CHAPMAN & HALL/CRC
APPLIED ENVIRONMENTAL STATISTICS

DANIEL MANDALLAZ

# SAMPLING TECHNIQUES FOR FOREST INVENTORIES

$$\mathbb{V}(\hat{Y}) = \frac{1}{n_2 \mathbb{E}_x M(x)} V_s \left(1 + \rho\left(\mathbb{E}_x M(x) - 1\right) + \rho \frac{\mathbb{V}_x M(x)}{\mathbb{E}_x M(x)}\right)$$

$$\mathbb{E}_\omega \mathbb{V}(\hat{Y}) = \frac{1}{n_2 \lambda^2(F)} \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} + \frac{1}{n_2} \beta_1^2$$

Chapman & Hall/CRC
Taylor & Francis Group

CHAPMAN & HALL/CRC
APPLIED ENVIRONMENTAL STATISTICS

# Sampling Techniques for Forest Inventories

# CHAPMAN & HALL/CRC APPLIED ENVIRONMENTAL STATISTICS

Series Editor
Richard Smith, Ph.D.

## Published Titles

*Timothy G. Gregoire and Harry T. Valentine*, Sampling Strategies for Natural Resources and the Environment

*Steven P. Millard and Nagaraj K. Neerchal*, Environmental Statistics with S Plus

*Michael E. Ginevan and Douglas E. Splitstone*, Statistical Tools for Environmental Quality

*Daniel Mandallaz*, Sampling Techniques for Forest Inventory

## Forthcoming Titles

*Thomas C. Edwards and Richard R. Cutler*, Analysis of Ecological Data Using R

*Bryan F. J. Manly*, Statistics for Environmental Science and Management, 2nd Edition

*Song S. Qian*, Environmental and Ecological Statistics with R

CHAPMAN & HALL/CRC
APPLIED ENVIRONMENTAL STATISTICS

# Sampling Techniques for Forest Inventories

DANIEL MANDALLAZ
ETH Zurich, Department of Environmental Sciences

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an **informa** business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742


Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-58488-976-2 (Hardcover)



**Library of Congress Cataloging-in-Publication Data**

Mandallaz, Daniel.
Sampling techniques for forest inventories / Daniel Mandallaz.
p. cm. -- (Applied environmental statistics ; 4)
Includes bibliographical references and index.
ISBN 978-1-58488-976-2 (alk. paper)
1. Forest surveys. I. Title. II. Series.

SD387.S86M36 2007
634.9'285--dc22 2007028682

**Visit the Taylor & Francis Web site at**
**http://www.taylorandfrancis.com**

**and the CRC Press Web site at**
**http://www.crcpress.com**

**A Léa Marine**

**Le hasard fait bien les choses**
French adage and statistical paradigm

# Contents

# Preface

Inventories are the bases for forest management planning, with the goal being the optimal utilization of resources under given constraints. To accomplish this, managers must collect, summarize, and interpret information – that is perform statistical work. The development and improvement of forest management practices, which began toward the end of the Middle Ages, have strongly depended on the parallel evolution of inventory techniques and statistical methodology, in particular sampling schemes. Without these, current forest inventories would be impossible to conduct.

Over the last 80 years, the number of techniques, the demand for more and better information, and finally the mere complexity of their incumbent investigations seem to have grown exponentially. Furthermore, the increased importance of related problems in landscape research and ecology (keywords e.g. biomass, carbon sequestration, and bio-diversity) as well as their interactions with the sociological and economic environment have required specialized procedures for data collection and statistical inference. However, their accompanying economic constraints have necessitated cost-efficient approaches in performing all of these tasks.

The objective of this textbook is to provide graduate students and professionals with the up-to-date statistical concepts and tools needed to conduct a modern forest inventory. **This exposition is as general and concise as possible**. Emphasis has been placed deliberately on the mathematical-statistical features of forest sampling to assess classical dendrometrical quantities. It is assumed that the reader has a sufficient understanding of elementary probability theory, statistics, and linear algebra. More precisely, one must be able to calculate unconditional and conditional probabilities and understand the concepts of random variables, distributions, expectations, variances (including their conditional versions as derived and summarized in Appendix B), central limit theorem and confidence intervals, as well as utilize the least-squares estimation technique in linear models (using matrix notation). The standard notation of naive set theory (e.g. $A \cup B$, $A \cap B$, $A \setminus B$, $A \subset B$, $A \supset B$, $x \in A$, $A \ni x$, $x \notin A$) is presented throughout. Likewise, the reader will ideally have some prior knowledge of the general economic-political background of forest inventories and aspects of mensuration (e.g. the handling of instruments), plus skills in remote sensing and geographical information systems (GIS). MSc and PhD students in Forestry, and particularly in Forest Management, will almost

surely have had introductory courses in all of these topics. This book will also be useful to experienced forest biometricians who wish to become rapidly acquainted with a modern approach to sampling theory for inventories, as well as some recent developments not yet available in book form.

