{N}S_{yU}^2} \\
&= \sqrt{\frac{N - n_B}{n_A}\frac{N - n_A}{n_B}}\frac{1}{N}S_{yU}^2
\end{aligned}$$

and together with the above displayed expression, it follows that

$$\mathrm{Corr}(\hat{y}_A, \hat{y}_B) = -\sqrt{\frac{n_A n_B}{(N - n_B)(N - n_A)}}. \qquad \square$$

4.4. Sampling with Replacement

In this section sampling with replacement is considered, and the estimation of the population total is analyzed. The following example introduces the Hurwitz-Hansen estimator.

Example 4.4.1. Let $\mathcal{U} = \{U_1, U_2, \ldots, U_N\}$ be a population of size N and suppose that $y_i = y(U_i)$ is the quantity of interest associated with the unit U_i. An ordered sample $\tilde{s} = (u_{i_1}, u_{i_2}, \ldots, u_{i_m})$ is selected with replacement in such a way that, in each draw, the probability of selecting unit U_i is p_i, $i = 1, 2, \ldots, N$. Consider the Hansen-Hurwitz estimator of the total $t = y_1 + y_2 + \cdots + y_N$ (or p-expanded estimator) $\hat{t}_{pwr}$ which is given by

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{j=1}^{m} \frac{y_{i_j}}{p_{i_j}}.$$

Show that

(a) $E[\hat{t}_{pwr}] = t$;

(b) The variance of $\hat{t}_{pwr}$ is given by

$$\text{Var}\left[\hat{t}_{pwr}\right] = \frac{1}{m} V_1, \quad \text{where} \quad V_1 = \sum_{k=1}^{N} p_k \left(\frac{y_k}{p_k} - t \right)^2.$$

(c) V_1 has estimator

$$\hat{V}_1 = \frac{1}{m-1} \sum_{j=1}^{m} \left(\frac{y_{i_j}}{p_{i_j}} - \hat{t}_{pwr} \right)^2$$

(d) Show that

$$V_1 = \sum_{k=1}^{N} \frac{y_k^2}{p_k} - t^2, \quad \text{and} \quad \hat{V}_1 = \frac{1}{m-1} \left[\sum_{j=1}^{m} \left(\frac{y_{i_j}}{p_{i_j}} \right)^2 - m\, \hat{t}_{pwr}^2 \right]. \tag{4.4.1}$$

□

Solution. (a) Let N_i be the random number of times that unit U_i appears in the sample, and observe that

$$N_k \sim B(m, p_k), \quad k = 1, 2, 3, \ldots, N,$$

as well as

$$\sum_{j=1}^{m} \frac{y_{i_j}}{p_{i_j}} = \sum_{k=1}^{N} \frac{y_k}{p_k} N_k.$$

Therefore, $E[N_k] = m p_k$ and

$$E\left[\sum_{j=1}^{m} \frac{y_{i_j}}{p_{i_j}} \right] = E\left[\sum_{k=1}^{N} \frac{y_k}{p_k} N_k \right] = \sum_{k=1}^{N} \frac{y_k}{p_k} E\left[N_k\right] = \sum_{k=1}^{N} \frac{y_k}{p_k} m p_k = mt,$$

so that

$$E[\hat{t}_{pwr}] = E\left[\frac{1}{m} \sum_{j=1}^{m} \frac{y_{i_j}}{p_{i_j}} \right] = t.$$

(b) Let Z_j be defined by

$$Z_j = \frac{y_{i_j}}{p_{i_j}}, \quad j = 1, 2, \ldots, m.$$

Since in each selection the unit U_r is selected with probability p_r, it follows that, for each j, the variable Z_j attains value y_k/p_k with probability p_k, so that

$$E[Z_j] = \sum_{k=1}^{N} \frac{y_k}{p_k} p_k = \sum_{k=1}^{N} y_k = t \tag{4.4.2}$$

and

$$\mathrm{Var}\,[Z_j] = E[(Z_j - t)^2] = \sum_{k=1}^{N} p_k \left(\frac{y_k}{p_k} - t\right)^2 =: V_1. \tag{4.4.3}$$

Observe now that $Z_1, Z_2, \ldots, Z_m$ are independent and identically distributed, and that

