PhD THESIS DECLARATION

The undersigned

SURNAME *Missiroli*

FIRST NAME *Silvia*

PhD Registration Number *1511815*

Thesis title: *Adaptive Sequential Sampling for Finite Populations with Applications in Agricultural and Agro-Environmental Statistics*

PhD in *Statistics*

Cycle *28*

Candidate's tutors: *Professor Elisabetta Carfagna*

DECLARES

Under *her* responsibility:

1) that, according to Italian Republic Presidential Decree no. 445, 28th December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30th April 1999, to keep a copy of the thesis on deposit at the "Bilioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary embargo protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

3) that the Bocconi Library will file the thesis in its "Archivio istituzionale ad accesso aperto" (institutional registry) which permits online consultation of the complete text (except in cases of a temporary embargo);

4) that, in order to file the thesis at the Bocconi Library, the University requires that the thesis be submitted online by the candidate in unalterable format to Società NORMADEC (acting on behalf of the University), and that NORMADEC will indicate in each footnote the following information:

   - thesis: *Adaptive Sequential Sampling for Finite Populations with Applications in Agricultural and Agro-Environmental Statistics*;
   - by *Missiroli Silvia*;
   - defended at Università Commerciale "Luigi Bocconi" – Milano in *2017*;
   - the thesis is protected by the regulations governing copyright (Italian law no. 633, 22th April 1941 and subsequent modifications). The exception is the right of Università Commerciale "Luigi Bocconi" to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis submitted online to NORMADEC is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal, and shall be exempt from any requests or claims from third parties;

7) that the thesis is not subject to "embargo", i.e. that it is not the result of work included in the regulations governing industrial property; it was not written as part of a project financed by public or private bodies with restrictions on the diffusion of the results; is not subject to patent or protection registrations.

Date *November 15, 2016*

SURNAME *Missiroli*
FIRST NAME *Silvia*

To my grandmother
Assunta Vandi.

Lo studio e, in generale, la ricerca della verità e della bellezza
sono una sfera di attività nella quale
ci è consentito di rimanere bambini per tutta la vita.
Albert Einstein

The pursuit of truth and beauty
is a sphere of activity in which we are permitted
to remain children all our lives.
Albert Einstein

Contents

List of Figures

List of Tables

Acknowledgements

This PhD thesis has been a long journey and as the most beautiful and challenging journeys, it would not be possible without the help of some Allies.

First of all, I would like to express my deep gratitude to my advisor, Professor Elisabetta Carfagna, who never stopped to follow and support me, from the most basic task to the hardest one. With patience, kindness and precision, she has guided me through all the fundamental passages towards my first marked research activity, voicing my ideas and letting me structure my own personality as a researcher. Moreover, with her experience and her smart ideas she has shaped several parts of this work, transmitting me the light through which she is able to look wisely at scientific issues in order to find solutions.

I am also very grateful to my co-advisor Professor Sonia Petrone for the time and patience she has devoted to me, supporting from the beginning my idea to choose this thesis project. She has always provided smart and interesting hints and she has strongly helped me to face difficulties during my research path, giving me wise suggests.

I would like to thank my professors of Decision Sciences Department at Bocconi University for their interesting courses and for their refined approach to the statistical subjects. Especially, I would like to thank Professor Pietro Muliere, for his generous and helpful comments and for his encouragement.

Thanks to all the secretaries and the members of the PhD Administrative Centre, particularly to Mr Gualtiero Valsecchi and Mrs Barbara Contaldo for having always mixed sweetness and support with their efficient job activities.

My PhD colleagues Sumeda Siriwardena, Sajid Alì and Alì Noorikhajavi have been my second family during these years. I'm very grateful to them for their precise and frequent explanations on technical aspects, for their support and their deep friendship. Especially, Sumeda has been my first Ally in this journey, I will never stop to thank her.

Many thanks to Isadora Antoniano and Giulia Marcon for their sweet affection and support since the beginning. Similarly, I would like to thank the PhD students of other cycles for their nice company and challenging conversations.

A special thank goes to Linda, Domenica and Gemma, for having transformed Milan in a warm house and to all my lifelong friends, particularly to Linda, Valentina, Cristiana, Elisa, Esmeralda, Mattia to have devoted themselves to me, simply and deeply, since a

life.

Annamaria, Carlo and Massimo supported me with big affection as a second family, far but always close. I thank them greatly.

A huge thank to my parents and to my brother Marco for their infinite love and continual support that has made this dream become true. Since we are a team, this is also their victory.

Finally, thank you granny for your protection. This is for you.

Abstract

In stratified sampling for finite populations, several authors have discussed the problem of sample allocation and selection in absence of previous information on the variability inside the strata, suggesting various kinds of two steps sampling or sequential sampling strategies. However, the proposed methods either do not allow design unbiased estimates of the population parameters or are not optimal or do not take into consideration budget constraints.

In this thesis we propose a group sequential adaptive procedure with permanent random numbers (AGSPRN) fo stratified finite populations. It generates a sample allocation very close to Neyman's optimal one and design unbiased estimates of the population mean or total, taking into consideration budget constraints and a linear cost function.

Among all the AGSPRN procedures characterized by different number of steps K and units q added at each step, this work aims at finding the optimal pair (K, q) that minimizes: 1) the variance of the estimator given a cost function and a budget constraint, 2) the total cost given a threshold on the estimator variance, 3) a risk function obtained as a combination of cost and estimator variance. Since these problems are analytically intractable because of the prohibitive form of the distribution of the estimator variance, we proceed through a Monte Carlo investigation. First of all, we provide a simulation study in order to show some properties of the optimal AGSPRN procedure. Especially, the study shows that the optimal AGSPRN procedure tends not to coincide with some of the sampling designs proposed in the literature, which are less efficient. Moreover, focusing on the cost function, we assess the impact of various values of the cost components on the optimal group sequential sampling technique.

Then, we set up a methodology that allows to obtain the optimal AGSPRN procedure when the population values are not known, which is the usual case. The proposed method is adopted for a real application in the field of territory management, in order to obtain the optimal group sequential procedure and provide efficient estimates for the quality control index of a land cover database, in presence of a cost function.

Some convergence properties are also proved for a slightly modified version of the AGSPRN procedure in the context of infinite populations, using martingales arguments.

Chapter 1

Introduction

Adaptive sequential sampling for finite populations is the core topic of this thesis. The need of implementing efficient sampling methods that hinge on the information gained along the procedure arises in many practical situations, where it is fundamental to produce efficient analyses with time and money constraints. In agro-environmental and agricultural contexts this is more than true. A real issue in agro-environmental statistics has led us to consider the ambitious goal of finding an optimal adaptive sequential sampling procedure that offers efficient estimates with the least amount of money.

1.1 Motivation

In front of practical problems, statistical analyses are required to be reliable and efficient in terms of cost, time and precision. Resources are often limited and solutions have to be timely, aiming at saving money while preserving efficiency. One of the most critical point for a statistical analysis is the collection of a sample from a finite population. Adding flexibility to a sampling procedure for finite populations is one of the challenges of this thesis. The main tool we are going to use to reach this crucial aim is *adaptive sequential sampling*. Four different elements become parts of our complex framework: sampling for finite populations, adaptive procedure, costs and sequential setting with various stopping rules. These are all important components that real applications should require from a statistical analysis. Particularly, what has led us to consider all these aspects are some procedures in territory management that require a very efficient finite population sampling frame, which seeks also to save costs and time.
Among all the finite population sampling designs, *stratified sampling* is the most widely used procedure in applied contexts. Hence, we have chosen it as the basic sampling technique for this thesis. Further important details about this design will be presented in Chapter 2.

In a finite populations sampling context, it is a well known practice to introduce costs,

assuming a linear cost function which plays an important role in the search of the optimal sample (see, for instance, Cicchitelli et al. [1992, pp 96–99], Cochran [1997, pp 318–322], Thompson [2012, pp 147]). The traditional approaches to cost issues are integrated in conventional sample designs, in adaptive cluster sampling (Thompson and Seber [1996, pp 152–153]), in stratified sampling with adaptive allocation (Carfagna et al. [2012]). One of our goals is to associate a cost function to a *group* sequential stratified sampling design with adaptive allocation in order to generate the most efficient estimators both in terms of cost and precision. A design with *adaptive allocation* denotes a sampling procedure where the sample units are selected in each stratum along more steps and the current allocation may depend on the previously collected data (Thompson and Seber [1996, pp 192–199]). Moreover, we will consider a *sequential* framework with different stopping rules that take into account the estimator variance, the total cost including the *step* cost and the total risk obtained as a combination between cost and estimator variance. Our sampling procedure will stop when the estimator variance will be under a threshold v or the total cost and the risk will be minimized. Stein [1945], Chow and Robbins [1965], Ray [1957] and Liu [1997] investigated a similar sequential setting, but they refer to infinite population and normally distributed data, which are not divided into strata. Kadane [2005] showed how an optimal *dynamic* sample allocation among strata can be computed in the presence of a cost function and estimator variance constraint. However, he did not consider the adaptive rule and the cost per step. Hardwick and Stout [1995, 2009] proposed computational algorithms for adaptive designs in a simpler setting. Rosenberger and Sriram [1997], Jennison and Turnbull [1999], Melfi and Page [1998, 2000], Melfi et al. [2001], Muliere et al. [2006a], Muliere et al. [2006b], Rosenberger et al. [2001], Rosenberger and Lachin [2015], Antognini and Giovagnoli [2015] proposed adaptive designs for *infinite* population, mainly in the context of clinical trials. They also studied consistency and asymptotic normality of the estimators for a wide class of designs, often utilizing martingales arguments. In Chapter 6 we will extend our design to an infinite population framework, studying consistency and asymptotic normality of the proposed estimator, using the same tools of the cited authors and of Etoré and Jourdain [2010].

In the finite populations context, Carfagna [2007] introduced a two steps adaptive procedure with permanent random numbers which is extended by Carfagna and Marzialetti [2009b] to a sequential setting, giving rise to an adaptive sequential procedure with permanent random numbers. Since the methods introduced by Carfagna [2007], Carfagna and Marzialetti [2009b] deal with finite populations, adaptive allocation and sequential framework, we use them as a starting point.

Therefore, we will propose a group sequential stratified sampling procedure with adaptive allocation for finite populations in the presence of a cost function that considers the cost per each selected unit, the cost per step and a fixed cost. The aim will be to find the *optimal* adaptive procedure in terms of minimum cost, maximum precision of the estima-

tor, minimum risk computed as a combination between cost and estimator variance. The search of this optimal adaptive design needs to know the distribution of the estimator variance, which is analytically prohibitive. Hence, we will proceed through a Monte Carlo study, which requires as inputs some precise information about the target population. We will propose a methodology that allows to update this information at each step of the sampling procedure, in order to obtain the optimal adaptive technique for a population that is very similar to the target one.

1.2 Motivating application

This thesis has a very concrete and important motivation. It originates from the need of evaluating the quality and validate land cover databases produced through photo-interpretation of remote-sensing data.
Aerial or satellite remote-sensing data are widely used for monitoring land cover and for testing the effect of land use policies. Satellite images are given as a set of measures of electromagnetic radiation reflected by the earth surface. Each individual measure corresponds to an area unit (pixel) and a certain interval of wavelength (channel), generally ranging from the visible spectrum to the thermal infrared. The most important characteristics of an optical satellite sensor are the number of channels, the wavelength of each of them and the spatial resolution, that is, the size of the pixel; for civil use images, it can range from less than 1 m to 5 km, while the most widely used images for land cover monitoring have medium resolution, for example, 30m (Figure 1.1). Remote-sensing data are photo-interpreted according to a land cover legend in which each class (or label) represents a land cover type. Generally, the legend is defined before starting the photo-interpretation. The result of this process is a database, whose main elements are the polygons (homogeneous land areas with regular borders) of specific land cover types. The database is created in a Geographic Information System (GIS) that allows many kinds of operations on the polygons, such as dividing a polygon into two pieces, merging the polygons, overlaying different information layers and so on. Further details about GIS and remotely sensed data analysis can be found in Benedetti et al. [2015].

Public administrations are the main customers of land cover databases for various purposes; but we have to remember that a prerequisite for making a reliable analysis using a land cover database is the quality of the evaluated database, that should be high. The scale of remote-sensing data used in the photo-interpretation represents only the level of detail of the basic material and cannot be considered as the quality of the land cover database. The quality should be analysed from two viewpoints: the quality of the photo-interpretation and the level of agreement between the database and reality (validation). Owing to cost and time, the quality control of the photo-interpretation as

Figure 1.1: Example of satellite images with various pixel sizes.

well as the validation can be performed only on the basis of a sample of polygons in the methodological framework of statistical sampling from finite population.

Quality control should be done by repeating the photo-interpretation process with the same basic material. In our case, a very expert photo-interpreter (the controller) repeats the photo-interpretation on a sample of polygons.

Validation is a comparison of the land cover database with another representation of reality, which is considered more reliable. We compare a sample of polygons with the corresponding ground truth, in case the scale of remote-sensing data is compatible with the ground truth; otherwise, the comparison is made with other remote-sensing data with a more detailed scale.

Generally, in photo-interpretation projects, very few resources are devoted to quality control; thus, a very cost-effective sample design is required. Since we want to use quality control for continuously improving the database production process, it must be performed in a very short time, during the photo-interpretation process. This is the reason that has led us to adopt an adaptive sequential sampling procedure for quality control as well as for validation: it allows to reach high precision of estimates with the smallest sample size and in the shortest time, especially when a small amount of information is available. Moreover, the costs can be controlled step by step, ending up with a sample that is the most efficient in terms of estimates precision and costs.

It is worth mentioning that sometimes the desire of homogeneity suggests to use a common legend or satellite-data even though they are inappropriate for specific projects. The use of an adaptive sequential sampling procedure during the photo-interpretation not only allows quality control and validation, but it is also a way to improve the legend according to customer's needs.

1.2.1 Adaptive sequential sampling for quality control

Let us focus on the quality control operation of a land cover database. A photo-interpreter classifies the satellite images according to a legend of different land cover types. However, during this procedure, he can make mistakes concerning the border of the polygons as well as the land cover type. Therefore, another expert (the controller) repeats the photo-interpretation on a sample of polygons in order to test if some mistakes have been made by the previous photo-interpreter. This procedure is called quality control of land cover databases and it is measured through some parameters such as the percentage of area correctly photo-interpreted and the percentage of the polygons correctly photo-interpreted. In Chapter 5 we will focus on estimating the first index by an estimator developed through an adaptive sequential sampling for finite populations. Indeed, the most important aim in this operation for territory management is to obtain efficient estimates of the quality indexes, saving time and resources, modifying the sampling procedure according to the stakeholders needs and to the results obtained along the way.
The use of stratified finite population is required in this context since the kind of land cover type and the size of the polygons affect the probability of making mistakes in the photo-interpretation. However, the layers under stratification must be chosen carefully. In our application, we will choose the land cover type and the size of the polygons as stratifying factors. We will discuss in details in Chapter 5 all the aspects of stratification and the sampling procedures for our data-set.

1.2.2 Adaptive sequential sampling for validation

In order to assess the agreement of the database with the ground truth and verify the capability of the database to satisfy the client's needs, an adaptive sequential sampling can be adopted also for validating the database. Indeed, a continuous sequential validation during the production process allows:

- timely detection of discrepancies between the database and reality, which makes the product inappropriate for the customer's needs;
- drawing the characteristics of the database progressively closer to the client's needs; the customer himself often is not aware of his requirements until he starts using the database;
- changing the data source in order to cut some costs if, for instance, during the procedure the adopted remote-sensing data are realized to be more detailed than those required by the user's needs, producing unjustified costs;
- timely testing (and in case changing) the legend for photo-interpretation during the

photo-interpretation process in order to identity the most appropriate legend for the specific area and project.

1.2.3 Data

In 1999, the Italian Statistical Institute (ISTAT) carried out an experiment funded by Eurostat and produced a land-cover/land-use database with a detailed CORINE [1] (Carfagna and Marzialetti [2009a]) legend and a scale of 1:25,000 for the Arezzo province.
We will analyse, in Chapter 5, the classification of 110 polygons that became part of the dataset produced by ISTAT. Since we have the results of both classifications made, respectively, by the photo-interpreter and by the controller during the quality control operation, we will compute the quality control index of the area correctly photo-interpreted. Then we will evaluate the performance of its estimate obtained through the adaptive group sequential procedure which uses as input the entire data-set or only a sample of it.

1.3 Outline of thesis

Many aspects of adaptive sequential sampling for finite populations are examined in this thesis. In Chapter 2 we review the main concepts of finite populations sampling and conventional designs (simple random sampling, stratified random sampling), proceeding towards adaptive cluster sampling and sampling with adaptive allocation (multiple steps sampling, two steps sampling with permanent random numbers, adaptive sequential sampling with permanent random numbers). An excursus about two steps and sequential estimation in the infinite population context is also presented.
In Chapter 3 we propose an adaptive group sequential procedure with permanent random

[1] Between 1985 and 1996 the first Corine land-cover (CLC) inventory for the EU-15 and most of the new member states was implemented. It was a project carried out in order to characterize the land surface. A uniform nomenclature across Europe at a scale of 1:100,000 was used. The CLC nomenclature mostly included land-cover items, though land-use elements could also be found. The CLC Technical Team (under the responsibility of the European Topic Centre on Terrestrial Environment) carried out a validation at the end of the project as well as a quality control during the production process (see European Environment Agency 2006). Validation was not performed by acquiring new ground data. The LUCAS 2001/2002 survey originally carried out for agro-environmental purposes was used instead. The accuracy of the CLC database was assessed by reinterpreting the LU- CAS field photographs (in combination with IMAGE2000 and other LUCAS statistics), which were provided for 8,231 locations in the 18 x 18 km sampling grid. The total percent correct was 87.0 $\pm$ 0.8. However, since LUCAS was not originally intended to validate Corine Land Cover, 22 of the 44 CLC classes could not be validated due to low representativeness in LUCAS; thus, the reliability of CLC for half of the classes could not be evaluated by LUCAS. Moreover, the LUCAS survey was available only for 18 of the 29 countries where CLC was created. The quality control during the production process was meant to monitor and provide guidelines about where to improve the production of the CLC database in the different countries. The feedback given at this stage was qualitative and its overall objective was to realize a homogeneous and comparable database at European level.

numbers (AGSPRN) in presence of a linear cost function. We aim at finding an optimal adaptive stratified sampling procedure in terms of:

- minimum variance of the estimator given a cost function and a budget constraint (Case 1);
- minimum total cost given a threshold on the estimator variance (Case 2);
- minimum risk obtained as a combination of estimator variance and cost (Case 3).

These problems are analytically unsolvable because of the prohibitive distribution of the estimator variance. Hence, in order to solve them, we propose a Monte Carlo study, which requires as inputs some precise information about the target population. This is a very delicate point. In Chapter 3 the Monte Carlo method is applied directly to a specific target population, under particular values of the cost components. In this case, we use the entire target population to find the optimal sampling design. This step seems quite ineffective, but it is useful to show some properties of the optimal AGSPRN procedure and to assess the effects of the cost components on it. The main results of this chapter are reported in Missiroli and Carfagna [2016].
In Chapter 4 we set up a methodology in order to find the optimal AGSPRN procedure when only a pilot sample of the population is available, that is usually the case in theoretical and practical situations. The results are then compared with those obtained in Chapter 3.
Chapter 5 deals with the search and the application of the optimal AGSPRN procedure in order to estimate the quality control index for the data produced by ISTAT (see Section 1.2.3).
In Chapter 6, we try to connect a slightly modified version of our AGSPRN procedure to the infinite population context, proving consistency and asymptotic normality of the AGSPRN estimator.
Finally, we provide a final discussion and directions for future research.

Chapter 2

Literature review

2.1 Introduction

This chapter deals with the basic theory of *survey sampling* for finite populations and in particular it is focused on some results from the literature about *adaptive sequential sampling*, which is the topic under investigation. An excursus about different sampling designs is presented. In the finite populations sampling context, it is critical to highlight the distinction between *'model-based' sampling* and *'design-based' sampling*; the latter is, in its adaptive framework, our main area of interest. The most important results for infinite population adaptive sampling are also reported.

2.2 Different approaches in survey sampling

As we underlined, it is important to distinguish between a 'design-based' approach and a 'model-based' sampling. Thompson and Seber [1996] can be a good guide to understand these important concepts. A finite populations sampling context is characterized by a population of N units $(u_1, ..., u_N)$ indexed by their labels $(1, ..., N)$ and a variable of interest y_i is associated to unit i, for $i = 1, ..., N$. The value y of the variable can be nominal, ordinal or scalar, unidimensional or vector valued. The vector of the population y-values will be denoted by $\mathbf{y} = (y_1, ..., y_N)'$. In the 'design based' approach, $\mathbf{y}$ is considered a fixed set of unknown constant. In the 'model based' view (Chambers and Clark [2012]), the vector $\mathbf{y}$ is considered a realization of a random vector $\mathbf{Y}$ having a joint distribution F which may depend on unknown parameters $\boldsymbol{\phi}$ ranging in a parameter space Φ. In a Bayesian setting $\boldsymbol{\phi}$ has a known prior distribution.
An unordered sample of size n is a set $s = \{i_1, ..., i_n\}$ of n of the N labels (some of which may be the same, as in sampling with replacement), which forms, together with the associated y-values, denoted by $\mathbf{y}_s$, the set of data $D = (s, \mathbf{y}_s)$. The last term D is a random variable, since each sample s is characterized by a probability of being selected.

A realization of D is denoted by d.
An ordered sample of size n is a sequence $s_0 = (i_1, ..., i_n)$ of n of the N labels. The association of the ordered sample y-values, denoted by $\mathbf{y}_0$, with their units labels is represented by the random variable $D_0 = (s_0, \mathbf{y}_0)$. S indicates the set of all possible samples with fixed or variable size of a given population and $\mathcal{Y}$ is the set of all possible values of $\mathbf{y}$. The aim is to select a sample, observe the y-values and estimate some function $\eta(\mathbf{y})$ of the population y-values. The population total $\eta(\mathbf{y}) = \sum_{i=1}^{N} y_i = \tau$, the population mean $\eta(\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{Y}$ and the population variance $\eta(\mathbf{y}) = \sum_{i=1}^{N}(y_i - \mu)^2/(N-1) = \sigma^2$ are examples of population functions $\eta(\mathbf{y})$ that can be estimated.
Different procedures, called *sampling designs*, can be used to select a sample. The key of the selection is the probability $p(s|\mathbf{y})$ of getting sample s, for each possible sample $s \in S$. It may depend on the configuration of $\mathbf{y}$, the y-values of the population. The design probabilities satisfy $p(s|\mathbf{y}) \geq 0$ for all $s \in S$ and $\sum_{s \in S} p(s|\mathbf{y}) = 1$ for all $\mathbf{y} \in \mathcal{Y}$.
If the selection probabilities do not depend on the value of the variable of interest or on any parameter values, although they may depend on the value of auxiliary variables $\mathbf{x}$ that may be known in the population, the selection procedures are known as *conventional designs*. Inside this class $p(s|\mathbf{y}) = p(s)$ holds. They include simple random sampling, stratified random sampling, systematic, cluster sampling.
If the procedure for selecting units depends on the values of the variable of interest, but only through the units included in the sample denoted by $\mathbf{y}_s$, that is $p(s|\mathbf{y}) = p(s|\mathbf{y}_s)$, we are in presence of *adaptive designs*. This class includes random sampling with a sequential stopping rule, adaptive cluster sampling designs, adaptive allocation in stratified designs and many others.
Design of the form $p(s|\mathbf{y})$, in which the selection probabilities are influenced by values of units not included in the sample or by unknown parameter values, will be termed *nonstandard*.

2.3 The efficiency of an estimator

Starting from the classical context of statistical inference for infinite populations, let us denote with Z a random variable taking value in $\mathcal{Z}$, according to a distribution P_θ, with the unknown parameter θ laying in Θ. In decision theory, an estimator represents a *decision rule* δ that is a function from $\mathcal{Z}$ to A, the space of actions. The unknown parameter θ to be estimated is called *state of nature* and takes values in Θ. In estimation problems with frequentist approach, the loss associated to a decision rule is expressed by a *loss function* $L(\theta, \delta(z))$: $\Theta \times A \to R$, which represents the divergence of the estimate $\delta(z)$, given a realization z of the random variable Z, from the parameter θ. The *risk function* is defined as the expected value of the loss function with respect to the distribution of

Z, P_θ, that is: $R_\theta(\theta, \delta(Z))$=$E_\theta[L(\theta, \delta(Z))]$. A decision rule δ_1 is R-better than another decision rule δ_2 if $R_\theta(\theta, \delta_1(Z)) \leq R_\theta(\theta, \delta_2(Z))$ for all $\theta \in \Theta$, with strict inequality for some θ. A decision rule is said to be admissible if there exists no R-better decision rule.
Different kinds of loss functions can be used. In an estimation setting, the common choice is the *quadratic loss* function $L(\theta, \delta(Z)) = (\theta - \delta(Z))^2$, and the risk function corresponds to the mean square error (MSE):

$$MSE_\theta(\theta, \delta(Z)) = E_\theta[(\delta(Z) - \theta)^2] = var_\theta[\delta(Z)] + [E_\theta[\delta(Z)] - \theta]^2,$$

where the second term is the bias. When $\delta(Z)$ is unbiased, the $MSE_\theta(\theta, \delta(Z))$ is the variance of $\delta(Z)$. Among the unbiased estimators, the one that has uniformly the minimum variance for each value of θ is called UMVUE (uniform minimum variance unbiased estimator). Particularly, an unbiased estimator $\overline{\delta}(Z)$ for θ is UMVUE if $\forall\theta \in \Theta$ the following holds:

$$var_\theta[\overline{\delta}(Z)] \leq var_\theta[\tilde{\delta}(Z)] \tag{2.1}$$

for any unbiased estimator $\tilde{\delta}(Z)$. Not always an UMVUE exists, since inequality (2.1) may not hold for all values θ in Θ. Fixing the target θ, an estimator is more efficient than another if its MSE is smaller for that target.
Moving to the finite populations context an estimator $\delta(D)$ of θ is a function of the data D and it is unbiased for θ if its expected value is equal to θ, for any possible values of θ, that is:

$$E(\delta(D)) = \sum_{s_i \in S} \delta(d_i)p(s_i) = \theta, \quad \forall\theta,$$

where the sum is extended to all possible samples in the sample space S provided by the plan, $p(s_i)$ is the probability of extracting the specific sample s_i and $d_i = (s_i, \mathbf{y}_{s_i})$ are the observed data.
The MSE in a finite population context is equal to:

$$MSE_\theta(\delta(D)) = E(\delta(D) - \theta)^2 = \sum_{s_i \in S} [\delta(d_i) - \theta]^2 p(s_i) = V(\delta(D)) + [E[\delta(D)] - \theta]^2,$$

where $V(\delta(D))$ is the variance of the estimator, that is $V(\delta(D)) = [\delta(D) - E(\delta(D))]^2$ and $[E[\delta(D)] - \theta]^2 = B^2$ is the square of the bias. Comparing estimators and increasing their precision will become fundamental in this thesis, since one of our aim is to find the optimal sampling design which generates the estimator with the highest precision.
Now let us focus on the different kinds of sampling designs.

2.4 Conventional designs

With the term *conventional designs* we indicate a sampling procedure in which the selection probabilities of the units do not depend on the values of the variable of interest or on any other parameter. Designs of this type are simple random sampling, stratified random sampling, systematic sampling, cluster sampling, double sampling (see Cicchitelli et al. [1992], Cochran [1997], Conti and Marella [2012]).

2.4.1 Simple random sampling

Simple random sampling (Cochran [1997, pp 18-30]) is a sampling design which assigns to each unit of the population of size N an equal probability of being drawn. Similarly, according to this sampling plan, each of distinct subset of size n has an equal chance of being chosen for the sample. If the unit that has been drawn is removed from the population for all subsequent draws, this method is called random sampling *without replacement* (SRSWOR). In this case, two samples are considered distinct if they differ for at least one unit. Hence, the number of samples is equal to all the possible combinations of N elements taken n at a time without repetition, that are computed through the binomial coefficient $\binom{N}{n}$. According to SRSWOR, the probability of selecting a sample of size n from a population of size N is equal to $1/\binom{N}{n}$ for all the possible samples.
Simple random sampling *with replacement* (SRSWR) is instead a design where all the N elements of the population are associated to an equal and constant probability of being drawn, that is equal to $\frac{1}{N}$, without counting the times they have already been drawn. Two samples are considered distinct if they contain one or more different units or the same units but with a different selection order. Hence, the sample space is composed by N^n samples, where N^n is the number of combinations of N elements taken n at a time with repetition. The probability of selecting a sample according to SRSWR is $1/N^n$.
Denote with $\overline{Y}$ the population mean, with $\overline{y}$ the sample mean, with $\tau = N\overline{Y}$ the population total that is estimated by $\widehat{\tau} = N\overline{y}$. It is well known that the sample mean based on simple random sampling is an unbiased estimate of $\overline{Y}$ and $\widehat{\tau}$ is unbiased for the population total τ.
The variance of the mean $\overline{y}$ from a simple random sample without replacement is

$$V(\overline{y}) = \frac{S^2}{n}\frac{N-n}{N} = \frac{S^2}{n}(1-\frac{n}{N}) = \frac{S^2}{n}(1-f), \tag{2.2}$$

where $S^2 = \frac{\sum_1^N (y_i - \overline{Y})^2}{N-1}$ is the population variance and $f = \frac{n}{N}$ is the sampling fraction.
An unbiased estimate of $V(\overline{y})$ is $\widehat{V}(\overline{y}) = \frac{s^2}{n}\frac{N-n}{N} = \frac{s^2}{n}(1-f)$ where $s^2 = \frac{\sum_1^n (y_i - \overline{y})^2}{n-1}$ is an unbiased estimate of S^2. The variance of the sample total and its unbiased estimate are the same of the sample mean ones multiplied by N^2.

The variance of the mean estimator for a simple random sampling with replacement is the same of expression (2.2) multiplied by $(N-n)/(N-1)$.
Throughout this thesis we will focus on simple random sampling without replacement, since it is often used as a base for the configuration of other sampling designs.

2.4.2 Stratified random sampling

In *stratified sampling* the population of size N is first divided into H subpopulations of $N_1, N_2, ..., N_H$ units so that $N_1+N_2+...+N_H = N$. The subpopulations are called *strata*. A simple random sample without replacement is taken in each stratum. We are interested in studying the variable Y, knowing that at each unit i of the population is associated a value of Y equal to y_i, for $i = 1, ..., N$.

