UNIVERSITÉ DE NEUCHÂTEL

DOCTORAL THESIS

Random sampling with repulsion

Author:
Matthieu WILHELM

Supervisor:
Prof. Yves TILLÉ

Examiners:
Prof. Michel BENAÏM, Université de Neuchâtel
Dr. Guillaume CHAUVET, ENSAI
Prof. Arnaud DOUCET, University of Oxford
Prof. Philip B. STARK, University of California, Berkeley

A thesis submitted in fulfillment of the requirements
for the degree of Doctorat ès Sciences

in the

Institut de Statistique
Faculté des Sciences

September 25, 2017

UNIVERSITÉ DE NEUCHÂTEL

Faculté des Sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel – Suisse
Tél : + 41 (0)32 718 21 00
E-mail : secretariat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Monsieur Matthieu WILHELM

Titre:

“Random Sampling with Repulsion”

sur le rapport des membres du jury composé comme suit:

- Prof. Yves Tillé, Université de Neuchâtel, Suisse
- Prof. Michel Benaïm, Université de Neuchâtel, Suisse
- Prof. Arnaud Doucet, Oxford University, UK
- Prof. Philip B. Stark, University of California, Berkeley, USA
- Dr Guillaume Chauvet, ENSAI-INSEE, Rennes, France

Neuchâtel, le 21 septembre 2017

Le Doyen, Prof. R. Bshary

A ceux qui m'ont précédé et qui me précèdent:
Angelo e Elisa Mozzi-Strik,
François et Françoise Wilhelm,
Luc et Luisa Wilhelm.

UNIVERSITÉ DE NEUCHÂTEL

Abstract

Faculté des Sciences

Doctorat ès Sciences

Random sampling with repulsion

by Matthieu WILHELM

In this thesis, we explore the concept of repulsion and exploit it to develop new sampling designs.

The first chapter is a review of the literature about sampling in finite population and consists of two distinct parts. In the first one, we suggest three principles as guidelines for designing a survey: randomization, overrepresentation and restriction. In the second part, we review the literature about balanced sampling and its spatial versions. We intend to give the reader insights about some models and the corresponding optimal sampling design.

The second chapter of this thesis relates to sampling in finite population. By modelling the distribution of the number of units between two selected units, we obtain a family of sampling designs with a tunable repulsion. The cases of fixed and random sample size are considered separately. In both cases, the first order inclusion probabilities are equal and the second order inclusion probabilities are known under closed-form.

The third chapter is about repulsive sampling in an interval of the real line. In continuous spaces, a sampling design is a point process. We aim at imposing the process to be stationary which is similar to imposing equal first order inclusion probabilities in finite population. We focus on the intervals between two successive occurrences of the process, called spacings. If spacings between units are independent, the point process is a renewal process and the sample size is random. We can impose a fixed size sample but the spacings are only exchangeable and not independent. In both cases, the proposed family of point processes encompasses the most basic processes, that is, binomial in the case of fixed sample size and Poisson in the case of binomial sample size; it also includes the systematic sampling as a limiting case. Thus we obtain a family of processes depending on a parameter that allows us to continuously tune the repulsion between selections.

Finally, we consider the problem of developing a repulsive spatial point process with varying prescribed intensity. In particular, we propose a repulsive point process based on Determinantal processes from which we can easily sample on a geometrically complex domain. The first and second order inclusion densities are known under closed-form and are strictly positive. The Horvitz-Thompson estimator is unbiased and we can unbiasedly estimate its variance.

Keywords: sampling in finite population; repulsion; point processes; determinantal point processes;

UNIVERSITÉ DE NEUCHÂTEL

Résumé

Faculté des Sciences

Doctorat ès Sciences

Échantillonnage aléatoire avec répulsion

par Matthieu WILHELM

Cette thèse vise à explorer le concept de répulsion et à l'exploiter afin de développer de nouveaux plans de sondages.

Le premier chapitre est une revue de la littérature qui porte sur l'échantillonnage en population finie. Il est constitué de deux parties distinctes. Dans la première, trois principes visant à guider les praticiens dans leur conception d'un plan de sondage sont préconisés : la maximisation de l'entropie, la surreprésentation et la restriction. Dans la seconde partie, on fait une revue de la littérature concernant l'échantillonnage équilibré ainsi que de ses différentes variantes spatiales.

Le deuxième chapitre de cette thèse porte sur l'échantillonnage en population finie. En cherchant à modéliser l'écart entre deux unités sélectionnées, on obtient une famille de plans de sondage dont on peut ajuster la répulsion. Les cas de tailles fixes et aléatoires sont traités séparément et dans les deux cas, on propose des plans dont les probabilités d'inclusion d'ordre un sont constantes et celles d'ordre deux sont connues sous forme close.

Le troisième chapitre traite un problème similaire mais dans le cas d'un intervalle de la droite réelle. Le plan d'échantillonnage est en fait un processus ponctuel que l'on désire stationnaire et on modélise les distances entre deux occurrences du processus. Lorsque les distances entre deux unités successives sont indépendantes, le processus est de renouvellement et la taille de l'échantillon est en général aléatoire. On peut imposer la taille fixe et, dans ce cas, les intervalles entre deux unités ne sont plus indépendants mais seulement échangeables. Dans les deux cas, de taille fixe et de taille aléatoire, la famille de processus dévelopée englobe les processus les plus basiques, respectivement de Poisson et binomial ainsi que le tirage systématique comme cas limite. Ainsi, on dispose d'une famille de processus dépendant d'un paramètre qui permet de faire varier continument la répulsion.

Finalement, on traite du problème de l'échantillonnage spatial à intensité variable. En particulier, on développe un processus ponctuel répulsif basé sur les processus déterminantaux, que l'on peut échantillonner sur un domaine géométriquement complexe et dont on connaît analytiquement les intensités jointes d'ordre un et deux. L'estimateur d'Horvitz-Thompson est sans biais et on peut estimer sans biais sa variance.