The fundamental concepts and techniques, as used primarily in sociological and economics studies, are presented in chapter 2 and can be summarized as **design-based survey sampling and inference for finite populations** (of e.g. geographical areas, enterprises, households, farms, employees or students), usually so large that a full survey (census) is neither feasible nor even meaningful. **Inclusion probabilities** and the **Horvitz-Thompson estimator** form the cornerstone of this chapter and are also essential to a forest inventory. More advanced topics are addressed in chapter 3. Excellent classical works at the intermediate mathematical level include those by Cochran (1977) and Särndal et al. (2003), and in French by Gourieroux (1981) and Tillé (2001). Likewise, Cassel et al. (1977), Chaudhuri and Stenger (1992), Chaudhuri and Vos (1988), and Tillé (2006) describe more complicated mathematical and statistical themes.

Key references (in English) for sampling theory in forest inventories are from de Vries (1986) and Schreuder et al. (1993). Those compiled by Kangas and Maltamo (Eds, 2006) and Köhl et al. (2006), the latter containing an extensive bibliography, give broad and up-to-date introductions to this subject, but without proofs of the exhibited statistical techniques. Johnson (2000) provides an elementary and encyclopedic (900 pp!) review of standard procedures, while the writing of Gregoire and Valentine (2007) is an excellent introduction to modern concepts in sampling strategies with interesting chapters on some specific problems. Pardé and Bouchon (1988) and Rondeux (1993), both writing in French, as well as Zöhrer (1980), in German, present basic overviews with emphases on practical work. Unfortunately, none of these authors, except Gregoire and Valentine (2007) in some instances, utilize the so-called **infinite population or Monte Carlo approach** that is much better suited to forest inventories and, in many ways easier to understand. Therefore, this formalism for inventories, within a **design-based** framework, is developed here in chapter 4 (foundations and one-phase sampling schemes) and in chapter 5 (two-phase sampling schemes). It rests upon the concept of **local density**, which is essentially an adaptation of the Horvitz-Thompson estimator. These two chapters give a full treatment of **one-phase** and **two-phase** sampling schemes at the point (plot) level, under both **simple random sampling** and **cluster random sampling**, with either **one-stage** or **two-stage** selection procedures at the tree level. These techniques usually suffice for most routine inventories or serve as building blocks for more complex ones. The treatment of cluster-sampling differs markedly from the classical setup, being simpler and easier from both a theoretical and a practical point of view. Simulations performed on a small real forest with full census illustrate the techniques discussed in chapters 4 and 5. Those results are then displayed and critiqued in

Appendix A. More advanced topics, such as **model-dependent** inference and its interplay with **model-assisted** techniques (g-weights), as well as **small area estimations** and **analytical studies**, are dealt with in chapter 6. **Geostatistics** and the associated **Kriging** procedures are presented in chapter 7. Using a case study, chapter 8 describes various estimation procedures. Chapter 9 tackles the difficult problem of **optimal design** for forest inventories from a modern point of view relying on the concept of anticipated variance. The resulting optimal schemes are illustrated in chapter 10 with data from the Swiss National Forest Inventory. Chapter 11 outlines the essential facts pertaining to the estimation of growth and change. Finally, chapter 12 provides a short introduction to **transect-sampling** based on the stereological approach. A small number of exercises are also proposed in selected chapters.

It is worth mentioning that the formalism developed in chapters 4, 5 and 6 can be used to estimate the integral of a function over a spatial domain – a key problem in such fields as soil physics, mining or petrology. This is a simpler alternative to the geostatistical techniques developed in chapter 7, which are usually more efficient, particularly for local estimations.

This book is based partly on the writings of C.E. Särndal, as adapted to the context of a forest inventory. In addition, references are made to research, both recent and older, by outstanding forest inventorists, including B. Matérn and T.G. Gregoire, as well as to the author's own work and lectures at ETH Zurich. Whenever feasible, proofs are given, in contrast to most books on the subject. These occasionally rely on heuristic arguments to minimize the amount of mathematics to a reasonable level of sophistication and spacing. It cannot be overemphasized that readers should not only have a good command of definitions and concepts but also have at least a sufficient understanding of the proofs for the main results.

The scope of this book is restricted when compared to the seemingly unlimited field of applications for sampling techniques within environmental and sociological-economic realms. Nevertheless, the average reader will need time and endurance to master all of the topics covered. Many sections are therefore intended for either further reading or specific applications on an as-needed basis, or they will facilitate one's access to more specialized references. Readers who desire to familiarize themselves quickly with the key aspects of a forest inventory can in a first perusal focus their attention on the following topics: chapters 1, 2 (sections 2.1 to 2.6), 4 and 5, plus a brief glance at the case study in chapter 8. This should suffice for tackling standard estimation problems (without the planning aspects). Courageous readers who persevere through this entire tome should be able to consult all of the current literature on forest sampling (and partly on general survey sampling) and, why not, eventually contribute their own solutions to existing and oncoming challenges?