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{j=1}^{m} Z_j = \overline{Z}_m.$$

These two last displays immediately lead to

$$\mathrm{Var}\left[\hat{t}_{pwr}\right] = \frac{V_1}{m}.$$

(c) V_1 is the variance of the common distribution of the variables Z_j, which are independent and identically distributed. Thus, an unbiased estimator of V_1 is the (corrected) sample variance

$$\begin{aligned}\hat{V}_1 &= \frac{1}{m-1} \sum_{j=1}^{m} (Z_j - \overline{Z}_m)^2 \\ &= \frac{1}{m-1} \sum_{j=1}^{m} \left(\frac{y_{i_j}}{p_{i_j}} - \hat{t}_{pwr}\right)^2\end{aligned}$$

(d) Observe that $V_1 = \mathrm{Var}\,[Z_j] = E[Z_j^2] - (E[Z_j])^2$. Thus, since Z_j attains the values y_k/p_k with probability p_k, $k = 1, 2, 3, \ldots, N$, it follows from (4.4.2) that

$$V_1 = \sum_{k=1}^{N} p_k \left(\frac{y_k}{p_k}\right)^2 - t^2 = \sum_{k=1}^{N} \frac{y_k^2}{p_k} - t^2,$$

establishing the first equality in (4.4.1). As for the second one, recall that for $a_1, a_2, \ldots, a_m \in \mathbb{R}$,

$$\sum_{k=1}^{m} (a_k - \bar{a})^2 = \sum_{k=1}^{m} a_i^2 - m\,\bar{a}$$

Now set $a_k = y_{i_k}/p_{i_k}$ and note that $\bar{a} = m^{-1} \sum_{k=1}^{m} (y_{i_k}/p_{i_k}) = \hat{t}_{pwr}$. Thus, the above display yields

$$\hat{V}_1 = \frac{1}{m-1} \sum_{k=1}^{m} \left(\frac{y_{i_k}}{p_{i_k}} - \hat{t}_{pwr}\right)^2 = \frac{1}{m-1} \left[\sum_{k=1}^{m} \left(\frac{y_{i_k}}{p_{i_k}}\right)^2 - m\,\hat{t}_{pwr}^2\right]$$

which is the second equality in (4.4.1). $\square$

4.5. An Example: Income per Household

In this section an example about the average income per household is analyzed. There are two interesting features in this context: The sampling scheme is *with replacement*, and the sampling units are not the elements of the population (the households) but the inhabitants.

Example 4.5.1. To estimate the average income per household ($\sum_U y_k/N$) for a population of $N = 200$ households, a listing of the 600 individuals that belong to the 200 households was used as follows: A simple random sample with replacement of $m = 10$ persons was drawn. The households of the selected persons were identified, and information on the average income in the household (y_i/x_i) was collected, where y_k is the total household income in dollars, and x_k is the number of persons in the households. The results are as follows:

Draw j	Average household income (y_{i_j}/x_{i_j})
1	7000
2	8000
3	6000
4	5000
5	9000
6	4000
7	7000
8	8000
9	4000
10	2000

Compute an estimate of the average income per household based on the pwr estimator as well as the corresponding estimated coefficient of variation. □

Solution. The population consists of $N = 200$ households, whereas the sampling scheme is done on the class of all 600 inhabitants of the households. Once a person is selected, the corresponding household is fully analyzed to determine the total income (y_i). Thus, the scheme selects household i with probability $p_i = x_i/600$, where x_i is the number of inhabitants of household i. Tthe p-expanded estimator of the total $t = \sum_U y_k$, based on a sample with replacement of size $m = 10$ is

$$\hat{t}_{pwr} = \frac{1}{10}\sum_{k=1}^{10}\frac{y_{i_k}}{p_{i_k}} = 600\frac{1}{10}\sum_{k=1}^{10}\frac{y_{i_k}}{x_{i_k}}$$

and

$$\tilde{t} = \frac{1}{200}\hat{t}_{pwr}$$

is an unbiased estimator of the average income per household $\sum_U y_k/200$. Note that

$$\hat{V}(\tilde{t}) = \frac{1}{200^2}\hat{V}(\hat{t}_{pwr}) = \frac{1}{200^2}\frac{1}{10}\frac{\sum_{k=1}^{10}[(y_{i_k}/p_{i_k}) - \hat{t}_{pwr}]^2}{10-1}$$

and then

$$\hat{V}(\tilde{t}) = \frac{1}{10}\frac{\sum_{k=1}^{10}[3(y_{i_k}/x_{i_k}) - \tilde{t})^2}{10-1}$$