Mots-clés: échantillonnage en population finie; répulsion; processus ponctuels; processus ponctuels déterminantaux;

Contents

Ad maiorem Dei gloriam

Chapter 1

Introduction

This thesis is dedicated to the study and to the development of some new sampling designs, in various settings. We consider three different spaces: a finite population, an interval of the real line and a two dimensional surface. We have deliberately adopted a design-based approach, avoiding to make any modelling assumption on the variable of interest.

This thesis is the collection of papers (or preprints) that are the fruit of the following collaborations: Chapter 2 has been jointly written with Yves Tillé, Chapters 3 and 4 are a joint work with Lionel Qualité and Yves Tillé, and Chapter 6 is a common work with Lionel Wilhelm. I benefited from many discussions and much advice from Arnaud Doucet during my visit at the statistics department of the University of Oxford. I also acknowledge the support of Luca Dedè from which I learned a lot in writing my first paper which is not part of this thesis.

In Chapter 2, we introduce some principles that are in our opinion important in designing a sampling. These theoretical principles are randomization, overrepresentation and restriction. We discuss these principles and present some examples where they can be applied. Moreover, we review balanced sampling and its connection with the model-assisted framework. Balanced sampling is an optimal sampling design for some basic models. This chapter is essentially a reprint of Tillé and Wilhelm (2017), with some slight modifications.

The guideline of the following chapters is the concept of repulsion. If we assume that a variable of interest is a smooth function of some known auxiliary variables, then two units that are "close" one from the other in respect of their auxiliary variables are likely to have similar values for their variable of interest. Therefore, for a given fixed sample size, it is natural to try and select units that are far from others in order to maximize the quantity of observed information. Note that in the case where the information is uncorrelated with the variable of interest, the use of such designs has an accuracy similar to Simple Random Sampling (SRS) design (see for instance Tillé, 2006).

This idea becomes clearer when we consider a continuous universe. If one wants to estimate a smooth function of interest defined over a surface, for instance over the territory of a country, it is reasonable to consider that the value of the variable of interest measured at two neighbouring locations will be close too. Indeed, the information contained in two close measurements would be essentially the same. It is thus natural to use sampling design that spreads the units. We call *repulsive* such a sampling design. This idea is explored in various ways in this thesis, and is the core idea of Chapters 3, 4 and 6.

Chapters 3 and 4 are closely related. These two chapters are based on the same idea, being applied to a finite population and to a one dimensional continuous population respectively. We first draw an initial sample of a basic sampling design and we then subsample among it with a systematic design. This very simple idea has been developed to cover the cases of fixed and random sample size. The solutions we found to circumvent the problem of the boundaries

(edge effects) are applicable to both cases of finite and continuous population. In particular, we focus on the concept of *spacings*, which are the number of units (resp. the length of the interval in the continuous case) between two successive selected units in the sample. We use the classical theory of renewal processes in both chapters and their structure is somewhat similar.

In Chapter 3, we develop sampling designs for finite populations with tunable repulsion between units. We cover the case of fixed and random size, both with equal inclusion probabilities. In the case of random size, we use the theory of stationary discrete renewal process to derive the distribution of the first spacing, and the other spacings are independently drawn. In the case of fixed sample size, the spacings are not independent since their sum must be equal to the size of the population. However, their joint distribution is exchangeable. For all the discussed distributions of spacings, we describe the corresponding sampling designs and express the first and second order inclusion probabilities in closed-form. This chapter is a reprint of Tillé et al. (2017).

In Chapter 4, we consider a continuous population on an interval. Again, we aim at developing sampling processes that exhibit some repulsion. In the first part, we consider sampling processes, i.e. sampling designs in a continuous universe, with random size that are special cases of stationary renewal processes. In the second part, we develop sampling processes that have fixed size and that are stationary. In this case, the spacings have exchangeable joint densities. We also show how to modify these sampling processes in order to deal with non-stationarity, which is, roughly speaking, the continuous analogue of unequal inclusion probability. This chapter is a reprint of Wilhelm et al. (2017).

In Chapter 5, we give some known results about point processes and about the continuous version of the Horvitz-Thompson estimator. The results in this chapter are not new and must be considered as preliminaries for the following chapter.

In Chapter 6, we tackle the problem of finding a sampling design having the following properties:

1. one can sample from it on complex domains and control the effect of the boundaries on the process;
2. it must cope with a varying density of points;
3. it must be repulsive;
4. the Horvitz-Thompson estimator of the total and its variance estimator are unbiased and tractable.

Stationary Determinantal Point Processes (DPP), first introduced by Macchi (1975) as *Fermion processes*, satisfy all those properties but one: the varying density of points, which we will refer to as intensity hereafter. We use an independent thinning in order to modify the intensity. The other properties are still satisfied by doing this slight modification. We illustrate the relevance of such a point process in the context survey sampling by drawing a sample on a wind turbine blade. The variable of interest is the pressure from which we can deduce the aerodynamic force acting on the blade. The final aim is to draw a sample which could be used to conduct a wind tunnel experiment.

Chapter 2

Principles for Choice of Design and Balanced Sampling

Abstract

The aim of this paper is twofold. First, three theoretical principles are formalized: randomization, overrepresentation and restriction. We develop these principles and give rationale for their use in choosing the sampling design in a systematic way. In the model-assisted framework, knowledge of the population is formalized by modeling the population and the sampling design is chosen accordingly. We show how the principles of overrepresentation and of restriction naturally arise from the modeling of the population. The balanced sampling then appears as a consequence of the modeling. Second, a review of probability balanced sampling is presented through the model-assisted framework. For some basic models, balanced sampling can be shown to be an optimal sampling design. Emphasis is placed on new spatial sampling methods and their related models. An illustrative example shows the advantages of the different methods. Throughout the paper, various examples illustrate how the three principles can be applied in order to improve inference. [a]

[a] This chapter is essentially a reprint of Tillé and Wilhelm (2017), up to the addition of the Section 2.7.6 and some modifications in Sections 2.6.3 and 2.7.3.