The following quantities can be defined for stratum h, with $h = 1, ..., H$:

N_h: total number of units in stratum h;

n_h: number of sample units in stratum h;

$W_h = \frac{N_h}{N}$: weight of stratum h;

$f_h = \frac{n_h}{n}$: sampling fraction in stratum h;

y_{hi}: value of the variable of interest for unit i in stratum h;

$\overline{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$: mean in stratum h;

$\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$: sample mean in stratum h;

$S_h^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi}-\overline{Y}_h)^2$: variance in stratum h;

$s_h^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\overline{y}_h)^2$: unbiased estimate of the variance S_h^2 in stratum h.

An estimator of the population mean, according to the stratified random sampling design, is given by:

$$\overline{y}_{st} = \sum_{h=1}^{H} \frac{N_h \overline{y}_h}{N} = \sum_{h=1}^{H} W_h \overline{y}_h,$$

that is unbiased since in every stratum the sample mean $\overline{y}_h$ is unbiased for the stratum mean $\overline{Y}_h$, with $h = 1, ..., H$. Indeed, a simple random sampling without replacement is implemented in each stratum and it ensures unbiased mean estimator.
The **stratified mean estimator** $\overline{y}_{st}$ and the **stratified estimator** for **the total** $\widehat{\tau} = \sum_{h=1}^{H} N_h \overline{y}_h$ are the core elements of our investigation.
The variance of $\overline{y}_{st}$ is the following:

$$V(\overline{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n_h} \frac{N_h - n_h}{N_h}. \tag{2.3}$$

An unbiased estimator of $V(\overline{y}_{st})$ is

$$\widehat{V}(\overline{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{s_h^2}{n_h} \frac{N_h - n_h}{N_h}.$$

The allocation of the sample units into the strata, i.e. the choice of the values $n_1, ..., n_H$, is an important step in stratified sampling. It depends on the aim of the analysis, on the available information, on the examined context and on the variable of interest. A popular choice is to allocate the sample units $n_1, ..., n_H$ into the strata *proportionally* to the stratum weights $W_1, ..., W_H$, that is:

$$\frac{n_h}{n} = \frac{N_h}{N} \quad \texttt{or} \quad \frac{n_h}{N_h} = \frac{n}{N}.$$

This allocation scheme is called **proportional allocation** and it provides the equality between the stratified mean estimator $\overline{y}_{st}$ and the sample mean estimator $\overline{y} = \frac{\sum_{h=1}^{H} n_h \overline{y}_h}{n}$. In the following section we will focus on the allocation schemes which minimize some objective functions, such as the variance of the stratified mean estimator.

2.4.3 Simple and multipurpose allocation

Also in the finite population sampling context we can be interested in obtaining the sample size n that minimizes an objective function $\Psi(n)$, given a sampling plan. Particularly, focusing on stratified random sampling, we aim at finding the allocations $n_1, ..., n_H$ that minimize an objective function $\Psi(n_1, ..., n_H)$. Starting from the simplest case, the allocations are chosen in order to minimize a single criterion such as the variance of the stratified mean estimator. If $\Psi(n_1, ..., n_H) = V(\overline{y}_{st})$ the following optimization problem has to be solved for a fixed sample size n:

$$min_{n_1,...,n_H} V(\overline{y}_{st}) = min_{n_1,...,n_H} \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n_h} \frac{N_h - n_h}{N_h}. \tag{2.4}$$

The solution is called **optimal allocation** or **Neyman allocation** and it is:

$$n_h = n \frac{N_h S_h}{\sum N_h S_h} \quad h = 1, ..., H \tag{2.5}$$

If we substitute (2.5) in expression (2.3) we derive the minimum value of $V(\overline{y}_{st})$ for a fixed n, that is:

$$V(\overline{y}_{st})_{min} = \frac{(\sum_{h=1}^{H} W_h S_h)^2}{n} - \frac{\sum_{h=1}^{H} W_h S_h^2}{N}.$$

Neyman allocation is a key element of this thesis and it will have a fundamental role

among the following chapters.
In more complex situations we can be interested in finding the allocations which optimize a good trade-off between more experimental criteria. This target allocation scheme should be obtained by two traditional approaches, that Antognini and Giovagnoli [2015] reclaimed in a similar context:

- *combined optimization*: the function $\Psi(\cdot)$ that has to be optimized is a combination of the different experimental goals by means of suitable weights. For combination purposes, all criteria have to be standardized to put them into a comparable scale. For instance, considering two criteria $\Psi_1(\cdot)$ and $\Psi_2(\cdot)$, the standardization consists in choosing them to lie in $[0;1]$, namely Ψ_1, Ψ_2: $[0;1] \rightarrow [0;1]$. The criterion that is minimized is a convex combination between the two different criteria:

$$\Psi(\cdot) = \omega\Psi_1(\cdot) + (1-\omega)\Psi_2(\cdot),$$

  where the weights ω and $1-\omega$ in $(0,1)$ represent the relative importance of Ψ_1 and Ψ_2. If both criteria are strictly convex, the solution is unique.

- *constrained optimization*: one criterion is optimized under suitable conditions on the other criterion.

The first approach is used in Section 3.5 of Chapter 3, where a risk function is obtained as a combination of cost and estimator variance criteria.
Let us focus on stratified random sampling and consider as experimental goals the variance of the stratified mean estimator $V(\overline{y}_{st})$ and the total sampling cost. The latter criterion is chosen to be represented by a linear cost function:

$$C(n_1, .., n_H) = C_0 + \sum_{h=1}^{H} c_h n_h,$$

where C_0 is the fixed cost and c_h is the variable cost for each unit in stratum h, with $h = 1, ..., H$. *Constrained optimization* is applied in order to obtain the optimal allocation. The estimator variance $V(\overline{y}_{st})$ is minimized for a specified sampling cost or the total cost $C(n_1, .., n_H)$ is minimized for a specified value of $V(\overline{y}_{st})$, that is:

$$min_{n_1,...,n_H} V(\overline{y}_{st}) \quad \texttt{subject to} \quad C(n_1, .., n_H) = C_0 + \sum_{h=1}^{H} c_h n_h = C \tag{2.6}$$

and

$$min_{n_1,...,n_H} C_0 + \sum_{h=1}^{H} c_h n_h \quad \texttt{subject to} \quad V(\overline{y}_{st}) = v, \tag{2.7}$$

where C and v are two positive constants.
Both problems (2.6) and (2.7) lead to the same solution ([Cochran, 1997, pp 97-98]), that is:

$$n_h = n\frac{W_h S_h/\sqrt{c_h}}{\sum_{l=1}^{H} W_l S_l/\sqrt{c_l}}. \tag{2.8}$$

This result is similar to expression (2.5), indeed it represents Neyman's allocation when costs are also considered. Substituting (2.8) in the constraint of problem (2.6) we obtain the following overall sample size:

$$n = (C - C_o)\frac{1}{\sum_{h=1}^{H} \frac{N_h S_h \sqrt{c_h}}{\sum_{l=1}^{H} W_l S_l/\sqrt{c_l}}};$$

whereas if we replace (2.8) in the constraint of problem (2.7) we get:

$$n = \frac{(\sum_{h=1}^{H} N_h S_h \sqrt{c_h}) \sum_{h=1}^{H} W_h S_h/\sqrt{c_h}}{v + \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2}.$$

These passages are fundamental for the development of our topic. The framework we are focusing on is more complex, but the underlying idea is very similar: we aim at finding the optimal sampling procedure that minimizes the variance of the estimator given a budget constraint or the cost function given a precision of the estimator, or the risk function obtained as a combination of cost and estimator variance according to the *combined optimization approach* described previously.
If we are interested in finding an optimal sampling procedure for stratified sampling it is necessary to know the *strata variances* in formula (2.5) to compute the optimal allocation that leads to the minimum value of the estimator variance. If they are not know, we have to move to an *adaptive sampling* context, which allows to estimate them proceeding by steps. We will see some adaptive designs in Section 2.5 and our sampling proposals in Chapter 3 and 4.

2.5 Adaptive designs

Both *adaptive sampling* and *sequential sampling* belong to the class of *adaptive designs*. In such designs, the units are added to the sample along different steps and the information gained at each step plays an important role to build the later steps. In *sequential sampling* one unit (*purely sequential sampling*) or a bunch of units (*group sequential sampling*) is added at each step and provides information to decide whether or not stopping sampling, whereas in *adaptive sampling* the information gained at each step is essential to decide which units are chosen in the next step and/or how they are allocated in the strata

(*adaptive stratified sampling*). In both sampling procedures the sample size is not fixed a priori. Further details about group sequential sampling with adaptive allocation rule for infinite population can be found in Denne and Jennison [2000], Müller and Schäfer [2001], Posch et al. [2005], Morgan and Stephen Coad [2007] who have explored them in the context of clinical trials.

2.5.1 Adaptive sampling in the 'design based' approach

Adaptive cluster sampling

In this section we briefly introduce adaptive cluster sampling. A complete reference is Thompson and Seber [1996].
For this sampling design the idea of neighborhood is very important. A neighborhood is defined for each unit in a population of size N in such a way that it has objective characteristics (i.e. if the units are land areas a neighborhood of a unit consists of all the units adjoining its borders). A simple random sample s_0 including n_0 units is selected. If the value y_i of the variable of interest for unit $i \in s_0$ satisfies a condition C, the units in its neighborhood are selected. If any other units of that neighborhood satisfies C also their neighborhoods are included in the sample. We continue this procedure until a cluster of units is obtained; the units in the border on this cluster do not satisfy condition C and they are called *edge* units. The final sample consists of n_0 clusters, not necessary different. If a unit in the initial sample does not satisfy condition C, its neighborhood is not selected, and it forms a cluster of size 1. The following quantities are defined:
A_i: cluster generated by unit i with its *edge* units removed (a selection of any unit in A_i leads to the selection of all A_i);
m_i: number of units in A_i;
a_i: number of units in networks where unit i is an edge unit;
α_i: probability that unit i is included in the sample;
α_{ij}: probability that both units i and j are included in the sample.
The probability α_i is defined as follows:

$$\alpha_i = 1 - \left[\binom{N - m_i - a_i}{n_1} / \binom{N}{n_1} \right]$$

The adaptive cluster estimator $\widehat{\mu}$ for the population mean is

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{\nu} \frac{y_i}{\alpha_i} = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{y_k^*}{\alpha_k},$$

where $y_1, y_2, ..., y_\nu$ are the y-values of the ν distinct units in the final sample, y_k^* is the sum of the y-values for the k_{th} network, κ is the number of distinct networks and α_k is

the probability that the initial sample intersects the k-th network. If there are x_k units in the k_{th} network:

$$\alpha_k = 1 - \left[\binom{N - x_k}{n_1} / \binom{N}{n_1} \right]$$

$$\alpha_{jk} = 1 - \left[\binom{N - x_j}{n_1} + \binom{N - x_k}{n_1} - \binom{N - x_j - x_k}{n_1} \right] / \binom{N}{n_1}$$

$$\widehat{var}(\widehat{\mu}) = \frac{1}{N^2} \left[\sum_{j=1}^{\kappa} \sum_{k=1}^{\kappa} \frac{y_j^* y_k^*}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \right],$$

where $\widehat{var}(\widehat{\mu})$ is the estimated variance of the mean estimator $\widehat{\mu}$.
Adaptive cluster sampling is an example of a pure adaptive design: the selected units give information on the units to be chosen consequently. If the selected units provides information only on the numbers of units to be later selected we are dealing with a sampling design with adaptive allocation.

Adaptive allocation in stratified sampling

In conventional stratified sampling, the optimal allocation (Neyman allocation) depends on population variances (see formula (2.5)) that are generally unknown. A practical recommendation is to substitute them with sample variances computed from past data or from a pilot survey. Another approach is to develop the survey in steps and estimate the variances from initial steps, thus approximating the optimal allocation based on prior knowledge of the actual variances.
Cox [1952], Sandvik et al. [1996], Mukhopadhyay [2005] are some of the authors that have considered the problem of sample allocation with unknown strata variances and proposed different kinds of two steps sampling. However, these procedures do not admit design unbiased estimates of the parameters.
Thompson and Seber [1996, pp 189-191] suggested a sequential approach in k phases (phases are considered sampling steps). At the kth phase, a complete stratified random sample is selected, with sample sizes possibly depending on data from previous phases. Then the traditional stratified estimator for the total, call it $\widehat{\tau}_k$, based on the data from the kth phase, results to be unbiased for the population total τ. Therefore the weighted average

$$\widehat{\tau}_w = \sum_{k=1}^{K} w_k \widehat{\tau}_k,$$

where $w_k \geq 0$ denotes a given weight, such that $\sum_{k=1}^{K} w_k = 1$, is an unbiased estimator of the population total τ.
$\widehat{\tau}_w$ is unbiased because each of the estimators $\widehat{\tau}_k$ is design unbiased and the weights w_k are fixed in advance, i.e. they do not depend on observations during the survey. However,

stratum boundaries as well as sample sizes can change from phase to phase, being based on observations in the previous phase of the survey. Thompson and Seber [1996] also proved that the variance of the total $\widehat{\tau}_w$ can be estimated by using the conventional estimator. Carfagna et al. [2008] applied the approach proposed by Thompson and Seber with just two steps, i.e. K=2. In the first phase, they selected a complete stratified random sample of polygons with probability proportional to stratum size. The number of sample units n_1 selected in the first phase must be fairly small because the unique aim of this first sample is to estimate the standard deviation of the strata necessary to compute the Neyman's allocation. At the same time, n_1 has not to be too small since the estimates of the standard deviations must be reliable, in fact they are not updated like it is done in the sequential procedure. Once the standard deviations in the different strata are estimated, the Neyman's allocation with sample $n_1 + n_2$ is computed, where $n_1 + n_2$ is the sample size required for reaching a specified standard deviation S of the estimate of the total (Cochran [1997], formula 5.50):

$$n_1 + n_2 = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{S^2 + \sum_{h=1}^{H} N_h s_h^2}, \tag{2.9}$$

where s_h^2 is the estimate of the population variance for stratum h computed at the first step, H is the total number of strata and N_h is the population units in stratum h.
First of all, Carfagna et al. [2008] assigned the same weights to the estimators in the two phases (weights independent of the observations), which result into an unbiased estimator of the total. Next, they assigned weights proportional to the sample size (n_1 and n_2). Finally, they found the optimal weights that minimize the variance of $\widehat{\tau}_k$ under the assumption that sampling is independent in the two phases. These weights are

$$w_{1opt} = \frac{Var(\widehat{\tau}_2)}{Var(\widehat{\tau}_1) + Var(\widehat{\tau}_2)}, \quad w_{2opt} = \frac{Var(\widehat{\tau}_1)}{Var(\widehat{\tau}_1) + Var(\widehat{\tau}_2)}.$$

Therefore, it turns out that the higher the variance of the estimator of phase one, the higher is the weight assigned to phase two estimator. In this way the weights depend on the observations, making the estimator biased but efficient.

Since the Thompson and Seber's method requires the selection of a complete stratified random sample at each phase, it results not to be very efficient. Hence, Carfagna [2007] proposed a two steps selection procedure with permanent random number (TSPRN) in order to develop a more efficient design which allocates the sample units of the second step only in those strata where supplementary selection is needed. Indeed, the TSPRN assigns a random number to each unit in each stratum, then, the units are ordered according to the associated number and this order corresponds to the selection order. At the first

step, a subset of units is selected in each stratum. At the second step, a group of units in each stratum is added to the sample, following the selection order fixed at the beginning of the procedure. The size of the group of units selected at the second step depends on the results obtained during the first selection. Thus, only one selection is performed and the set of selected units in each stratum can be considered as a random sample without replacement (Ohlsson [1995]); clearly, there is no need to select at least two units from each stratum at each step. Consequently, the TSPRN results to be more efficient than the procedure proposed by Thompson and Seber, as shown in Carfagna [2007]. At the first step the sample units n_1 are allocated proportionally to each stratum getting n_{h1}, for $h = 1, ..., H$; the standard deviations of the strata are then calculated to compute Neyman's allocation with total sample size $n_1 + n_2$ obtained by formula (2.9). In some strata, the optimum sample size (n_h) can be less than or equal to the number of units already controlled (n_{h1}) In such cases, no other sample units are sampled; otherwise, n_{h2} sampling units are controlled: $n_{h2} = n_h - n_{h1}$ if $n_h - n_{h1} \geq 0$, otherwise $n_{h2} = 0$. This implies that the effective sample size will be larger that the one computed by formula (2.9).
Now suppose that we are interested in estimating the population mean $\overline{Y}$ through the stratified mean estimator $\overline{y}_{stT}$ generated by the TSPRN procedure, that is

$$\overline{y}_{stT} = \sum_{h=1}^{H} \frac{N_h}{N} \sum_{i=1}^{n_{h1}+n_{h2}} \frac{y_{hi}}{n_{h1} + n_{h2}}. \tag{2.10}$$

In the TSPRN procedure the total sample size is not based on the values assumed by the estimator $\overline{y}_{stT}$, but only on the standard deviation of this estimator, hence $\overline{y}_{stT}$ is unbiased. Indeed, using permanent random numbers, the selection of the units at the second step is independent from the results obtained in the first step and the selection is considered a simple random sampling without replacement. The sample size of each stratum in the second step is the unique element that depends on the standard deviations of the strata calculated in the previous steps. Thus, the conditional expectation of $\overline{y}_{stT}$ at the second selection, given the data d_1 collected at the first step, is:

$$E[\overline{y}_{stT}|d_1] = E[\overline{y}_{stT}|n_{11} + n_{12}, ..., n_{H1} + n_{H2}],$$

which is unbiased for $\overline{Y}$. Thus the unconditional expectation is also unbiased for $\overline{Y}$.
Likewise, the unconditional variance of $\overline{y}_{stT}$ is:

$$Var[\overline{y}_{stT}] = E[Var(\overline{y}_{stT}|n_{11} + n_{12}, ..., n_{H1} + n_{H2})] + Var[E(\overline{y}_{stT}|n_{11} + n_{12}, ..., n_{H1} + n_{H2})],$$

but $E(\overline{y}_{stT}|n_{11} + n_{12}, ..., n_{H1} + n_{H2})$ is the constant $\overline{Y}$, having variance equal to zero, so that $Var[\overline{y}_{stT}] = E[Var(\overline{y}_{stT}|n_{11} + n_{12}, ..., n_{H1} + n_{H2})]$, which, for stratified random

sampling, is:

$$E[Var(\overline{y}_{stT}|n_{11}+n_{12},...,n_{1H}+n_{2H})] = E\left[\sum_{h=1}^{H}\frac{N_h}{(n_{h1}+n_{h2})}\frac{N_h-(n_{h1}+n_{h2})}{N^2}S_h^2\right],$$

where S_h^2 is the population variance for stratum h. The term inside the expected value is estimated by:

$$\widehat{Var}(\overline{y}_{stT}|n_{11}+n_{12},...,n_{H1}+n_{H2}) = E\left[\sum_{h=1}^{H}\frac{N_h}{(n_{h1}+n_{h2})}\frac{N_h-(n_{h1}+n_{h2})}{N^2}s_h^2\right],$$

where s_h^2 is the sample variance for stratum h computed at the second step.
Therefore, an unbiased estimator of $Var(\overline{y}_{stT})$ is:

$$\widehat{Var}(\overline{y}_{stT}) = \sum_{h=1}^{H}\frac{N_h}{n_{h1}+n_{h2}}\frac{N_h-(n_{h1}+n_{h2})}{N^2}\frac{\sum_{i=1}^{n_{h1}+n_{h2}}\left(y_{hi}-\frac{\sum_{i=1}^{n_{h1}+n_{h2}}y_{hi}}{n_{h1}+n_{h2}}\right)^2}{(n_{h1}+n_{h2})-1}.$$

Adaptive sequential stratified sampling with adaptive allocation

Carfagna and Marzialetti [2009b] went further, proposing an adaptive sequential procedure with the use of permanent random numbers (ASPRN). When the population is divided into strata it is possible to assign a random number to each unit in each stratum and order the units according to the associated number; this order corresponds to the selection order. The procedure continues selecting a first stratified random sample with probability proportional to stratum size, including at least two sample units per stratum and estimating the variability inside each stratum. In case in one stratum the estimated variability is zero, the ASPRN assigns to this stratum the variance estimated in the stratum with the lowest positive variance. Then, Neyman's allocation is computed with sample size n+1 and one sample unit in the stratum with the maximum difference between actual and Neyman's allocation is selected. Then the parameter of interest and its precision are estimated. If the precision is acceptable, the process stops; otherwise, Neymas' s allocation is computed with sample size n+2, and so on, until the precision considered acceptable is reached. Due to the use of permanent random numbers, the sample size per stratum depends on the previously selected units, whereas the sample selection does not; thus the ASPRN allows design unbiased and efficient estimates of the parameters of interest.
Suppose we are interested in estimating the population mean $\overline{Y}$ through the stratified

mean estimator $\overline{y}_{stA}$, obtained by the AGSPRN procedure, that is

$$\overline{y}_{stA} = \sum_{h=1}^{H} \frac{N_h}{N} \sum_{i=1}^{n_{hk}} \frac{y_{hi}}{n_{hk}}, \tag{2.11}$$

where n_{hk} is the sample size in stratum h at the k-th selection, for $h = 1, ..., H$. As in TSPRN, the estimator $\overline{y}_{stA}$ in (2.11) is unbiased for $\overline{Y}$ since, in the proposed sequential procedure, the stopping rule is not based on values assumed by $\overline{y}_{stA}$ but only on its standard deviation. Moreover, the estimates of the strata variances that are updated at each step affect only the sample size of the different strata, not the sample selection in each stratum. This is due to the use of the permanent random numbers (Ohlsson [1995]). Let d_{k-1} denotes the data collected through the sequential procedure before the k-th selection. The data d_{k-1} affect only the strata allocations $n_{1k}, ..., n_{Hk}$ at the k-th step, they do not affect the selection of the units which is independent because of the use of permanent random numbers. Thus, the conditional expectation of $\overline{y}_{stA}$ at the k-th selection, given the information collected in previous selections, is:

$$E\left[\overline{y}_{stA}|d_{k-1}\right] = E\left[\overline{y}_{stA}|n_{1k}, ..., n_{Hk}\right].$$

Given the sample size n_{hk}, $\overline{y}_{stA}$ at the k-th selection is, under stratified random sampling, unbiased for $\overline{Y}$. Therefore the unconditional expectation of $\overline{y}_{stA}$ is also unbiased for $\overline{Y}$. The stopping rule is not based on the value of $\overline{y}_{stA}$ but on its standard deviation, thus the last selection of the sequential procedure does not differ from the k-th selection. This can allow us to conclude that $\overline{y}_{stA}$, under stratified random sampling, is an unbiased estimator for $\overline{Y}$.
The unconditional variance of $\overline{y}_{stA}$ is:

$$Var[\overline{y}_{stA}] == E[Var(\overline{y}_{stA}|n_{1k}, ..., n_{Hk})] + Var[E(\overline{y}_{stA}|n_{1k}, ..., n_{Hk})],$$

but $E(\overline{y}_{stA}|n_{1k}, ..., n_{Hk})$ is the constant $\overline{Y}$, having variance equal to zero, so that $Var[\overline{y}_{stA}] = E[Var(\overline{y}_{stA}|n_{1k}, ..., n_{Hk})]$, which, with stratified random sampling, is:

$$E[Var(\overline{y}_{stA}|n_{1k}, ..., n_{Hk})] = E\left[\sum_{h=1}^{H} \frac{N_h}{n_{hk}} \frac{N_h - n_{hk}}{N^2} S_h^2\right],$$

where S_h^2 is the population variance for stratum h. An estimator for the variance of $\overline{y}_{stA}$ is:

$$\widehat{Var}(\overline{y}_{stA}) = \sum_{h=1}^{H} \frac{N_h}{n_{hk}} \frac{N_h - n_{hk}}{N^2} \frac{\sum_{i=1}^{n_{hk}} \left(y_{hi} - \frac{\sum_{i=1}^{n_{hk}} y_{hi}}{n_{hk}}\right)^2}{n_{hk} - 1}.$$

The permanent random numbers (PRN) technique

Ohlsson [1995] proves that the use of Permanent Random Numbers (PNR) in sampling from a population yields to a simple random sample without replacement. Indeed, if we associate to each unit of the population a random number, uniformly distributed over the interval $(0, 1)$ and if we order them in an ascending order, the first n units of the list constitute a simple random sampling without replacement of size n.
Following Ohlsson [1995], let us denote with $Pr(s)$ the probability of extracting the sample s of size n from a population of size N through the PRN technique. We shall prove that

$$Pr(s) = \frac{1}{\binom{N}{n}}. \tag{2.12}$$

Let X_n be the largest of the PRN corresponding to the last unit nth of the sample s and let $f(x)$ be the pdf of X_n. Conditioning on the outcome of $X_n = x$ we get:

$$Pr(s) = \int_0^1 P(s|X_n = x)f(x)dx.$$

The probability of s conditioned on the value x of X_n is equal to the probability of observing $N - n$ units with values greater than x, hence we obtain:

$$Pr(s) = \int_0^1 (1-x)^{N-n} n x^{n-1} dx = nB(n, N-n+1) = n\frac{(n-1)!(N-n)!}{N!}, \tag{2.13}$$

where $B(.,.)$ is the Beta function. Result in (2.13) is equal to the right term in (2.12).

Comparison between different adaptive sampling procedures

Carfagna and Marzialetti [2009b] introduced a comparison between TSPRN, ASPRN and other sampling procedures, dealing with the estimator of A, that is the percentage of the area correctly photo-interpreted, a quality index introduced in Section 1.2.1. It is the main element of our application described in Chapter 5 and it will be defined in details in Section 5.4. They showed that for the AGSPRN procedure, the standard deviation of the estimate of A decreases as the sample size increases, but this decrease is not strictly monotone. They also reported that as the sample size increases, the estimator tends to converge to the real value in the population. Moreover, the ASPRN procedure is more efficient than the stratified sampling with proportional allocation and fixed sample size, requiring less sample units to achieve the same standard deviation of the estimate.
Carfagna and Marzialetti [2008] analysed also the behavior of Cohen's Kappa estimator with an adaptive sequential procedure (ASPRN). Cohen's Kappa is an agreement index which discounts the total proportion of agreement (p_0) by the level of agreement expected by chance (p_c). It is based on a matrix in which, in the case of land quality control, we

insert the area of polygons classified to class i by the photo-interpreter and to class j by the controller, divided by the total area.
Cohen's Kappa is:

$$\kappa = \frac{p_o - p_c}{1 - p_c} = \frac{\sum_i p_{ii} - \sum p_{i+} \cdot p_{+i}}{1 - \sum p_{i+} \cdot p_{+i}}$$

where p_{ii} is the percentage of the total area photo-interpreted as the same land cover type "i", while p_{i+} and p_{+i} are the percentages of the total area classified in the class "i"respectively by the photo-interpreter and the controller.
When the sample size is sufficiently large, it is possible to compute the standard deviation of Cohen's Kappa using a large sample variance of the estimate:

$$Var(\kappa) = \frac{1}{n(1-p_c)^2} \left(\sum_i p_{ii} \left\{ 1 - (p_{i+} + p_{+i})(1-\kappa) \right\}^2 + (1-\kappa)^2 \sum_{i \neq j} p_{ij} (p_{i+} + p_{+j})^2 - \left\{ \kappa - p_c(1-\kappa) \right\}^2 \right).$$

Carfagna and Marzialetti [2008] found out that the behaviour of the estimate of Cohen's Kappa for increasing sample size is less volatile with ASPRN than with proportional allocation. The ASPRN and the proportional allocation procedures result into very similar values of the standard deviation as the sample size increases. The same holds for the coefficient of variation. The reason can be that the formula for computing the standard deviation of Cohen's Kappa cannot take the advantage of optimal allocation of the sample units in the strata, since it does not take into account the stratification. Moreover, the standard deviation of Cohen's Kappa is based on large sample properties, hence a high pilot sample size is required to increase the performance of the ASPRN procedure. Carfagna and Marzialetti [2008] analyzed the behavior of the Cohen's Kappa's standard deviation for large sample size, hence the conclusion should be reliable.
Carfagna and Marzialetti [2009b] compared the ASPRN with TSPRN finding out that the ASPRN is more efficient than TSPRN because the former allows to obtain the same precision of the estimates sampling less units. However, if a budget constraint is introduced, the ASPRN becomes less efficient than TSPRN because the selection of each sample unit is a step in the process, with the consequent cost (Carfagna et al. [2012]). This motivates the study developed in Chapter 3, where we will investigate the possibility of defining an adaptive procedure with a optimum number of steps K, that is a compromise between the TSPRN and the ASPRN. It should reduce the costs incurred by the ASPRN and, at the same time, it should preserve the ASPRN capability to generate a sample allocation very close to Neyman's one.