With the above data, direct calculations yield that $\hat{t}_{pwr} = 3600,000$ and then

$$\tilde{t} = 18,000.$$

On the other hand

$$\hat{V}(\tilde{t}) = \frac{1}{10}\frac{\sum_{k=1}^{10}[3(y_{i_k}/x_{i_k}) - \tilde{t})^2}{10-1} = 4400,000$$

so that

$$cve(\tilde{t}) = \frac{(\hat{V}(\tilde{t}))^{1/2}}{\tilde{t}} = \frac{2,097.618}{18,000} = 0.1165343.$$

is the estimated coefficient of variation. □

Example 4.5.2. In the general with-replacement sampling of size m, show that the first and second order inclusion probabilities are

$$\pi_k = 1 - (1-p_k)^m$$

and

$$\pi_{jk} = 1 - (1-p_k)^m - (1-p_j)^m + (1-p_j-p_k)^m.$$

□

Solution. Recall that p_k is the probability of drawing unit k in any extraction. Thus, in m extractions the probability that the unit k is not present is $(1-p_k)^m$, and then

$$\pi_k = P[I_k = 1] = P[\text{Unit } k \text{ appears in the sample}] = 1 - (1-\pi_k)^m.$$

On the other hand,

$$\begin{aligned} P[I_k = 0 \text{ or } I_j = 0] &= P[I_k = 0] + P[I_j = 0] - P[I_j = 0 \text{ and } I_k = 0] \\ &= (1-\pi_k)^m + (1-p_j)^m - (1-p_j-p_k)^m, \end{aligned}$$

and then $\pi_{j,k} = P[I_j = 1 \text{ and } i_k = 1] = 1 - P[I_k = 0 \text{ or } I_j = 0]$, so that $\pi_{j,k} = 1 - (1-\pi_k)^m - (1-p_j)^m + (1-p_j-p_k)^m$. □

4.6. Multivariate Hypergeometric Distribution

Let the population $\mathcal{U}$ of size N be the union of k subpopulations $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_k$ of sizes $N_1, N_2, \ldots, N_k$. A simple random sample of size n is taken form $\mathcal{U}$ and X_i denotes the number of elements in the sample that belong to $\mathcal{U}_i$, $i = 1, 2, \ldots, k$. The distribution of the vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ is the Hypergeometric distribution $\mathcal{H}_k(n, N; N_1, N_2, \ldots, N_k)$ and is determined by

$$P[\mathbf{X} = (n_1, n_2, \ldots, n_k)] = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N}{n}} \tag{4.6.1}$$

where $n_1, n_2 \ldots n_k$ are nonnegative integers adding up to n. Note that

$$\sum_{\substack{n_1 \geq 0, n_2 \geq 0, \ldots, n_k \geq 0 \\ n_1 + n_2 + \cdots + n_k = n}} \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N}{n}} = 1 \tag{4.6.2}$$

The mean and variance matrix of $\mathbf{X}$ will be now determined: The identity

$$\binom{a}{b} = \frac{a}{b}\binom{a-1}{b-1}, \quad a \geq b > 0 \tag{4.6.3}$$

will be used (Dudewicz and Mishra, 2008).