2.1 Introduction

Very early in the history of statistics, it appeared that censuses were unachievable in many practical situations. Thus the idea of using a subset of the target population to infer certain characteristics of the entire population naturally appeared. This idea can be traced back at least to Pierre-Simon Laplace (Laplace, 1847). In the first half of the XX$^{\text{th}}$ century, it became clear that only random sampling can provide an unbiased estimate. Kruskal and Mosteller (1980) provide a concise review of the history of probability sampling.

Classical references for sampling designs include Sukhatme (1954), Cochran (1977), Jessen (1978), and Brewer and Hanif (1983), who gave a list of 50 methods to select a sample with unequal inclusion probabilities. More modern textbooks include Särndal et al. (1992), Tillé (2006), Lohr (2009) and Thompson (2012). The more recent developments in survey sampling have been mainly motivated by new applications and new types of data, as for instance functional data (Cardot and Josserand, 2011).

For Hájek (1959), a survey is always characterized by a strategy composed of a sampling design and of an estimator of the parameter of interest. In the present paper, we focus on the

choice of sampling design while we restrict attention to the Narain-Horvitz-Thompson (NHT) estimator of the total (Narain, 1951; Horvitz and Thompson, 1952). In the case of a design-based approach, apart from the estimator, the practitioner can only choose how the sample is selected, i.e. she/he has to determine a sampling design. This is the core of the theory of design-based survey sampling. This choice is driven by both theoretical and practical aspects.

In this paper, three important principles are introduced: randomization, overrepresentation and restriction. The relevance of these principles is justified. We are probably not the first to highlight that those principles are desirable, but we would like to introduce and discuss them in a comprehensive and systematic way.

The randomization principle states that the sampling designs must be as random as possible. Indeed, the more random the sample is, the better the asymptotic approximations are (Berger, 1998b,a). Since most of the quantification of uncertainty is carried out using asymptotic results, this is a very important aspect. Another point is that very random sampling designs (which will be clarified later) are more robust (Grafström, 2010b). The principle of overrepresentation suggests to preferentially select units where the scattering is larger. The principle of restriction excludes very particular samples such as samples with empty categories or samples where the NHT-estimators of some auxiliary variables are far from the population total. In this way, samples that are either non-practical or known to be inaccurate are avoided. The restrictions could consist of only choosing fixed size samples for example.

When auxiliary information is available, it is desirable to include it in the sampling design in order to increase the precision of the estimates. In the design-based approach, the auxiliary information should be used when choosing the sampling design. A balanced sample is such that the estimated totals of the auxiliary variables are approximately equal to the true totals. Intuitively, this can be seen as an a priori calibration (Deville and Särndal, 1992). The cube method (Deville and Tillé, 2004) is a way to implement a probability sampling design which is balanced with equal or unequal first-order inclusion probabilities. The cube method is then a direct implementation of the principles of overrepresentation and restriction since it enables us to select samples with given inclusion probabilities and at the same time balanced on totals of auxiliary variables. Special emphasis is also placed on balanced sampling with spatial applications.

The suggested principles cannot be the only foundation for the choice of sampling design. Many other aspects are important such as the simplicity of the procedure, the quality of the data frame or a low rate of non-response. Thus, these general principles are not always applicable because of practical constraints. However, we recommend adopting an approach where general principles should be considered in order to improve the quality of a survey.

There is no intention to be exhaustive in the enumeration of all the recent advances that have contributed to survey sampling. Our intention is more to highlight that taking into account the aforementioned principles can be a motivation for both theoretical and practical advances and that this is well illustrated by balanced sampling.

The paper is organized as follows. In Section 2.2, definitions and the notation are given. In Section 2.3, the most basic sampling designs are briefly described. In Section 2.4, some principles of sampling are proposed. Section 2.5 describes balanced sampling and briefly present the cube method. In Section 2.6, we propose a model-assisted selection of sampling designs in light of those principles. In Section 2.7, we present new methods for spatial sampling. An illustrative example presented in Section 2.8 enables us to compare these methods. Finally, a discussion concludes the paper in Section 2.9.

2.2 Probability sampling and estimation

In the following, a list sampling frame is supposed to be available. Consider a population U composed of N units that are denoted by their order numbers so it can be written $U = \{1, \ldots, k, \ldots, N\}$. Let us denote by $\mathcal{S}$ the set of the subsets of U, which has cardinality 2^N. A sample without replacement is simply an element $s \in \mathcal{S}$, that is a subset of the population. Note that the empty set is a possible sample. A sampling design $p(.)$ is a probability distribution on $\mathcal{S}$

$$p(s) \geq 0 \text{ and } \sum_{s \in \mathcal{S}} p(s) = 1.$$

A random sample S is obtained by selecting a sample s with probability $p(s)$. Thus $\Pr(S = s) = p(s)$, for all $s \in \mathcal{S}$. Hence, S denotes the random variable and s the realization of it. The set $\{s \in \mathcal{S} : p(s) > 0\} \subset \mathcal{S}$ is called the support of the sampling design. For instance, one can consider $\mathcal{S}_n = \{s \in \mathcal{S} | \#s = n\}$ for a sampling design of fixed sample size n.

The first-order inclusion probability π_k is the probability of selecting the unit k. The joint (or second order) inclusion probability $\pi_{k\ell}$ is the probability that two different units k, ℓ are selected together in the sample. They can be derived from the sampling design:

$$\pi_k = \sum_{s \ni k} p(s) \text{ and } \pi_{k\ell} = \sum_{s \supset \{k,\ell\}} p(s).$$

The aim is to estimate a total

$$Y = \sum_{k \in U} y_k$$

of the values y_k taken by the variable of interest on all the units of the population.