# Acknowledgments

I would like to express my thanks to Professor em. P. Bachmann, former chair of Forest Inventory and Planning at ETH, and to Professor H.R. Heinimann, chair of Land Use Engineering at ETH, for their continuous support as well as for the working environment they have succeeded in creating. Thanks are also due to Professor H.R. Künsch, Department of Mathematics at ETH, for scrutinizing some mathematical aspects as well as to Dr. A. Lanz and E. Kaufmann from the FSL Institute in Birmensdorf, to Dr. R. Ye from the Chinese Academy of Forest Sciences in Beijing and to H.P. Caprez at ETH, all for their technical support, and to Priscilla Licht for editing the manuscript. Last but not least, I thank Professor T.G. Gregoire, Yale University, for encouraging me to write this book, and the publisher for his assistance and patience.

CHAPTER 1

# Introduction and terminology

We now proceed to define the terminology and notation that will be used throughout this work. A particular population $\mathcal{P}$ of $N$ individuals (sometimes also called elements or units) $\{u_1, u_2, \ldots u_n\}$ are identified by their labels $i = 1, 2, \ldots N$. $\mathcal{P}$ may consist of all the students at ETH, of all the trees of Switzerland (where, in this case, $N$ is unknown), of all the employees older than 18 years on August 1st 2007 in Switzerland. In the set theoretical sense **it must be clear whether something belongs to the population $\mathcal{P}$ or not**. Surprisingly, this seemingly simple requirement can be the source of great problems in applications (what is a tree, an unemployed person, etc. ?). Defining the population under study is a key task at the planning stage, often requiring intensive discussions and frustrating compromises, a matter we shall not discuss any further in this book. For each individual $i$ in $\mathcal{P}$ one is interested in $p$ response variables with numerical values $Y_i^{(m)}, m = 1, \ldots p, i = 1 \ldots N$, which can be measured at a given time point in an error-free manner. Note that any qualitative variable can always be coded numerically with a set of 0/1 indicator variables. Whenever ambiguity is excluded we shall drop the upper index that identifies the response variable. An error-free assumption can be problematic even when dealing with physical quantities (e.g. the volume of a tree) and can also be a source of great difficulties in the case of non-response during interviews. Usually the quantities of primary interest are population totals, means and variances. That is

$$Y^{(m)} = \sum_{i=1}^{N} Y_i^{(m)} \tag{1.1}$$

$$\bar{Y}^{(m)} = \frac{Y^{(m)}}{N} \tag{1.2}$$

$$S^2_{Y^{(m)}} = \frac{\sum_{i=1}^{N} (Y_i^{(m)} - \bar{Y}^{(m)})^2}{N-1} \tag{1.3}$$

Sometimes, more complicated statistical characteristics of the population are needed, such as ratios, covariances, or correlations

$$R_{l,m} = \frac{Y^{(m)}}{Y^{(l)}} \tag{1.4}$$

$$C_{l,m} = \frac{\sum_{i=1}^{N}(Y_i^{(l)} - \bar{Y}^{(l)})(Y_i^{(m)} - \bar{Y}^{(m)})}{N-1} \tag{1.5}$$

$$\rho_{l,m} = \frac{C_{l,m}}{\sqrt{S^2_{Y^{(l)}} S^2_{Y^{(m)}}}} \tag{1.6}$$

In any case, the estimation of totals will be a key issue. In pursuing a forest inventory the spatial mean of additive quantities is frequently more important than the population total. Suppose that a forested area $F$ with a surface area $\lambda(F)$ in ha contains a well-defined population of $N$ trees. Moreover, say that all trees have at least a $12cm$ diameter at $1.3m$ above the ground (diameter at breast height, or $DBH$) and that the response variables of interest are $Y_i^{(1)} \equiv 1$ and $Y_i^{(2)}$ =volume in $m^3$. Then the spatial mean $\bar{Y}_s^{(m)} = \frac{Y^{(m)}}{\lambda(F)}$ represents the number of stems per ha ($m = 1$) and the volume per ha ($m = 2$) respectively. Note that $N$ is usually unknown and will have to be estimated via the variable $Y_i^{(1)}$. Likewise, the mean volume per tree can be obtained by estimating the ratio $R_{2,1}$.

In practice $N$ can be very large, making a complete evaluation of the entire population impossible (and usually not even meaningful, not to mention the illusion of almost unlimited resources). For this reason, one must restrict one's investigation to a subset $s$ of the population $\mathcal{P}$, also known as a "sample". An element $u \in s$ is then called a sampling unit. The problem is to draw conclusions for the entire population based solely on the sample $s$. The next question is how to choose that sample. Essentially two ways are possible: by expert judgement (purposive sampling) or by some random mechanism. The criterion used here is that the sample should be representative of the population (this is of course rather vague because if one considers a sample to be representative, one presumably knows roughly what the population looks like already). It is now widely (but not universally!) accepted that representativeness can be insured only by introducing at least partial random selection. Therefore, in the next chapter we shall define and analyze some of the most important sampling schemes (i.e. procedures for sample selection) as well as the estimation techniques that allow us to make inferences from the sample at hand to the entire population.