2.5.2 Adaptive designs in the 'model based' approach

Optimal two phases strategy for known population model

Thompson and Seber [1996, pp 237-239] showed that when the population model is known, the optimal adaptive procedure is always better or as good as the best nonadaptive method. In the 'model based' approach, $\mathbf{y}$ is a realization of a random vector $\mathbf{Y}$ having a density function $f(\mathbf{y}, \boldsymbol{\phi})$ which is assumed to be known exactly; $\boldsymbol{\phi}$ can have a prior distribution. The ideal adaptive two-phase design is obtained in two steps. The first step presents a selection of units s_1, whereas in the second steps units s_2 are sampled in such a way that the following function is minimized:

$$g_{s_2}(s_1, \mathbf{y}_{s_1}) = E\left\{\left[t(s_1, \mathbf{y}_{s_1}, s_2, \mathbf{Y}_{s_2}) - \eta(\mathbf{Y})\right]^2 | s_1, \mathbf{y}_{s_1}\right\} = \int (t-\eta)^2 f(\mathbf{y}_{\bar{s}_1} | s_1, \mathbf{y}_{s_1}; \boldsymbol{\phi}) d\mathbf{y}_{\bar{s}_1}$$

where $\bar{s}_1$ is the vector of units not in the first sample, $\mathbf{y}_{s_1}$ are the y-values of the units s_1, $f(\mathbf{y}_{\bar{s}_1} | s_1, \mathbf{y}_{s_1}; \boldsymbol{\phi})$ is the conditional density, given the initial sample, of the components of $\mathbf{y}$ not in s_1, t is an estimator and $\eta = \eta(\mathbf{Y})$ is the population target we want to estimate. The conditional mean-square error for the optimal choice is then

$$min_{s_2} g_{s_2}(s_1, \mathbf{y}_{s_1}) = min_i E\left\{\left[t(s_1, \mathbf{y}_{s_1}, s_2, \mathbf{Y}_{s_2}) - \eta)\right]^2 | s_1, \mathbf{y}_{s_1}, s_2 = i\right\}$$

where i refers to the integer that identifies one of the possible sample in S_2, the countable set of all possible second-phase samples.
Thus, starting with initial sample s_1 and taking the optimal choice for s_2 (say $s_2 = j$), the overall mean-square error is

$$E[(T-\eta)^2 | s_1, s_2 = j] = \int min_{s_2} g_{s_2}(s_1, \mathbf{y}_{s_1}) f(\mathbf{y}; \boldsymbol{\phi}) d\mathbf{y}.$$

The best nonadaptive design, on the other hand, will select s_2 to minimize the mean-square error without taking the first-phase observations $\mathbf{y}_{s_1}$ into account. If this occurs at $s_2 = k$, the mean-square error for the optimal conventional design is

$$E[(T-\eta)^2 | s_1, s_2 = k] = min_{s_2} \int g_{s_2}(s_1, \mathbf{y}_{s_1}) f(\mathbf{y}; \boldsymbol{\phi}) d\mathbf{y}.$$

We can compare this with the adaptive design as given above by noting that, by a basic property of integration,

$$\int min_{s_2} g_{s_2}(s_1, \mathbf{y}_{s_1}) f(\mathbf{y}; \boldsymbol{\phi}) d\mathbf{y} \leq min_{s_2} \int g_{s_2}(s_1, \mathbf{y}_{s_1}) f(\mathbf{y}; \boldsymbol{\phi}) d\mathbf{y}.$$

Hence, the optimal adaptive procedure will always be as good as or better than the best nonadaptive procedure.

2.5.3 Adaptive designs in the infinite population context

Fixed-width confidence intervals and two-stage procedures

In an infinite population context, we assume that $y_1, y_2, ...$ is a sequence of independent and identically distributed (i.i.d) $N(\mu, \sigma^2)$ random variables where both parameters are unknown with $\mu \in (-\infty, +\infty), \sigma \in (0, +\infty)$. The aim is to construct an interval I for μ of fixed length $2d$ and such that $P_{\mu,\sigma}(\mu \in I) \geq 1 - \alpha$, with $\alpha \in (0, 1), d(> 0)$. Ghosh et al. [1997, Section 3.7] showed that any fixed sample size procedure does not exist as a solution for this problem. Stein [1945] referred to the two-stage procedures giving more mathematical and statistical foundations.
Given the sample $\mathbf{y}_s = (y_1, ..., y_n)'$, the confidence interval for μ is $I_n = [\overline{y}_n \pm d]$ where $\overline{y}_n$ is the sample mean. $P_{\mu,\sigma}(\mu \in I) = 2\Phi(n^{1/2}d/\sigma) - 1$, which would be at least $(1 - \alpha)$ if we choose n to be the smallest integer greater than or equal to $a^2\sigma^2/d^2$, where $\Phi(a) = 1 - (1/2)\alpha$. If σ would be known, $C = a^2\sigma^2/d^2$ is also known and the solution would be easy. But in this case it is not and we require a two-stage procedure: in the first phase $m(\geq 2)$ units, $y_1, ..., y_m$, are drawn and let $\overline{y}_m = m^{-1}\sum_{i=1}^{m} y_i$, $S_m^2 = (m-1)^{-1}\sum_{i=1}^{m}(y_i - \overline{y}_m)^2$. We define:

$$N = N(d) = max\left\{m, \left[\frac{a_{m-1}^2 S_m^2}{d^2}\right]^* + 1\right\},$$

where a_{m-1} is the upper $(\alpha/2)$th quantile of the Student t distribution with $(m-1)$ degrees of freedom and $[x]^*$is the larger integer smaller than x.
If $N = m$ the process stops and no more units will be added, otherwise $(N - m)$ units are sampled at the second stage. The interval $I_N = [\overline{y}_N \pm d]$ based on all N units sampled is proposed for μ and the following properties are satisfied:
(i) $P_{\mu,\sigma}\{\mu \in I_n\} \geq 1 - \alpha \ for \ all \ \mu \ and \ \sigma^2$;
(ii) $\frac{a_{m-1}^2\sigma^2}{d^2} \leq E_{\mu,\sigma}(N) \leq m + \frac{a_{m-1}^2\sigma^2}{d^2}$;
(iii) $\lim_{d\to 0} E_{\mu,\sigma}(\frac{N}{C}) = \frac{a_{m-1}^2}{a^2}$;
(iv) $\lim_{d\to 0} P_{\mu,\sigma}\{\mu \in I_n\} = 1 - \alpha \qquad for \ all \ \mu \ and \ \sigma^2$.

Sequential fixed-width interval estimation

The two-stage procedure proposed by Stein was extended by Ray [1957] who was the first to suggest a sequential procedure to obtain a fixed-width interval estimator for estimating the mean of a normal distribution when σ is unknown. Assuming the same setting of the previous paragraph, where $y_1, y_2, ...$ is a sequence of independent and identically distributed (i.i.d) $N(\mu, \sigma^2)$ random variables with both parameters unknown, $\mu \in (-\infty, +\infty), \sigma \in (0, +\infty)$. If σ^2 is known the optimal sample size to reach the

prefixed value $1-\alpha$ for the confidence interval of μ would be the smallest integer such that $n \geq \sigma^2 a^2/d^2 = n^*$, where a is the upper $100(\alpha/2)\%$ quantile of $N(0,1)$ and d is the length of the semi-confidence interval for μ.
Ray proposed a stopping rule, when σ^2 is not known:

$$N^* = \texttt{smallest odd integer } n \geq n_0 (\geq 3) \texttt{ for which } n \geq \frac{a^2 S_n^2}{d^2},$$

where S_n^2 is the sample variance and n_0 is the initial sample size. Let us define $\lambda = \sigma^2/d^2$, $C^*(\lambda) = P_{\mu,\sigma^2}(\overline{y}_{N^*} - d \leq \mu \leq \overline{y}_{N^*} + d)$, and $D^*(\lambda) = E_{\mu,\sigma^2}(N^*)$, where $\overline{y}_{N^*}$ is the sample mean when N^* is the sample size. Ray's computations show that as λ becomes large, $C^*(\lambda)$ tends to a value smaller than 0.95.
Chow and Robbins [1965] define another stopping rule:

$$N = \texttt{inf}\{n \geq n_0 : n \geq \frac{a^2 S_n^2}{d^2}\},$$

where n_0 is the initial sample size. This sequential procedure possesses the following properties, if $0 \leq \sigma^2 \leq +\infty$:
(i) $P_{\sigma^2}(N < \infty) = 1$;
(ii) $E_{\sigma^2}(N) \leq n_0 + 1 + n^*$;
(iii) $E_{\sigma^2}(N^2) \neq (n_0 + 1 + n^*)^2 - 2$;
(iv) $N \equiv N(d) \downarrow$ a.s. in d; $N \to \infty$ a.s. as d $\to 0$; $E(N) \to \infty$ as $d \to 0$;
(v) $N/n^* \to 1$ a.s. as d $\to 0$;
(vi) $\lim_{d\to 0} E_{\sigma^2}(N/n^*) = 1$;
(vii) $\lim_{d\to 0} P_{\mu,\sigma}(\overline{y}_N - d \leq \mu \leq \overline{y}_N + d) = 1 - \alpha$.

2.6 The connection between infinite population approach and 'design based' approach in two-steps and sequential adaptive estimation

In this section the properties of the estimators generated by two-steps and sequential adaptive procedures in infinite population context are extended to those estimators obtained through the same procedures with a finite population approach. The main tool that allows this connection is the classical *central limit theorem*(CLT). It usually requires independent and identically distributed random variables, conditions that are usually not satisfied in sampling without replacement. Hence, it is not easy to verify the validity of CLT for the majority of the finite populations sampling schemes. Each case needs a proper analysis and a typical approach to demonstrate the validity of the finite population

central limit theorem is through simulations (Bellhouse [2001]).

2.6.1 Simple random sampling with replacement (SRSWR)

In simple random sampling the values $y_1, y_2, ..., y_n$ of the units in the sample are i.i.d. and they have common distribution $F_N(x) = 1/N \sum_{i=1}^{N} I\{y_i \leq x\}$ which is unknown. The aim is to estimate the population mean $\overline{Y}$. The conditions to apply the classical central limit theorem are here satisfied. Let us suppose that $\overline{y}_n$ is the mean of a sample of size n and a confidence interval for $\overline{Y}$ is $I_d(\overline{y}_n) = (\overline{y}_n - d, \overline{y}_n + d)$. By the CLT the coverage probability of $I_d(\overline{y}_n)$ is approximately

$$CP \cong 2\Phi\left(\frac{d\sqrt{n}}{\sigma_N}\right) - 1,$$

where σ_N is the variance of F_N. Thus, if

$$n \geq \frac{\sigma_N^2 a^2}{d^2},$$

the coverage probability is approximately 1-α, with a the upper (1-$\alpha/2$) quantile of a standardized Normal distribution.
When σ_N^2 is not known, the Chow-Robbins sequential procedure can be used, starting from a sample size $m(\geq 2)$ and then calculating the sample size at each step through the sample variance based on n observations, that is:

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y}_n)^2.$$

The sequential stopping variable is:

$$N(d) = min\left\{n : n \geq m, n \geq \frac{s_n^2 a^2}{d^2}\right\}.$$

The Chow-Robbins sequential procedure is asymptotically consistent and efficient, as $d \to 0$. Zacks [2009, pp 82–83] showed by a simulation study that this holds also in a finite population context.
Stein's two-stage procedure may often be more convenient: at stage one, m units are sampled with a SRSWR procedure, computing s_m^2 and

$$N_m = \left[\frac{a^2 s_m^2}{d^2}\right]^* + 1.$$

If $N_m \leq m$ the procedure stops and the confidence interval of $\overline{Y}$ is $(\overline{y}_m - d, \overline{y}_m + d)$; however, if $N_m > m$, we go to the second step where $N_m - m$ units are sampled and, combining the two samples, we can compute the confidence interval $(\overline{y}_{N_m} - d, \overline{y}_{N_m} + d)$.

2.6.2 Simple random sampling without replacement (SRSWOR)

In SRSWOR the random variables $y_1, y_2, ..., y_n$ are neither independent nor identically distributed, but if the sample size n is very small respect to the population size N ($n/N < 0.1$), some properties of the sample mean like convergence in distribution to the normal might still approximately hold. Erdős and Rényi [1959] and Hájek [1960] established necessary and sufficient conditions for the validity of central limit theorem in simple random sampling without replacement from finite populations.
In SRSWOR the variance of $\overline{y}_n$ is $(S^2/n)(1 - n/N) = S^2(1/n - 1/N)$, where S^2 is the population variance. Thus if we want to apply Stein's two-stage procedure with s_m^2 the variance of the sample drawn in the first step, the stopping variable is:

$$N_m = min\left\{\left[N\frac{a^2 s_m^2}{Nd^2 + a^2 s_m^2}\right]^* + 1, N\right\}.$$

In the second step, we draw $(N_m^* - m)$ units following a SRSWOR procedure and we calculate $\overline{y}_n$ averaging the two combined samples:

$$\overline{y}_n = \frac{m\overline{y}_m^{(1)} + (N_m - m)\overline{y}_{N_m - m}^{(2)}}{max\{m, N_m\}},$$

where $\overline{y}_m^{(1)}$ is the mean of the first sample and $\overline{y}_{N_m-m}^{(2)}$ is the mean of the second sample. The conditional expectation of $\overline{y}_n$ given the first sample y_{s_1} is:

$$E[\overline{y}_n | y_{s_1}] = \overline{Y} P\{N_m > m\} + E[\overline{y}_m^{(1)} I\{N_m \leq m\}].$$

If

$$P\{N_m \leq m\} = P\left\{s_m^2 \leq \frac{Nmd^2}{(N-m)a^2}\right\}$$

is small, then $\overline{y}_n$ is approximately unbiased.
An adaptive sequential procedure can be also applied by starting from a sample size equal to m and then adding one unit at each step to calculate the sample variance of the combined sample. The stopping variable is:

$$N^*(d) = min\left\{n : n \geq m, \frac{Na^2 s_n^2}{Nd^2 + a^2 s_n^2} \leq n\right\}.$$

Zacks [2009, pp 86] showed by a simulation study that the two-steps procedure with SRSWOR is more efficient than that with SRSWR. Moreover, he also showed that the sequential procedure is not more efficient that the two-steps one.

2.6.3 Stratified simple random sampling

It is possible to apply Stein's two-stage procedure to stratified simple random sampling. Bickel and Freedman [1984] studied the conditions required to apply the classical central limit theorem to stratified random sampling from finite populations.
Knowing the population weights W_h of the strata ($h = 1, ..., H$), at step one m units are allocated proportionally to the strata, that is $n_{h1} = mW_h$. The sample variances of each stratum $s^2_{n_{h1}}$ are then calculated; they are unbiased estimates of the strata population variances S^2_h. The number of units that have to be sampled to obtain a fixed length $2d$ of the confidence interval for the population mean $\overline{Y}$ is:

$$N_m = \left[N \frac{a^2 \sum_{h=1}^{H} W_h s^2_{n_{h1}}}{Nd^2 + a^2 \sum_{h=1}^{H} W_h s^2_{n_{h1}}} \right]^* + 1.$$

If $N_m \leq m$ the process stops and the population mean $\overline{Y}$ is estimated by $\overline{y}_m \pm d$, where $\overline{y}_m = \sum_{h=1}^{H} W_h \overline{y}_{n_{h1}}$. However, if $N_m > m$ the remaining units $N_m - m$ are allocated among the strata according to the Neyman's rule:

$$n_h^{(2)} = (N_m - m) \frac{W_h^{(2)} s^2_{n_{h1}}}{\sum_{j=1}^{H} W_j^{(2)} s^2_{n_{j1}}}$$

where $W_h^{(2)} = (N_h - n_{h1})/(N - m)$. The estimate of $\overline{Y}$ is $\widehat{\overline{y}}_n \pm d$, where

$$\widehat{\overline{y}}_n = \sum_{h=1}^{H} W_h \frac{n_{h1} \overline{y}_{n_h^{(1)}} + n_{h2} \overline{y}_{n_h^{(2)}}}{n_{h1} + n_{h2}}. \tag{2.14}$$

$\overline{y}_{n_{h1}}$ and $\overline{y}_{n_{h2}}$ are respectively the mean of stratum h of the first-step and the second-step samples, with $h = 1, ..., N$.

Theorem 2.6.1. *Zacks [2009, pp 89]*
Under the two-steps stratified sampling $\widehat{\overline{y}}_n$ is unbiased, that is:

$$E[\widehat{\overline{y}}_n] = \overline{Y} \quad \texttt{for all values of} \quad \overline{Y}.$$

Proof: Denote with $\mathcal{F}_1$ the σ-field generated by $(\overline{y}_{n_{h1}}, S^2_{n_{h1}}, h = 1, ..., H)$. Notice that, given n_{h2}, the second step sample is a simple random sample $(N_h - n_{h2})$ for each stratum.

The conditional expected value of $\overline{y}_{n_{h2}}$ given $\mathcal{F}_1$ is:

$$E[\overline{y}_{n_{h2}}|\mathcal{F}_1] = \frac{N_h\overline{Y}_h - n_{h1}\overline{y}_{n_{h1}}}{N_h - n_{h1}}, \tag{2.15}$$

where $\overline{Y}_h$ is the population mean in stratum h. Substituting the quantity in (2.15) in (2.14), we obtain:

$$E[\widehat{\overline{y}}_n|\mathcal{F}_1] = \sum_{h=1}^{H} W_h\overline{Y}_{n_{h1}} \quad .$$

Finally, since $E[\overline{Y}_{n_{h1}}] = \overline{Y}_h$, $h = 1, ..., H$, we get:

$$E[\widehat{\overline{y}}_n] = \sum_{h=1}^{H} W_h\overline{Y}_h = \overline{Y}, \quad \texttt{for all values of } \overline{Y}.$$

Zacks [2009, pp90] showed by simulations that the coverage probability (CP) of $I_d(\widehat{\overline{y}}_n)$ is approximately equal to $1-\alpha$ and there is a significant advantage in adopting a stratified random sampling with respect to a simple random sampling.
This is the framework we want to expand in Chapter 3 to a group sequential procedure (slightly different from the Chow and Robbin's one) with the addiction of budget and cost function.

2.7 Summary

In this chapter we have discussed the main aspects of the topic underlying this thesis. We have introduced the finite population sampling situation, presenting two of the most used conventional designs: simple random sampling and stratified random sampling, focusing on the research of the optimal sample size when costs are considered and a prefixed level of the estimator variance is desired. We have explored some adaptive designs, e.g. adaptive cluster sampling, dealing particularly with adaptive allocation in stratified sampling. We discussed the k-phases adaptive sampling procedure proposed by Thompson and Seber [1996], a two steps adaptive procedure with permanent random numbers (TSPRN) introduce by Carfagna [2007] and a sequential adaptive procedure with permanent random numbers (ASPRN) presented by Carfagna and Marzialetti [2009b]. The contribution that we will present in Chapter 3 is an extension of these methods. Moreover, in this chapter we have discussed the adaptive sequential methods for an infinite population context. The two steps procedure of Stein [1945] and the sequential one of Ray [1957] are the main results in this setting. Zacks [2009] have investigated the connection between infinite population approach and finite population one in two steps and adaptive estimation. In Chapter 3 we will extend the existing methods to a group sequential context and we will

add the costs of sampling, proposing an optimal *adaptive group sequential procedure* for finite population, in the presence of a cost function.

Chapter 3

The adaptive group sequential procedure with permanent random numbers (AGSPRN)

3.1 Introduction

In Chapter 2 we have introduced the adaptive sequential procedure with permanent random numbers (ASPRN) and the two steps sequential procedure with permanent random numbers (TSPRN), which are two special and opposite cases of adaptive procedure. The ASPRN procedure consists in selecting one unit per step until the stopping rule is satisfied. The TSPRN procedure involves only two steps, adding, after having selected the pilot sample, all the necessary units at the second step. Moreover, Section 2.5.1 shows that if a cost function with a relevant step cost is introduced, the TSPRN results more efficient than the ASPRN, a point that reshuffles all the conclusions reached before. The intuition at the base of this chapter is that it should exist a compromise procedure between the TSPRN and the ASPRN, able to generates a more efficient estimator in presence of a cost function. Therefore, in the next section we will propose an adaptive group sequential procedure with the use of permanent random numbers (AGSPRN), from which we can derive, as a particular case, the compromise solution we are looking for. The AGSPRN proceeds along K steps, adding a bunch of q units at each step and following the same adaptive allocation rule of TSPRN and ASPRN procedures, which can be derived as particular cases.
We will repeatedly apply the AGSPRN sampling scheme with different q and K to a known simulated population in the presence of a linear cost function and we will show that for some combinations (K, q) it gives rise to more efficient estimators than those generated by the TSPRN, the ASPRN and by other classical sampling procedures. The efficiency is measured in terms of:

1. variance of the generated estimator given a cost function and a budget constraint (Section 3.3);

2. total cost given a threshold on the estimator variance (Section 3.4);

3. a risk function, obtained through a combination of cost and estimator variance (Section 3.5).

Particularly, the results will show that, for each of these three criteria, there is a most efficient estimator generated by the AGSPRN procedure characterized by an optimal number of steps K_{opt} and an optimal number of units q_{opt} added at each step (optimal AGSPRN procedure).
Three Monte Carlo algorithms are provided in order to find the optimal AGSPRN procedure according to criteria 1, 2 and 3. However, these algorithms require as inputs some precise information about the target population. We can use the pilot sample to derive some assumptions, even though it may be not reliable, particularly if its size is small. Hence, in Chapter 4 we will develop a method which allows to update, at each step, the information about the distribution of the variable of interest in each stratum, in order to obtain the optimal AGSPRN procedure. In this chapter we aim at showing some properties of the optimal sampling technique, using as inputs of the Monte Carlo algorithms directly the target population.

3.2 The adaptive group sequential procedure with permanent random numbers (AGSPRN)

In Chapter 2 we have pointed out that if a budget constraint and a cost function with a relevant step cost are introduced, the ASPRN procedure may become less efficient than the TSPRN procedure. Therefore, the precision of the estimator under budget constraint can be increased adding more units per step. This is the reason that leads us to propose an adaptive group sequential procedure with permanent random numbers (AGSPRN). Our objective is to find the optimal number of steps K_{opt} for the AGSPRN procedure and the optimal number of units added per step q_{opt} that allow to preserve both the cost efficiency of the TSPRN procedure and a sample allocation close to Neyman's one that results from the ASPRN procedure. Indeed, the closer the stratum allocation is to the optimal one the higher will be the precision of the estimator.
Therefore, our idea is to preserve the same adaptive sequential scheme of the ASPRN, adding at each step a number of unit q in the strata where the difference between the Neyman allocation and the sequential allocation is positive, in a proportional way.

The adaptive group sequential procedure with permanent random numbers (AGSPRN) is a group sequential procedure which generates a stratified random sample, with an adaptive strata allocation at each step. Given a population of size N that is stratified into H strata of size $N_1, ..., N_H$, for any integer $q \in [1, N-n_0]$, where n_0 is the preliminary or first step sample size, the AGSPRN procedure is developed as following:

(i) assign a random number to each unit in each stratum, then order the units according to the associated number;

(ii) at the first step $[k = 1]$ select a first stratified random sample of size n_0 with proportional allocation, selecting at least two sample units per stratum and estimate the variance inside each stratum;

(iii) compute Neyman's allocation with sample size $n = n_0 + q$ and select q sample units only in the strata with positive difference between Neyman's allocation and the actual one (the allocation is proportional to this difference). Then estimate the parameter of interest and its precision;

(iv) if the stopping rule is satisfied, or the units of the population are all drawn, stop the process; otherwise estimate the strata variances and start again from step (iii) using a sample size equal to $n + q$.

Let us suppose that the parameter of interest is the population mean $\overline{Y}$ of some variable Y. Let K denote the step at which the stopping rule is satisfied. Due to the use of the permanent random numbers, we can adopt the direct expansion stratified estimator in K, that is $\overline{y}_{stK} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_{hK}$, where h refers to the stratum and $\overline{y}_{hK}$ is the sample mean in stratum h after K steps. One can derive some interesting properties of the estimator $\overline{y}_{stK} = \sum_{h=1}^{H} \frac{N_h}{N} \sum_{i=1}^{n_{hK}} y_{ihK}$, where n_{hK} are the units allocated to the stratum h at step K and y_{ihK} is the y-value of unit i in stratum h after K steps, with $i = 1, ..., n_{hK}$ and $h = 1, ..., H$.
Let d_{K-1} denote the data collected through the AGSPRN procedure between the first and the $(K-1)$-th selection. The permanent random numbers assign the selection order to all the sampling units in each stratum before the adaptive selection procedure starts, thus d_{K-1} affects only the allocations $n_{1K}, ..., n_{HK}$ at step K, not the selection of the units. Hence, the conditional expectation of $\overline{y}_{stK}$ at the K-th selection, given the information collected in previous selections, is:

$$E\left[\overline{y}_{stK} | d_{K-1}\right] = E\left[\overline{y}_{stK} | n_{1K}, ..., n_{HK}\right] = \overline{Y}.$$

Given the sample size n_{hK}, $\overline{y}_{stK}$ at the K-th selection is, under stratified random sampling, unbiased for $\overline{Y}$. Therefore, the unconditional expectation of $\overline{y}_{stK}$ is also unbiased for $\overline{Y}$.

Note that the stopping rule is not based on the value of $\overline{y}_{stK}$ but on its standard deviation, thus the last selection of the sequential procedure does not differ from the K-th selection. This allows us to conclude that $\overline{y}_{stK}$, under stratified random sampling, is an unbiased estimator for $\overline{Y}$:

$$E\left[\overline{y}_{stK}\right] = E\{E\left[\overline{y}_{stK}|d_{K-1}\right]\} = E\{E\left[\overline{y}_{stK}|n_{1K},...,n_{HK}\right]\} = E[\overline{Y}] = \overline{Y}. \tag{3.1}$$

The unconditional variance of $\overline{y}_{stK}$ is:

$$V(\overline{y}_{stK}) = E[V(\overline{y}_{stK}|n_{1K},...,n_{HK})] + V[E(\overline{y}_{stK}|n_{1K},...,n_{HK})],$$

but $E(\overline{y}_{stK}|n_{1K},...,n_{HK})$ is the constant $\overline{Y}$, having variance equal to zero, so $V[\overline{y}_{stK}] = E[V(\overline{y}_{stK}|n_{1K},...,n_{HK})]$, which, for stratified random sampling, is:

$$V(\overline{y}_{stK}) = E[V(\overline{y}_{stK}|n_{1K},...,n_{HK})] = E\left[\sum_{h=1}^{H} \frac{N_h^2}{n_{hK}N^2}\frac{N_h - n_{hK}}{N_h}S_h^2\right], \tag{3.2}$$

where S_h^2 is the population variance for stratum h. It is prohibitive to solve expression (3.2) analytically, since it depends on the all possible realizations of the allocations $n_{1K},...,n_{HK}$ at the K-th step.
An estimator of $V(\overline{y}_{stK})$ is:

$$\widehat{V}(\overline{y}_{stK}) = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\frac{N_h - n_{hK}}{N_h n_{hK}}\frac{\sum_{i=1}^{n_{hK}}(y_{ihK} - \overline{y}_{hK})^2}{n_{hK} - 1} \tag{3.3}$$

This result is an extension of that derived by Thompson and Seber [1996, pp 191] for the k-phases adaptive estimator.
Since an AGSPRN procedure is defined for each integer $q \in [1, N - n_0]$, let us denote with $V(\overline{y}_{stK}; K, q)$ and $\widehat{V}(\overline{y}_{stK}; K, q)$ respectively the quantities in (3.2) and (3.3) computed at the K-th step with q units added at each step.
Moreover, if we aim at estimating the percentage P of units that posses some characteristic, the estimator is

$$p_{stK} = \sum_{h=1}^{H}\frac{N_h}{N}\sum_{i=1}^{n_{hK}}\frac{y_{ihK}}{n_{hK}} = \sum_{h=1}^{H}\frac{N_h}{N}\overline{y}_{hK} = \sum_{h=1}^{H}\frac{N_h}{N}p_{hK},$$

where y_{ihK} is the variable which assumes value 1 if the unit i in stratum h possesses the characteristic and 0 otherwise, p_{hK} is the percentage of sample units in stratum h which

possesses the characteristic. Given $n_{1K}, ..., n_{HK}$ an estimate of $V(\overline{y}_{stK})$ is given below:

$$\widehat{V}(\overline{y}_{stK}; K, q) = \widehat{V}(p_{stK}; K, q) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_{hK}}{N_h n_{hK}} s_{hK}^2; \tag{3.4}$$

where $s_{hK}^2 = \frac{n_{hK}}{n_{hK}-1} p_{hK}(1 - p_{hK})$ is the sample variance in stratum h at step K.

To introduce the cost in AGSPRN procedure, let us consider the following linear cost function:

$$C = C(K, q) = C_0 + c_n[n_0 + q(K - 1)] + c_k K, \tag{3.5}$$

where C_0 is the fixed cost, c_n is the cost per unit, c_k is the cost per step and K is the total number of steps performed by the AGSPRN procedure before stopping.

Another important element to be considered is the risk function which combines both criteria of estimator variance and cost for each pair (K, q) characterizing each AGSPRN procedure stopped at the K-th step. Its expression is given by the sum of the expected loss plus cost and when a square loss function is chosen it is equal to:

$$\begin{aligned} R(\overline{y}_{stK}, \overline{Y}_N; K, q) &= E[L(\overline{y}_{stK}, \overline{y}_N; K, q)] + C(K, q) = \lambda MSE(\overline{y}_{stK}, \overline{y}_N; K, q) + C(K, q) = \\ &= \lambda V(\overline{y}_{stK}; K, q) + C(K, q), \end{aligned} \tag{3.6}$$

where λ is a constant used to combine the two criteria $V(\overline{y}_{stK}; K, q) = \Psi_1(\cdot)$ and $\Psi_2(\cdot) = C(K, q)$ according to a comparable scale. For each pair (K, q), $R(\overline{y}_{stK}, \overline{y}_N; K, q)$ takes a defined value. We aim at minimizing it with respect to all the possible combinations (K, q)'s in order to find the optimal AGSPRN procedure. However, this optimization problem is analytically intractable. Hence, we will proceed through a Monte-Carlo study and we will apply the *compound optimization approach* (Section 2.4.3) with the standardized versions $\Psi_{1s}(\cdot)$ and $\Psi_{2s}(\cdot)$ of $\Psi_1(\cdot)$ and $\Psi_2(\cdot)$:

$$\omega\Psi_{1s}(\cdot) + (1 - \omega)\Psi_{2s}(\cdot),$$

where the weights ω and $1 - \omega$ in $(0; 1)$ represent the relative importance of Ψ_{1s} and Ψ_{2s}, with Ψ_{1s}, Ψ_{2s}: $[0; 1] \to [0; 1]$.

The final aim is to find the optimal AGSPRN procedure with an optimal number of steps K_{opt} and units q_{opt} which minimizes:

Case 1. the variance of the estimator in (3.2) given the linear cost function in (3.5) and a

budget constraint (C in (3.5) is given);

Case 2. the cost function in (3.5) given a threshold v on the estimator variance in (3.2);

Case 3. the risk function in (3.6), given thresholds v and c respectively on the estimator variance and on the total cost.

In the following sections we will present the optimization problems related to the three cases and we will provide algorithms in order to find their Monte Carlo solution since they are analytically intractable. However, these algorithms require a quite precise knowledge of the population. First of all, we will apply them directly to a normal population in order to illustrate some properties of the optimal AGSPRN procedure in comparison with TSPRN, ASPRN and other classical sampling procedures. In the next chapter we will provide a reliable and useful method in order to obtain the optimal AGSPRN procedure when precise knowledge of the population is absent, that is the most frequent situation.