The total Y can be estimated by the Narain-Horvitz-Thompson estimator (Narain, 1951; Horvitz and Thompson, 1952)

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

If $\pi_k > 0$ for all $k \in U$, this estimator is unbiased, i.e. $\mathrm{E}_p(\widehat{Y}) = Y$, where $\mathrm{E}_p(.)$ is the expectation under the sampling design $p(.)$.

Define

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell & \text{if } k \neq \ell \\ \pi_k(1 - \pi_k) & \text{if } k = \ell. \end{cases}$$

The design variance $\mathrm{var}_p(.)$ of the NHT-estimator is equal to:

$$\mathrm{var}_p\left(\widehat{Y}\right) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.$$

When the sample size is fixed, this variance simplifies to

$$\mathrm{var}_p\left(\widehat{Y}\right) = -\frac{1}{2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ k \neq \ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell}.$$

Estimators can be derived from these two expressions. For the general case,

$$\widehat{\text{var}}\left(\widehat{Y}\right) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\frac{\Delta_{k\ell}}{\pi_{k\ell}},$$

where $\pi_{kk} = \pi_k$. When the sample size is fixed, the variance estimator (Sen, 1953; Yates and Grundy, 1953) is given by:

$$\widehat{\text{var}}\left(\widehat{Y}\right) = -\frac{1}{2}\sum_{k\in S}\sum_{\substack{\ell\in S\\ k\neq\ell}}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2\frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

These estimators are both unbiased provided that $\pi_{k\ell} > 0, k \neq \ell \in U$.

Provided that the first-order inclusion probabilities are positive, the NHT estimator is unbiased and the variance and the mean squared error are equal. Provided that the first and the second order inclusion probabilities are positive, the variance estimators give an unbiased estimation of the mean-squared error. It is usual to assume a normal distribution to quantify the uncertainty. In many sampling designs, the normality assumption is asymptotically valid. The rate of convergence depends on the entropy (Berger, 1998b,a), which is roughly speaking, a measure of randomness. We further discuss the concept of entropy in Section 2.4.1.

2.3 Some basic designs

In the following, a list sampling frame is supposed to be available. In some situations, this may not be the case, as for instance in spatial sampling where the sampling frame can be a geographical region and the units a subdivision of this region. The sampling designs presented in this section are all implemented in various R packages (R Development Core Team, 2015). Valliant et al. (2013, chap. 3.7) provide a review of the current R and SAS packages for survey sampling.

2.3.1 Bernoulli sampling design

In Bernoulli sampling, the units are independently selected according to independent Bernoulli random variables with the same inclusion probabilities π. Then,

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s}\ , \text{ for all } s \in \mathcal{S},$$

where n_s is the sample size of sample s. The sample size is random and has a binomial distribution, i.e. $n_s \sim Bin(N, \pi)$. The sample size expectation is $N\pi$. The first-order inclusion probability is $\pi_k = \pi$ and the second order inclusion probability is equal to $\pi_{k\ell} = \pi^2$ for $k \neq \ell$.

2.3.2 Poisson sampling design

When the inclusion probabilities π_k are unequal, the sampling design obtained by selecting the units with independent Bernoulli random variables with parameter π_k is called Poisson sampling. The sampling design is

$$p(s) = \prod_{k\in s}\pi_k\prod_{k\notin s}(1-\pi_k)\ , \text{ for all } s \in \mathcal{S}.$$

The inclusion probabilities are π_k and $\pi_{k\ell} = \pi_k \pi_\ell$, for all $k \neq \ell \in U$. The sample size is random and has a Poisson binomial distribution (Hodges and Le Cam, 1960; Stein, 1990; Chen, 1993).

2.3.3 Simple random sampling

In simple random sampling (SRS) without replacement, the sample size is fixed and denoted by n. All the samples of size n have the same probability of being selected. The sampling design is then

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \text{for all } s \in \mathcal{S}_n \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathcal{S}_n = \{s \subset U | \# s = n\}$. The inclusion probabilities are $\pi_k = n/N$ and $\pi_{k\ell} = n(n-1)/[N(N-1)]$, for all $k \neq \ell \in U$.

2.3.4 Conditional Poisson sampling

The problem of selecting a sample with given unequal inclusion probabilities π_k and with fixed sample size is far from being simple. Several dozen methods have been proposed (see Brewer and Hanif, 1983; Tillé, 2006). Conditional Poisson sampling (CPS) is a sampling design of fixed size and with prescribed unequal inclusion probabilities. The sampling design is given by

$$p(s) = \frac{\sum_{k \in S} \exp \lambda_k}{\sum_{s \in \mathcal{S}_n} \sum_{k \in S} \exp \lambda_k},$$

where the λ_k are obtained by solving

$$\sum_{s \in \{s \in \mathcal{S}_n | s \ni k\}} p(s) = \pi_k, k \in U. \tag{2.1}$$

The implementation is not simple. The complexity comes from the sum over $s \in \mathcal{S}_n$ in Expression (2.1) that is so large that shortcuts must be used. However, several solutions have been proposed by Chen et al. (1994) and Deville (2000) in order to implement this sampling design by means of different algorithms (see also Tillé, 2006). The joint inclusion probabilities can easily be computed. CPS is also called maximum entropy sampling because it maximizes the entropy as defined in Section 2.4.1 subject to given inclusion probabilities and fixed sample size.