CHAPTER 2

# Sampling finite populations: the essentials

## 2.1 Sampling schemes and inclusion probabilities

Here we consider a population $\mathcal{P}$ of $N$ individuals and their associated response variables $Y_i^{(m)}$. A sampling scheme is a procedure that involves one or more random mechanisms to select a subset $s \subset \mathcal{P}$ of the population, i.e. the sample. The set of all possible samples s is denoted by $\mathcal{S}$, which is a subset of the set of all subsets (the power set) of $\mathcal{P}$. A well-known example might be a lottery machine that may choose 6 balls out of 45. In that case the set $\mathcal{S}$ consists of the $\binom{45}{6}$ potential outcomes, which are all equally possible with a probability $\binom{45}{6}^{-1}$. In a survey one usually needs a **sampling frame**, i.e. a list of all individuals in the population, which are identified by a key in the data base (e.g. the social security number of Swiss residents). In this book the identifying key is an integer number called the **label** and is simply denoted by $i = 1, 2 \ldots N$. Again, a forest inventory is peculiar in that no such list can exist, but this difficulty can be circumvented as we shall see. Using pseudo-random numbers (e.g. generated by a computer program and not by a physical mechanism of some kind) one can draw, in most instances sequentially, the individuals forming the sample. At this point it is not necessary to describe the practical implementation of such schemes; these will be discussed later.

We introduce the **indicator variables** $I_i$ which for each individual informs us whether it belongs to the sample or not:

$$I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

The probability that the sample $s$ will be selected is denoted by $p(s)$. We emphasize the fact that in this setup the same individual may be drawn many times (i.e. sampling with replacement) and the number of distinct individuals in that sample also is generally a random variable. The order in which individuals are drawn can be important. Therefore, the set theoretical interpretation of the sample $s$ is not quite appropriate in the general case, see Cassel et al. (1977) for details. The probability that a given individual will be included in

the sample is then given as

$$\pi_i = \mathbb{P}(I_i = 1) = \mathbb{E}(I_i) = \sum_{s \ni i} p(s) \tag{2.2}$$

The symbols $\mathbb{P}$ and $\mathbb{E}$ denote probability and expectation with respect to the sampling schemes. Note that the inclusion probabilities are not assumed to be constant over all individuals and that the most efficient procedures precisely rest upon unequal $\pi_i$. The number $n_s = \sum_{i=1}^{N} I_i$ of **distinct** elements in the sample satisfies the relationships given in the next two theorems

**Theorem 2.1.1.** *The effective sample size $n = \mathbb{E}n_s$ is given as*

$$\mathbb{E}(n_s) = \mathbb{E}\left(\sum_{i=1}^{N} I_i\right) = \sum_{i=1}^{N} \pi_i = n$$

Calculating the variance requires knowledge of the so-called pair-wise inclusion probabilities defined according to

$$\pi_{ij} = \mathbb{P}(I_i = 1, I_j = 1) = \mathbb{E}(I_i I_j) \tag{2.3}$$

Note that $\pi_{ii} = \pi_i$. The variances and covariances of indicator variables are

$$\mathbb{V}(I_i) = \pi_i(1 - \pi_i) = \Delta_{ii}, \;\; \mathbb{COV}(I_i, I_j) = \pi_{ij} - \pi_i \pi_j = \Delta_{ij} \tag{2.4}$$

The following properties are important:

**Theorem 2.1.2.** *The pair-wise inclusion probabilities satisfy the relationships*

$$\mathbb{E}\left(n_s(n_s - 1)\right) = \sum_{i,j \in \mathcal{P}, i \neq j} \pi_{ij}$$

*and*

$$\sum_{j \in \mathcal{P}, j \neq i} \pi_{ij} = \pi_i \Big( \mathbb{E}(n_s \mid I_i = 1) - 1 \Big)$$

*In particular for a fixed sample size $n_s \equiv n$ one has*

$$\sum_{j \in \mathcal{P}, j \neq i} \pi_{ij} = \pi_i (n - 1)$$

The first equality follows by calculating $\mathbb{E}n_s^2$ and using Theorem 2.1.1, the second equality by noting that

$$\mathbb{E}(n_s \mid I_i = 1) = 1 + \sum_{j, j \neq i} \mathbb{E}(I_j \mid I_i = 1) = 1 + \sum_{j, j \neq i} \frac{\pi_{ij}}{\pi_i}$$

## 2.2 The Horvitz-Thompson estimator

We can now define what is probably the most important estimator used in sampling theory, introduced in 1952 by D.G. Horvitz and D.J Thompson, now

simply called the Horvitz-Thompson (HT) estimator or also the $\pi$-estimator:

$$\hat{Y}_{\pi}^{(m)} = \sum_{i \in s} \frac{Y_i^{(m)}}{\pi_i} = \sum_{i \in \mathcal{P}} \frac{I_i Y_i^{(m)}}{\pi_i} \tag{2.5}$$