3.3 Case 1: minimization of the estimator variance given a budget constraint

To find the optimal AGSPRN procedure which minimizes the estimator variance given a budget constraint the following optimization problem has to be solved:

$$min_{(K,q)} V(\overline{y}_{stK}; K, q) \tag{3.7}$$

subject to

$$C(K, q) = C \quad (\leq C) \tag{3.8}$$

The equality is considered in the cost constraint (3.8); indeed we are supposed to spend all the budget: more units are selected higher is the precision of the estimates. Given C, for each integer $K \in [2, \frac{C-C_0-c_n n_0+c_n}{c_n+c_k}]$ there is a corresponding integer $q = \lfloor (C - C_0 - c_n n_0 - c_k K)/(C_n K - c_n) \rfloor$, where $\lfloor \cdot \rfloor$ indicates the integer part, rounding to the floor.

Let $\mathcal{H}_{(K,q)_c}$ denote the set of constrained pairs of integers:

$$\{(K, q) : q = \lfloor (C - C_0 - c_n n_0 - c_k K)/(C_n K - c_n) \rfloor, \forall \texttt{ integer } K \in [2, \frac{C - C_0 - c_n n_0 + c_n}{c_n + c_k}]\}.$$

Hence, (3.7) reduces to:

$$min_{(K,q) \in \mathcal{H}_{(K,q)_c}} V(\overline{y}_{stK}; K, q)$$

As we explained previously, it is not easy to compute analytically the value of $V(\overline{y}_{stK}; K, q)$ for each $(K, q) \in H_{(K,q)_c}$, therefore the solution should be investigated through Monte

Carlo simulations. Morrison et al. [2008], Salehi and Brown [2010], Salehi et al. [2010] are some of the authors who compared the efficiency of adaptive designs by simulations.

3.3.1 Monte Carlo study

The search of the optimal AGSPRN procedure, the optimal sampling procedure supposed to be offered to the stakeholder, requires to apply a Monte Carlo study to a population that has to be very close to the one under study. However, neither the population values of the variable of interest Y are given nor the distribution of Y is known; just a pilot sample of size n_0 is available. The assumptions about the distribution of Y in each stratum, i.e. $f(y_i|y_i \in h) = f_h(y, \boldsymbol{\theta}_h)$ for $h = 1, ..., H$, are the inputs of Algorithm 1, which is applied to find the optimal AGSPRN procedure in presence of budget constraint. They are derived from the pilot sample or from the study of the phenomenon and they are not consequently updated after the selection of new units. For this reason Algorithm 1 presents some limitations that will be overcome in the next chapter. Indeed, in Chapter 4 we will develop an implementable method to make the search of the optimal AGSPRN more reliable and useful, updating the inputs $f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$, at each sampling step. In fact, if the assumptions on the distribution of Y in each stratum are very close to the real distributions, the application of a sampling procedure is quite senseless, since we would not need to estimate the target population parameters. In contrast, if $f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$, are not precise the need of updating them is relevant.

Algorithm 1: Monte Carlo algorithm to find the optimal AGSPRN procedure in terms of minimum estimator variance given a budget constraint.

Input: $C, C_0, c_n, c_k, H, R, n_0, f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$
Input: N_h, for $h = 1, ..., H$, s.t. $\sum_{h=1}^{H} N_h = N$
for $h = 1, ..., H$ **do**
  draw N_h values from $f_h(y, \boldsymbol{\theta}_h)$
end
foreach $(K, q) \in \mathcal{H}_{(K,q)_c}$ **do**
  for $r = 1, ..., R$ **do**
    apply the AGSPRN procedure and compute $\widehat{V}^r(\overline{y}_{stK}; K, q)$
  end
  compute the Monte Carlo mean
  $\langle V^R(\overline{y}_{stK}; K, q)\rangle = \frac{1}{R}\sum_{r=1}^{R} \widehat{V}^r(\overline{y}_{stK}; K, q) \simeq E(\widehat{V}(\overline{y}_{stK}; K, q)) = V(\overline{y}_{stK}; K, q)$
end
choose (K,q) $\in \mathcal{H}_{(K,q)_c} = argmin_{(K,q)\in\mathcal{H}_{(K,q)_c}} \langle V^R(\overline{y}_{stK}; K, q)\rangle$.

3.3.2 Normal distribution case

In this section we apply the Monte Carlo study to a known population, where the variable of interest is distributed according to a Normal with different parameters in each stratum. This means that the inputs $f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$, of Algorithm 1 are the real distributions and they are not derived from a pilot sample. This seems to make the sampling procedure quite useless. However, the aim of this section is to explore some properties of the optimal AGSPRN procedure and to compare the results with those reported in Chapter 4, where the population is supposed unknown.
The following setting is considered:

(i.) $C = 500, C_0 = 80, c_n = 2, c_k = 4, H = 10, n_0 = 40, W_h \overset{iid}{\sim} U[450, 500]$;

(ii.) $f_h(y, \boldsymbol{\theta}_h) = N[h \times 25 + 250, a_h \times 0.9]$, with h the stratum indicator, a_h the hth element of the vector $a = [600\ 130\ 320\ 250\ 40\ 150\ 100\ 180\ 74\ 400]'$;

(iii.) $R = 1000$. The AGSRPN procedure is applied 1000 times for each pair $(K, q) \in \mathcal{H}_{(K,q)_c}$ where $\mathcal{H}_{(K,q)_c}$ denote the set of constrained pairs of integers:

$$\{(K, q) : q = \lfloor (170 - 2K)/(K - 1) \rfloor, \forall K \texttt{integer} \in [2, 57]\}.$$

These are the inputs of Algorithm 1, that is applied in order to derive the optimal AGSPRN procedure and some of its properties.
Figure 3.1 shows the overall behaviour of the Monte Carlo variance of the mean estimator $\langle V^R(\overline{y}_{stK}; K, q)\rangle$ as the constrained pair $(K, q) \in \mathcal{H}_{(K,q)_c}$ varies. In the horizontal axe only the number of steps is reported, since a unique q is associated to each value of K. It is clear that $\langle V^R(\overline{y}_{stK}; K, q)\rangle$ reaches a minimum in correspondence of a pair that is denoted by (K_{opt}, q_{opt}). The optimal AGSPRN procedure in presence of budget constraint is the one that adds, at each step, q_{opt} units and proceeds until the step K_{opt}.

Some results for normally distributed data are given in Table 3.1, where the Monte Carlo error (MCE) is also reported. One can see from Table 3.1 that the optimal AGSPRN lasts 4 steps and adds 54 units per step, excluding the pilot sample selected at the first step. A comparison with TSPRN and ASPRN is also presented, since they are two extreme cases of the AGSPRN procedure with $(K, q) = (2, 166)$ and $(K, q) = (61, 1)$ respectively. Table 3.1 shows that the optimal AGSPRN procedure tends not to coincide with the TSPRN and ASPRN procedures which give rise to an estimator whose variance is higher. This confirms our intuition that exists a compromise procedure between TSPRN and ASPRN which generates more efficient estimators in presence of a cost function. A quite high value of the M. C. estimator variance is also reached using a stratified random sampling (STRS), i.e., drawing all the units obtained with the available budget in just one step with proportional allocation. Using the optimal AGSPRN procedure we gain a

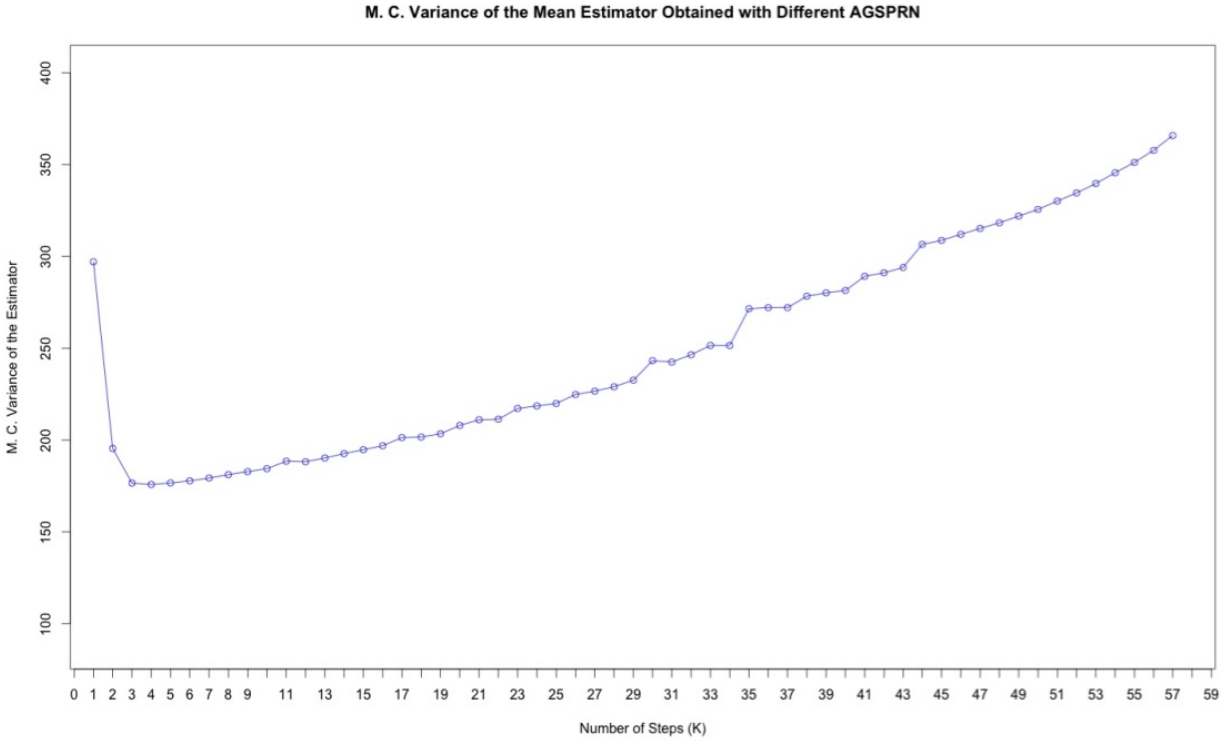

Figure 3.1: Monte Carlo variance of different AGSPRN mean estimators, $\langle V^R(\overline{y}_{stK}; K, q)\rangle$, for normally distributed data ($\overline{y}_N = 391.35$) and with budget constraint, $c_n = 2$, $c_k = 4$.

variance reduction for the mean estimator of 40% with respect to STRS, with the same budget. The true value of the mean in the simulated finite population is equal to 391.35 and we observe that all the procedures generate estimates very close to the real value. However, among all the estimators, the optimal AGSPRN estimator is the closest to the true value.

Table 3.1: Comparison of different adaptive estimators assuming Normal population with $C = 500$, $C_0 = 80$, $c_n = 2$, $c_k = 4$. The first row presents the optimal solution with the value of $\overline{y}_{stK}$, its variance, the MCE, the sample size n and the pilot sample size n_0. The consecutive rows show the comparisons with other sampling procedures: TSPRN, ASPRN and STRS. Here, $\overline{Y} = 391.35$.

	K	q	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	MCE	n	n_0
Optimal AGSPRN	4	54	391.53	175.80	0.221	202	40
TSPRN	2	166	391.71	195.43	0.233	206	40
ASPRN	57	1	391.84	365.76	0.500	96	40
STRS	1	0	391.72	297.04	0.300	207	207

Table 3.2 shows the comparison between different allocation methods. Using Neyman allocation with the real values of the strata variances, we obtain the minimum value of the

estimator variance. Hence, the allocations generated by TSPRN, ASPRN and AGSPRN procedures should be compared with those reported in the last column of Table 3.2. The AGSPRN procedure gives rise to the closest allocations to the optimal ones (the differences are reported in brackets). The ASPRN performs relatively good but it allows to sample only 96 units because of the step cost, thus it is generally inefficient.

Table 3.2: Comparison between different allocation methods. The differences with Neyman's allocation are reported in brackets.

STRATA	N_h	n_h proport. allocation	n_h TSPRN	n_h ASPRN	n_h optimal AGSPRN	n_h Neyman's allocation
Stratum 1	493	21	47 (-10)	28	49 (-8)	57
Stratum 2	495	21	10 (-2)	4	15 (+3)	12
Stratum 3	496	21	37 (+7)	13	33 (+3)	30
Stratum 4	479	21	21 (-3)	9	21 (-3)	24
Stratum 5	453	20	4 (+1)	4	4 (+1)	3
Stratum 6	466	20	16 (+3)	4	12 (-1)	13
Stratum 7	463	20	4 (-4)	4	7 (-1)	8
Stratum 8	495	21	9 (-7)	8	16 (0)	16
Stratum 9	481	21	7 (0)	4	7 (0)	7
Stratum 10	487	21	51 (+14)	18	38 (+1)	37
Total	4808	207	206	96	202	207

Unit cost and step cost can vary considerably in applied problems. We believe that taking into account the impact of the cost components on the results is critical. Hence, we have performed extensive simulation to assess the impact of different values of the unit cost on the optimum combination of K and q and on the variance of the the mean estimator. As Table 3.3 shows, an increase of c_n under a fixed budget causes a decrease of q_{opt} and an increase of K_{opt}. Obviously the number of total sample units decreases and, consequently, the variance of the estimator increases. The optimal AGSPRN procedure tends to maintain a high number of steps K_{opt} as much as possible, since a decrease of K_{opt} inflates the estimator variance more than a decrease of q_{opt}. On the other hand, if the cost per step c_k increases, the number of optimal steps K_{opt} decreases with a relatively low effect on the total sample size and, consequently, on the variance of the estimator.

Table 3.3: Effect of c_n and c_k for normally distributed data in presence of budget constraints.

c_n	c_k	K_{opt}	q_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	MCE	n
2	4	4	54	391.53	175.80	0.221	202
2.5	4	5	30	391.42	221.07	0.277	160
3	4	5	23	391.54	268.12	0.336	132
4	4	5	15	391.87	357.63	0.473	100
c_n	c_k	K_{opt}	q_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	MCE	n
2	2	5	41	391.59	173.05	0.216	204
2	4	4	54	391.53	175.80	0.221	202
2	6	3	80	391.52	180.23	0.219	200
2	8	3	79	391.57	181.87	0.222	198

3.4 Case 2: minimization of the total cost given a threshold on the estimator variance

We are interested in the search of the optimal AGSPRN procedure which minimizes the total cost C as expressed in (3.5) given a constraint v on the estimator variance in (3.2).

Let us consider all the possible pairs of integers $(K, q) \in \mathcal{H}_{(K,q)}$, where the set $\mathcal{H}_{(K,q)}$ is defined as following:

$$\{(K,q) : K \texttt{ integer } \in [2, \frac{N-n_0}{q}], q \texttt{ integer } \in [1, \frac{N-n_0}{K}]\}.$$

We aim at solving the following optimization problem:

$$min_{(K,q)\in\mathcal{H}} C(K,q) = min_{(K,q)\in\mathcal{H}}[C_0 + c_n[n_0 + q(K-1)] + c_k K]$$

subject to

$$V(\overline{y}_{stK}; K, q) \leq v$$

Since the expression in (3.2) is analytically intractable, we proceed through a Monte Carlo study, applying for each integer q the AGSPRN procedure until step K, with $(K, q) \in \mathcal{H}$, and computing $\widehat{V}(\overline{y}_{stK}; q, K)$ at the K-th step.

3.4.1 Monte Carlo study

The Monte Carlo algorithm (Algorithm 2) requires as inputs to derive the distribution of the variable of interest in each stratum. Indeed the AGSPRN procedure has to be applied to a population very close to the one under study. Thus the inputs $f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$, are some distributions very close to the real ones, derived from the study of the phenomenon or from the pilot sample. However, this is a very crucial point in a

finite population setting and a more refined method is proposed in Chapter 4 to update $f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$.

Algorithm 2: Monte Carlo algorithm to find the optimal AGSPRN procedure in terms of minimum cost given a threshold v on the estimator variance.

Input: $C_0, c_n, c_k, H, R, n_0, v, f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$
Input: N_h, for $h = 1, ..., H$, s.t. $\sum_{h=1}^{H} N_h = N$
for $h = 1, ..., H$ **do**
  draw N_h values from $f_h(y, \boldsymbol{\theta}_h)$
end
foreach $(K, q) \in \mathcal{H}_{(K,q)}$ **do**
  for $r = 1, ..., R$ **do**
    apply the AGSPRN procedure and compute $\widehat{V}^r(\overline{y}_{stK}; K, q)$
  end
  compute the Monte Carlo mean:
  $\langle V^R(\overline{y}_{stK}; K, q)\rangle = \frac{1}{R}\sum_{r=1}^{R} \widehat{V}^r(\overline{y}_{stK}; K, q) \simeq E(\widehat{V}(\overline{y}_{stK}; K, q)) = V(\overline{y}_{stK}; K, q)$;
  compute the cost C(K, q)
end
choose (K,q) $\in \mathcal{H}_{(K,q)}$ s.t. $min_{(K,q)\in\mathcal{H}_{(K,q)}} C(K, q)$ and $\langle V^R(\overline{y}_{stK}; K, q)\rangle \leq v$

3.4.2 Results for normally distributed data

In this section, the optimal AGSPRN procedure that minimizes the cost function given a constraint on the variance of the estimator is obtained for a known population. The aim is to show some of its properties and to assess the impacts of the cost components on it. Algorithm 2 is applied using the same population and the same setting of Section 3.3.2:

(i.) $C_0 = 80, c_n = 2, c_k = 4, H = 10, n_0 = 40, v = 200, W_h \overset{iid}{\sim} U[450, 500]$;

(ii.) $f_h(y, \boldsymbol{\theta}_h) = N[h \times 25 + 250, a_h \times 0.9]$, with h the stratum indicator, a_h the hth element of the vector $a = [600\ 130\ 320\ 250\ 40\ 150\ 100\ 180\ 74\ 400]'$;

(iii.) $R = 1000$. The AGSRPN procedure is applied 1000 times for each pair $(K, q) \in \mathcal{H}_{(K,q)}$.

From Figure 3.2 it is clear that if the number of steps K of an AGSPRN procedure increases, than it would be enough to add less units q at each step to allow the Monte Carlo estimator variance to be under the threshold v. Specifically, an increase of K results into a rapid decrease of the estimator variance. This is obvious since at each step the variances inside each stratum are estimated more precisely and, consequently, the allocations get

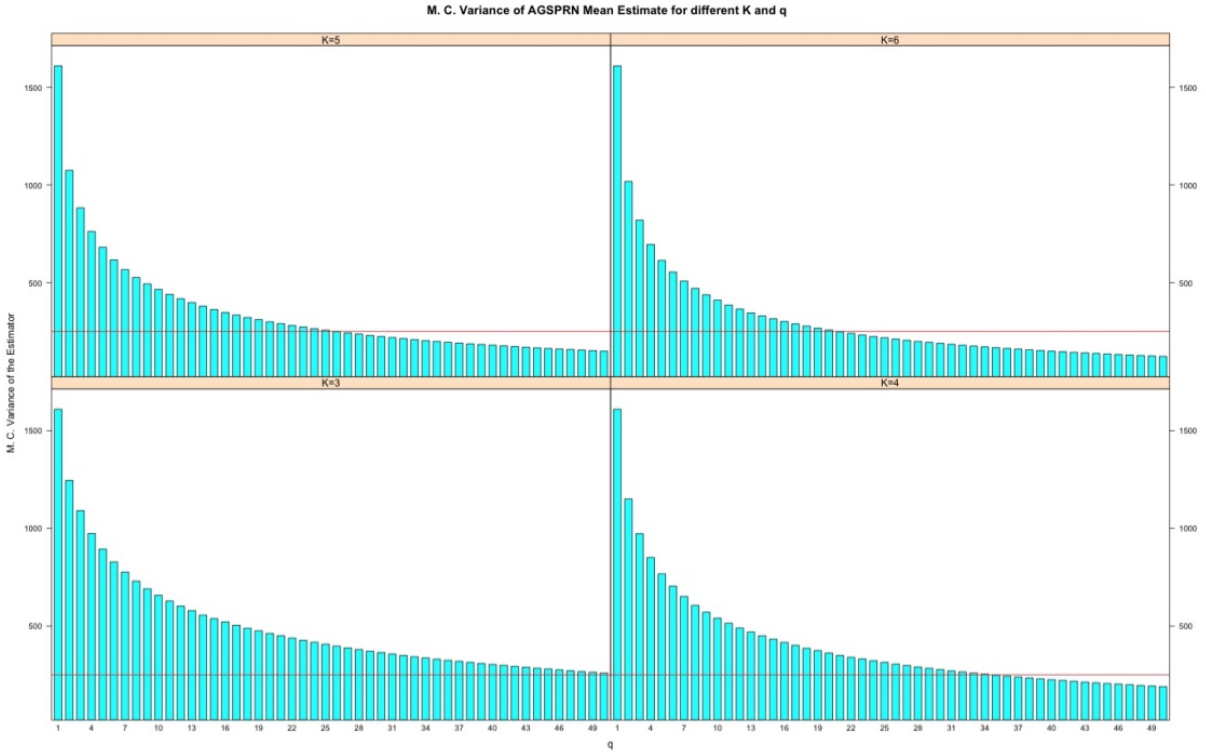

Figure 3.2: Monte Carlo variance of the mean estimator generated by AGSPRN procedures with different value of K and q. The red line represents the threshold v.

closer to the optimal ones.

Table 3.4 shows the results related to Normal distributed data. The Monte Carlo error (MCE) has also been reported in the second last column. The optimal AGSPRN procedure when $c_n = 2$, $c_k = 4$ and the threshold $v = 200$ consists of 4 steps and 46 units per step for a total cost of 452. This result is very similar to that observed previously when the budget constraint C was equal to 500, $c_n = 2$, $c_k = 4$ and $v = 175.8$.

If the threshold v increases to 250, 3 steps with 52 units per step are enough to guarantee the estimated variance to be lower than the constraint. Obviously, the higher the threshold v, the lower will be sample size, cost and number of steps.

Let us investigate what happens if the cost components are modified. A decrease of c_n produces a decrease of K_{opt} from 4 to 3 and an increase of q_{opt} from 46 to 71. If the cost decreases, then the sample size is higher to balance the decrease of the number of steps. If c_n increases, then K_{opt} increases and q_{opt} decreases as we expected; the sample size also decreases, whereas the cost increases.

The last three rows of the table shows the impact of the cost per step c_k. If c_k increases, keeping fixed the cost per unit c_n, the number of steps K_{opt} obviously decreases, with a small increase of the sample size and of the total cost.

By comparing rows 3-5 to 6-8 of Table 3.4, we notice a symmetric pattern. The optimal AGSPRN procedure seems to rely on the ratio c_n/c_k that is a core element from a practical

point of view.

Table 3.4: AGSPRN procedure for different values of v, c_n and c_k for normally distributed data and $n_0 = 40$, $C_0 = 80$, $R = 10^3$, $H = 10$. Here, $\overline{Y} = 391.35$.

v	c_n	c_k	K_{opt}	q_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	C(K,q)	MCE	n
200	2	4	4	46	391.84	199.14	452	0.240	178
250	2	4	3	52	392.20	248.31	380	0.308	144
200	1	4	3	71	392.17	199.07	280	0.248	182
200	3	4	5	34	391.88	199.52	628	0.239	176
200	4	4	6	27	391.81	199.76	804	0.239	175
200	2	2	6	27	391.81	199.76	442	0.239	175
200	2	2.6	5	34	391.88	199.52	445	0.239	176
200	2	8	3	71	392.17	199.07	468	0.248	182

3.5 Case 3: minimization of the risk given constraints on the budget and on the estimator variance

In this section we focus on the optimal AGSPRN procedure which minimizes the risk function given in (3.6), that is obtained as a convex combination of two standardized criteria: estimator variance and total cost. This is particularly important in practical situations since the stakeholders have limited budget and they do not want to compromise efficiency. Therefore, considering the trade off between estimator variance and cost is crucial.

Let us consider all the possible pairs of integers $(K, q) \in \mathcal{H}_{(K,q)}$, where the set $\mathcal{H}_{(K,q)}$ is defined as following:

$$\{(K,q) : K \texttt{ integer } \in [2, \frac{N-n_0}{q}], q \texttt{ integer } \in [1, \frac{N-n_0}{K}]\}.$$

The optimal AGSPRN procedure is the solution of the following optimization problem:

$$min_{(K,q)\in\mathcal{H}} R(\overline{y}_{stK}, \overline{y}_N; K, q) = min_{(K,q)\in\mathcal{H}}[\omega(\lambda V(\overline{y}_{stK}; K, q)) + (1-\omega)C(K,q)] \quad (3.9)$$

subject to

$$\begin{cases} V(\overline{y}_{stK}; K, q) \leq v \\ C(K,q) \leq c \end{cases}$$

The constant λ in (3.9) is used to put $V(\overline{y}_{stK}; K, q)$ and $C(K,q)$ in a comparable scale. We have chosen the following value:

$$\lambda = \frac{max_{(K,q)\in\mathcal{H}} C(K,q) - min_{(K,q)\in\mathcal{H}} C(K,q)}{max_{(K,q)\in\mathcal{H}} V(\overline{y}_{stK}; K, q) - min_{(K,q)\in\mathcal{H}} V(\overline{y}_{stK}; K, q)} \tag{3.10}$$

Moreover, we subtract the values $min_{(K,q)\in\mathcal{H}} V(\overline{y}_{stK}; K, q)$ and $min_{(K,q)\in\mathcal{H}} C(K,q)$ respectively, in order to obtain the standardized versions of $V(\overline{y}_{stK}; K, q)$ and $C(K,q)$. Since the optimization problem expressed by (3.9) is analytically prohibitive, we proceed through a Monte Carlo study, applying for each integer q the AGSPRN procedure until step K, with $(K,q) \in \mathcal{H}$, and computing $\widehat{V}(\overline{y}_{stK}; q, K)$ at the K-th step.

3.5.1 Monte Carlo study

The limitation of the proposed Monte Carlo study is that it requires to make some realistic assumptions about the distribution of Y in each stratum. Similarly to Algorithms 1 and 2, the inputs of Algorithm 3 includes the assumed strata distributions, i.e. $f(y_i|y_i \in h) = f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$, that generate the population to which the AGSPRN procedure is applied for each integer q. They can be derived from the study of the phenomenon or from the pilot sample.

Algorithm 3: Monte Carlo algorithm to find the optimal AGSPRN procedure in terms of minimum risk obtained as a combination of the estimator variance and the cost function.

Input: $c, c_n, c_k, H, R, n_0, v, \omega, f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$
Input: N_h, for $h = 1, ..., H$, s.t. $\sum_{h=1}^{H} N_h = N$
for $h = 1, ..., H$ **do**
  draw N_h values from $f_h(y, \boldsymbol{\theta}_h)$
end
foreach $(K,q) \in \mathcal{H}_{(K,q)}$ **do**
  for $r = 1, ..., R$ **do**
    apply the AGSPRN procedure and compute $\widehat{V}^r(\overline{y}_{stK}; K, q)$
  end
  compute the Monte Carlo mean:
  $\langle V^R(\overline{y}_{stK}; K, q)\rangle = \frac{1}{R}\sum_{r=1}^{R} \widehat{V}^r(\overline{y}_{stK}; K, q) \simeq E(\widehat{V}(\overline{y}_{stK}; K, q)) = V(\overline{y}_{stK}; K, q)$;
  compute the cost C(K, q) and the risk R(K, q)
end
choose (K,q) $\in \mathcal{H}_{(K,q)}$ s.t. $min_{(K,q)\in\mathcal{H}_{(K,q)}} R(K,q)$ given $\langle V^R(\overline{y}_{stK}; K, q)\rangle \leq v$ and $C(K,q) \leq c$.

3.5.2 Normal distribution case

We are interested in showing some properties of the optimal AGSPRN procedure which minimizes the risk function. To reach this aim we apply the Monte Carlo study directly to a given simulated population, where the strata distributions $g_h(y, \boldsymbol{\beta}_h)$, for $h = 1, ..., H$, are Normal with different parameters in each stratum. Hence, the inputs $f_h(y, \boldsymbol{\theta}_h)$, for $h = 1, ..., H$, of Algorithm 3 coincide here with the true strata distributions $g_h(y, \boldsymbol{\beta}_h)$, for $h = 1, ..., H$; they are not estimated from the pilot sample or from the study of the phenomenon. Algorithm 3 is then applied to the same setting of Sections 3.3.2 and 3.4.2:

(i.) $C_0 = 80, c_n = 2, c_k = 4, H = 10, n_0 = 40, v = 300, c = 600, W_h \overset{iid}{\sim} U[450, 500]$;

(ii.) $f_h(y, \boldsymbol{\theta}_h) = N[h \times 25 + 250, a_h \times 0.9]$, with h the stratum indicator, a_h the hth element of the vector $a = [600\ 130\ 320\ 250\ 40\ 150\ 100\ 180\ 74\ 400]'$;

(iii.) $R = 1000$. The AGSRPN procedure is applied 1000 times for each pair $(K, q) \in \mathcal{H}_{(K,q)}$.

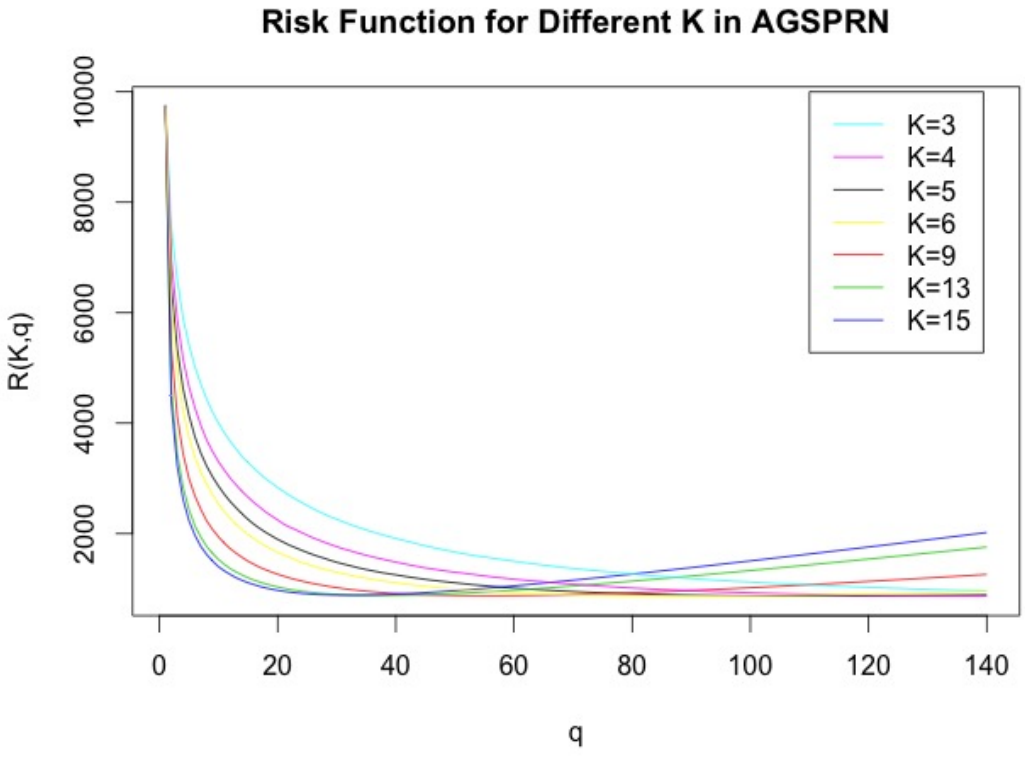

Figure 3.3: Risk functions as q varies for different K with AGSPRN procedure.