2.3.5 Stratification

The basic stratified sampling design consists in splitting the population into H nonoverlapping strata $U_1, \ldots, U_H$, of sizes $N_1, \ldots, N_H$. Next in each stratum, a sample of size n_h is selected with SRS. The sampling design is

$$p(s) = \begin{cases} \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1} & \text{for all } s \text{ such that } \#(U_h \cap s) = n_h,\ h = 1, \ldots, H, \\ 0 & \text{otherwise.} \end{cases}$$

The inclusion probabilities are $\pi_k = n_h/N_h$ for $k \in U_h$ and

$$\pi_{k\ell} = \begin{cases} \frac{n_h(n_h-1)}{N_h(N_h-1)} & k, \ell \in U_h \\ \frac{n_h n_i}{N_h N_i} & k \in U_h, \ell \in U_i, i \neq h. \end{cases}$$

There are two basic allocation schemes for the sample sizes:

- In proportional allocation, the sample sizes in the strata are proportional to the stratum sizes in the population, which gives $n_h = nN_h/N$. Obviously n_h must be rounded to an integer value.
- Neyman (1934) established the optimal allocation by searching for the allocation that minimizes the variance subject to a given total sample size n. After some algebra, we obtain the optimal allocation:

$$n_h = \frac{nN_hV_h}{\sum_{\ell=1}^{H} N_\ell V_\ell}, \tag{2.2}$$

where

$$V_h^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} \left(y_k - \overline{Y}_h\right)^2, \text{ and } \overline{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k,$$

for $h = 1, \ldots, H$. Again, n_h must be rounded to an integer value. When the population is skewed, Equation (2.2) often gives values $n_h > N_h$, which is often the case in business statistics. In this case, all the units of the corresponding stratum are selected (take-all stratum) and the optimal allocation is recomputed on the other strata. In cases where a list sampling frame is not available, the proportional and the optimal stratification might be slightly adapted.

2.4 Some sampling principles

The main question is how to select a sample or, in other words, what sampling method one should use. Survey statisticians know that designing a survey is an intricate question that requires experience, a deep knowledge of the sampling frame and of the nature of variables of interest. Most sampling design manuals present a list of sampling methods. However, the choice of the sampling design should be the result of the application of several principles. In what follows, we try to establish some theoretical guidelines. Three principles can guide the choice of sample: the principle of randomization, the principle of overrepresentation and the principle of restriction.

2.4.1 The principle of randomization

In design-based inference, the extrapolation of the sample estimators to the population parameters is based on the sampling design, i.e. on how the sample is selected. The first principle not only consists in selecting a sample at random but as random as possible.

A sampling design should assign a positive probability to as many samples as possible and should tend to equalize these probabilities between the samples. This enables us to avoid null joint inclusion probabilities and produces an unbiased estimator of the variance of the NHT estimator. The common measure of randomness of a sampling design is its entropy given by

$$I(p) = -\sum_{s \in \mathcal{S}} p(s) \log p(s),$$

with $0 \log 0 = 0$.

Intuitively, the entropy is a measure of the quantity of information and also a measure of randomness. High entropy sampling designs generate highly randomized samples, which in turns make the design more robust. A discussion about high entropy designs and its relationship with robustness can be found in Grafström (2010b). The convergence towards asymptotic normal distributions of the estimated totals also depends on entropy. The higher the entropy is, the higher the rate of convergence is (Berger, 1998b,a). Conversely, if the support is too small, then the distribution of the estimated total is rarely normal.

For complex sampling designs, second-order inclusion probabilities are rarely available. However, when considering high-entropy sampling designs, the variance can be estimated by using formulae that do not depend on the second-order inclusion probabilities (Brewer and Donadio, 2003). Those estimators are approximate but are of common use. It is worth mentioning that methods for quantifying the uncertainty of complex sampling designs have been developed (Antal and Tillé, 2011; Berger and De La Riva Torres, 2016).

2.4.2 The principle of overrepresentation

Sampling consists in selecting a subset of the population. However, there are no particular reasons to select the units with equal inclusion probabilities. In business surveys, the establishments are generally selected with very different inclusion probabilities that are in general proportional to the number of employees. To be efficient, the choice of units is intended to decrease uncertainty. So it is more desirable to overrepresent the units that contribute more to the variance of the estimator.

The idea of "representativity" is thus completely misleading and is based on the false intuition that a sample must be similar to the population to perform an inference because the sample is a "scale copy" of the population (see among others Kruskal and Mosteller, 1979a,b,c). In fact, the only requirement for the estimator to be unbiased consists of using a sampling design with non-null first-order inclusion probability for all units of the population, which means that the sampling design does not have coverage problems (see Särndal et al., 1992, p. 8). Unequal probability sampling can be used to estimate the total Y more efficiently. The main idea is to oversample the units that are more uncertain because the sample must collect as much information as possible from the population, which was already the basic idea of the seminal papers of Jerzy Neyman (1934, 1938) on optimal stratification. In general, the principle of overrepresentation implies that a sampling design should have unequal inclusion probabilities if prior information is available. There exist different ways to deduce the inclusion probabilities from a working model as we will see in Section 2.6.

2.4.3 The principle of restriction

The principle of restriction consists in selecting only samples with a given set of characteristics, for instance, by fixing the sample size or the sample sizes in categories of the population (stratification). There are many reasons why restrictions should be imposed. For instance, empty categories in the sample might be avoided, which can be very troublesome when the aim is to estimate parameters in small subsets of the population. It is also desirable that the estimates from the sample are coherent with some auxiliary knowledge. So only samples that satisfy such a property can be considered. By coherent, we mean that the estimate from the sample of an auxiliary variable should match a known total. Such samples are said to be balanced. Balanced

sampling is discussed Section 2.5. More generally, restrictions can reduce or even completely remove the dispersion of some estimators.

At first glance, the principle of restriction seems to be in contradiction with the principle of randomization because it restricts the number of samples with non-null probabilities. However the possible number of samples is so large that, even with several constraints, the number of possible samples with non-null probabilities can remain very large. It is thus still reasonable to assume a normal distribution for the estimates. Balanced sampling enables us to avoid the "bad" samples, which are those that give estimates for the auxiliary variables that are far from the known population totals.

2.5 Balanced sampling

A sample without replacement from a population of size N can be denoted by a vector of size N such that the kth component is equal to 1 if the kth unit is selected and 0 if it is not. Following this representation, a sample can be interpreted as a vertex of the unit hypercube of dimension N. This geometrical interpretation of a sample is central in the development of some sampling algorithms (Tillé, 2006).