An estimator $\hat{T}(s)$ is considered **unbiased** for a population quantity $\theta$ (mean, total, variance, etc.) if its expected value under the random mechanism that generates the samples $s$ is equal to the true value $\theta$, that is if $\mathbb{E}_s T(s) = \sum_s p(s)T(s) = \theta$. The HT estimator then yields an unbiased point estimate of the population total as long as $\pi_i > 0$ for all $i$

$$\mathbb{E}(\hat{Y}_{\pi}^{(m)}) = Y^{(m)} \tag{2.6}$$

The proof is immediate because according to Eq. 2.2 the $\mathbb{E}(I_i)$ and the $\pi_i$ cancel each other. Note that $\hat{Y}_{\pi}^{(m)}$ is a random variable because the indicator variables $I_i$ are random. In this model the response variables $Y_i^{(m)}$ are fixed. This is the so-called **design-based approach**. Under hypothetical repeated sampling we know that the point estimates will be distributed around the true unknown value of the population total in such a way that the expected value of the point estimates is precisely the quantity we want to predict. We can say that, in some sense, the randomization procedure allows us to draw conclusions from the observed values in the available sample and apply them to the unobserved values of the remaining individuals of the population under study. Again, we can drop the upper index $(m)$. To calculate the variance we recall the simple fact that for random variables $X_i$ and real numbers $a_i$ one has

$$\mathbb{V}(\sum_i a_i X_i) = \sum_i a_i^2 \mathbb{V}(X_i) + \sum_{i \neq j} a_i a_j \mathbb{COV}(X_i, X_j)$$

Using 2.4 we obtain for the theoretical variance:

**Theorem 2.2.1.**

$$\mathbb{V}(\hat{Y}_{\pi}) = \sum_{i=1}^{N} \frac{Y_i^2(1-\pi_i)}{\pi_i} + \sum_{i=1, j=1, i \neq j}^{N} \frac{Y_i Y_j (\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j}$$

In practice one also needs an estimate of the variance. To do so let us note that the first sum in 2.2.1 can be predicted by considering the new variable $\frac{Y_i^2(1-\pi_i)}{\pi_i}$, and then estimating it by HT with the $\pi_i^{-1}$ weights. Likewise, for the second sum, we estimate by HT over the population of all pairs $i \neq j$ with the weights $\pi_{ij}^{-1}$. Hence, the following is an unbiased point estimate of the theoretical variance, provided that $\pi_{ij} > 0$ for all $i, j \in \mathcal{P}$

**Theorem 2.2.2.**

$$\hat{\mathbb{V}}(\hat{Y}_{\pi}) = \sum_{i=1}^{N} \frac{I_i Y_i^2(1-\pi_i)}{\pi_i^2} + \sum_{i=1, j=1, i \neq j}^{N} \frac{I_i I_j Y_i Y_j (\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}}$$

**The condition** $\pi_{ij} > 0 \ \forall \ i, j \in \mathcal{P}$ **is crucial**. Of course $\pi_{ij} > 0 \ \forall \ i, j \in s$

but this does not suffice. In systematic sampling this condition is violated and Theorem 2.2.2 can be totally misleading.

We introduce the following notation and terms which occur repeatedly in formulae: $\check{Y}_i = \frac{Y_i}{\pi_i}$ which is called the expanded value. Likewise, we define $\check{\Delta}_{ij} = \frac{\Delta_{ij}}{\pi_{ij}}$. Then one can rewrite

$$\hat{Y}_\pi = \sum_{i \in s} \check{Y}_i = \sum_{i \in \mathcal{P}} I_i \check{Y}_i \tag{2.7}$$

$$\mathbb{V}(\hat{Y}_\pi) = \sum_{i,j \in \mathcal{P}} \check{Y}_i \check{Y}_j \Delta_{ij} \tag{2.8}$$

$$\hat{\mathbb{V}}(\hat{Y}_\pi) = \sum_{i,j \in s} \check{Y}_i \check{Y}_j \check{\Delta}_{ij} \tag{2.9}$$

Although the above general formulae are useful for theoretical considerations, calculating those double sums can be prohibitive in practice. Instead, one usually obtains, as we shall see, computationally simpler expressions for specific sampling schemes.

We say that an estimator $\hat{T}(s)$ of the population parameter $\theta$ is **consistent** if its expected mean square error $\mathbb{E}_s(\hat{T}(s) - \theta)^2$ tends to zero as the sample size increases. Let us note that the mere concept of asymptotic is rather difficult to define in finite populations. We must consider an increasing sequence of population and samples, i.e. $N,\ n_s \to \infty$ (for a short introduction see Särndal et al. (2003)). In practice, this means large samples in much greater populations.