(i) The compute $E[X_i]$ observe that, by symmetry, it is sufficient to find $E[X_1]$:

$$\begin{aligned}
E[X_1] &= \sum_{\substack{n_1 \geq 0, n_2 \geq 0, \ldots, n_k \geq 0 \\ n_1 + n_2 + \cdots + n_k = n}} n_1 \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N}{n}} \\
&= \sum_{\substack{n_1 > 0, n_2 \geq 0, \ldots, n_k \geq 0 \\ n_1 + n_2 + \cdots + n_k = n}} n_1 \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N}{n}} \\
&= n\frac{N_1}{N} \sum_{\substack{n_1 > 0, n_2 \geq 0, \ldots, n_k \geq 0 \\ n_1 + n_2 + \cdots + n_k = n}} \frac{\binom{N_1-1}{n_1-1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N-1}{n-1}} \\
&= n\frac{N_1}{N} \sum_{\substack{k_1 \geq 0, n_2 \geq 0, \ldots, n_k \geq 0 \\ k_1 + n_2 + \cdots + n_k = n-1}} \frac{\binom{N_1-1}{k_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N-1}{n-1}} \\
&= n\frac{N_1}{N}
\end{aligned}$$

where (4.6.3) was used to set the the third equality and (4.6.2) (with $N_1 - 1$ and $N - 1$ instead of N_1 and N, respectively) was used in the last step. Therefore,

$$E[X_i] = n\frac{N_i}{N}, \quad i = 1, 2, \ldots, N. \tag{4.6.4}$$

(ii) Now the expectation of $E[X_i(X_i - 1)]$ will be determined. As before, it is sufficient to consider the case $i = 1$.

$$\begin{aligned}
E[X_1(X_1-1)] &= \sum_{\substack{n_1\ge 0, n_2\ge 0,\ldots,n_k\ge 0\\ n_1+n_2+\cdots+n_k=n}} n_1(n_1-1)\frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N}{n}}\\
&= \sum_{\substack{n_1> 1, n_2\ge 0,\ldots,n_k\ge 0\\ n_1+n_2+\cdots+n_k=n}} n_1(n_1-1)\frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N}{n}}\\
&= n(n-1)\frac{N_1(N_1-1)}{N(N-1)} \sum_{\substack{n_1> 1, n_2\ge 0,\ldots,n_k\ge 0\\ n_1+n_2+\cdots+n_k=n}} \frac{\binom{N_1-2}{n_1-2}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N-2}{n-2}}\\
&= n(n-1)\frac{N_1(N_1-1)}{N(N-1)} \sum_{\substack{k_1\ge 0, n_2\ge 0,\ldots,n_k\ge 0\\ k_1+n_2+\cdots+n_k=n-2}} \frac{\binom{N_1-2}{k_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N-2}{n-2}}\\
&= n(n-1)\frac{N_1(N_1-1)}{N(N-1)}
\end{aligned}$$

where a double application of (4.6.3) lead to the third equality and (4.6.2) (with the appropriate parameters) was used in the last step. Therefore,

$$E[X_1^2] = E[X_1(X_1-1)] + E[X_1] = n(n-1)\frac{N_1(N_1-1)}{N(N-1)} + n\frac{N_1}{N}$$

and then

$$\begin{aligned}
\mathrm{Var}\,[X_1] &= E[X_1^2] - (E[X_1])^2\\
&= n(n-1)\frac{N_1(N_1-1)}{N(N-1)} + n\frac{N_1}{N} - \left(n\frac{N_1}{N}\right)^2\\
&= n\frac{N_1}{N}\left((n-1)\frac{N_1-1}{N-1} - n\frac{N_1}{N} + 1\right)\\
&= n\frac{N_1}{N}\left(\frac{(n-1)(N_1-1)N - nN_1(N-1) + (N-1)N}{N(N-1)}\right)\\
&= n\frac{N_1}{N}\frac{(N-n)(N-N_1)}{N(N-1)}\\
&= n\frac{N_1}{N}\left(1-\frac{N_1}{N}\right)\frac{N-n}{N-1}
\end{aligned}$$

Therefore,

$$\mathrm{Var}\,[X_i] = n\frac{N_i}{N}\left(1-\frac{N_i}{N}\right)\frac{N-n}{N-1}, \quad i = 1, 2, \ldots, k. \tag{4.6.5}$$

(iii) Finally, the covariance between X_i and X_j will be determined. As usual, it is sufficient to find

$\text{Cov}\,(X_1, X_2)$.