Figure 3.3 shows how the risk varies for different values of K and q. More specifically, when q is low the risk is also lower for higher values of K, but the contrary holds when q

is large. This happens because of the increase of the cost, which impacts more than the decrease of the estimator variance.

Table 3.5: AGSPRN procedure for different values of v, c, c_n, c_k and ω with normally distributed data and $n_0 = 40$, $C_0 = 80$.

ω	v	c	c_n	c_k	K_{opt}	q_{opt}	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	C(K,q)	R(K,q)	MCE	n
0.5	300	600	2	4	5	52	142.40	596	1480.98	0.17	248
0.5	300	600	3	4	6	25	211.29	599	2387.24	0.25	165
0.5	300	600	4	4	5	21	281.07	596	3505.14	0.37	124
0.5	300	600	2	2	5	53	140.20	594	1046.28	0.16	252
0.5	300	600	2	8	4	68	146.15	600	2379.87	0.17	244
0.5	200	500	2	4	4	54	176.54	500	1736.28	0.21	202
0.1	300	600	2	4	4	47	195.78	458	612.43	0.23	182
0.9	300	600	2	4	5	52	142.40	596	2320.16	0.17	248

Results for different values of v, c, c_n, c_k and ω are shown in Table (3.5). The optimal AGSPRN procedure for $c_n = 2$ and $c_k = 4$ involves 5 steps and 52 units per step, a result very similar to those obtained in Case 1 and Case 2. If the cost per unit c_n increases to 3, the number of steps increases, the total sample size decreases, and the risk will be double. An additional increase of c_n reduces both K_{opt} and q_{opt}, producing a high increment of the variance and consequently, high increment of risk. An increase of c_n has a deep negative impact on the risk.
If c_k increases, the number of steps K_{opt} decreases, but the overall impact on the sample size and on the estimator variance is very small.

Moreover, a decrease of the values of the constraints makes K_{opt} and n smaller, whereas the risk larger. Since the risk function inflates more the estimator variance because of the large value of λ, we choose ω to be equal to 0.2, in order to weigh more the cost component. The last row of Table (3.5) shows that 4 steps and 65 units per steps characterize the optimal AGSPRN procedure when the cost of the sampling process gets higher importance. Indeed, a lower number of steps and a smaller sample size will result into a cheap procedure, but with less precise estimates.

3.6 Discussion

In Chapter 2 we have analysed the literature concerning adaptive and sequential procedures, devoting particular attention to the two steps adaptive procedure with permanent random numbers (TSPRN, Carfagna [2007]) and the adaptive sequential procedure with permanent random numbers (ASPRN, Carfagna and Marzialetti [2009b]). Carfagna et al. [2012] showed that, when a cost function is introduced and the step cost is high, the ASPRN could be less efficient than the TSPRN. Therefore, it has arisen the need to

introduce a compromise solution between TSPRN and ASPRN, which reduces the costs to be suffered by ASPRN when the cost per step c_k becomes relatively high, preserving the advantages that an adaptive sequential procedure has in terms of efficiency of the estimators. Thus, in this chapter, we have addressed the problem of identifying an adaptive sampling procedure with an optimum number of steps and of sample units allocated at each step in order to reach the maximum efficiency of the estimator in terms of variance, cost or risk. We have proposed an Adaptive Group Sequential Procedure with permanent random numbers (AGSPRN) from which the TSPRN and the ASPRN can be derived as particular cases.

We have introduced the approach based on the minimization of a risk function which is a convex combination of two standardized criteria: the cost of the procedure involving K steps and q units per step, and the variance of the estimator generated by the procedure. This approach can be useful in applied problems, where it is important to take into account the precision of the results but also the cost of reaching that precision. A procedure that balances the precision of the estimates and the cost to reach it, assigning to the two criteria different levels of importance, is essential.

Through a Monte Carlo study applied to a simulated normal population, we have showed that, when the cost per step is not negligible, the optimal AGSPRN procedure tends not to coincide with ASPRN, TSPRN and also it is more efficient than a simple stratified random sampling that is applied in just one step, since the technique of proceeding along more steps allows to generate a sample allocation very close to Neyman's one.

A key role is played by the cost function and by the values of its components. Although, we have chosen a linear cost function, the impact of different functions on the optimal procedure should be analyzed in future works. In particular, we have seen that the the ratio c_n/c_k is relevant in choosing the optimal number of q and K. Hence, in applied problems, reducing some costs in favor of others can be fundamental to gain efficiency in the estimates.

In all the three minimized criteria (variance, total cost and risk), we have noticed that an increase of the unit cost, under a fixed budget and a linear cost function, causes an increase of the number of optimal steps K_{opt} and a decrease of the number of units per step q_{opt}. Obviously, the total number of sample units decreases and consequently, the variance of the estimator of the mean increases. The optimal AGSPRN procedure tends to maintain a high number of steps, since a decrease of the number of steps inflates the estimator variance more than a decrease of the number of sampling units per step. On the other hand, if the cost per step increases, the number of optimal steps decreases, with a relatively low effect on the total sample size and, consequently, on the variance of the estimator.

Moreover, when the optimal AGSPRN procedure is obtained through the minimization

of the risk function, it is important to choose carefully the values of the thresholds v, c and the value of the weight ω according to the customers needs.

To find the optimal AGSPRN procedure, one has to make some strong assumptions about the distribution of the variable of interest inside each stratum. Usually they can be derived by the study of the phenomenon or by the information collected in the pilot sample. This is not very practical and applicable unless we are aware that the proposed distribution can suite perfectly the analyzed data. In this chapter a Monte Carlo study is applied directly to the target population, in order to study some properties of the estimator generated by the optimal AGSPRN procedure. However, applying the Monte Carlo study directly to the target population in order to find the optimal sampling procedure is quite useless. In fact, it would be probably more convenient to estimate directly the population parameter of interest. Hence, to make the research of the optimal AGSPRN procedure more flexible and applicable to real problems, the estimate of the distribution of the variable of interest derived from the pilot sample can be updated at each step, when additional units are selected. We will develop this idea in the next chapter.

Chapter 4

The search of the optimal AGSPRN procedure

4.1 Introduction

In Chapter 3 we have proposed the adaptive group sequential procedure with permanent random numbers (AGSPRN). A Monte Carlo simulation study was suggested to investigate the optimal AGSPRN procedure in terms of:

1. minimum variance of the estimator given a cost function and a budget constraint (Case 1, Section 3.3);

2. minimum cost given a threshold on the estimator variance (Case 2, Section 3.4);

3. minimum risk, obtained through a combination of cost and estimator variance (Case 3, Section 3.5).

In these three cases, the simulation procedures require as inputs some accurate assumptions about the distributions of the variable of interest inside each stratum. They are usually unknown, but they can be derived from the study of the phenomenon or they can be estimated through the data collected in the pilot sample. However, in the proposed Monte Carlo procedures (see Algorithms 1, 2, 3 of Chapter 3), the estimates of these distributions are not updated when new units are selected. The aim of this chapter is to overcome this limitation, proposing a method to obtain the optimal AGSPRN procedure updating the distribution of Y at each step k, when more information is gained. It is an extension of the bootstrap technique for adaptive designs proposed by Rosenberger and Hu [1999] who applied it to infinite population with Bernoulli distribution in the clinical trials context.

4.2 The AGSPRN procedure in practice

In this section, we present in details the search of the optimal AGSPRN procedure when the distribution of the variable of interest Y is updated at each step after the selection of new units. We have the following steps:

(i) at the first step $[k = 1]$ select a first stratified random sample of size n_0 with proportional allocation, selecting at least two sample units per stratum and estimate the variance inside each stratum;

(ii) make some assumptions on the distribution form of Y inside each stratum using the selected units, estimate the parameters and generate from that distribution $N_h - \lfloor n_0^h \rfloor$ values, $h = 1, ..., H$, such that all the finite population of size N is obtained; here $\lfloor n_0^h \rfloor$ is the integer part rounding to the floor of the n_0 units allocated to stratum h;

(iii) Using the estimated population, simulate R times the application of the AGSPRN procedure with different values of q and choose the optimal pair (K_{opt}, q_{opt}) according to the three different criteria (minimum estimator variance given a budget constraint, minimum cost given a threshold on the estimator variance, minimum risk);

(iv) compute Neyman's allocation with sample size $n = n_0 + q_{opt}$ and select q_{opt} sample units only in the strata with positive difference between Neyman's allocation and the actual one (the allocation is proportional to this difference). Then estimate the parameter of interest and its precision;

(v) if the stopping rule is satisfied and $K_{opt} = 2$ stop the process, otherwise start again from step (ii) fixing $n_0 = n_0 + q_{opt}$ and $C_0 = C_0 + c_k$;

The optimal AGSPRN procedure obtained through this method is characterized by a number of units added at each step that can vary from step to step, depending on the updated population. This procedure allows to obtain a sample allocation as close as possible to Neyman's one, by computing, at each step, the allocation and the combination of number of steps and of units per step, given the linear cost function in (3.5).
The method is formally described in Algorithm 4.

Algorithm 4: Algorithm for the optimal AGSPRN procedure when only a pilot sample of size n_0 is available and the population is updated at each selection of new units.

Input: C_0, c_n, c_k, H, n_0, C (Case 1), v(Case 2), c and v(Case 3)
Input: N_h, for $h = 1, ..., H$, s.t. $\sum_{h=1}^{H} N_h = N$
for $m = 1, ..., N - n_0$ **do**
  for $h = 1, ..., H$ **do**
    from $\lfloor n_0^h \rfloor$ make assumptions on the distribution $f_h^m(y, \boldsymbol{\theta}_m)$ and estimate its parameters;
    generate $N_h - \lfloor n_0^h \rfloor$ values from $f_h^m(y, \boldsymbol{\theta}_m)$;
    compute the variance S_h^m using the population value in stratum h;
  end
  if *m=1 and the stopping rule with the sample of size n_0 is satisfied* **then**
    stop the process and estimate the parameter of interest;
  else
    find the optimal pair (K_{opt}^m, q_{opt}^m) of the AGSPRN procedure according to the chosen criterion (Case 1, Case 2, Case 3);
    select q_{opt}^m units among the strata proportionally to the positive difference between the Neyman's allocation (computed with sample size $n = n_0 + q_{opt}^m$ and S_h^m as variances for $h = 1, ..., H$) and the actual one;
    if *the stopping rule is satisfied (and for Case 1,3 $K_{opt}^m = 2$)* **then**
      stop the process and estimate the parameter of interest from the sample of size $n_0 + q_{opt}^m$;
    else
      set $C_0 = C_0 + c_k$ and $n_0 = n_0 + q_{opt}^m$;
    end
  end
end

4.3 Case 1: minimization of the estimator variance given a budget constraint

The aim of this section is to find the optimal AGSPRN procedure in term of minimum variance of the mean estimator (formula (3.2)), given a budget constraint and the linear cost function in (3.5).
We also assume the same finite population and setting as those used for simulations is Chapter 3 (Sections 3.3.2, 3.4.2, 3.5.2), i.e.,

(i) $H = 10, N_h \overset{iid}{\sim} U[450, 500]$;

(ii) $y_{ih} \overset{iid}{\sim} N[h \times 25 + 250, a_h \times 0.9]$, with h the stratum indicator, a_h the hth element of the vector $a = [600\ 130\ 320\ 250\ 40\ 150\ 100\ 180\ 74\ 400]'$;

(iii) $\overline{Y} = 391.35$.

At m=1, we suppose to have a pilot sample of size n_0 selected from the population designed by (i) and (ii). A density estimation is obtained for each stratum and $N - n_0$ values are generated from it. This can be done by applying a normal kernel density estimation for each stratum from the data collected in the pilot sample or by making assumptions on the distribution of Y in each stratum and estimating the parameters. The last procedure is recommended when the pilot size in each stratum is small, that is usually the case. Then, Algorithm 1 is applied to find the optimal AGSPRN procedure for the finite population composed by n_0 units selected from the real population and $N - n_0$ units generated by the estimated density, with $R = 1000$, $C = 500$, $c_0 = 80$, $c_n = 2$, $c_k = 4$.

Table 4.1: Optimal AGSPRN procedure at m=1, with $C = 500$, $C_0 = 80$, $c_n = 2$, $c_k = 4$. The optimal pair (K^1_{opt}, q^1_{opt}), the value of $\overline{y}_{stK}$, its variance, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=1	K^1_{opt}	q^1_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	4	54	364.60	303.93	0.341	202	40
Optimal AGSPRN with model assumption	4	54	349.98	178.33	0.201	202	40

Table 4.1 shows the optimal solutions obtained by applying a kernel density estimation (first row) and a parameters estimation after having made assumptions on the distribution form of Y (second row). Since for this simulation experiment we choose a pilot sample of size $n_0 = 40$ and 10 strata, the second method is more accurate. In this case, after an exploratory analysis of the pilot sample units, we assume that in each stratum Y is normally distributed with different parameters that are estimated. Now we set $C_0 = C_0 + c_k = 80 + 4 = 84$ and $n_0 = n_0 + q^1_{opt} = 40 + 54 = 94$, we estimate the variances inside each stratum and calculate Neyman's allocation with the new size n_0. Then, q^1_{opt} units are selected only in those strata with positive difference between Neyman's allocation and the actual one (the selection is proportional to this difference). From the sample of size n_0 we compute again a kernel density estimation or we make assumptions on the distribution of Y estimating its parameters in each stratum in order to generate the new population values. Algorithm 1 is again applied to find the optimal AGSPRN solutions that are showed in Table 4.2.

By repeating the same steps at $m = 3$, i.e., $C_0 = C_0 + c_k = 84 + 4 = 88$ and $n_0 = n_0 + q^2_{opt} = 94 + 54 = 148$, the optimal AGSPRN procedure is shown in Table 4.3.

Table 4.2: Optimal AGSPRN procedure at m=2, with $C = 500$, $C_0 = 84$, $c_n = 2$, $c_k = 4$. The optimal pair (K^2_{opt}, q^2_{opt}), the value of $\overline{y}_{stK}$, its variance, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=2	K^2_{opt}	q^2_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	3	54	365.53	167.45	0.174	202	94
Optimal AGSPRN with model assumption	3	54	370.00	123.77	0.127	202	94

Finally, since at m=3 the optimal solution consists of two steps, we stop the process and select 54 units from the real population. The estimate $\overline{y}_{stK}$ is equal to 394.85 with a variance $\widehat{V}(\overline{y}_{stK}; K, q)$ equal to 179.50 if we use a kernel estimation of the density. The values of $\overline{y}_{stK}$ and $\widehat{V}(\overline{y}_{stK}; K, q)$ are respectively equal to 395.41 and 170.73 if during the procedure we have made assumptions on the distribution of Y, estimating its parameters. In both cases we observe that $\overline{y}_{stK}$ is close to the real value $\overline{Y}$=391.5. Moreover, if we compare these solutions to the optimal AGSPRN procedure (Table 3.1) applied when the finite population is known from the beginning (not estimated step by step), the results are very similar: the optimal pair is the same ($K_{opt} = 4, q_{opt} = 54$) and the final estimate is slightly less precise when the population is estimated at each step. This small loss in precision is a low price to pay for making the search of the optimal AGSPRN procedure applicable in practice.

Table 4.3: Optimal AGSPRN procedure at m=3, with $C = 500$, $C_0 = 88$, $c_n = 2$, $c_k = 4$. The optimal pair (K^3_{opt}, q^3_{opt}), the value of $\overline{y}_{stK}$, its variance, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=3	K^3_{opt}	q^3_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	2	54	388.12	192.65	0.192	202	148
Optimal AGSPRN with model assumption	2	54	389.49	158.24	0.168	202	148

4.4 Case 2: minimization of the total cost given a threshold on the estimator variance

The second aim of this chapter is to find the optimal AGSPRN procedure that minimizes the linear cost function in (3.5), given a threshold v on the variance of the mean estimator in (6.1).

Our real finite population is the same of that used in the previous section and for simulations in Chapter 3 (Sections 3.3.2, 3.4.2, 3.5.2):

(i.) $H = 10, N_h \overset{iid}{\sim} U[450, 500]$;

(ii.) $y_{ih} \overset{iid}{\sim} N[h \times 25 + 250, a_h \times 0.9]$, with h the stratum indicator, a_h the hth element of the vector $a = [600\ 130\ 320\ 250\ 40\ 150\ 100\ 180\ 74\ 400]'$;

(iii.) $\overline{Y} = 391.35$.

Algorithm 2 of Chapter 3 is applied, with $R = 1000$, $v = 200$, $C_0 = 80$, $c_n = 2$, $c_k = 4$, $n_0 = 40$. At each m, the optimal pair (K^m_{opt}, q^m_{opt}) is obtained, finding the AGSPRN procedure which minimizes the linear cost function, given a threshold v on the variance of the mean estimator. We stop when $K^m_{opt} = 2$ and $\widehat{V}(\overline{y}_{stK}; K, q) \leq v$, with a sample composed of $n_0 + q^m_{opt}$ units. At each m, n_0 is set to be equal to the sum of q^{m-1}_{opt} and the previous n_0, whereas $C_0 = C_0 + c_k$. Moreover, a density estimation is obtained for each stratum and $N - n_0$ values are generated from it. This can be done by using the units of the pilot sample to compute a kernel density estimation for each stratum or to make assumptions on the distribution of Y in each stratum, estimating the parameters. We favour the latter procedure since the pilot sample size n_0 is small with respect to the number of strata $H = 10$. From the information gained through the pilot sample, we assume that Y is normally distributed with different parameters in each stratum.
Tables 4.4 - 4.6 show the results of the optimal AGSPRN procedure at each m obtained through a kernel density estimation of the finite population and by model assumptions with estimated parameters, before the stopping rule is satisfied and $K^m_{opt} = 2$.

Table 4.4: Optimal AGSPRN procedure at m=1, with $v = 200$, $C_0 = 80$, $c_n = 2$, $c_k = 4$. The optimal pair (K^1_{opt}, q^1_{opt}), the value of $\overline{y}_{stK}$, its variance, the cost, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=1	K^1_{opt}	q^1_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	$C(K, q)$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	3	122	363.37	199.99	660	0.46	284	40
Optimal AGSPRN with model assumption	4	45	349.74	199.53	446	0.464	175	40

At $m = 2$, K^2_{opt} is equal to 2 for both estimation methods, thus in order to stop the procedure, we have to verify if the stopping rule is satisfied. After the strata variances and Neyman's allocations are computed, we select q^2_{opt} units only in those strata with positive difference between Neyman's allocation and the actual one (the selection is proportional to this difference). Then, we calculate $\overline{y}_{stK}$ and its variance. They are respectively equal to 392.61 and 231.36 if the kernel density estimation method is applied, whereas they

assume the values of 386.53 and 326.40 if model assumptions method is employed. In both cases $\widehat{V}(\overline{y}_{stK}; K, q)$ does not satisfy the constraint v. Hence, we proceed to $m = 3$, we update the estimation of the distribution of Y and we apply Algorithm 2 of Chapter 3.

Table 4.5: Optimal AGSPRN procedure at m=2, with $v = 200$, $C_0 = 80$, $c_n = 2$, $c_k = 4$. The optimal pair (K^2_{opt}, q^2_{opt}), the value of $\overline{y}_{stK}$, its variance, the cost, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=2	K^2_{opt}	q^2_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	$C(K, q)$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	2	23	388.47	199.83	462	0.408	185	162
Optimal AGSPRN with model assumption	2	49	364.82	199.80	360	0.457	134	85

Table 4.6: Optimal AGSPRN procedure at m=3, with $v = 200$, $C_0 = 80$, $c_n = 2$, $c_k = 4$. The optimal pair (K^3_{opt}, q^3_{opt}), the value of $\overline{y}_{stK}$, its variance, the cost, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=3	K^3_{opt}	q^3_{opt}	$\overline{y}_{stK}$	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	$C(K, q)$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	2	21	390.38	198.52	510	0.41	206	185
Optimal AGSPRN with model assumption	2	30	391.62	197.07	424	0.443	164	134

For $m = 3$, the optimal AGSPRN procedure results are given in Table 4.6. According to the kernel density estimation method, we select 21 units and we compute the values of $\overline{y}_{stK}$ and $\widehat{V}(\overline{y}_{stK}; K, q)$ that are equal to 394.20 and 193.91 respectively. Since $K^3_{opt} = 2$ and $\widehat{V}(\overline{y}_{stK}; K, q) \leq v$ we stop the process and obtain an optimal adaptive sequential procedure that considers 4 steps with 40, 122, 23 and 21 units added respectively at each step and a cost of 510. It is quite more expensive than the optimal AGSPRN procedure obtained when the population is totally known (first row of Table 3.4), since the final sample contains 28 more units, despite the fact that it is achieved through the same number of steps. This is the cost of gaining precision when the population y-values are not totally known. However, if we use the model assumptions method to derive the distribution of Y, the selection of additional 29 units (q^3_{opt}) does not allow the variance of the estimator to be under the threshold v. Hence, we repeat again the procedure at $m = 4$ and obtain an optimal pair consisting of 2 steps and 15 units per step. The values of $\overline{y}_{stK}$ and $\widehat{V}(\overline{y}_{stK}; K, q)$ are respectively equal to 390.44 and 199.85, satisfying the constraint v. The optimal adaptive sequential procedure for an unknown population consists of 5

steps and 40, 45, 49, 30, 15 units added respectively at each step, for a total cost of 458. We found that the number of sampled units is close to that generated by the optimal AGSPRN procedure with known population, but the latter involves one step less and is slightly cheaper, as we expected.

4.5 Case 3: minimization of the risk given constraints on the budget and on the estimator variance

In this case we are interested in finding the optimal AGSPRN procedure which minimizes the risk function in expression (3.6) given thresholds c and v respectively on the cost and the estimator variance. Here, the real population is not supposed to be totally known in advance, contrary to what we have assumed in Section 3.5. However, the units are selected from the same population and the same setting is applied:

(i.) $H = 10, N_h \overset{iid}{\sim} U[450, 500]$;

(ii.) $y_{ih} \overset{iid}{\sim} N[h \times 25 + 250, a_h \times 0.9]$, with h the stratum indicator, a_h the hth element of the vector $a = [600\ 130\ 320\ 250\ 40\ 150\ 100\ 180\ 74\ 400]'$;

(iii.) $\overline{Y} = 391.35$.

We fix $v = 300$, $c = 600$, $C_0 = 80$, $c_n = 3$, $c_k = 4$ and apply Algorithm 4.
At $m = 1$, a pilot sample of size $n_0 = 40$ is selected among the 10 strata and the distribution of the variable of interest Y is derived in each stratum through a kernel estimation (method 1) or assuming a distribution form and estimating its parameters from the selected units (method 2). Then, the y-values of the population are generated from the estimated distributions, assuming that the strata sizes N_h are known, for $h = 1, ..., H$. The number of repetitions R is chosen to be equal to 1000 and Algorithm 3 is applied to the generated population, whose n_0 values are substituted by the pilot sample selected from the real population.
Table 4.7 shows the optimal AGSPRN procedure at $m = 1$. The first row refers to the results obtained when the distribution of the variable of interest is estimated through the kernel method, whereas the second row reports the optimal AGSPRN procedure when model assumptions are used to derive the distribution of Y inside each stratum. We notice that the latter method is more appropriate since the units of the pilot sample selected in each stratum are few. Particularly, after some exploratory analyses on the data of the pilot sample, normal distribution is assumed for Y, with different parameters in each stratum.
For each m, the weight ω is set equal to 0.5 and the value of λ in expression (3.10) is equal to 10.12 if model assumptions are considered, with $max_{(K,q)\in\mathcal{H}} C(K, q) = C(K = N -$

$n_0, q = 1) = 28772$, $min_{(K,q)\in\mathcal{H}} C(K,q) = C(K = 1, q = 0) = 164$, $max_{(K,q)\in\mathcal{H}} V(\overline{y}_{stK}; K, q) = V(\overline{y}_{stK}; K = 1, q = 0) = 2826.46$ and $min_{(K,q)\in\mathcal{H}} V(\overline{y}_{stK}; K, q) = 0$.
However, $max_{(K,q)\in\mathcal{H}} V(\overline{y}_{stK}; K, q)$ is equal to 5720.32, when a kernel estimation for the density of Y is used, with a consequent λ equal to 5.01. For the same values of total cost and variance estimator, different lambda's produce different values of the risk in absolute term, but not in relative term. Indeed, as Tables 4.7 - 4.9 show, the optimal pair is the same for both estimation methods, even though the value of λ is different.
We compute the strata variances and then we select $q^1_{opt} = 107$ units from the real population, only in those strata with positive difference between Neyman's allocation and the actual one (the selection is proportional to this difference). The allocations are the same for both estimation methods, giving rise to the same values of $\overline{y}_{stK}$ and $\widehat{V}(\overline{y}_{stK}; K, q)$, respectively equal to 372.96 and 204.22. However, the values of the risk do not coincide because of the different λ's: 979.70 and 1142.50 are the risk amounts respectively related to the kernel density estimation method and the model assumption one.

Table 4.7: Optimal AGSPRN procedure at m=1, with $v = 300$, $c = 600$, $C_0 = 80$, $c_n = 2$, $c_k = 4$. The optimal pair (K^1_{opt}, q^1_{opt}), the estimated variance of $\overline{y}_{stK}$, the cost $C(K,q)$, the risk $R(K,q)$, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=1	K^1_{opt}	q^1_{opt}	$\langle V^R(\overline{y}_{stK}; K, q)\rangle$	$C(K,q)$	$R(K,q)$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	3	107	227.41	600	786.65	0.254	254	40
Optimal AGSPRN with model assumption	3	107	124.52	600	848.18	0.142	254	40

At $m = 2$, we set $n_0 = q^1_{opt} + n_0$ and $C_0 = C_0 + 4 = 80 + 4 = 84$. Moreover, we estimate more precisely the distribution of Y in each stratum after the selection of q^1_{opt} units and we generate from it $N - n_0$ units in order to apply Algorithm 3. The optimal AGSPRN procedure for both estimation methods is shown in Table 4.8. It consists of other 3 steps and 52 units per step. We proceed computing the variances inside each stratum and Neyman's allocations with size $n_0 + q^2_{opt} = 147 + 52 = 199$. Supplementary units are selected only in those strata with positive difference between Neyman's allocation and the actual one, for a total amount of 52 units. Then, the mean and its variance are calculated, obtaining a value of 400.04 and 200.22 respectively. The risk is 662.66 for the kernel estimation method and 1176.27 if we use the assumption of normality for the distribution of Y. We argue that this gap between risk measures is due to the different values of λ.

At $m = 3$, we set $C_0 = C_0 + 4 = 88$, $n_0 = n_0 + q^2_{opt} = 147 + 52 = 199$ and Algorithm 3 is applied to the population generated by the updated estimated distribution of the variable of interest. As Table 4.9 shows, we obtain the same optimal AGSPRN procedure by using

Table 4.8: Optimal AGSPRN procedure at m=2, with $v = 300$, $c = 600$, $C_0 = 84$, $c_n = 2$, $c_k = 4$. The optimal pair (K^2_{opt}, q^2_{opt}), the estimated variance of $\overline{y}_{stK}$, the cost $C(K,q)$, the risk $R(K,q)$, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=2	K^2_{opt}	q^2_{opt}	$\langle V^R(\overline{y}_{stK};K,q)\rangle$	$C(K,q)$	$R(K,q)$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	3	52	111.56	598	495.96	0.130	251	147
Optimal AGSPRN with model assumption	3	52	94.06	598	693.05	0.10	251	147

both kernel estimation or model assumptions for the density of Y. Since $K^3_{opt} = 2$ we stop the process and choose as the optimal adaptive sequential procedure the one which proceeds for 4 steps adding at each step 40, 107, 52, 53 units respectively.
Finally, we calculate the strata variances and we compute Neyman's allocations with sample size $n_0 + 53 = 252$. We select 53 units from the real population only in those strata with positive difference between Neyman's allocation and the actual one. We obtain the same final strata sample sizes by using both methods of density estimation for Y. Hence, the values of $\overline{y}_{stK}$ and $\widehat{V}(\overline{y}_{stK};K,q)$ are the same, respectively equal to 402.65 and 154.13. However, the risk assumes different values, i.e., 601.41 when the kernel density estimation method is applied and 996.01 if we use the assumption of normality for the distribution of Y. Both values are smaller than those obtained at the previous steps and the constraints on cost and estimator variance are satisfied. Hence, it is safe to conclude that estimating the distribution of Y by steps when it is unknown leads to a correct AGSPRN procedure which is optimal for the minimization of the risk. Moreover, it is a very close result to the one obtained with known population (first row of Table 3.5), which consists in a procedure that samples just 4 units less, using one more step.