By definition, a balanced sample satisfies

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}=\sum_{k\in U}\mathbf{x}_k. \tag{2.3}$$

where $\mathbf{x}_k=(x_{k1},\ldots,x_{kp})^\top$ is a vector of p auxiliary random variables measured on unit k. Vectors $\mathbf{x}_k$ are assumed to be known for each unit of the population, i.e. a register of population is available for the auxiliary information. The choice of the first-order inclusion probabilities is discussed in Section 2.6 and is a consequence of the principle of overrepresentation. Balanced sampling designs are designs whose support is restricted to samples satisfying (or approximately satisfying) Equation (2.3). In other words, we are considering sampling designs of prescribed first-order inclusion probabilities $\pi_1,\ldots,\pi_N$ and with support

$$\left\{s\in\mathcal{S}:\sum_{k\in s}\frac{\mathbf{x}_k}{\pi_k}=\sum_{k\in U}\mathbf{x}_k\right\}.$$

More generally, an approximately balanced sample s is a sample satisfying

$$\left\|\mathbf{D}^{-1}\left(\sum_{k\in U}\mathbf{x}_k-\sum_{k\in s}\mathbf{x}_k/\pi_k\right)\right\|\leq c, \tag{2.4}$$

where $\mathbf{D}$ is a $p\times p$ matrix defined by $\mathbf{D}=\operatorname{diag}(\sum_{k\in U}\mathbf{x}_k)$, c is a positive constant playing the role of a tolerance from the deviation of the balancing constraints and $\|\cdot\|$ denotes any norm on $\mathbb{R}^p$. Balanced sampling thus consists in selecting randomly a sample whose NHT-estimators are equal or approximately equal to the population totals for a set of auxiliary variables. In practice, exact balanced sampling designs rarely exist. The advantage of balanced sampling is that the design variance is null or almost null for the NHT-estimators for these auxiliary variables. Thus, if the variable of interest is strongly correlated with these auxiliary variables, the variance of the NHT-estimator of the total for the variable of interest is also strongly reduced.

Sampling designs with fixed sample size (SRS, CPS) and stratification are particular cases of balanced sampling. Indeed in sampling with fixed sample size the only balancing variable is the first-order inclusion probability $\mathbf{x}_k = \pi_k$ for all $k \in U$. In stratification, the H balancing variables are $\mathbf{x}_k = (\pi_k I(k \in U_1), \ldots, \pi_k I(k \in U_h), \ldots, \pi_k I(k \in U_H))^\top$, where $I(k \in U_h)$ is the indicator variable of the presence of unit k in stratum U_h.

The first way to select a balanced sample could consist in using a rejective procedure, for instance by generating samples with SRS or Poisson sampling until a sample satisfying Constraint (2.4) is drawn. However, a conditional design does not have the same inclusion probabilities as the original one. For instance, Legg and Yu (2010) have shown that a rejective procedure fosters the selection of central units. So the units with extreme values have smaller inclusion probabilities. Well before, Hájek (1981) already noticed that if samples with a Poisson design are generated until a fixed sample size is obtained, then the inclusion probabilities are changed. This problem was solved by Chen et al. (1994) who described the link between the Poisson design and the one obtained by conditioning on the fixed sample size (see also Tillé, 2006, pp. 79-96). Unfortunately, the computation of conditional designs seems to be intractable when the constraint is more complex than fixed sample size. Thus the use of rejective methods cannot lead to a sampling design whose inclusion probabilities are really computable.

The cube method (Deville and Tillé, 2004) allows us to select balanced samples at random while preserving the possibly unequal prescribed first order inclusion probabilities. The method starts with the prescribed vector of inclusion probabilities. This vector is then randomly modified at each step in such a way that at least one component is changed to 0 or 1 and such that this transformation respects the prescribed first order inclusion probabilities. Thus the cube algorithm sequentially selects a sample in at most N steps. At each step, the random modification is realized while respecting the balancing constraints and the inclusion probabilities. The algorithm has two distinct phases: the first is the flight phase, where the balanced equations are exactly satisfied. At some point, it is possible that the balancing equations can only be approximated. In the second phase, called landing phase, the algorithm selects a sample that nearly preserves the balancing equation while still exactly satisfying the prescribed inclusion probabilities.

It is not possible to fully characterize the sampling design generated by the cube method. In particular, second order inclusion probabilities are intractable. In order to compute the variance, Deville and Tillé (2005) gave several approximations using only first order inclusion probabilities. Breidt and Chauvet (2011) suggest using a martingale difference approximation of the values of $\Delta_{k\ell}$ that takes into account the variability of both the flight and the landing phase, unlike the estimators proposed by Deville and Tillé (2005).

The cube method has been extended by Tillé and Favre (2004) to enable the coordination of balanced samples. Optimal inclusion probabilities are studied in the perspective of balanced sampling by Tillé and Favre (2005) and are further investigated by Chauvet et al. (2011). Deville (2014, in French) sketches a proof of the conditions that must be met on the inclusion probabilities and on the auxiliary variables in order to achieve an exact balanced sample. In this case, the cube algorithm only has a flight phase. From a practical standpoint, several implementations exist in the R language (Tillé and Matei, 2015; Grafström and Lisic, 2016) and in SAS (Rousseau and Tardieu, 2004; Chauvet and Tillé, 2005).

2.6 Model-assisted choice of the sampling design and balanced sampling

2.6.1 Modeling the population

The principles of overrepresentation and restriction can be implemented through a modeling of the links between the variable of interest and the auxiliary variables. This model may be relatively simple, for instance a linear model:

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \tag{2.5}$$

where $\mathbf{x}_k = (x_{k1}, \ldots, \ldots, x_{kp})^\top$ is a vector of p auxiliary variables, $\boldsymbol{\beta}$ is a vector of regression coefficients, and ε_k are independent random variables with null expectation and variance $\sigma^2_{\varepsilon k}$. The model thus admits heteroscedasticity. The error terms ε_k are supposed to be independent from the random sample S. Let also $\mathrm{E}_M(.)$ and $\mathrm{var}_M(.)$ be respectively the expectation and variance under the model.