$$
\begin{aligned}
E[X_1X_2] &= \sum_{\substack{n_1\geq 0, n_2\geq 0,\ldots,n_k\geq 0\\ n_1+n_2+\cdots+n_k=n}} n_1n_2\frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N}{n}}\\
&= \sum_{\substack{n_1>0, n_2>0, n_3\geq 0\ldots,n_k\geq 0\\ n_1+n_2+\cdots+n_k=n}} n_1n_2\frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N}{n}}\\
&= n(n-1)\frac{N_1N_2}{N(N-1)}\sum_{\substack{n_1>0, n_2>0, n_3\geq 0,\ldots,n_k\geq 0\\ n_1+n_2+\cdots+n_k=n}} \frac{\binom{N_1-1}{n_1-1}\binom{N_2-1}{n_2-1}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N-2}{n-2}}\\
&= n(n-1)\frac{N_1N_2}{N(N-1)}\sum_{\substack{k_1\geq 0, k_2\geq 0, n_3\geq 0,\ldots,n_k\geq 0\\ k_1+n_2+\cdots+n_k=n-2}} \frac{\binom{N_1-1}{k_1}\binom{N_2-1}{k_2}\binom{N_3}{n_3}\cdots\binom{N_k}{n_k}}{\binom{N-2}{n-2}}\\
&= n(n-1)\frac{N_1N_2}{N(N-1)}
\end{aligned}
$$

where, as before, a double application of (4.6.3) lead to the third equality and (4.6.2) (with the appropriate parameters) was used to set the last equality. Thus,

$$\text{Cov}\,(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2] = n(n-1)\frac{N_1N_2}{N(N-1)} - n\frac{N_1}{N}n\frac{N_2}{N}$$

and then

$$\text{Cov}\,(X_1, X_2) = \frac{nN_1N_2}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) = \frac{nN_1N_2}{N}\left(\frac{n-N}{N(N-1)}\right)$$

so that

$$\text{Cov}\,(X_1, X_2) = -n\frac{N_1}{N}\frac{N_2}{N}\frac{N-n}{N-1} \tag{4.6.6}$$

The above discussion is summarized in the following theorem

Theorem 4.6.1. Suppose that $\mathbf{X} \sim \mathcal{H}_k(n, N; N_1, \ldots, N_k)$ is a random vector with the k-dimensional hypergeometric distribution; see (4.6.1). Set

$$p_i = \frac{N_1}{N}, \quad i = 1, 2, \ldots, k$$

so that $\sum_{i=1}^k p_i = 1$, and define the row vector $\mathbf{p}$ and the $k \times k$ matrix $\mathbf{V}$ by

$$\mathbf{p} := (p_1, p_2, \ldots, p_k), \tag{4.6.7}$$

and

$$\mathbf{V} := diag(\mathbf{p}) - \mathbf{p}'\mathbf{p}.$$

In this case

$$E[\mathbf{X}] = n\mathbf{p}, \qquad \text{and Var}\,[\mathbf{X}] = n(1-\tilde{f})\mathbf{V},$$

where

$$\tilde{f} = \frac{n-1}{N-1}$$

is (a form of) the finiteness correction term.

The assertions in this theorem follow directly combining (4.6.4) –(4.6.6) with (4.6.7). Observe the following interesting points:

(i) $\mathbf{p}$ and $\mathbf{V}$ are the mean and variance of a multidimensional Bernoulli random vector $\mathbf{Y}$ with parameter $\mathbf{p}$. Hence, $n\mathbf{p}$ and $n\mathbf{V}$ are the mean and variance of a vector with multinomial distribution $\mathcal{M}(n, \mathbf{p})$ with parameters n and $\mathbf{p}$.

(ii) As $n/N \to 0$, the correction term $\tilde{f}$ goes to 0, and then the variance of $\mathbf{X}$ approximates the variance of $\mathcal{M}(n, \mathbf{p})$. The reason for this convergence is that, as $\tilde{f}$ goes to 0, the hypergeometric distribution $\mathcal{H}_k(n, N; N_1, N_2, \ldots, N_k)$ approximates $\mathcal{M}(n, \mathbf{p})$.