To obtain a $1 - \alpha$ confidence interval one can rely for large samples on the central limit theorem:

$$CI_{1-\alpha}(\hat{Y}_\pi) = \left[ \hat{Y}_\pi - z_{1-\alpha} \sqrt{\hat{\mathbb{V}}(\hat{Y}_\pi)}\ ,\ \hat{Y}_\pi + z_{1-\alpha} \sqrt{\hat{\mathbb{V}}(\hat{Y}_\pi)} \right] \tag{2.10}$$

where $z_{1-\alpha}$ is the two-sided $1-\alpha$ quantile of the standard normal distribution. Recall that, for example, $\alpha = 0.05$, i.e. 95 percent confidence intervals, $z_{1-\alpha} = 1.96 \approx 2$.

Under hypothetical repeated sampling 95 percent of these random intervals will contain the true unknown total. Note that **this does not mean** that the true value has a 95% chance of lying within the confidence interval calculated with the survey data, because the true value is either in or out. Although we do not know which alternative is correct, we have a statistical certainty. Out of the thousands of surveys conducted each year, roughly 95% of them will give a confidence interval containing the true unknown total (if the job has been done properly!) but we will not know for which surveys this holds. This is the classical frequentist interpretation. Of course, other philosophical approaches exist, such as the Bayesian school; which, very roughly speaking, contends that prior to the survey the true value could be anywhere and that the a posteriori probability (given the data) for the true value to be in $CI_{1-\alpha}(\hat{Y}_\pi)$ is approximatively $1 - \alpha$.

Using Theorems 2.1.1 and 2.1.2 and after tedious but simple algebraic manipulation, one arrives at the following result:

$$\mathbb{V}(\hat{Y}_\pi) = -\frac{1}{2}\sum_{i,j\in\mathcal{P}} \Delta_{ij}(\check{Y}_i - \check{Y}_j)^2 + \sum_{i\in\mathcal{P}} \frac{Y_i^2}{\pi_i}\Big(\mathbb{E}(n_s \mid I_i = 1) - \mathbb{E}(n_s)\Big) \quad (2.11)$$

A similar but equivalent form of Eq. 2.11 has been described by Ramakrishnan (1975b). Under a fixed sample size the second term vanishes and one obtains the so-called Yates-Grundy formula:

$$\begin{aligned} \mathbb{V}(\hat{Y}_\pi) &= -\frac{1}{2}\sum_{i,j\in\mathcal{P}} \Delta_{ij}(\check{Y}_i - \check{Y}_j)^2 \\ \hat{\mathbb{V}}(\hat{Y}_\pi) &= -\frac{1}{2}\sum_{i,j\in s} \check{\Delta}_{ij}(\check{Y}_i - \check{Y}_j)^2 \end{aligned} \quad (2.12)$$

It is worth noting that the above theoretical variance is, under that fixed sample size, the same as in Theorem 2.2.1, even though the point estimates of variances from Theorem 2.2.2 and Eq. 2.12 will generally differ. The Yates-Grundy formula tells us that if the inclusion probabilities $\pi_i$ are proportional to the response variables $Y_i$ then $\check{Y}_i$ is constant and therefore the variance is zero, thereby making this the ideal sampling scheme! Nevertheless, this is no longer true with random sample sizes. To implement such a sampling scheme would usually require us to know the $Y_i$ for the entire population, which, of course, defeats the point. However, if prior auxiliary information is available in the form of a response variable $X_i$ that is known for all individuals (from, say, a previous census) and if one can expect a strong correlation between the $X_i$ and the $Y_i$ then one should sample with a probability proportional to the $X_i$. Such an approach is called **Probability Proportional to Size** (**PPS**) sampling. This is intuitively obvious: suppose that you have to estimate the total weight of a population consisting of 5 elephants and 10'000 mice, then you evaluate the elephants and consider the mice less so! We shall later see that this technique is fundamental in optimizing sample surveys. In practice one usually must investigate many response variables with the same survey. It is clear that a sampling scheme efficient for one variable may be inefficient for another one. In other words, one has to choose a design based on priorities while respecting the objectives.

Estimating the population mean is straightforward. If $N$ is known, one simply sets

$$\hat{\bar{Y}} = \frac{\hat{Y}_\pi}{N} \quad (2.13)$$

$$\mathbb{V}(\hat{\bar{Y}}_\pi) = \frac{\mathbb{V}(\hat{Y}_\pi)}{N^2} \quad (2.14)$$

$$\hat{\mathbb{V}}(\hat{\bar{Y}}_\pi) = \frac{\hat{\mathbb{V}}(\hat{Y}_\pi)}{N^2} \quad (2.15)$$

However, even if $N$ is known (which is almost never the case in a forest inventory), it is rather surprising that one can construct estimates than can be better than that from Eq. 2.13 in some circumstances. The so-called weighted sample mean is such an example, defined as

$$\tilde{Y}_s = \frac{\hat{Y}_\pi}{\hat{N}}$$

where

$$\hat{N} = \sum_{i=1}^{N} \frac{1}{\pi_i}$$

is an estimate of the population size. We shall revisit this point in the section on the estimation of ratios. For now we will consider the most important schemes used in applications and we will examine the previous general results in these particular situations.