Under model (2.5), the anticipated variance of the NHT-estimator is

$$\mathrm{AVar}(\widehat{Y}) = \mathrm{E}_p \mathrm{E}_M (\widehat{Y} - Y)^2 = \mathrm{E}_p \left(\sum_{k \in S} \frac{\mathbf{x}_k^\top \boldsymbol{\beta}}{\pi_k} - \sum_{k \in U} \mathbf{x}_k^\top \boldsymbol{\beta} \right)^2 + \sum_{k \in U} (1 - \pi_k) \frac{\sigma_k^2}{\pi_k}.$$

The second term of this expression is called the Godambe-Joshi bound (Godambe and Joshi, 1965).

Considering the anticipated variance, for a fixed sample size n, the sampling design that minimizes the anticipated variance consists in

- using inclusion probabilities proportional to $\sigma_{\varepsilon k}$,
- using a balanced sampling design on the auxiliary variables $\mathbf{x}_k$.

The inclusion probabilities are computed using

$$\pi_k = \frac{n \sigma_{\varepsilon k}}{\sum_{\ell \in U} \sigma_{\varepsilon \ell}}, \tag{2.6}$$

provided that $n\sigma_{\varepsilon k} < \sum_{\ell \in U} \sigma_{\varepsilon \ell}$ for all $k \in U$. If it is not the case, the corresponding inclusion probabilities are set to one and the inclusion probabilities are recomputed according to Expression (2.6).

If the inclusion probabilities are proportional to $\sigma_{\varepsilon k}$ and the sample is balanced on the auxiliary variables $\mathbf{x}_k$, the anticipated variance becomes (Nedyalkova and Tillé, 2008)

$$N^2 \left[\frac{N - n}{N} \frac{\bar{\sigma}_\varepsilon^2}{n} - \frac{\mathrm{var}(\sigma_\varepsilon)}{N} \right],$$

where

$$\bar{\sigma}_\varepsilon = \frac{1}{N} \sum_{k \in U} \sigma_{\varepsilon k} \text{ and } \mathrm{var}(\sigma_\varepsilon) = \frac{1}{N} \sum_{k \in U} (\sigma_{\varepsilon k} - \bar{\sigma}_\varepsilon)^2 .$$

Applying the randomization principle would result in a maximum entropy sampling design under the constraint of minimizing the anticipated variance. However, except for the very particular cases given in Table 2.1, there is no known general solution to this problem. When

existing, we refer to such sampling design as "optimal". All the designs presented in Section 2.3 are an application of this optimal design for particular cases of Model (2.5) and are explicitly described in Table 2.1.

TABLE 2.1: Particular cases of Model (2.5) with the corresponding sampling design for the optimal design

Underlying Model	Design	Model Variance	π_k
$y_k = \beta + \varepsilon_k$	SRS	σ^2	n/N
$y_k = \varepsilon_k$	Bernoulli sampling	σ^2	$\pi = \mathrm{E}(n_S)/N$
$y_k = x_k\beta + \varepsilon_k$	CPS	$x_k^2\sigma^2$	$\pi_k \propto x_k$
$y_k = \varepsilon_k$	Poisson sampling	$x_k^2\sigma^2$	$\pi_k \propto x_k$
$y_k = \beta_h + \varepsilon_k, k \in U_h,$	Proportional stratification	σ^2	n/N
$y_k = \beta_h + \varepsilon_k, k \in U_h,$	Optimal stratification	σ_h^2	$\pi_k \propto \sigma_h$

Maximizing entropy tends to equalize the probabilities of selecting samples. Under SRS and stratification, all the samples with a non-null probability have exactly the same probability of being selected, even for optimal stratification. For Bernoulli sampling all the samples of the same size have the same probability of being selected. When the inclusion probabilities are unequal, it is not possible to equalize the probabilities of the samples. However in CPS, all the samples of size n have positive probabilities of being drawn and in Poisson sampling all the samples (of any size) have non-null probabilities.

The common sampling designs presented in Table 2.1 correspond to very simple models. For SRS, the model only assumes a parameter β and homoscedasticity. In stratification, the means β_h of the variable of interest can be different in each stratum. Moreover for optimal stratification, the variances σ_h^2 of the noise ε_k are presumed to be different in each stratum. Unfortunately, there is no general algorithm that enables us to implement an unequal probability balanced sampling design with maximum entropy for the general case of Model (2.5).

The cube method is still not a fully complete optimal design for the general Model (2.5) because the entropy is not maximized. The cube method however gives a solution to a general problem, which involves SRS, unequal probability sampling with fixed sample size and stratification that are all particular cases of balanced sampling. Even if it is not possible to maximize the entropy with the cube method, it is possible to randomize the procedure, for instance, by randomly sorting the units before applying the algorithm.

2.6.2 A link with the model-based approach

An alternative literature is dedicated to the model-based approach. In this framework, the problem of estimating a total is seen as a prediction problem and the population is modelled. The model-based approach assumes a super-population model and the inference is carried out using this model (Brewer, 1963; Royall, 1970b,a, 1976b,a, 1992; Royall and Herson, 1973a,b). The Best Linear Unbiased Predictor (BLUP) under Model (2.5) is given by

$$\widehat{Y}_{BLU} = \sum_{k\in S} y_k + \sum_{k\in U\setminus S} \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}},$$

where

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{k \in S} \frac{\mathbf{x}_k^\top \mathbf{x}_k}{\sigma_{\varepsilon k}^2} \right)^{-1} \sum_{k \in S} \frac{\mathbf{x}_k^\top y_k}{\sigma_{\varepsilon k}^2},$$

see alsoValliant et al. (2000).