(iii) If the vector $\mathbf{p}$ has been loaded in the R environment as `p`, then the matrix $\mathbf{V}$ is easily obtained using the code

```
V <- diag(p) - crossprod(rbind(p)).
```

4.7. The Bernoulli Sampling Design: Properties

The Bernoulli sampling design (BE) is implemented via the following draw sequential selection method: Let N be the population size and let $X_1, X_2, \ldots, X_N$ be N independent random variables with uniform distribution in $[0, 1)$. The units of the population are considered one by one from U_1 to U_N, and U_k is included in the sample if and only if $X_k < \pi$ where $\pi \in (0, 1)$ is a constant fixed before starting the selection process. Hence, the indicator function of the event [U_k belongs to the sample] is $I_k = I[X_k < \pi]$, so that

$$I_1, I_2, \ldots, I_N \text{ are independent and identically distributed,}$$

and

$$\pi_k = P[I_k = 1] = \pi = 1 - P[I_k = 0], \quad \pi_{j,k} = P[I_k = 1, I_j = 1] = \pi^2, \quad j \neq k.$$

Thus, $E[I_k] = \pi$, $\mathrm{Var}\,[I_k] = \pi(1-\pi)$ and $\mathrm{Cov}\,(I_j, I_k) = 0$ when $j \neq k$. The π-expanded unbiased estimator of the total is

$$\hat{t} = \sum_S \check{y}_k = \sum_U \check{y}_k \, I_k = \frac{1}{\pi} \sum_U y_k \, I_k \text{ where } \check{y}_k = \frac{y_k}{\pi_k} = \frac{y_k}{\pi}. \tag{4.7.1}$$

Theorem 4.7.1. (i) $\mathrm{Var}\left[\hat{t}\right] = \left(\frac{1}{\pi} - 1\right)\sum_U y_k^2$.

(ii) $\hat{V}(\hat{t}) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_s y_k^2$ is an unbiased estimator of $\mathrm{Var}\left[\hat{t}\right]$.

Proof. Since the variables I_k are independent and identically distributed with common *Bernoulli* (π) distribution it follows that

$$\mathrm{Var}\left[\hat{t}\right] = \mathrm{Var}\left[\sum_U \check{y}_k I_k\right] = \sum_U \pi(1-\pi)\check{y}_k^2 = \left(\frac{1}{\pi} - 1\right)\sum_U y_k^2.$$

Observe that $\mathrm{Var}\left[\hat{t}\right]$ is the population total for the variable

$$w_k = \left(\frac{1}{\pi} - 1\right) y_k^2,$$

which admits the following (π-expanded) unbiased estimator

$$\hat{V}(\hat{t}) = \sum_s \check{w}_i = \sum_s \frac{w_i}{\pi} = \left(\frac{1}{\pi} - 1\right)\sum_s \frac{y_k^2}{\pi} = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_s y_k^2,$$

completing the argument. □

Remark 4.7.1. A remarkable fact of the formula in Theorem 4.7.1 is that the variance of $\hat{t}$ is a *positive definite quadratic form*, in contrast with other sampling designs where the variance for $\hat{t}$ is a quadratic vanishing on the the space of constant vectors. Now set

$$n = N\pi$$

(the expected sample size) and note that

$$\mathrm{Var}\left[\hat{t}\right] = \left(\frac{N}{n} - 1\right)\sum_U y_k^2 = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N}\sum_U y_k^2.$$

Combining this relation with $\sum_U y_k^2 = \sum_U (y_k - \bar{Y})^k + N\bar{Y}^2 = (N-1)S_{yU}^2 + N\bar{Y}^2$, it follows that

$$\begin{aligned} \mathrm{Var}\left[\hat{t}\right] &= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N}\left((N-1)S_{yU}^2 + N\bar{Y}^2\right) \\ &= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)S_{yU}^2\left(1 - \frac{1}{N} + CV_{yU}^{-2}\right). \end{aligned} \tag{4.7.2}$$

For the *SI* design with sample size (approximately) n the variance of $\hat{t}$ is

$$\mathrm{Var}_{SI}(\hat{t}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)S_{yU}^2$$

and the efficiency of the *SI* plan with respect to the *BE* design, is

$$\frac{\mathrm{Var}\left[\hat{t}\right]}{\mathrm{Var}_{SI}(\hat{t})} = 1 - \frac{1}{N} + CV_{yU}^{-2}$$

Thus, essentially, the SI plan is always more efficient than the BE plan, and is substantially better when CV_{yU} is “small’. □