Table 4.9: Optimal AGSPRN procedure at m=3, with $v = 300$, $c = 600$, $C_0 = 88$, $c_n = 2$, $c_k = 4$. The optimal pair (K^3_{opt}, q^3_{opt}), the estimated variance of $\overline{y}_{stK}$, the cost $C(K,q)$, the risk $R(K,q)$, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=3	K^3_{opt}	q^3_{opt}	$\langle V^R(\overline{y}_{stK};K,q)\rangle$	$C(K,q)$	$R(K,q)$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	2	53	157.07	600	606.29	0.161	252	199
Optimal AGSPRN with model assumption	2	53	122.56	600	831.19	0.122	252	199

4.6 Discussion

In this chapter we have proposed a method useful to obtain the optimal AGSPRN procedure when the values of the variable of interest Y are not known for all the units in the population. Since the population is generally unknown, the presented method is very useful to estimate a population parameter timely and efficiently, especially when only a pilot sample is available and no previous studies on the phenomenon are provided.
At each selection of new units, the distribution of Y is derived through a kernel density estimation or making some assumptions on it and estimating its parameters. The two methods give rise to optimal AGSPRN procedures that are very similar and also very close to those obtained in Chapter 3 when the population is completely known.
It is worthy to say that when the pilot sample size is small as compared to the number of strata, estimating the distribution of Y on the basis of assumptions about its form is more appropriate than the kernel density estimation method.
We observe that a moderate decrease of precision with respect to the optimal AGSPRN procedure with known population is detected particularly in Case 2 and Case 3 and it is often compensated adding one step or some more units to the sampling process.
In the next chapter we will investigate the search of the optimal AGSPRN procedure for a real application, when the population values of the variable of interest are supposed to be both known and not totally known.

Chapter 5

Application: quality control of a land cover database

5.1 Introduction

Adaptive sequential sampling plays a significant role in many practical applications. It may be an important tool used by territory management to evaluate the quality of a land cover database.

Land cover databases are essential instruments for the territory management activities. Since they give information about land cover and land cover dynamics (when their application is repeated over time), they are used to assess the impact of alternative policies in a region in order to increase the potential of a territory. They are produced through aerial photos of high, medium and coarse resolution satellite data, depending also on the kind of utilization required, ranging from a local to a global scale. Satellite images are given as a set of measures of electromagnetic radiation reflected by a unit area of the Earth's surface. These unit areas are called pixels and they can range in size from less than 1 m to 5 km. The size of the pixels represents the spatial resolution of an optical satellite sensor and it is one of its main characteristics, together with the number of channels and the wavelength of each of them. The most widely used images for land cover monitoring have medium resolution, for example, 30 m.

The images are photo-interpreted or semi-automatically classified in order to produce the land cover databases, whose main users are public administrations and the scientific community. Semi-automatic classification is performed pixel by pixel or by continuous groups of pixels; it can be supervised or unsupervised. Photo-interpretation and supervised classification of remote sensing data require to define in advance a legend of different cover types according to which the pixels or the polygons (land areas with regular borders) are classified. The result is a database or a land cover map created in a geographic information system (GIS) that allows different kinds of operations on the pixels or polygons, such

as their subdivision, merger or overlay.

Evaluating the *quality* of a land cover database is a fundamental step in the production process and its quality should be high to make the database reliable. Sometimes, reliability is confused with the scale of the images which represents the level of detail of the basic material, not the quality of the database that is evaluated by the degree of precision in the classification procedure. Indeed, it is not an easy task to classify correctly pixels or polygons since the variability among pixels of the same land cover type in different conditions (soil, humidity, phenological phase etc.) can be much higher than the variability between different land cover types. The result of the classification, i.e., a land cover map or a database, has generally some missing data and a certain proportion of errors.

Quality control of the photo-interpretation and its **validation** are the procedures through which the errors are detected and the overall quality of the database is measured. Quality control is the procedure that consists in repeating the photo-interpretation on a sample of polygons by a very expert photo-interpreter (the controller).

Validation is a comparison of the land cover database with another representation of reality, which is considered more reliable. A sample of polygons is usually compared with the corresponding ground truth or with other remote-sensing data compatible with the source of images.

Strahler et al. [2006] underlined the importance of validation: "As a guideline, producing a global land cover map should consist of three more-or-less equal parts: data preparation, classification and validation. Without proper validation, any land cover map, whether at global, regional or local scale, remains an untested hypothesis."

However, in photo-interpretation projects, very few resources are generally devoted to quality control and validation; thus, these procedures are performed on a sample of polygons or points in the methodological framework of design-based statistical inference. The choice of a design-based sample is due to the minimal assumptions required to justify the validity of the quality estimators and their precision for different kind of applications and users. Strahler et al. [2006] again stated: "An inference framework heavily dependent on a model or other assumptions would require the cumbersome task of not only explicitly identifying these assumptions and model structures, but also justifying that they were satisfied for the particular application. The multitude of uses and users of a global map would suggest that validating assumptions may be even more difficult because of the large number of different analyses to which the data would be subject. Lastly, the objectivity provided by the randomization protocol of probability sampling provides assurance that the sample has not been selected, either consciously or unconsciously, to produce favorable accuracy results".

Due to the limited resources devoted to quality control and validation of land cover databases, a very cost-effective sample design is required. Moreover, it is necessary to have timely estimates in order to improve the database production, during the photo-

interpretation process. Thus, an adaptive sequential sampling is the best procedure to adopt for quality control as well as for validation: it allows to reach high precision of estimates with the smallest sample size and in the shortest time, especially when a small amount of information is available. Moreover, the costs can be controlled step by step, ending up with a sample that is optimal in terms of estimates precision and costs.
A key element of the production of a database is the legend, which is usually defined in advance. Due to the desire of homogeneity different projects share a common legend, even though it is not always appropriate. Thus, an adaptive sequential procedure is also an optimal tool to improve the legend during the process, according to customer's needs.
In this chapter we will present an application of the adaptive group sequential procedure with permanent random numbers (AGSPRN) to the quality control of a land cover database, in order to estimate the quality indexes efficiently in terms of precision and costs.

5.2 Adaptive sequential sampling for quality control and validation of a land cover database

The use of an adaptive sequential procedure for validation and quality control of a land cover database has many advantages. Traditionally, one of the oldest approach to quality control is acceptance sampling, which was widely used during the 1930s and 1940s. It consists in inspecting a sample of items from a given lot, in order to decide whether to accept or reject the whole lot. Different sampling schemes are applied to select the sample of items. Here, the number of items to be selected is fixed in advance. Wald [1947] introduced the idea of sampling *sequentially*, adding a unit (*item-by-item sequential sampling*) or a group of units (*group sequential sampling*) at each step, deciding whether or not stopping sampling according to a decision stopping rule and to the corresponding *probability ratio test*. The total sample size is not predetermined and, theoretically, the procedure may continue indefinitely. Wald [1947] showed that the sequential probability ratio test allows to save 50% of the total sampled units as compared with the most powerful classical test with the same errors of the first and the second kinds.
In the quality control and validation of a land cover database, the use of an adaptive sequential sampling has several advantages. Indeed, in order to assess the agreement of the database with the ground truth (validation) and verify the capability of the database to satisfy the client's needs, selecting a sample sequentially and adaptively during the production process allows to detect timely the discrepancies between the database and reality, which make the product inappropriate for customer's needs. These are not often clear at the beginning of the database production, hence proceeding step by step allows to draw the characteristics of the database and the legend progressively closer respectively

to the client's need and to the specific geographic area of interest. It is also possible to save some costs if, for instance, during the procedure, the adopted remote-sensing data are realized to be more detailed than those required by the user's needs, producing unjustified costs.

In the quality control procedure, instead, a photo-interpreter classifies the satellite images according to a legend of different land cover types. During this procedure he can make mistakes concerning the border of the polygons as well as the associated land cover type. Therefore, it is necessary that another expert (the controller) repeats the photo-interpretation on a sample of polygons in order to detect the mistakes and the discrepancies between the two classifications. The level of accuracy is then measured through some parameters such as the *percentage of area correctly classified*, the *percentage of the polygons correctly classified* by the first photo-interpreter and *pixel counting* (Gallego et al. [2010]).

In this chapter we are interested in estimating the percentage of area correctly classified, assuming that the sample given to the controller is selected through an optimal AGSPRN procedure, in order to obtain more precise and timely estimates of the quality parameters, saving costs. Usually, 85 % of polygons correctly classified represents an acceptable level of accuracy, even though some applications can require a higher value (Carfagna and Gallego [2005]).

In the field of classification of land cover images, stratifying the population is necessary because the kind of land cover type and the size of polygons affect the probability of making mistakes in the photo-interpretation. Hence, in our application we will choose these factors as layers for the stratification. Usually stratification is derived from the first photo-interpretation of remote sensing images, even though using existing land cover maps can be cheaper. For instance, CORINE Land Cover (CLC) map [1] has been used as a basis for stratification in Spain with satisfactory results (Gallego et al. [1999]). In some cases, strata are associated to specific land cover type or groups of land cover types (summer or winter crops for example). Some examples can be found in FAO [1996], FAO [1998], Cotter and Tomczak [1994]. The relative efficiency of stratification is the ratio between the variance that would have been obtained without stratification and the estimated stratified variance (Cochran [1997, pp 99-101]). Landscape complexity deeply affects efficiency. Where landscapes present a strong mixture of different crops, as in most western European countries, the efficiency is generally low. It can be increased if a stratum is characterized by a very dominant crop (in Catalonia the efficiency of stratification reached a value of 10 for rice in 1992). In the Mississippi area, relative efficiency in 1999 was around 3 for dominant crops and around 1 (no gain) for other crops.

Since the majority of applications deal with finite population, sampling designs for finite

[1] see footnote 1, pp 5-6

populations are the most required tools. In the next section, we will discuss in details all the aspects related to the stratification and the application of the AGSPRN procedure for our data-set.

5.3 Dataset

In 1999, the Italian Statistical Institute (ISTAT) carried out an experiment funded by Eurostat that produced a land-cover/land-use database with a detailed CORINE legend (Carfagna and Marzialetti [2009a]) and a scale of 1:25,000 for the Arezzo province. The resulted database has required a strong coherence between its data in order to be an efficient tool of territorial analysis, management and planning.
In this chapter we analyse the classification of 110 polygons that became part of the dataset produced by ISTAT. Four land cover types are analysed. The following information are available for each polygon:

- the land cover type associated to the polygon by the first photo-interpreter (type 1, 2, 3 or 4);
- the land cover type associated to the polygon by the controller (type 1, 2, 3 or 4);
- the size of the polygon.

The stratification is computed according to two factors: the land cover type associated to the polygons by the first photo-interpreter and the size of the polygon (small-large size, the median for each cover type is used as the cut-off). The result is the following:

- stratum 1: land cover type 1 and small size (18 polygons);
- stratum 2: land cover type 2 and small size (21 polygons);
- stratum 3: land cover type 3 and small size (14 polygons);
- stratum 4: land cover type 4 (6 polygons);
- stratum 5: land cover type 1 and big size (17 polygons);
- stratum 6: land cover type 2 and big size (21 polygons);
- stratum 7: land cover type 3 and big size (13 polygons);

For the fourth land cover type, the two strata identified by polygon's size have been collapsed because of their very small sizes. The result is a partition in seven strata.
We are interested in one of the main quality control indexes: the percentage of the area correctly photo-interpreted, knowing that the analysed area (X) covers totally 272.228 m^2.

A polygon is correctly photo-interpreted if the first photo-interpreter and the controller classify it according to the same cover type. Since the dataset provides information for all the population of polygons, we can directly compute the values of this index, that is 83.67 %. Indeed, the controller assigns a different land cover type to 28 out of 110 polygons, specifically, to 7, 12, 7, 2 polygons of the 1st, 2nd, 3rd and 4th land cover types, respectively.
The aim of this chapter is to show the properties of the estimator of the area correctly photo-interpreted generated by the optimal AGSPRN procedure, finding the optimal pair (K_{opt}, q_{opt}) in terms of minimum variance of the estimator given a cost function and a budget constraint (Case 1, Section 3.3), minimum cost given a threshold on the estimator variance (Case 2, Section 3.4), minimum risk obtained as a combination of estimator variance and cost (Case 3, Section 3.5). The results are obtained by assuming to known the entire population or just a sample of it, in order to make a comparison. In the latter case we apply the procedure described in Chapter 4.

5.4 Estimation of the percentage of the area correctly photo-interpreted through the AGSPRN procedure

We are interested in estimating the percentage (A) of the area correctly photo-interpreted through the optimal AGSPRN procedure. The value of A is equal to 83.67% as we mentioned in the previous section. Mathematically, it can be expressed by the following quantity:

$$A = \frac{\tau}{X} \times 100,$$

where X is the area of the photo-interpreted region that is known and equal to 272.228 m^2, whereas τ is the area correctly photo-interpreted and it is supposed to be estimated, even though we know the population. Indeed, the aim is to investigate the general performance of the optimal AGSPRN procedure. The estimator of A at the Kth step is:

$$a_{stK} = \frac{\widehat{\tau}_{stK}}{X} \times 100 = \frac{\sum_{h=1}^{H} N_h \sum_{i=1}^{n_{hK}} \frac{y_{ih}}{n_{hK}}}{X} \times 100,$$

where H denotes the number of strata, y_{ih} is a variable that takes value equal to the area of the polygon i in stratum h if it is correctly photo-interpreted (its classification by the first photo-interpreter and the controller coincides), 0 otherwise, N is the size of the population, N_h denotes the size of stratum h, for $h = 1, ..., H$, n_{hK} is the sample size of stratum h at step K. For our dataset we have: $N = 110$, $H = 7$, $N_1 = 18, N_2 = 21, N_3 = 14, N_4 = 6, N_5 = 17, N_6 = 21, N_7 = 13$.

The estimator a_{stK} is unbiased for A, since the estimator of the total $\widehat{\tau}_{stK}$ is unbiased in a stratified sampling design and X is a constant. Moreover, as we reported in Section 3.2 for the mean estimator $\overline{y}_{stK}$, if the stopping rule is independent from $\widehat{\tau}_{stK}$, the use of permanent random numbers in the AGSPRN procedure allows (3.1) to hold for a_{stK}.
The variance of a_{stK} is given by:

$$V(a_{stK}) = E[V(a_{stK}|n_{1K}, ..., n_{HK})] = E\left[\sum_{h=1}^{H} \frac{N_h^2}{n_{hK}} \frac{N_h - n_{hK}}{N_h} S_h^2\right] \times \frac{100^2}{X^2}, \tag{5.1}$$

where S_h^2 is the population variance for stratum h. Given a realization of the allocations $n_{1K}, ..., n_{HK}$, the estimate of $V(a_{stK})$ is:

$$\widehat{V}(a_{stK}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_{hK}}{N_h n_{hK}} \frac{\sum_{i=1}^{n_{hK}} (y_{ihK} - \overline{y}_{hK})^2}{n_{hK} - 1} \times \frac{100^2}{X^2}, \tag{5.2}$$

where $\overline{y}_{hK}$ is the mean estimate in stratum h after K steps. In the next sections the values of $\widehat{V}(a_{stK})$ are always scaled by 10000.
In the case under study, the linear cost function takes the expression of (3.5), where the cost components C_0, c_n, c_k are chosen to be respectively equal to 30, 2, 4 and the sample size n_0 is 32. Let us see in details the solutions of the optimal AGSPRN which minimizes:

Case 1. the variance of the estimator in (5.1) given the linear cost function in (3.5) and a budget constraint (C in (3.5) is given);

Case 2. the cost function in (3.5) given a threshold v on the estimator variance in (5.1);

Case 3. the risk function in (3.6), given thresholds v and c respectively on the estimator variance and on the total cost.

In the next section, the optimal AGSPRN procedure is obtained by a Monte Carlo study applied to the whole population. In Section 5.4 the results will be compared with the optimal AGSPRN procedure derived from just a part of population, using the method described in Chapter 4. Indeed, using the whole population to obtain an optimal sampling procedure is a contradiction; it is just an expedient useful to compare the proposed method with the most precise solution.

5.4.1 The optimal AGSPRN procedure with known population

The entire population showed in Section 5.3 is here totally used to derive the optimal AGSPRN procedure according to *Case* 1, 2 or 3. The results will be used to show some properties of the optimal AGSPRN procedure and they will be compared with those obtained in Section 5.4.

Case 1

The AGSPRN procedure in presence of budget constraint is applied to the dataset described in section 5.3 in order to verify the performance of the estimator a_{stK} for the percentage of the area correctly photo-interpreted. Algorithm 1 of Chapter 3 is implemented by fixing $C = 180$, $C_0 = 30$, $c_n = 2$, $c_k = 4$, $n_0 = 32$ and $R = 1000$. The y-values are known for each unit in the population, that means this is basically an exploratory study about the AGSPRN estimator. The true value A of the percentage of the area correctly photo-interpreted is 83.67%.

Table 5.1: Comparison of different adaptive estimators for the percentage of the area correctly photo-interpreted, with $C = 180$, $C_0 = 30$, $c_n = 2$, $c_k = 4$. The first row presents the optimal solution with the value of a_{stK}, its variance, the MCE, the sample size n and the pilot sample size n_0. The consecutive rows show the comparisons with other sampling procedures: TSPRN, ASPRN and STRS. Here, $A = 83.67\%$.

	K	q	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	MCE	n	n_0
Optimal AGSPRN	3	18	0.820	0.0013	2.53×10^{-6}	68	32
TSPRN	2	39	0.818	0.0015	2.59×10^{-6}	71	32
ASPRN	14	1	0.787	0.0052	9.86×10^{-6}	45	32
STRS	1	0	0.816	0.0049	5.15×10^{-6}	73	73

Table 5.1 shows that selecting 18 units at the second and third step allows to get a reduction of 73 % in the estimator variance with respect to the variance of the estimator obtained through a simple stratified random sample with the selection performed in just one step, using the same amount of budget. We also note that the bias of the estimator is not high in both cases. Moreover, the optimal AGSPRN procedure is shown not to coincide with the TSPRN and ASPRN procedures which produce a higher estimator variances. Table 5.2 shows the effect of c_n and c_k on the optimal AGSPRN procedure. Since the population size and the budget are not very high, an increase of c_n under a fixed budget causes a decrease of K_{opt} and an increase of q_{opt}, in order to maintain a quite high sample size with a low increase of the estimator variance. As c_n increases, q_{opt} and the sample size decrease, with a consequent reduction of the estimator variance. Contrary to Table 3.3, the optimal number of steps K_{opt} does not increase, because when the budget and the population size are low, it is preferable to reduce K_{opt} in order to maintain a

quite high sample size. On the other hand, if the cost per step c_k increases, the number of optimal steps K_{opt} decreases with a relatively low effect on the total sample size and, consequently, on the variance of the estimator also reduced.

Table 5.2: Effect of c_n and c_k for the percentage of the area correctly photo-interpreted, with $C = 180$, $C_0 = 30$, $c_n = 2$, $n_0 = 32$.

c_n	c_k	K_{opt}	q_{opt}	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	MCE	n
2	4	3	18	0.820	1.3×10^{-3}	2.53×10^{-6}	68
2.5	4	2	24	0.807	2.9×10^{-3}	4.50×10^{-6}	56
3	4	2	15	0.800	5.1×10^{-3}	7.69×10^{-6}	47
c_n	c_k	K_{opt}	q_{opt}	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	MCE	n
2	2	4	13	0.823	1.0×10^{-3}	1.80×10^{-6}	71
2	4	3	18	0.820	1.3×10^{-3}	2.53×10^{-6}	68
2	8	2	35	0.818	1.7×10^{-3}	2.59×10^{-6}	67

Case 2

Case 2 deals with the search of the optimal AGSPRN procedure which minimizes the cost function in (3.5) given a threshold v on the estimator variance in (5.1). Algorithm 2 of Chapter 3 is applied to the dataset described in Section 5.3, for $n_0 = 32$, $C_0 = 30$, $R = 1000$ and for different values of v, c_n and c_k. Since the y-values are available for all the units in the population, our objective is to conduct an exploratory analysis to evaluate the performance of the optimal AGSPRN estimator and assess the impacts of the cost components on the optimal procedure.

Table 5.3: AGSPRN procedure for different values of v, c_n and c_k for the percentage of the area correctly photo-interpreted and $n_0 = 32$, $C_0 = 30$, $R = 10^3$, $H = 7$. Here, $A = 83.67\%$.

v	c_n	c_k	K_{opt}	q_{opt}	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	C(K,q)	MCE	n
0.00125	2	4	3	17	0.819	1.23×10^{-3}	174	2.18×10^{-6}	66
0.00096	2	4	4	12	0.819	0.95×10^{-3}	182	1.77×10^{-6}	68
0.00125	1	4	2	37	0.819	1.21×10^{-3}	107	2.06×10^{-6}	69
0.00125	4	4	3	17	0.819	1.23×10^{-3}	306	2.18×10^{-6}	66
0.00125	4.2	4	4	11	0.816	1.23×10^{-3}	319.6	2.02×10^{-6}	65
0.00125	2	1.9	4	11	0.816	1.23×10^{-3}	167.6	2.02×10^{-6}	65
0.00125	2	2	3	17	0.819	1.23×10^{-3}	168	2.18×10^{-6}	66
0.00125	2	8	2	37	0.819	1.21×10^{-3}	184	2.06×10^{-6}	69

Table 5.3 shows the optimal AGSPRN procedure when $c_n = 2$, $c_k = 4$ and the threshold is $v = 1.25 \times 10^{-3}$. It consists of 3 steps and 17 units per step for a total cost of 174. This result is very similar to that observed in Table 5.2 when the budget constraint C was equal to 180, $c_n = 2$, $c_k = 4$ and $v = 1.3 \times 10^{-3}$.

If the threshold v decreases to 0.96×10^{-3}, more steps are necessary to guarantee the estimated variance to be lower than the constraint. The optimal pair (K_{opt}, q_{opt}) is equal to (4, 12), i.e., the lower is the threshold v, the higher will be the sample size, the cost and the number of steps.
Now we assess the effects of the different cost components. From our study, we observe that a decrease of c_n produces a decrease of K_{opt} from 4 to 2 and an increase of q_{opt} from 17 to 37. Consequently, we also observe a decrease of the total cost and an increase of the sample size to balance the reduction of the number of steps. If c_n increases, then K_{opt} increases and q_{opt} decreases as we expected, producing a reduction of the sample size and a conspicuous increase of the total cost.
Moreover, if c_k increases, keeping fixed the cost per unit c_n, the number of optimal steps K_{opt} obviously decreases, whereas q_{opt} increases, with a small increase of the sample size and of the total cost.
By comparing rows 3-5 to 6-8 of Table 3.4, we notice symmetry in the results. This confirms the conclusion shown in Table 3.4: the optimal AGSPRN procedure seems to rely on the ratio c_n/c_k that is a core element from a practical point of view.
Moreover, the bias of a_{stK} is not high, when a sample size that ranges from 55% to 65% of the population size is considered.

Case 3

We are now interested in finding the optimal AGSPRN procedure which gives rise to an estimator for the percentage of the area correctly photo-interpreted in such a way that the risk function in (3.6) is minimized, given thresholds v and c respectively on the estimator variance and on the total cost.
Algorithm 3 of Chapter 3 is applied to the land cover dataset, fixing $n_0 = 32$, $C_0 = 30$, $R = 1000$. Table 5.4 shows the result for different values of c_n, c_k, the thresholds v and c, the weight ω.

Table 5.4: AGSPRN procedure for different values of v, c, c_n, c_k and ω for the percentage of the area correctly photo-interpreted and $n_0 = 32$, $C_0 = 30$, $R = 10^3$, $H = 7$. Here, $A = 83.67\%$.

ω	v	c	c_n	c_k	K_{opt}	q_{opt}	$\langle V^R(a_{stK}; K, q)\rangle$	C(K,q)	R(K,q)	MCE	n
0.5	0.005	230	2	4	4	20	0.21×10^{-3}	230	93.09	4.66×10^{-7}	92
0.5	0.005	230	3	4	3	15	1.71×10^{-3}	228	372.09	2.85×10^{-6}	62
0.5	0.005	230	4	4	2	16	4.84×10^{-3}	230	1241.7	7.84×10^{-6}	48
0.5	0.005	230	2	2	7	10	0.17×10^{-3}	228	87.43	4.08×10^{-7}	92
0.5	0.005	230	2	8	3	28	0.30×10^{-3}	230	104.38	6.46×10^{-7}	88
0.5	0.0012	180	2	4	3	18	1.03×10^{-3}	178	171.62	1.92×10^{-6}	68
0.1	0.005	230	2	4	3	17	1.24×10^{-3}	174	88.53	2.02×10^{-6}	66

When the cost and estimator variance criteria are weighted equally ($\omega = 0.5$), the

optimal AGSPRN procedure for $v = 0.005$, $c = 230$, $c_n = 2$ and $c_k = 4$ consists of 4 steps and 20 units per step. For each unitary increment of the cost c_n, the number of steps and the total sample size decrease, whereas the risk quadruples.
If c_k increases, the number of steps K_{opt} decreases, with a small impact on the sample size, on the estimator variance and, consequently on the risk.
Moreover, a decrease in the values of the thresholds reduces K_{opt} and n, with a negative impact on the risk, which increases. This is obvious: higher constraints require more units or a higher estimator variance to be satisfied.
When ω is equal to 0.5, we notice that the optimal AGSPRN procedure selects as many units as possible, until the threshold c of the budget is reached. This is due to the large value of λ and to the small value of the constraint c with respect to the term $max_{(K,q)\in\mathcal{H}}C(K,q)$ in expression (3.10). These factors contribute to inflate more the estimator variance than the cost in the risk function. In order to balance the criteria, we used ω to be equal to 0.1. The last row of Table (5.4) shows that a decrease of ω produces a decrease of K_{opt} and q_{opt}, with a consequent decrement of the sample size and of the risk. This means that when the cost of the sampling process gets higher importance, the optimal AGSPRN procedure will be cheaper, but with less precise estimators.

5.4.2 The optimal AGSPRN procedure with unknown population

In this section we assume that the y-values are known only for a sample of polygons. We apply Algorithm 4 of Chapter 4 to find the optimal AGSPRN procedure when the population is not known. Moreover, we compare the results to the previous sections, where the population y-values in the dataset are completely used, giving rise to the most precise AGSPRN estimators. Let us illustrate the solutions of the optimal AGSPRN procedure in terms of minimum variance of the estimator given a cost function and a budget constraint (Case 1), minimum cost given a threshold on estimator variance (Case 2), minimum risk obtained as a combination of cost and estimator variance (Case 3).

Case 1

To find the optimal AGSPRN procedure that minimizes the variance of the percentage of polygons correctly photo-interpreted given a budget constraint and only a pilot sample we apply Algorithm 4, as given in Section 4.3.
Let us fix $C_0 = 30$, $C = 180$, $c_n = 2$, $c_k = 4$ and the seven strata weights as described in the dataset of Section 5.3.
At $m = 1$, a sample of size $n_0 = 32$ is selected from this dataset, observing the y-values relative to the area correctly photo-interpreted for each sampled polygon. This pilot

sample is used to estimate the distribution of the variable of interest, generating $N - n_0$ values. Since the strata sizes in the pilot sample are not high, it is preferable to use the pilot units to make assumptions about the distribution of Y in each stratum and to estimate the parameters. The area of a polygon is a continuous and positive variable, let us denote it with S. A realization of S for the polygon i is denoted with s_i. We assume a Gamma distribution for each stratum. The Shapiro-Wilk test confirms this assumption for each stratum. The parameters of each Gamma distribution are then estimated and $N - n_0$ values are overall generated. The correct classification of a polygon can be instead described by a Bernoulli random variable, let us call it Z. The reason of Bernoulli distribution selection is that Z takes value 1 if the polygon is correctly classified, 0 if it is not. The dataset described in Section 5.3 includes this information for each polygon, hence it is possible to estimate in each stratum the probability of classifying correctly the polygon, using the units selected in the pilot sample. Then, $N_h - n_{h0}$ units are generated from the h-th bernoulli distribution, where n_{h0} are the units of the pilot sample selected in stratum h, for $h = 1, ..., H$. The y-values of the area correctly photo-interpreted for each polygon are obtained multiplying the values of Z and S, in such a way that y_i is equal to s_i if the polygon i is correctly photo-interpreted, 0 otherwise.
Once we have generated the required data, we can apply Algorithm 1 of Chapter 3 to find the optimal AGSPRN procedure in the presence of a cost function and a budget constraint, fixing $R = 1000$.

Table 5.5: Optimal AGSPRN procedure at m=1, with $C = 180$, $C_0 = 30$, $c_n = 2$, $c_k = 4$. The optimal pair (K^1_{opt}, q^1_{opt}), the value of a_{stK}, its variance, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=1	K^1_{opt}	q^1_{opt}	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	3	18	1.083	2.17×10^{-3}	3.52×10^{-6}	68	32
Optimal AGSPRN with model assumption	3	18	1.153	2.59×10^{-3}	4.09×10^{-6}	68	32

Table 5.5 shows the optimal solutions. We present the optimal AGSPRN procedure obtained deriving the distribution of Y from a kernel density estimation in the first row of Table 5.5, whereas the results for model based assumption method are reported in the second row. The latter method is more appropriate to estimate the density of Y, since the pilot sample strata sizes are small.

Following the steps of Algorithm 4, at $m = 2$ we set $C_0 = C_0 + c_k = 30 + 4 = 34$ and $n_0 = n_0 + q^1_{opt} = 32 + 18 = 50$. The variances inside each stratum are then estimated and Neyman's allocations are computed with size n_0, selecting q^1_{opt} units only in those strata with positive difference between Neyman's allocation and the actual one (the

selection is proportional to this difference). From the sample of size n_0 we compute again a kernel density estimation (supplementary method) and we estimate in each stratum the parameters of the Gamma and Bernoulli distributions, in order to generate the updated population values. Algorithm 1 is again applied to find the optimal AGSPRN procedure for both methods. The solutions are showed in Table 5.6.

Table 5.6: Optimal AGSPRN procedure at m=2, with $C = 180$, $C_0 = 34$, $c_n = 2$, $c_k = 4$. The optimal pair (K^2_{opt}, q^2_{opt}), the value of a_{stK}, its variance, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=2	K^2_{opt}	q^2_{opt}	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	2	19	0.867	2.19×10^{-3}	3.05×10^{-6}	69	50
Optimal AGSPRN with model assumption	2	19	0.781	1.74×10^{-3}	2.86×10^{-6}	69	50

We found that, at $m = 2$, the optimal number of steps of the AGSPRN procedure is equal to 2 and we have to verify whether or not the stopping rule is satisfied. We compute the variances inside each stratum and Neyman's allocations. Other 19 units are selected only in those strata with positive difference between Neyman's allocation and the actual one (the selection is proportional to this difference).
Finally, we calculate a_{stK} and $\widehat{V}(a_{stK}; K, q)$ that are respectively equal to 0.821 and 1.12×10^{-3} for both methods used to estimate the distribution of Y. This means that both methods lead to the same allocations and to a value of a_{stK} very close to the real one (A=0.836).
Moreover, the comparison of these solutions with those obtained through the optimal AGSPRN procedure applied when the population is totally known (Table 5.1) shows that the optimal pair is the same ($K_{opt} = 3, q_{opt} = 18$), with just a unit less in the second step. The final estimate is even slightly more precise when the population is estimated, because of the addition of one extra unit that is allowed by the flexibility of these methods which let q_{opt} vary along the steps. The most important point is that the optimal AGSPRN procedure, when the population is known, leads to the same strata allocations than that applied to an estimated population.