## 2.3 Simple random sampling without replacement

This scheme has constant inclusion probabilities with a fixed sample size $n$. One example would be the common lottery machines. This type of sampling is frequently used as a building block for more complicated schemes. Because of Theorem 2.1.1 this implies that $\pi_i \equiv \frac{n}{N}$. All samples $s$ have the same probability of being chosen, i.e.,

$$p(s) = \frac{1}{\binom{N}{n}} = \frac{(N-n)!n!}{N!}$$

Also

$$\pi_i = \sum_{s \ni i} p(s) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

This combinatorial argument follows from the fact that if a particular individual $i$ is included in the sample we can choose the remaining $n-1$ in the sample only out of the remaining $N-1$ in the population. The rest is simple algebra. Likewise, one can obtain with $\pi_{ij} = \mathbb{P}(I_i = 1 \mid I_j = 1)\mathbb{P}(I_j = 1)$ the pair-wise inclusion probabilities according to

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}$$

Tedious but elementary algebra can be applied to the general results from the previous section for this particular case (Note that it is a good exercise to

write down the proofs). One then obtains

$$\begin{aligned} \hat{Y}_\pi &= N\frac{1}{n}\sum_{i\in s} Y_i = N\bar{Y}_s \\ \mathbb{V}(\hat{Y}_\pi) &= N^2(1-\frac{n}{N})\frac{1}{n}\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{(N-1)} \\ \hat{\mathbb{V}}(\hat{Y}_\pi) &= N^2(1-\frac{n}{N})\frac{1}{n}\frac{\sum_{i\in s}(Y_i-\bar{Y}_s)^2}{(n-1)} \end{aligned} \tag{2.16}$$

**Remarks**:

- $\bar{Y}_s$ is the ordinary mean of the observations in the sample. It is obviously equal to the unbiased estimate $\hat{\bar{Y}}_\pi$ from Eq. 2.13. Its theoretical and estimated variances can be obtained from the above equations by dropping $N^2$.
- With unequal probability sampling the ordinary sample mean does not generally estimate the mean of the population and can, therefore, be totally misleading. This is also true for haphazard sampling where the inclusion probabilities are unknown (e.g. interviews carried out with students sampled in the cafeteria, where heavy coffee drinkers will have a much higher probability of being sampled).
- In elementary textbooks the notation $\frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_s$ is occasionally used. This is misleading because the labels for individuals in the sample are almost never precisely those of the first $n$ individuals in the population.
- $\frac{n}{N}$ is the so-called sampling fraction. For a census it is 1, in which case we logically obtain zero for the theoretical and estimated variance.
- $S_Y^2 = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}{(N-1)}$ is the population variance and can be estimated without bias by the sample variance $\frac{\sum_{i\in s}^{N}(Y_i-\bar{Y}_s)^2}{(n-1)} = s_Y^2$.

To implement this scheme with a sampling frame such as with a list of the $N$ individuals in a file, one can proceed as follows:

- Step 1: Generate for each individual a random variable $U_i$ that is uniformly distributed on the interval $[0, 1]$. Most statistical packages provide this facility.
- Step 2: Rank in increasing order the individual according to their $U_i$ values to obtain the sequence $U_{(1)}, U_{(2)}, \ldots U_{(N)}$.This gives a random permutation of the initial ordering of the individuals in the list.
- Step 3: Select the first $n$ individuals from that permutated list. These form the sample $s$.

The above algorithm is easy to perform. In contrast, the implementation of sampling schemes can be difficult with unequal inclusion probabilities and with

fixed sample sizes (e.g. **PPS**), such that calculation of the $\pi_{ij}$ is cumbersome, see (Särndal et al., 2003) for details and further references. Therefore, the next two sections will present simple procedures for conducting unequal probability sampling with random sample sizes. As we shall see, in a forest inventory, unequal probability sampling is not too difficult to implement, even though the number of trees selected will almost always be random during practical applications.

### 2.4 Poisson sampling

Given a set of inclusion probabilities $\pi_i$ the **Poisson sampling** design has a simple list-sequential implementation. Let $\epsilon_i, \ldots \epsilon_N$ be $N$ independent random variables distributed uniformly on the interval $[0, 1]$. If $\epsilon_i < \pi_i$ the individual $i$ is selected, otherwise not, which by definition occurs with the required probability $\pi_i$. Poisson sampling is a scheme without replacement, that is a selected individual occurs only once in the sample. The sample size $n_s$ is obviously random with mean $\mathbb{E}(n_s) = \sum_{i=1}^{N} \pi_i$ and variance $\mathbb{V}(n_s) = \sum_{i=1}^{N} \pi_i(1 - \pi_i)$. Because of the independence of the $\epsilon_i$ our pair-wise inclusion probabilities satisfy $\pi_{ij} = \pi_i \pi_j > 0$ and consequently $\Delta_{ij} = \check{\Delta}_{ij} = 0$. In this special case, the general results provide the following formulae:

$$\hat{Y}_\pi = \sum_{i \in s} \check{Y}_i \tag{2.17}$$

$$\mathbb{V}(\hat{Y}_\pi) = \sum_{i=1}^{N} \pi_i(1 - \pi_i)\check{Y}_i^2 \tag{2.18}$$

$$\hat{\mathbb{V}}(\hat{Y}_\pi) = \sum_{i \in s} (1 - \pi_i)\check{Y}_i^2 \tag{2.19}$$