An alternative estimator under BE is given by

$$\hat{t}_{\text{alt}} = N\frac{1}{n_S}\sum_S y_k \text{ if } n_S \neq 0, \quad \hat{t}_{\text{alt}} = 0 \quad \text{ if } n_S = 0. \tag{4.7.3}$$

Recall that given $n_S = k$ the sample S is uniformly distributed on the samples of size k (as if S has been selected under the SI design). Thus, on the event $n_S > 0$,

$$\begin{aligned} E[\hat{t}_{\text{alt}}|n_S] &= N\,E\left[\frac{1}{n_S}\sum_S y_k\right] = N\bar{Y} = Y \\ \text{Var}\left[\hat{t}_{\text{alt}}|n_S\right] &= N^2\left(\frac{1}{n_S} - \frac{1}{N}\right)S_{yU}^2 \end{aligned} \tag{4.7.4}$$

Next, observe that

$$\begin{aligned} E\left[\left(\frac{1}{n_S} - \frac{1}{\pi N}\right)^2 \middle| n_S > 0\right] &= E\left[\left(\frac{n_S - N\pi}{n_S N\pi}\right)^2 \middle| n_S > 0\right] \\ &\leq E\left[\left(\frac{n_S - N\pi}{N\pi}\right)^2 \middle| n_S > 0\right] \\ &\leq \frac{1}{N^2\pi^2}E[(n_S - N\pi)^2|n_S > 0] \\ &= \frac{1}{N^2\pi^2}N\pi(1-\pi) \\ &\leq \frac{1-\pi}{\pi N P[n_S > 0]}. \end{aligned}$$

Setting $n_S = 1$ when S is empty, it follows that

$$\begin{aligned} E\left[\left|\frac{1}{n_S} - \frac{1}{\pi N}\right|\right] &\leq E\left[\left(\frac{1}{n_S} - \frac{1}{\pi N}\right)^2 \middle| n_S > 0\right]^{1/2} + \left(1 - \frac{1}{N\pi}\right)P[n_S = 0] \\ &\leq \left(\frac{1-\pi}{\pi N(1-(1-\pi)^N)}\right)^{1/2} + \left(1 - \frac{1}{N\pi}\right)(1-\pi)^N \end{aligned}$$

Thus, if

$$N \text{ is large and } (1-\pi)^N \approx 0 \tag{4.7.5}$$

then

$$E\left[\left|\frac{1}{n_S} - \frac{1}{\pi N}\right|\right] \approx 0$$

and then

$$E\left[\frac{1}{n_S}\right] \approx \frac{1}{\pi N}.$$

Combining this fact with (4.7.4) it follows from the formula for the variance in terms of the conditional expectation and variance that, under (4.7.5),

$$\begin{aligned}\text{Var}\left[\hat{t}_{\text{alt}}\right] &= \text{Var}\left[E[\hat{t}_{\text{alt}}|n_S]\right] + E[\text{Var}\left[\hat{t}_{\text{alt}}|n_S\right]] \\ &\approx N^2\left(E\left[\frac{1}{n_S}\right] - \frac{1}{N}\right)S^2_{yU} \\ &= N^2\left(\frac{1}{N\pi} - \frac{1}{N}\right)S^2_{yU} \\ &= N^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2_{yU},\end{aligned}$$

where $n = E[n_S] = N\pi$.

Theorem 4.7.2. Under the *BE* design, let $\hat{t}_{\text{alt}}$ be the estimator of the total Y defined in (4.7.3). With this notation, in the context of condition (4.7.5),

$$E[\hat{t}_{\text{alt}}] \approx Y \quad \text{and} \quad \text{Var}\left[\hat{t}_{\text{alt}}\right] \approx N^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2_{yU},$$

where $n = N\pi$ is the expected sample size. Consequently, the efficiency of $\hat{t}_{\text{alt}}$ with respect to π-expanded estimator $\hat{t}$ is

$$\frac{\text{Var}\left[\hat{t}\right]}{\text{Var}\left[\hat{t}_{\text{alt}}\right]} \approx 1 - \frac{1}{N} + CV^{-2}_{yU}.$$