Case 2

In this case we are interested in finding the optimal AGSPRN procedure that minimizes the linear cost function in (3.5), given a threshold v on the variance of the percentage of the area correctly photo-interpreted as expressed in (5.1).
Algorithm 4 is applied, with $v = 1.25 \times 10^{-3}$, $C_0 = 30$, $c_n = 2$, $c_k = 4$ and strata weights corresponding to the seven groups of polygons as described in the dataset of Section 5.3.

From this finite population, at $m = 1$, a pilot sample of size $n_0 = 32$ is selected and the optimal AGSPRN procedure is obtained, following Algorithm 2 of Chapter 3, with $R = 1000$.

Table 5.7: Optimal AGSPRN procedure at m=1, with 1.25×10^{-3}, $C_0 = 30$, $c_n = 2$, $c_k = 4$. The optimal pair (K^1_{opt}, q^1_{opt}), the value of a_{stK}, its variance, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=1	K^1_{opt}	q^1_{opt}	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	C(K,q)	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	4	16	1.075	1.23×10^{-3}	206	3.77×10^{-6}	80	32
Optimal AGSPRN with model assumption	4	14	1.121	1.17×10^{-3}	194	2.49×10^{-6}	74	32

Table 5.7 shows the results for both methods of estimating the distribution of Y in each stratum, where Y is the random variable associated to the area correctly photo-interpreted. The first row refers to the solution concerning the case in which a kernel density estimation is applied to the y-values of the pilot sample, in order to generate the overall y-values of the population, whereas the second row shows the optimal pair of the AGSPRN procedure applied to a population generated from a distribution whose form and parameters are derived from the pilot sample. Similarly to the previous section, the y-values are obtained by multiplying bernoulli and gamma random variables. We found that the estimation method based on model assumptions is more appropriate, since the pilot stratum sizes are not high. As Algorithm 4 outlines, for each m both methods of estimating the distribution of Y are applied to a pilot sample with updated size $n_0 = n_0 + q_{m-1}$ and the optimal AGSPRN procedure is obtained by following Algorithm 2, with $C_0 = C_0 + c_k$ and $R = 1000$.

Table 5.8: Optimal AGSPRN procedure at m=2, with 1.25×10^{-3}, $C_0 = 34$, $c_n = 2$, $c_k = 4$. The optimal pair (K^2_{opt}, q^2_{opt}), the value of a_{stK}, its variance, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=2	K^2_{opt}	q^2_{opt}	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	C(K,q)	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	3	16	0.682	1.16×10^{-3}	206	2.11×10^{-6}	80	48
Optimal AGSPRN with model assumption	3	14	0.776	1.21×10^{-3}	194	2.06×10^{-6}	74	46

Tables 5.8 and 5.9 show the results of the optimal AGSPRN procedure for $m = 2$ and $m = 3$ obtained through a kernel density estimation of the finite population and by model assumptions with estimated parameters, before the stopping rule is satisfied and $K^m_{opt} = 2$.

Table 5.9: Optimal AGSPRN procedure at m=3, with 1.25×10^{-3}, $C_0 = 38$, $c_n = 2$, $c_k = 4$. The optimal pair (K^3_{opt}, q^3_{opt}), the value of a_{stK}, its variance, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=3	K^3_{opt}	q^3_{opt}	a_{stK}	$\langle V^R(a_{stK}; K, q)\rangle$	C(K,q)	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	3	10	0.831	1.13×10^{-3}	218	2.31×10^{-6}	84	64
Optimal AGSPRN with model assumption	3	10	0.945	1.13×10^{-3}	210	2.37×10^{-6}	80	60

For each m, in order to verify whether or not the variance is below the threshold v, we select q^m_{opt} from the real population and we compute the estimator variance. At $m = 3$ the optimal solution consists of 3 steps and 10 units per step. We add 10 units to the sample and we compute the variance estimator, that is equal to 0.77×10^{-3} if we have used the kernel density estimation method to find the optimal pair, whereas it is equal to 1.07×10^{-3} by assuming a model for the distribution of Y. In both cases the estimate of $V(a_{stK}; K, q)$ is below the threshold $v = 1.25 \times 10^{-3}$ and we can stop the process. Hence, the optimal AGSPRN procedure with kernel density estimation consists of 4 steps: 32 units are selected at the first step, 16 at the second and at the third, 10 units at the fourth step, with a final estimate of the percentage of the area correctly photo-interpreted equal to 0.833 and a total cost equal to 194. Whereas the optimal AGSPRN procedure with model assumption for the distribution of Y involves 4 steps and 32 units selected at the first step, 14 at the second and the third steps, 10 at the fourth step, with $a_{st4} = 0.810$ and a total cost equal to 186.
If we compare these solutions with that shown in the first row of Table 5.3, where the optimal AGSPRN procedure is obtained using the real population, we notice some differences. The latter consists of 3 steps and 17 units per step, with an estimate of A equal to 0.819, a total sample size of 66 units and a total cost of 174. It requires less sample units to allow the estimator variance to be under the threshold and, consequently, it is less expensive, as we expected. However, the estimate of A is slightly less precise than that obtained through the optimal AGSPRN procedure with kernel estimation for the distribution of Y.

Case 3

Let us suppose that all the y-values of the population are not known, which is usually the case. Therefore, Algorithm 4 has been applied in order find the optimal AGSPRN procedure which gives rise to an estimator for the percentage of the area correctly photo-interpreted that minimizes the risk function in (3.6), given thresholds v and c respectively on the estimator variance and on the total cost. Algorithm 3 is not appropriate, because

the entire dataset is supposed not to be available, contrary to *Case* 3 of Section 5.4. We fix $n_0 = 32$, $C_0 = 30$, $v = 0.005$, $c = 230$, $c_n = 2$, $c_k = 4$. At $m = 1$, a sample of 32 polygons is selected from the land cover database with proportional allocation, assuming known the population sizes N_h, for $h = 1, ..., H$. Then, the distribution of the variable of interest Y is derived in each stratum, using a kernel estimation (method 1) or assuming a distribution form and estimating its parameters from the selected units (method 2). Since we have small size of the pilot sample in each stratum, the latter method is more appropriate. Successively, $N - n_0$ y-values are generated from the estimated distribution and then added to the pilot sample of size n_0 in order to assemble the entire population. Now it is possible to apply Algorithm 3 to the generated population, choosing $R = 1000$.

Table 5.10: Optimal AGSPRN procedure at m=1, with $v = 0.005$, $c = 230$, $C_0 = 30$, $c_n = 2$, $c_k = 4$. The optimal pair (K^1_{opt}, q^1_{opt}), the estimated variance of a_{stK}, the cost $C(K, q)$, the risk $R(K, q)$, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=1	K^1_{opt}	q^1_{opt}	$\langle V^R(a_{stK}; K, q)\rangle$	$C(K, q)$	$R(K, q)$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	3	18	1.70×10^3	178	57.00	2.99×10^6	68	32
Optimal AGSPRN with model assumption	3	19	1.38×10^3	182	59.20	2.38×10^6	70	32

Table 5.10 shows the optimal AGSPRN procedure at $m = 1$. We notice that if the distribution of Y is derived through a kernel density estimation the optimal adaptive sequential procedure consists of 3 steps and 18 units per step (first row of Table 5.10), whereas if assumptions are made on the distribution of Y the optimal pair considers one unit more at each step (second row of Table 5.10). For each m, in order to compute the risk, the component ω in expression (3.9) is set equal to 0.5 and the value of λ in formula (3.10) is equal to 19912.59, with $max_{(K,q)\in\mathcal{H}} C(K, q) = C(K = N - n_0, q = 1) = 566$, $min_{(K,q)\in\mathcal{H}} C(K, q) = C(K = 1, q = 0) = 98$, $max_{(K,q)\in\mathcal{H}} V(a; K, q) = V(a_{stK}; K = 1, q = 0) = 0.0235$ and $min_{(K,q)\in\mathcal{H}} V(\overline{y}_{stK}; K, q) = 0$. The term $max_{(K,q)\in\mathcal{H}} V(a; K, q)$ is equal to 0.0188 when model assumptions for the density of Y are used, with a consequent λ equal to 24823.81. Once we have computed the strata variances, Neyman's allocations are calculated with sample size $q^1_{opt} + n_0$. Successively, q^1_{opt} units are selected from the real population, only in those strata with positive difference between Neyman's allocation and the actual one (the selection is proportional to this difference). Then, we estimate the percentage of the area correctly photo-interpreted and its variance, that are equal to 0.845 and 6.3×10^3 if the kernel density estimation method is used and equal to 0.83871 and 5.86×10^3 if model assumptions are considered to derive the distribution of Y. The former method gives rise to a risk equal to 83.35, whereas the latter one generates a risk

of 93.84.

Table 5.11: Optimal AGSPRN procedure at m=2, with $v = 0.005$, $c = 230$, $C_0 = 34$, $c_n = 2$, $c_k = 4$. The optimal pair (K^2_{opt}, q^2_{opt}), the estimated variance of a_{stK}, the cost $C(K, q)$, the risk $R(K, q)$, the MCE, the sample size n and the pilot sample size n_0 are reported for the procedure involving kernel density estimation (first row) and density derivation with model assumptions and estimated parameters (second row).

m=2	K^2_{opt}	q^2_{opt}	$\langle V^R(a_{stK}; K, q)\rangle$	$C(K, q)$	$R(K, q)$	MCE	n	n_0
Optimal AGSPRN with kernel density estimation	2	28	1.51×10^3	198	65.05	2.37×10^6	78	50
Optimal AGSPRN with model assumption	2	23	1.53×10^3	190	65.07	2.71×10^6	74	51

At $m = 2$, we set $n_0 = q^1_{opt} + n_0$ and $C_0 = C_0 + 4 = 80 + 4 = 84$. Moreover, we estimate more precisely the distribution of Y in each stratum after the selection of q^1_{opt} units and we generate from it $N - n_0$ units in order to form the entire population. Then, we apply Algorithm 3 and results are given in Table 4.8, where it is showed that the two methods of density estimation give rise to similar optimal AGSPRN procedures, with final samples that differ for only four units. Since $K^2_{opt} = 2$, we verify if the stopping rule is satisfied by computing the variances inside each stratum and Neyman's allocations with size $n_0 + q^2_{opt}$. Supplementary units are selected only in those strata with positive difference between Neyman's allocation and the actual one, for a total amount of q^2_{opt} units. Then, the percentage of the area correctly interpreted, its variance and the risk are estimated through the final sample, obtaining values of 0.855, 0.42×10^{-3} and 54.25 respectively if the kernel estimation method is adopted and values equal to 0.843, 0.60×10^{-3} and 53.54 if model assumptions on the distribution of Y are used. The risk values in both cases are smaller than those obtained at the previous step. Furthermore, the constraints on cost and estimator variance are satisfied. This indicates that estimating the distribution of Y by steps when it is unknown leads, even in a real case, to a correct AGSPRN procedure which is optimal for the minimization of the risk. The results are slightly far from those presented in Section 5.4.1. Especially, the first row of Table 5.4 shows that when the population is known, the optimal AGSPRN procedure consists of 4 steps and 20 units per step, giving higher priority to decrease the variance of the estimator respect to the cost. However, when the y-values are generated from an estimated population, the optimal AGSPRN procedure considers just 3 steps and samples about 15 units less, giving more importance to the cost criterion. This can be due to the values of λ in the risk function and, particularly, to the ratio between the constraint c (or v) and the value of $max_{(K,q)\in\mathcal{H}} C(K, q)$ (or $max_{(K,q)\in\mathcal{H}} V(a; K, q)$).

5.5 Discussion

Practical problems require timely and efficient tools to achieve solutions with minimum amount of resources. This is the reason that has led us to apply the proposed AGSPRN procedure to the quality control operation of a land cover database. In the field of territory management, limited resources are devoted to measure the quality of a land cover database. Thus, adaptive sequential sampling can be an useful tool to provide efficient estimates, in terms of time and money. The aim of this chapter has been to show it.
In Section 5.4.1 we have showed that, given the same budget, the optimal AGSPRN procedure obtained through a Monte Carlo study applied to the known population provides estimates for the area correctly photo-interpreted which are more precise than those obtained with conventional standard designs. Moreover, it allows to reach the same precision of the estimator with a lower amount of money. These findings are very important from applied and operating points of view.
In this chapter we have also discussed the optimal AGSPRN procedure obtained when the population values of the area correctly photo-interpreted are not completely known. The results coincide with those derived in Section 5.4.2, with exception of some light differences in $Case$ 2 and $Case$ 3, confirming the validity of this procedure also in applied contexts.

Chapter 6

The AGSPRN procedure in an infinite population context: convergence properties of the estimator

6.1 Introduction

In a finite population context the definitions of consistency and convergence of an estimator are quite different respect to those related to an infinite population setting.
Following the definition of consistency for finite population stated by Cochran [1997, pp 21], an estimator is called *consistent* if the estimates become exactly equal to the population value when n=N, that is when the sample consists of the whole population. This definition has some limitations. For instance, the sample total is considered consistent according to this statement even though it is a very bad estimator for the population total. Hence, we present another definition of consistency [Cicchitelli et al., 1992, pp 57–58].
An estimator $\widehat{\theta}_n$ of θ is consistent if and only if the following holds:

$$lim_{\substack{n\to\infty \\ N\to\infty \\ \frac{n}{N}<1}} Pr(|\widehat{\theta}_n - \theta| < \epsilon) = 1,$$

where ϵ is a whatever small positive quantity, n is the sample size, N is the population size and θ does not change when N increases.
In order to study the convergence properties of the estimator generated by the optimal AGSPRN procedure, let us call it optimal AGSPRN estimator, we move to an infinite population context, i.e. supposing N, the finite population size, very big. Etoré and Jourdain [2010] found similar results for a stratified Monte Carlo estimator. Also Bélisle and Melfi [2008], Flournoy et al. [2012], Tymofyeyev et al. [2012] discussed some prop-

erties of estimators generated by adaptive sequential procedure in an infinite population framework, devoting particular attention to binary data.

6.2 Setting

Let Y be a R^d-valued random variable such that $E(Y^2) < \infty$. We are interested in estimating $c = E(Y)$ using a stratified sampling. Let $(A_h)_{1\leq h\leq H}$ be a partition of R^d into H strata such that $p_h = P(Y \in A_h)$. The values p_h are supposed to be known and positive for each integer $h \in [1, H]$. Given this partition, c is equal to $\sum_{h=1}^{H} p_h E(Y_h)$, where Y_h is distributed according to the conditional law of Y given $Y \in A_h$. The stratified estimator for c is:

$$\widehat{c} = \sum_{h=1}^{H} \frac{p_h}{N_h} \sum_{i=1}^{N_h} Y_h^i = \frac{1}{N} \sum_{h=1}^{H} \frac{p_h}{q_h} \sum_{h=1}^{q_h N} Y_h^i,$$

where N_h are i.i.d drawings of Y_h, Y_h^i indicates the unit i belonging to stratum h, $N = \sum_{h=1}^{H} N_h$ and $q_h = \frac{N_h}{N}$.
We have $E(\widehat{c}) = c$ and the variance of the estimator is equal to

$$V(\widehat{c}) = \sum_{h=1}^{H} \frac{p_h^2 \sigma_h^2}{N_h} = \frac{1}{N} \sum_{h=1}^{H} \frac{p_h^2 \sigma_h^2}{q_h} = \frac{1}{N} \sum_{h=1}^{H} (\frac{p_h^2 \sigma_h^2}{q_h})^2 q_h \geq \frac{1}{N} \sum_{h=1}^{H} (\frac{p_h^2 \sigma_h^2}{q_h} q_h)^2, \tag{6.1}$$

where $\sigma_h^2 = V(Y|Y \in A_h)$ for all $1 \leq h \leq H$. We consider that $\sigma_h > 0$ for at least one index h (*Condition 1*).
In our finite population context, we have $c = E(Y) = \overline{Y}$ and $W_h = p_h$, that is the probability of selecting one unit from stratum h. Moreover, $\widehat{c} = \overline{y}_{st}$, $V(\widehat{c}) = V(\overline{y}_{st})$, $N_h = n_h$, for $h \in 1, ..., H$ and $N = n$.
The lower bound of the variance in (6.1) is reached when the units are drawn from each stratum according to the *optimal* (Neyman) allocation:

$$q_h = \frac{p_h \sigma_h}{\sum_{j=1}^{H} p_j \sigma_j} =: q_h^*, \quad \forall 1 \leq h \leq H.$$

Substituting q_h^* in (6.1) we then have:

$$V(\widehat{c}) = \frac{1}{N} (\sum_{h=1}^{H} p_h^2 \sigma_h^2)^2 =: \frac{\sigma_*^2}{N}.$$

Usually both $E(Y_h)$ and σ_h are not known for all $h = 1, ..., H$, hence it is necessary to sample in steps. At each step we estimate the conditional variances and the allocations of the drawings in each stratum. This gives rise to an adaptive stratified estimator.

6.3 A modified version of the AGSPRN procedure for infinite population

A little modification of the AGSPRN procedure is fundamental to achieve some convergence properties. Let N_h^k denote the units selected in stratum A_h till the end of step k. The increments $M_h^k = N_h^k - N_h^{(k-1)}$ are obtained at each step k using the information contained in the N_{k-1} first drawings.
The modified AGSPRN procedure is applied as following:
At step **k=1** set $\widehat{\sigma}_h^0 = 1$ for all integer $h \in [1, H]$.
At step $\mathbf{k \geq 2}$ compute the strata standard deviations using the information gained in the previous steps:

$$\widehat{\sigma}_h^{k-1} = \sqrt{\frac{1}{N_h^{k-1}} \sum_{j=1}^{N_h^{k-1}} (Y_h^j)^2 - (\frac{1}{N_h^{k-1}} \sum_{j=1}^{N_h^{k-1}} Y_h^j)^2}.$$

Then we make at least one drawing in each stratum. This is the element that modifies the standard AGSRPN procedure. It ensures the convergence of the estimator and of the $\widehat{\sigma}_h^k$. In the AGSPRN procedure some strata can also be not sampled at step $k \geq 2$. This condition imposes to select at least one unit from each stratum and makes the AGSPRN procedure be less efficient.

We then have:

$$\forall 1 \leq h \leq H \quad M_h^k = 1 + \tilde{m}_h^k \qquad \texttt{with } \tilde{m}_h^k \in \mathcal{N}, \tag{6.2}$$

where $\sum_{h=1}^H \tilde{m}_h^k = N_k - N_{k-1} - H$, and possibly $\tilde{m}_h^k = 0$ for some indexes. In our AGSPRN procedure $q = N_k - N_{k-1}$ and we choose $\tilde{m}_h^k = \lfloor m_h^k \rfloor$, where $\lfloor \cdot \rfloor$ indicates the integer part rounding to the floor and

$$\begin{aligned} m_h^k &= \frac{\left[(N^{k-1} + q - H)\frac{p_h \widehat{\sigma}_h^{k-1}}{\sum_{j=1}^H p_j \widehat{\sigma}_j^{k-1}} - N_h^{k-1}\right]}{\sum_{i=1}^H \left[(N^{k-1} + q - H)\frac{p_i \widehat{\sigma}_i^{k-1}}{\sum_{j=1}^H p_j \widehat{\sigma}_j^{k-1}} - N_i^{k-1}\right]}(q - H) = \\ &= \frac{\left[(N^{k-1} + q - H)\frac{p_h \widehat{\sigma}_h^{k-1}}{\sum_{j=1}^H p_j \widehat{\sigma}_j^{k-1}} - N_h^{k-1}\right]}{[(N^{k-1} + q - H) - N^{k-1}]}(q - H) = \\ &= (N^{k-1} + q - H)\frac{p_h \widehat{\sigma}_h^{k-1}}{\sum_{j=1}^H p_j \widehat{\sigma}_j^{k-1}} - N_h^{k-1}, \quad \forall 1 \leq h \leq H, \end{aligned} \tag{6.3}$$

with the convention that $\tilde{m}_h^1 = \lfloor n_0 p_h \rfloor$, $h = 1, ..., H$, where n_0 is the pilot sample size of our AGSPRN procedure.

Proposition 1.1 If $E|Y| < \infty$, then

$$\widehat{c}^k \xrightarrow{k\to\infty} c \qquad a.s.$$

If moreover, $E(Y^2) < \infty$, then, a.s,

$$\forall 1 \le h \le H \quad \widehat{\sigma}_h^k \xrightarrow{k\to\infty} \sigma_h \qquad and \qquad \sum_{h=1}^{H} p_h \widehat{\sigma}_h^k \xrightarrow{k\to\infty} \sigma_*$$

The validity of the previous proposition is ensured by the strong low of large numbers and by the fact that the selection of at least one unit at each step in each stratum (condition 6.2) allows N_h^k to converge to infinity when $k \to \infty$.

6.4 Convergence Properties

Let us consider the following theorem.

Theorem 2.1

Given *Condition 1*, $E(Y^2) < \infty$ and $k/N_k \to 0$ as $k \to \infty$, then using the AGSPRN procedure (slightly modified) we have:

$$\sqrt{N_k}(\widehat{c}^k - c) \xrightarrow[k\to\infty]{inlaw} \mathcal{N}(0, \sigma_*^2)$$

To prove the theorem we need Proposition 2.1 and Proposition 2.2.

Proposition 2.1

If $E(Y^2) < \infty$ and

$$\forall 1 \le h \le H \quad \frac{N_h^k}{N^k} \xrightarrow[k\to\infty]{} q_h^* \quad a.s., \tag{6.4}$$

then

$$\sqrt{N_k}(\widehat{c}^k - c) \xrightarrow[k\to\infty]{inlaw} \mathcal{N}(0, \sigma_*^2).$$

Proposition 2.2

Given *Condition 1*, $E(Y^2) < \infty$ and $k/N_k \to 0$ as $k \to \infty$, then using the AGSPRN procedure (slightly modified) we get

$$\forall 1 \le h \le H \quad \frac{N_h^k}{N^k} \xrightarrow[k\to\infty]{} q_h^* \quad a.s..$$

Proof of Proposition 2.1

To prove the proposition 2.1 we need the CLT for martingales that we recall below.

Central limit theorem for martingales

Let $(\mu_n)_{n\in N}$ be a square-integrable $(\mathcal{F}_n)_{n\in N}$ vector martingale. Suppose for a deterministic sequence (γ_n) increasing to $+\infty$ we have,

i)
$$\frac{\langle\mu\rangle_n}{\gamma_n} \xrightarrow[n\to\infty]{\mathbb{P}} \Gamma$$

ii) The Lindeberg condition is satisfied, i.e. for all $\epsilon > 0$

$$\frac{1}{\gamma_n}\sum_{k=1}^{n} E\left[||\mu_k - \mu_{k-1}||^2 1_{\{||\mu_k-\mu_{k-1}||\geq\epsilon\sqrt{\gamma_n}\}}|\mathcal{F}_{k-1}\right] \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Then
$$\frac{\mu_n}{\sqrt{\gamma}_n} \xrightarrow[n\to\infty]{inlaw} \mathcal{N}(0,\Gamma).$$

In our case, we have:

$$\sqrt{N_k}(\widehat{c}^k - c) = \begin{pmatrix} p_1\frac{N^k}{N_1^k} \\ \vdots \\ p_H\frac{N^k}{N_H^k} \end{pmatrix}' \frac{1}{\sqrt{N^k}} \begin{pmatrix} \sum_{j=1}^{N_1^k}(Y_1^j - E(Y_1)) \\ \vdots \\ \sum_{j=1}^{N_H^k}(Y_H^j - E(Y_H)) \end{pmatrix}.$$

We can apply the CLT for martingales by setting $\mu_k := (\sum_{j=1}^{N_1^k}(Y_1^j - E(Y_1)), \ldots, \sum_{j=1}^{N_H^k}(Y_1^j - E(Y_H)))'$. If we define the filtration $(\mathcal{G}_k)_{k\in\mathcal{N}}$ by $(\mathcal{G}_k) = \sigma(1_{j\leq N_h^k}Y_h^j, 1\leq h\leq H, 1\leq j)$, it can be shown that (μ_k) is a $(\mathcal{G}_k)$-martingale. This is thanks to the fact that N_h^k are $(\mathcal{G}_{k-1})$ measurable. We can show that:

$$\frac{\langle\mu\rangle_k}{N^k} = diag(\frac{N_1^k}{N_k}\widehat{\sigma}_1^2, \ldots, \frac{N_H^k}{N_k}\widehat{\sigma}_H^2)$$

where diag(v) denotes the diagonal matrix with vector v on the diagonal. Thanks to (6.4), we thus obtain

$$\frac{\langle\mu\rangle_k}{N^k} \xrightarrow[k\to\infty]{a.s.} diag(q_1^*\sigma_1^2, \ldots, q_H^*\sigma_H^2)$$

This result with a use of the CLT for martingales and of Slutzky's theorem could lead to the desired result.

The Lindeberg's condition is not easy to prove for the sequence in k, hence Etoré and

Jourdain [2010] proved it for the sequence $(\tilde{c}^n)$ of estimators of c, such that $(\widehat{c}^k = \tilde{c}^{N^k})$. We will show that $(\widehat{c}^k)$ is a subsequence of $(\tilde{c}^n)$, therefore the results obtained for the latter can be extended to the former sequence. Let us see below some details.
Let $n \in \mathcal{N}$ and let $k \in \mathcal{N}$ such that $N^{k-1} < n \leq N^k$. The quantity ν_h^n is defined as the numbers of drawings in the h-th stratum among the first n drawings and we can write that $\sum_{h=1}^H \nu_h^n = n$. The following estimator is then defined:

$$\tilde{c}^n := \sum_{h=1}^{H} \frac{p_h}{\nu_h^n} \sum_{j=1}^{\nu_h^n} Y_h^j$$

Allocation rule for the single drawing

For $n = 0, \nu_h^n = 0$, for $h = 1, ..., H$.
1. For k> 0 set $r_h^k := \frac{N_h^k - N_h^{k-1}}{N^k - N^{k-1}}$ for $h = 1, ..., H$
2. For $N^{k-1} < n \leq N^k$, and given the ν_h^{n-1}'s find

$$h_n = \texttt{argmax}_{1 \leq h \leq H} \left(r_h^k - \frac{\nu_h^{n-1} - N_h^{k-1}}{n - N^{k-1}} \right)$$

If h_n is not unique, choose the index with the greatest r_h^k. If it is not enough to make h_n unique, choose the greatest h.
3. Set $\nu_{h_n}^n = \nu_{h_n}^{n-1} + 1$ and $\nu_h^n = \nu_h^{n-1}$ if $h \neq h_n$
There exist always an index h for which $r_h^k - \frac{\nu_h^{n-1} - N_h^{K-1}}{n - N^{k-1}} > 0$, since

$$\sum_{h=1}^{H} \frac{\nu_h^{n-1} - N_h^{K-1}}{n - N^{k-1}} = \frac{n - 1 - N^{k-1}}{n - N^{k-1}} < 1 = \sum_{h=1}^{H} r_h^k.$$

Moreover, for the first $n \in \{N^{k-1} + 1, ..., N^k\}$ such that $\nu_h^{n-1} = N_h^k$ in the h-th stratum, $r_h^k - \frac{\nu_h^{n-1} - N_h^{K-1}}{n - N^{k-1}} \leq 0$ and $\nu_h^{n'} = \nu_h^n = N_h^k$ for $n \leq n' \leq N^k$.
The consequence is that

$$\nu_h^{N^k} = N_h^k, \quad \forall 1 \leq h \leq H, \quad \forall k \in \mathcal{N},$$

and

$$\widehat{c}^k = \tilde{c}^{N^k}.$$

Proposition 2.3. Under the following assumptions of Proposition 2.1,

$$\sqrt{n}(\tilde{c}^n - c) \xrightarrow[n \to \infty]{inlaw} \mathcal{N}(0, \sigma_*^2).$$

To verify the Lindeberg condition in order to prove this proposition we need the following

lemma.

Lemma 2.1 When (6.4) holds, then

$$\forall 1 \leq h \leq H, \ \frac{\nu_h^n}{n} \xrightarrow[n\to\infty]{} q_h^* \ a.s.$$

Proof. For $x \in \mathcal{R}_+^*$ or $n \in \mathcal{N}$, the integer k is such that $N^{k-1} < x$ and $n \leq N^k$.
For any $n \in \mathcal{N}$ and $1 \leq h \leq H$ we have:

$$\frac{\nu_h^n}{n} = \frac{n - N^{k-1}}{n} \cdot \frac{\nu_h^n - N^{k-1}}{n - N^{k-1}} + \frac{N^{k-1}}{n} \cdot \frac{N_h^{k-1}}{N^{k-1}},$$

and we define for $x \in \mathcal{R}_+$ the following quantity:

$$f(x) := \frac{x - N^{k-1}}{x} \cdot \frac{N_h^k - N_h^{k-1}}{N^k - N^{k-1}} + \frac{N^{k-1}}{x} \cdot \frac{N_h^{k-1}}{N^{k-1}}.$$

We will see that, as n tends to infinity, $f(n)$ tends to q_h^* and $f(n) - \frac{\nu_h}{n}$ tends to zero.
The derivative of f on any interval $(N^{k-1}, N^k]$ shows that this function is monotonic on it. We know that $f(N^{k-1}) = \frac{N_h^{k-1}}{N^{k-1}}$ and $f(N^k) = \frac{N_h^k}{N^k}$. So if $\frac{N_h^k}{N^k}$ tends to q_h^* as k tends to infinity, the following holds:

$$f(n) \xrightarrow[n\to\infty]{} q_h^*. \tag{6.5}$$

As $r_h^k = \frac{N_h^k - N_h^{k-1}}{N^k - N^{k-1}}$ we have:

$$\frac{\nu_h^k}{n} - f(n) = \frac{n - N^{k-1}}{n} \left(\frac{\nu_h^n - N^{k-1}}{n - N^{k-1}} - r_h^k \right). \tag{6.6}$$

If we show that

$$r_h^k - \frac{H-1}{n - N_h^{k-1}} < \frac{\nu_h^n - N_h^{k-1}}{n - N^{k-1}} < r_h^k + \frac{1}{n - N^{k-1}}, \tag{6.7}$$

we can derive from (6.6) and (6.7) the following result:

$$-\frac{H-1}{n} < \frac{\nu_h^k}{n} - f(n) < \frac{1}{n},$$

which combined with (6.5) allows us to get the desired conclusion.
In order to check (6.7), we start showing that

$$\frac{\nu_h^n - N_h^{k-1}}{n - N^{k-1}} < r_h^k + \frac{1}{n - N^{k-1}}. \tag{6.8}$$

Two cases are discussed. Either no drawings at all are made in stratum h between N_h^{k-1} and n, that is $\nu_h^{n'} = N_h^{k-1}$ for all $N^{k-1} < n' \leq n$, then (6.8) is trivially verified.