The variances $\mathbb{V}(\hat{Y}_\pi)$ and $\hat{\mathbb{V}}(\hat{Y}_\pi)$ can be unduly large because of variability in the sample sizes. A better, but slightly biased, estimator can be obtained with model-assisted techniques:

$$\hat{Y}_{po} = N\tilde{Y}_s \tag{2.20}$$

$$\mathbb{V}(\hat{Y}_{po}) \approx \sum_{i \in \mathcal{P}} \frac{(Y_i - \bar{Y})^2}{\pi_i} - NS_Y^2 \tag{2.21}$$

$$\hat{\mathbb{V}}(\hat{Y}_{po}) \approx \left(\frac{N}{\hat{N}}\right)^2 \sum_{i \in s} \frac{(1 - \pi_i)}{\pi_i^2}(Y_i - \tilde{\bar{Y}}_s)^2 \tag{2.22}$$

where $\tilde{Y}_s = \frac{\hat{Y}_\pi}{\hat{N}}$ with $\hat{N} = \sum_{i \in s} \frac{1}{\pi_i}$. This estimate of the true population mean $\bar{Y}$ should be used even if $N$ is known.

To implement Poisson sampling with **PPS**, $\pi_i \propto X_i$, and the expected sample size $n = \mathbb{E}(n_s)$, it suffices to take

$$\pi_i = \frac{nX_i}{\sum_{k\in\mathcal{P}} X_k}$$

It is theoretically possible that for some individuals $\pi_i \geq 1$. All these units, say, $\{i_1, i_2, \dots i_k\}$ will have to be included in the sample. Then one considers the reduced population $\mathcal{P}^* = \mathcal{P} \setminus \{i_1, i_2 \dots i_k\}$ and iterates, if necessary, the procedure.

The special case of $\pi_i \equiv \pi = \frac{n}{N}$ (i.e. $X_i \equiv 1$) is called **Bernoulli sampling**. There, we would see that $\hat{Y}_{po} = \frac{n}{n_s}\hat{Y}_\pi$ and that $\mathbb{V}(\hat{Y}_{po})$ is very nearly the same as would be found for simple random sampling with a fixed size $n$. In contrast, the variance of the unmodified HT estimator is usually much larger. These examples demonstrate that a good strategy must consider both sampling schemes and estimators.

The next section presents a sampling scheme and an estimator that attempt to combine **PPS** and simplicity.

### 2.5 Unequal probability sampling with replacement

We now consider a population $\mathcal{P}$ with response variables $Y_i$ and an auxiliary variable $X_i$ known for all $i \in \{1, 2 \dots N\}$. The sampling frame is the list of all individuals ordered, without loss of generality, according to their labels $i$'s. We can then define the cumulative sums $S_0 = 0, S_1 = X_1, S_k = S_{k-1} + X_k, k = 2, 3 \dots N$. Note that $S_N = \sum_{k=1}^{N} X_k$.

The sampling procedure consists of $n$, fixed, consecutive identically but independently distributed draws of points $Z_l,\ l = 1, 2 \dots n$ that are uniformly distributed on the interval $[0, S_N]$ (i.e. $Z_l \sim S_N \times U[0, 1]$, with $U[0, 1]$ being a uniformly distributed random variable on the interval $[0, 1]$ ). The individual labeled $i$ is selected at the $l$-th draw if $S_{i-1} \leq Z_l < S_i$. This obviously occurs with probability $p_i = \frac{X_i}{S_N}$. Note that by construction $\sum_{i=1}^{N} p_i = 1$. The number of times $T_i \in \{0, 1, \dots n\}$ a given individual $i$ is included in the sample follows therefore a binomial distribution with parameter $n$ (number of draws) and $p_i$ (probability of success). This is a sampling procedure with replacement because the same individual can be selected more than once (maximum of $n$ times). The following facts are well known from elementary probability theory (although it is always a good exercise to prove them from scratch).

- The random vector $T_1, T_2, \dots T_N$ follows a multinomial distribution with parameter $p_i, p_2, \dots p_N$.
- $\mathbb{E}(T_i) = np_i$ and $\mathbb{V}(T_i) = np_i(1 - p_i)$
- Given $T_i = t_i$, $T_j$ follows a binomial distribution with parameter $n - t_i$ and $\frac{p_j}{1-p_i}$
- $\mathbb{E}(T_iT_j) = \mathbb{E}(T_i\mathbb{E}(T_j \mid T_i)) = n(n-1)p_ip_j$