Example 4.7.1. In a population of size $N = 1000$ a *BE* sample with $\pi = 0.40$ is selected. The observed sample size was $n_s = 300$ and he variable of interest is $y_i = 0$ or $y_i = 1$ for every i. It was observed that $\sum_s y_k = 200$. In this case

$$\hat{t} = \frac{1}{\pi}\sum_s y_k = 2.5(200) = 500, \quad \hat{V}(\hat{t}) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_s y_k^2 = 750.$$

A confidence interval with approximate confidence level of 95% is

$$\hat{t} \pm 1.96\sqrt{\hat{V}(\hat{t})} = 500 \pm 1.96\sqrt{750} = 500 \pm 53.7.$$

On the other hand, the estimator $\hat{t}_{\text{alt}}$ is given by

$$\hat{t}_{\text{alt}} = 1000\frac{1}{n_S}\sum_S y_k = 1000\frac{200}{300} = 666.66$$

and an (approximately unbiased) estimator of S^2_{yU} is

$$S^2_{yS} = \frac{1}{300-1}\left(\sum_s y_k^2 - n_s\bar{y}_s^2\right) = \frac{1}{300-1}(200 - 300(2/3)^2) = 0.2229654$$

and then

$$\hat{V}(\hat{t}_{\text{alt}})) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2_{yS} = 1000^2\left(\frac{1}{400} - \frac{1}{1000}\right)0.2229654 = 334.5$$

The (normal approximation) 95% confidence interval based on $\hat{t}_{\text{alt}}$ is

$$\hat{t}_{\text{alt}} \pm 1.96\sqrt{\hat{V}(\hat{t})} = 666.66 \pm 1.96 * \sqrt{334.5} = 666.66 \pm 35.8.$$

□

References

[1]. T. M. Apostol (1980), Mathematical Analysis, *Addison Wesley*, Reading, MA

[2]. A. A. Borovkov (1999), Mathematical Statistics, *Gordon and Breach*, New York

[3]. E. Dudewicz y S. Mishra (1998). Mathematical Statistics, *Wiley*, New York.

[4]. W. Fulks (1980), Cálculo Avanzado, *Limusa*, México, D. F.

[5]. F. A. Graybill (2000), Theory and Application of the Linear Model, *Duxbury*, New York.

[6]. F. A. Graybill (2001), Matrices with Applications in Statistics *Duxbury*, New York.

[7]. D. A. Harville (2008), Matrix Algebra Form a Statistician's Perspective, *Springer-Verlag*, New York.

[8]. M. H. Hansen, W. N. Hurwitz, W. G. Madow (2003), Sample Survey Methods and Theory : Volume I, Methods and Applications, Wiley (Classics Library), New York

[9]. A. I. Khuri (2002), Advanced Calculus with Applications in Statistics, *Wiley*, New York.

[10]. E. L. Lehmann and G. B. Casella, (1998), Theory of Point Estimation, Springer, New York.

[11]. E. L. Lehmann and J. P. Romano, (1999), Testing Statistical Hypothesis, Springer, New York.

[12]. M. Loève (1984), Probability Theory, I, Springer-Verlag, New York.

[13]. D. C. Montgomery (2011), Introduction to Statistical Quality Control, 6th Edition, Wiley, New York.

[14]. S. Lohr (2010), Sampling: Design and Analysis 2nd Ed. Cengage, San Francisco.

[15]. A. M. Mood, D. C. Boes and F. A. Graybill (1984), Introduction to the Theory of Statistics, McGraw-Hill, New York.

[16]. W. Rudin (1984), Real and Complex Analysis, *McGraw-Hill*, New York.

[17]. H. L. Royden (2003), Real Analysis, *MacMillan*, London.

[18]. J. Shao (2010), Mathematical Statistics, *Springer*, New York.

[19]. C. E. Sarndal, B. Swensson, J. Wretman (1992), Model Assisted Survey Sampling, Springer-Verlag, New York.

[20]. H. Tucker (2002), Mathenatical Methods in Sample Surveys, Word Scientific, Singapore.

[21]. D. Wackerly, W. Mendenhall y R. L. Scheaffer (2009), Mathematical Statistics with Applications, *Prentice-Hall*, New York.