Either some drawings are made between N_h^{k-1} and n. Let us denote by n' the index of the last one, i.e. $\nu_h^n = \nu_h^{n'} = \nu_h^{n'-1} + 1$. If a drawing is made at n' we see that $\frac{\nu_h^{n'} - N_h^{k-1}}{n' - N^{k-1}} < r_h^k$. Thus, the following holds:

$$\frac{\nu_h^{n'-1} - N_h^{k-1}}{n - N^{k-1}} \leq \frac{\nu_h^{n'-1} - N_h^{k-1}}{n' - N^{k-1}} < r_h^k$$

and

$$\frac{\nu_h^{n} - N_h^{k-1}}{n - N^{k-1}} = \frac{\nu_h^{n'-1} + 1 - N_h^{k-1}}{n' - N^{k-1}}$$

Hence (6.8) is showed for this case.
Knowing that $1 = \sum_{h=1}^H r_h^k = \sum_{h=1}^H \frac{\nu_h^n - N_h^{k-1}}{n - N^{k-1}}$ we get

$$\frac{\nu_h^n - N_h^{k-1}}{n - N^{k-1}} = r_h^k + \sum_{h \neq j} \left(r_j^k - \frac{\nu_h^n - N_h^{k-1}}{n - N^{k-1}} \right)$$

Using this result and expression (6.8) we obtain (6.7).

Proof of Proposition 2.3. For $n \geq N^1$, we have $\nu_h^n \geq 1$ for all $1 \leq h \leq H$ and we can write

$$\sqrt{n}(\widehat{c}^n - c) = \begin{pmatrix} p_1 \frac{n}{\nu_1^n} \\ \vdots \\ p_H \frac{n}{\nu_H^n} \end{pmatrix}' \frac{1}{\sqrt{n}} \mu_n, \tag{6.9}$$

where

$$\mu_n = \begin{pmatrix} \sum_{j=1}^{\nu_1^n} (Y_1^j - E(Y_1)) \\ \vdots \\ \sum_{j=1}^{\nu_H^n} (Y_H^j - E(Y_H)) \end{pmatrix}.$$

It can be shown that (μ_n) is a $(\mathcal{F}_n)$-martingale, where $(\mathcal{F}_n)_{n \in \mathcal{N}}$ is the filtration defined as $(\mathcal{F}_k) := \sigma(1_{j \leq \nu_h^n} Y_h^j, 1 \leq h \leq H, 1 \leq j)$. Indeed, for $n \in \mathcal{N}$, let $k \in \mathcal{N}$ such that $N^{k-1} < n \leq N^k$. For $h = 1, .., H$, N_h^{k-1} and N_h^k are respectively $\mathcal{F}_{N^{k-1}}$ and $(\mathcal{F}_{N^{k-2}})$ measurable. For each $1 \leq h \leq H$, the variable ν_h^k is $\mathcal{F}_{N^{k-1}}$ measurable since it depends on the N_h^{k-1}'s and the N_h^k's. Thus μ_n is $\mathcal{F}_n$-measurable and it can be shown easily that $\mathbb{E}[\mu_{n+1} | \mathcal{F}_n] = \mu_n$.
We wish to use Theorem 2.2 with $\gamma_n = n$. The term $\text{diag}(a_h)$ will denote the $H \times H$ matrix having null coefficients except the h-th element of the diagonal which is equal to a_h.
Let us verify the Lindeberg condition. By using the sequence (h_n) defined before, we

obtain the following result:

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{n}\mathbb{E}[||\mu_l-\mu_{l-1}||^2\mathbf{1}_{\{||\mu_l-\mu_{l-1}||>\epsilon\sqrt{n}}|\mathcal{F}_{l-1}]= \\
&=\frac{1}{n}\sum_{l=1}^{n}\mathbb{E}[|Y_{hl}^{\nu_{hl}^l}-\mathbb{E}(Y_{hl})|^2\mathbf{1}_{\{|Y_{hl}^{\nu_{hl}^l}-\mathbb{E}(Y_{hl})|>\epsilon\sqrt{n}}|\mathcal{F}_{l-1}]= \\
&\leq\frac{1}{n}\sum_{l=1}^{n}\texttt{sup}_{1\leq h\leq H}\mathbb{E}[|Y_h-\mathbb{E}(Y_h)|^2\mathbf{1}_{\{|Y_h-\mathbb{E}(Y_h)|>\epsilon\sqrt{n}}]= \\
&=\texttt{sup}_{1\leq h\leq H}\mathbb{E}[|Y_h-\mathbb{E}(Y_h)|^2\mathbf{1}_{\{|Y_h-\mathbb{E}(Y_h)|>\epsilon\sqrt{n}}].
\end{aligned} \tag{6.10}$$

As

$$\texttt{sup}_{1\leq h\leq H}\mathbb{E}[|Y_h-\mathbb{E}(Y_h)|^2\mathbf{1}_{\{|Y_h-\mathbb{E}(Y_h)|>\epsilon\sqrt{n}}]\xrightarrow[n\to\infty]{}0,$$

the Lindeberg condition is proved.
We now show point i) of the CLT for martingales. We have,

$$\begin{aligned}
\langle\mu\rangle_n&=\sum_{l=1}^{n}\mathbb{E}[(\mu_k-\mu_{k-1})(\mu_k-\mu_{k-1})'|\mathcal{F}_{k-1}]= \\
&=\sum_{l=1}^{n}\texttt{diag}(\mathbb{E}[|Y_{hk}^{\nu_{hk}^k}-\mathbb{E}(Y_{hk})|^2]) \\
&=\sum_{l=1}^{n}\texttt{diag}(\sigma^2_{h_k}).
\end{aligned} \tag{6.11}$$

Consequently, by using Lemma 2.1 we get:

$$\frac{\langle\mu\rangle_n}{n}=\texttt{diag}(\frac{\nu_1^n}{n}\sigma_1^2,\dots,\frac{\nu_H^n}{n}\sigma_H^2)\xrightarrow[n\to\infty]{}\texttt{diag}((q_1^*\sigma_1^2,...,q_H^*\sigma_H^2))\ \texttt{a.s.},$$

Theorems 2.2 implies that:

$$\frac{\langle\mu\rangle_n}{n}\xrightarrow[n\to\infty]{inlaw}\mathcal{N}(0,\texttt{diag}(q_1^*\sigma_1^2,\dots,q_H^*\sigma_H^2)) \tag{6.12}$$

By Lemma 2.1 the following holds:

$$(p_1\frac{n}{\nu_1^n},\cdots,p_H\frac{n}{\nu_H^n})\xrightarrow[n\to\infty]{}(\frac{p_1}{q_1^*},\cdots,\frac{p_h}{q_H^*})\ \texttt{a.s.} \tag{6.13}$$

Finally, by using Slutsky's Theorem 6.9, 6.12 and 6.13, we obtain:

$$\sqrt{n}(\tilde{c}^n-c)\xrightarrow[n\to\infty]{inlaw}\mathcal{N}(0,\sigma_*^2).$$

Proof of Proposition 2.2

In order to prove Proposition 2.2, we need to know that Proposition 1.1 implies convergence of $\rho_h^k = \frac{p_h \widehat{\sigma}_h^k}{\sum_{j=1}^H p_j \widehat{\sigma}_j^k}$ as $k \to +\infty$. The following Lemma is also fundamental.

Lemma 2.2. Under the assumptions of Theorem 2.1, the following holds

$$\forall 1 \leq h \leq H, \rho_h^k \xrightarrow[k\to\infty]{} q_h^* \quad a.s.$$

Now, we can proceed with the proof of Proposition 2.2.
For all $1 \leq h \leq H$ we have $\frac{N_h^k}{N^k} = \frac{k+\sum_{l=1}^k \tilde{m}_h^l}{N^k}$. Since we know that $m_h^l - 1 < \tilde{m}_h^l < m_h^l + 1$, then:

$$\frac{\sum_{l=1}^{k-1} \tilde{m}_h^l + m_h^k}{N^k} \leq \frac{N_h^k}{N^k} \leq \frac{2k}{N^k} + \frac{\sum_{l=1}^{k-1} \tilde{m}_h^l + m_h^k}{N^k}. \tag{6.14}$$

We get:

$$\frac{\sum_{l=1}^{k-1} \tilde{m}_h^l + m_h^k}{N^k} = \frac{N_h^{k-1} + (N^{k-1} + q - H)\frac{p_h \widehat{\sigma}_h^{k-1}}{\sum_{j=1}^H p_j \widehat{\sigma}_j^{k-1}} - N_h^{k-1}}{N^k} = \\ = \frac{(N^k - H)\rho_h^{k-1}}{N^k} = \rho_h^{k-1} - \frac{H\rho_h^{k-1}}{N^k},$$

where the sequence (ρ_h^{k-1}) defined by $\tilde{\rho}_h^n = \rho_h^n$ converges a.s. to q_h^* as n tends to infinity, by Lemma 2.2. The second term converges to 0 when k tends to infinity. Hence $\frac{\sum_{l=1}^{k-1} \tilde{m}_h^l + m_h^k}{N^k}$ converges to q_h^* as $k \to \infty$, and $\frac{N_h^k}{N^k} \to q_h^*$ by (6.14) when $k/N^k \to 0$.

6.5 Discussion

The aim of this chapter has been to show some important convergence properties about the stratified mean estimator generated by an adaptive sequential procedure that pursues Neyman's allocation in the infinite population context. More specifically, given a variable of interest Y, we have proved that the distribution of the estimator of $E(Y)$ converges to a Normal distribution having mean $E(Y)$ and variance equal to the minimum variance that the estimator can reach, the one achieved by Neyman's allocations. The convergence holds if either the number of steps k or the number of total drawings N tend to infinity. Hence, in the setting described for our AGSPRN procedure, the convergence would hold also as q tends to infinity if we work with infinite population framework. The key elements of this proof have been to use martingales arguments and to show that the sequence in k is a subsequence of that in n, the drawings.

These results are obtained for an adaptive sequential procedure that has the same setting and the same adaptive rule as our AGSPRN procedure. However, the latter has to be slightly modified in order to allow the convergence properties to hold, that is, providing the selection of at least one unit in each stratum at each step. This can become very cumbersome, especially in applied situations. Moreover, Carfagna [2007] introduced the permanent random technique to the adaptive sequential procedure in order to avoid to select at each step a complete stratified random sample, as in Thompson and Seber [1996, pp 189-191]. Hence, further research should investigate the behaviour of our AGSPRN estimator also in a *finite population context*, when both the numbers of units of the population N and the sampled units are large and when the sample units are selected not necessarily in all the strata. This is not a trivial task, as also Cicchitelli et al. [1992, pp 294-295] underlined. Thanks to the permanent random numbers technique, at each step k, the sample selected in each stratum can be considered as a simple random sample without replacement. If the approximation to the Normal holds in each stratum, the stratified mean estimator can be considered approximately normal, since it is a linear combination of the strata sample means. Zacks [2009, pp 84-86] investigated by simulations the approximation to the Normal of the mean estimator generated by a simple random sampling without replacement. Usually, a correct stratification gives rise to homogeneous strata and leads the strata means to be normally distributed.

When the strata sizes are small (as in several practical applications), it is necessary to use the Student's t-distribution, with the complication to compute its degrees of freedom. If the population size and the sample size are homogeneous among strata, if the distributions are approximately normal inside each stratum and the strata variances are similar, the degrees of freedom are set equal to $n - H$, where n is the total sample size and H is the number of strata. However, these regular conditions are particularly rare, hence the degrees of freedom have to be reduced in order to increase the confidence interval size. An approximated formula to compute the degree of freedom (gl) for general conditions is:

$$gl = \frac{[\sum_{h=1}^{H} \widehat{V}(W_h \overline{y}_h)]^2}{\sum_{h=1}^{H} \widehat{V}^2(W_h \overline{y}_h)/(n_h - 1)},$$

where $\widehat{V}(W_h \overline{y}_h) = W_h^2(1 - f_h)s_h^2/n_h$.

In stratified random sampling, the situation is more complex than in simple random sampling, because several factors affect the distribution of the sample mean such as the number of strata, the stratification procedure, the sample allocations. This complexity, that increases in an adaptive sequential setting, prevents to give some general instructions about the convergence properties of the AGSPRN estimator in a finite population context.

Chapter 7

Conclusions

One of the aim of this thesis has been to satisfy the need of finding a flexible sampling procedure for a *stratified finite population*, which allows to generate timely and efficient estimates of the parameter of interest, in the presence of a linear cost function. This need especially arises in real applications, where time and resources are often limited, and efficient estimates are generally requested.

Adaptive designs have been the first tools we have considered, since they allow more flexibility than the *conventional designs*. We have explored in Chapter 2 the literature concerning adaptive and sequential procedures, devoting particular attention to the finite population context. Stein's two step procedure, Ray's, Chow and Robbins's adaptive sequential procedures for infinite population have been briefly analysed. In the finite populations context, we have particularly focused on the two steps adaptive procedure with permanent random numbers (TSPRN, Carfagna [2007]) and the adaptive sequential procedure with permanent random numbers (ASPRN, Carfagna and Marzialetti [2009b]), which were proved to be more efficient than Thompson and Seber's method (Thompson and Seber [1996, pp 189-191]). Carfagna et al. [2012] showed that, when a cost function is introduced and the step cost is high, the ASPRN may be less efficient than the TSPRN. Thus, the presence of a *cost function* is a key element for the identification of the optimal adaptive procedure which can be seen as a compromise solution between TSPRN and ASPRN, able to reduce the costs to be suffered by ASPRN when the cost per step becomes relatively high, preserving the advantages that an adaptive sequential procedure has in terms of efficiency of the estimators. Determining this optimal adaptive sequential procedure has been the aim of our work. Hence, in Chapter 3 we have proposed an adaptive group sequential procedure for stratified sampling with the use of permanent random numbers (AGSPRN), which consists in adding q units at each step until the Kth step, where the stopping rule is satisfied. The TSPRN and the ASPRN can be derived from it as particular and usually less efficient cases. We have chosen different stopping rules, concerning the estimator variance, the total cost and the risk. Indeed, we have introduced

the approach based on the minimization of a risk function which is a convex combination of two standardized criteria: the cost of the procedure involving K steps and q units per step, and the variance of the estimator generated by the procedure. This approach can be useful in applied problems, where it is important to take into account the precision of the results but also the cost of reaching that precision. A procedure that balances the precision of the estimates and the cost to reach it, assigning to the two criteria different levels of importance, is essential.

Specifically, we have focused on the problem of determining the *optimal* AGSPRN procedure, that minimizes: 1) the variance of the estimator given a cost function and a budget constraint, 2) the total cost given a threshold on the estimator variance, 3) the risk function. These problems resulted to be analytically unsolvable, since the distribution of the variance of the estimator generated by the AGSPRN procedure is mathematically intractable. Hence, a Monte Carlo study has been performed to compute the value of the estimator variance for some AGSPRN procedures, with different values of q and K. The simulation experiment requires to have some inputs about the target population, which can be derived from the pilot sample. In Chapter 4 we have proposed an adaptive Monte Carlo, which allows to update, at each step, the distribution of the variable of interest in each stratum, in order to generate a population very close to the target one.

In Chapter 3 we have applied the Monte Carlo study directly to the target population, in order to show some properties of the optimal AGSPRN procedure. For instance, we have found out that, when the cost per step is not negligible, the optimal AGSPRN procedure is usually more efficient than ASPRN, TSPRN and also with respect to a simple stratified random sampling applied in just one phase, since the adaptive rule allows to generate a sample allocation very close to *Neyman's* one.

A key role is played by the cost function and by the values of its components. We have chosen a *linear cost function*, but the impact of different functions on the optimal procedure should be analyzed in future works. Especially, we have seen that the ratio c_n/c_k is relevant in choosing the optimal number of q and K. Hence, in applied problems, reducing some costs in favor of others can be fundamental to gain efficiency in the estimates.

Moreover, we have assessed the impact of different values of the cost components on the optimum number of steps and of sample units to be allocated at each step. We have noticed that an increase of the unit cost, under a fixed budget and a linear cost function, causes an increase of the number of steps and a decrease of the number of units per step. Obviously the total number of sample units decreases and, consequently, the variance of the estimator of the mean increases. The optimal AGSPRN procedure tends to maintain a high number of steps, since a decrease of the number of steps inflates the estimator variance more than a decrease of the number of sampling units per step. On the other side, if the cost per step increases, the number of optimal steps decreases, with a relatively

low effect on the total sample size and, consequently, on the variance of the estimator.

In Chapter 4, we have proposed a method useful to obtain the optimal AGSPRN procedure when the values of the variable of interest Y are not known for all the units in the population. This is the standard situations, thus the proposed technique can be convenient to estimate timely and efficiently a population target, when only a pilot sample is available. The optimal adaptive sequential procedure derived with this method is characterized by a variable number of units added at each step, that depends on the updated estimate of the population. Comparing the optimal AGSPRN procedure obtained when the population y-values are partially unknown with that found for a totally known population, we have noticed that they are very similar. This is an encouraging result. If a moderate decrease in precision arises, it is usually balanced by a small increase of the number of steps or units per step.

The proposed method can be implemented in many practical applications. In Chapter 5 we have applied it to find the optimal AGSPRN procedure in order to estimate the quality index of a land cover database, equal in this particular case to the percentage of the land area correctly photo-interpreted. Usually limited resources are devoted to the quality control operation of a land cover database, thus, timely and efficient estimates are required, in terms of cost, estimator precision and risk. The optimal AGSPRN procedure found when the population y-values are unknown is noticed to be close to that obtained for a totally known population. However, a particular attention has to be devoted to the formulation of the risk function, and, particularly to the value of λ in expression (3.9).
In the filed of territory management, when the cost of transport is high and the spatial autocorrelation of population units is moderate, sampling contiguous polygons can be cost efficient. Hence, adapting the proposed optimal AGSPRN procedure to cluster sampling of polygons can be a future development of this work, taking into account the spatial autocorrelation and the cost functions, which should incorporate also the cost of transport.
Additional efficiency may be gained adding q_k units at the kth step in such a way that a different number of units is added at each step according to the needs arisen on the way. Hence, a future development of this procedure is letting q varying along the steps. Moreover, the work can be extended considering the aim of getting information on more than one variable of interest, applying a multiple adaptive allocation along the steps (Bethel [1989], Ullah et al. [2014]).

Finally, some useful convergence properties are proved for a slightly modified version of the AGSPRN estimator, in the infinite population context. For instance, the distribution of the AGSPRN estimator is showed to converge to a Normal distribution having as expected value the real value of the parameter and as variance the minimum variance that a stratified mean estimator can reach, i.e. when the units are allocated to the strata according to the Neyman's rule. The convergence holds when either the number of steps

or the number of units per step go to infinity. In a finite population context proving the same properties is not easy, since when the strata initial sample sizes are small the normal approximation for the distribution of the mean estimator does not hold. Moreover, in a stratified random sampling the complexity is increased by the allocation rule and the stratification process. Showing convergence properties in a finite population context (supposing the population size is very big) for a mean estimator generated by a stratified adaptive sequential sampling procedure can be one of the challenge for future works.

In this thesis we have discussed a complex framework which considers sequential estimation with an adaptive allocation rule and continuous responses in a stratified finite population context and in the presence of a cost function. The intersection of all these aspects has not been explored yet; it is a complete open research topic, as also Hu and Rosenberger [2006, pp 158] stated for the clinical trials context, where a treatment corresponds to a stratum:

> For $K > 2$ treatments, we have discussed only binary responses. The optimization framework should apply to continuous responses, and this is a completely open problem.(...)
> Most larger clinical trials impose a sequential monitoring procedure to allow for early stopping. The basic statistical formulation requires determining the distribution of sequentially computed test statistics. Under response-adaptive randomization, this is a difficult task. Numerical studies have been performed (...).
> There has been little theoretical work done to this point, nor has there been any evaluation of sequential monitoring in the context of sequential estimation procedures such as the doubly-adaptive coin design. Regarding the intersection of sequential analysis and response-adaptive randomization, Rosenberger (2002) states: 'Surprisingly, the link between [response adaptive randomization] and sequential analysis has been tenuous as best, and this is perhaps the logical place to search for open research topics'.

Bibliography

A. Antognini and A. Giovagnoli. *Adaptive Designs for Sequential Treatment Allocation*. Chapman & Hall/CRC, 2015.

C. Bélisle and V. Melfi. Independence after adaptive allocation. *Statistics & Probability Letters*, 78(3):214–224, 2008.

D. R. Bellhouse. The central limit theorem under simple random sampling. *The American Statistician*, 55(4):352–357, 2001.

R. Benedetti, F. Piersimoni, and P. Postiglione. *Sampling Spatial Units for Agricultural Surveys*. Springer, 2015.

J. Bethel. Sample allocation in multivariate surveys. *Survey Methodology*, 15(1):47–57, 1989.

P. J. Bickel and D. A. Freedman. Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, pages 470–482, 1984.

E. Carfagna. Crop area estimates with area frames in the presence of measurement errors. In *Proceedings of ICAS-IV (56th Session, October 22-24)*, Advancing Statistical Integration and Analysis, pages 1–10. 2007.

E. Carfagna and F. J. Gallego. Using remote sensing for agricultural statistics. *International Statistical Review*, 73(3):389–404, 2005.

E. Carfagna and J Marzialetti. Quality improvement of land cover databases: a sequential approach via agreement measures. In *Proceeding of ENBIS8 (56th Session, October 22-24)*, European Network for Business and Industrial Statistics, pages 21–25. 2008.

E. Carfagna and J. Marzialetti. Continuous innovation of the quality control of remote sensing data for territory management. In *Statistics for Innovation*, pages 145–160. Springer, 2009a.

E. Carfagna and J. Marzialetti. Sequential design in quality control and validation of land cover data bases. *Journal of Applied Stochastic Models in Business and Industry (ASMBI)*, 25(2):195–205, 2009b.

E. Carfagna, J Marzialetti, and S. Maffei. Sequential and two phase sample designs for quality control. In *Proceeding XLIV Sci. (June 25-27)*, Meeting of the Italian Statistical Society. 2008.

E. Carfagna, P. Tassinari, M. Zagoraiou, S. Benni, and D. Torreggiani. Efficient statistical sample designs in a gis for monitoring the landscape changes. In Agostino Di Ciaccio, Mauro Coli, and Jose Miguel Angulo Ibanez, editors, *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Studies in Theoretical and Applied Statistics, pages 399–407. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-21036-5.

R. Chambers and R. Clark. *An Introduction to Model-Based Survey Sampling With Applications*, volume 37. OUP Oxford, 2012.

Y.S. Chow and H. Robbins. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statististics*, 36:457–462, 1965.

G. Cicchitelli, A. Herzel, and G. E. Montanari. *Il Campionamento Statistico.* Il mulino, 1992.

W.C. Cochran. *Sampling techniques.* Wiley, New York, 1997.

P.L. Conti and D. Marella. *Campionamento da Popolazioni Finite: il Disegno Campionario.* Springer Science & Business Media, 2012.

J. J. Cotter and C. Tomczak. An image analysis system to develop area sampling frames for agricultural surveys. *Photogrammetric Engineering and Remote Sensing*, 60(3): 299–306, 1994.

D.R. Cox. Estimation by double sampling. *Biometrika*, 39:pp. 217–227, 1952.

J. S Denne and C. Jennison. A group sequential t-test with updating of sample size. *Biometrika*, 87(1):125–134, 2000.

P. Erdős and A. Rényi. On the central limit theorem for samples from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Science*, 4: 49–57, 1959.

P. Etoré and B. Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, 12(3):335–360, 2010.

FAO. Vol. 1: current surveys based on area and list sampling methods. In *Multiple Frame Agricultural Surveys*, FAO statistical development series. 1996.

FAO. Vol. 2. In *Multiple Frame Agricultural Surveys*, FAO statistical development series. 1998.

N. Flournoy, C. May, and P. Secchi. Asymptotically optimal response-adaptive designs for allocating the best treatment: an overview. *International Statistical Review*, 80(2): 293–305, 2012.

J. Gallego, E. Carfagna, and S. Peedell. The use of corine land cover to improve area frame survey estimates. *Research in Official Statistics*, 2(2):99–122, 1999.

J. Gallego, E. Carfagna, and B. Baruth. Accuracy, objectivity and efficiency of remote sensing for agricultural statistics. *Agricultural Survey Methods*, pages 193–211, 2010.

M. Ghosh et al. *Sequential Estimation*. Wiley, New York, 1997.

J. Hájek. Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Science*, 5: 361–374, 1960.

J. Hardwick and Q. F. Stout. Algorithms for response adaptive sampling designs. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):118–122, 2009.

J. P. Hardwick and Q. F. Stout. Exact computational analyses for adaptive designs. *Lecture Notes-Monograph Series*, pages 223–237, 1995.

F. Hu and W. F. Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials*. John Wiley & Sons, 2006.

C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. CRC Press, 1999.

J. B. Kadane. Optimal dynamic sample allocation among strata. *Journal of Official Statistics*, 21(4):531, 2005.

Wei Liu. A k-stage sequential sampling procedure for estimation of normal mean. *Journal of Statistical Planning and Inference*, 65(1):109–127, 1997.

V. Melfi and C. Page. Variablility in adaptive designs for estimation of success probabilities. *Lecture Notes-Monograph Series*, pages 106–114, 1998.

V. Melfi and C. Page. Estimation after adaptive allocation. *Journal of Statistical Planning and Inference*, 87(2):353–363, 2000.

V. F Melfi, C. Page, and M. Geraldes. An adaptive randomized design with application to estimation. *Canadian Journal of Statistics*, 29(1):107–116, 2001.

S. Missiroli and E. Carfagna. Optimal adaptive group sequential procedure for finite populations in the presence of a cost function. In *Proceeding XLVIII Sci. (June 8-10)*, Meeting of the Italian Statistical Society. 2016.

C. C. Morgan and D. Stephen Coad. A comparison of adaptive allocation rules for group-sequential binary response clinical trials. *Statistics in medicine*, 26(9):1937–1954, 2007.

L. W. Morrison, D. R. Smith, C. C. Young, and D. W. Nichols. Evaluating sampling designs by computer simulation: a case study with the missouri bladderpod. *Population Ecology*, 50(4):417–425, 2008.

N. Mukhopadhyay. A new approach to determine the pilot sample size in two-stage sampling. *Communication in Statistics*, 34:1275–1295, 2005.

P. Muliere, A. M. Paganoni, and P. Secchi. Randomly reinforced urns for clinical trials with continuous responses. In *SIS-Proceedings of the XLIII Scientific Meeting*, pages 403–414, 2006a.

P. Muliere, A. M. Paganoni, and P. Secchi. A randomly reinforced urn. *Journal of Statistical Planning and Inference*, 136(6):1853–1874, 2006b.

H. Müller and H. Schäfer. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):886–891, 2001.

E. Ohlsson. Coordination of samples using permanent random numbers. In B. et al Cox, editor, *Business Survey Methods*, pages 153–169. John Wiley & Sons, Inc., 1995.

M. Posch, F. Koenig, M. Branson, W. Brannath, C. Dunger-Baldauf, and P. Bauer. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine*, 24(24):3697–3714, 2005.

W.D Ray. Sequential confidence intervals for the mean of a normal distribution with unknown variance. *Journal of the Royal Statistical Society, Series B*, 19:133–143, 1957.

W. F. Rosenberger and F. Hu. Bootstrap methods for adaptive designs. *Statistics in Medicine*, 18(14):1757–1767, 1999.

W. F. Rosenberger and J. M. Lachin. *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons, 2015.

W. F Rosenberger and T.N. Sriram. Estimation for an adaptive allocation design. *Journal of Statistical Planning and Inference*, 59(2):309–319, 1997.

W. F. Rosenberger, N. Stallard, A. Ivanova, C. N. Harper, and M. L. Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913, 2001.

M. Salehi and J. Brown. Complete allocation sampling: an efficient and easily implemented adaptive sampling design. *Population Ecology*, 52(3):451–456, 2010.

M. Salehi, M. Moradi, J. A. Brown, and D. Smith. Efficient estimators for adaptive stratified sequential sampling. *Journal of Statistical Computation and Simulation*, 80 (10):1163–1179, 2010.

L. Sandvik, J. Erikssen, P. Mowinckel, and E. A. Roedlan. A method for determining the size of internal pilot studies. *Statistics in Medicine*, 15(14):1587–1590, 1996.

C. Stein. A two sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16:243–258, 1945.

A. H. Strahler, L. Boschetti, G. M. Foody, M. A. Friedl, M. C. Hansen, M. Herold, P. Mayaux, J. T. Morisette, S. V. Stehman, and C. E. Woodcock. Global land cover validation: recommendations for evaluation and accuracy assessment of global land cover maps. Technical report, Office for Official Publications of the European Communities, Luxemburg, March 2006.

S. K. Thompson. *Sampling*. John Wiley & Sons, 2012.

S.K. Thompson and G.A.F. Seber. *Adaptive Sampling*. Wiley, New York, 1996.

Y. Tymofyeyev, W. F. Rosenberger, and F. Hu. Implementing optimal allocation in sequential binary response experiments. *Journal of the American Statistical Association*, 2012.

A. Ullah, J. Shabbir, Z. Hussain, and B. Al-Zahrani. Estimation of finite population mean in multivariate stratified sampling under cost function using goal programming. *Journal of Applied Mathematics*, 2014, 2014.

A Wald. *Sequential Analysis*. Wiley, New York, 1947.

S. Zacks. *Stage-Wise Adaptive Designs*. Wiley, New York, 2009.