Nedyalkova and Tillé (2008) show that, under the assumption that there are two vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ of $\mathbb{R}^p$ such that $\boldsymbol{\lambda}^\top \mathbf{x}_k = \sigma_{\varepsilon k}^2$ and $\boldsymbol{\gamma}^\top \mathbf{x}_k = \sigma_{\varepsilon k}$ for all $k \in U$, then, for the sampling design that minimizes the anticipated variance, the NHT-estimator is equal to the BLUP. In this case, both approaches coincide. So, in this respect, balanced sampling enables us to reconcile design-based and model-based approaches.

2.6.3 Beyond the linear regression model

Balanced sampling using the cube method allows us to minimize the anticipated variance under the linear Model (2.5). However, despite its general benefits, such a model is unlikely to hold. A much more flexible model is the linear mixed model (Jiang, 2007; Ruppert et al., 2003):

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \mathbf{z}_k^\top \boldsymbol{\gamma} + \varepsilon_k,$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^\top$,

$$\mathrm{E} \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{pmatrix} = 0 \text{ and } \mathrm{var} \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{pmatrix} = \sigma^2 \begin{pmatrix} \lambda^{-2}\mathbf{Q} & 0 \\ 0 & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ is a $N \times N$ identity matrix and $\mathbf{Q}$ is a $q \times q$ positive matrix and q is the dimension of vector $\mathbf{z}_k$. Such a model encompasses many widely used models like penalized splines, semiparametric models and multiple regression analysis of variance among others. They are extensively used in survey sampling, especially in the field of small area estimation.

Breidt and Chauvet (2012) investigate the application of balanced sampling in the case where the working model is a linear mixed model and they introduce penalized balanced sampling. Using a modified version of the cube algorithm, they draw a penalized balanced sample that can account for the mixed effect structure of the model. Such samples can reduce or eliminate the need for linear mixed model weight adjustments. Linear-mixed as well as non-parametric model-assisted approaches have been extensively studied in the context of survey sampling (Breidt and Opsomer, 2017).

2.7 Spatial balanced sampling

2.7.1 Modeling the spatial correlation

Spatial sampling is particularly important in environmental statistics. A large number of specific methods were developed for environmental and ecological statistics (see among others Marker and Stevens Jr., 2009; Thompson, 2012).

When two units are geographically close, they are in general similar, which induces a spatial dependency between the units. Consider the alternative model

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \tag{2.7}$$

where $\mathbf{x}_k = (x_{k1}, \ldots, \ldots, x_{kp})^\top$ is a set of p auxiliary variables, $\boldsymbol{\beta}$ is a vector of regression coefficients, and ε_k are random variables with $\mathrm{E}(\varepsilon_k) = 0$, $\mathrm{var}(\varepsilon_k) = \sigma_k^2$ and $\mathrm{cov}(\varepsilon_k, \varepsilon_\ell) = \sigma_{\varepsilon k}\sigma_{\varepsilon \ell}\rho_{k\ell}$. The model admits heteroscedasticity and autocorrelation. The error terms ε_k are supposed to be independent from the random sample S. Let also $\mathrm{E}_M(.)$ and $\mathrm{var}_M(.)$ be respectively the expectation and variance under the model.

Under model (2.7), Grafström and Tillé (2013) show that the anticipated variance of the NHT-estimator is

$$\mathrm{AVar}(\widehat{Y}) = \mathrm{E}_p\left(\sum_{k\in S}\frac{\mathbf{x}_k^\top\boldsymbol{\beta}}{\pi_k} - \sum_{k\in U}\mathbf{x}_k^\top\boldsymbol{\beta}\right)^2 + \sum_{k\in U}\sum_{\ell\in U}\Delta_{k\ell}\frac{\sigma_{\varepsilon k}\sigma_{\varepsilon\ell}\rho_{k\ell}}{\pi_k\pi_\ell}. \tag{2.8}$$

If the correlation $\rho_{k\ell}$ is large when the units are close, the sampling design that minimizes the anticipated variance consists in

- using inclusion probabilities proportional to $\sigma_{\varepsilon k}$,
- using a balanced sampling design on the auxiliary variables $\mathbf{x}_k$,
- and avoiding the selection of neighboring units, i.e. selecting a well spread sample (or spatially balanced).

If the selection of two neighboring units is avoided, the values of $\Delta_{k\ell}$ can be highly negative, which makes the anticipated variance (2.8) small.

The value of $\Delta_{k\ell}$ can be interpreted as an indicator of the spatial pairwise behavior of the sampling design. Indeed, if two units k and ℓ are chosen independently with inclusion probability π_k and π_ℓ respectively, then the joint inclusion probability is $\pi_k\pi_\ell$. Hence, if $\Delta_{k\ell} < 0$, respectively $\Delta_{k\ell} > 0$, the sampling design exhibits some repulsion, respectively clustering, between the units k and ℓ. In other words, the sign of $\Delta_{k\ell}$ is a measure of repulsion or clustering of the sampling design for two units k and ℓ. Similar ideas have been used in the literature on spatial point processes to quantify the repulsion of point patterns. In particular, the pair correlation function is a common measure of the pairwise interaction (Møller and Waagepetersen, 2004, chap. 4).

2.7.2 Generalized Random Tessellation Stratified Design

The *Generalized Random Tessellation Stratified* (GRTS) design was proposed by Stevens Jr. and Olsen (1999); Stenvens Jr. and Olsen (2004); Stevens Jr. and Olsen (2003). The method is based on the recursive construction of a grid on the space. The cells of the grid must be small enough so that the sum of the inclusion probabilities in a square is less than 1. The cells are then randomly ordered such that the proximity relationships are preserved. Next a systematic sampling is applied along the ordered cells. The method is implemented in the "spsurvey" R package (Kincaid and Olsen, 2015).

2.7.3 Local pivotal method

The pivotal method has been proposed by Deville and Tillé (2000) and consists in selecting two units (denoted by i and j) in the population at each step. Their inclusion probabilities (π_i, π_j)

are randomly transformed to $(\widetilde{\pi}_i, \widetilde{\pi}_j)$ using the following randomization:

$$(\widetilde{\pi}_i, \widetilde{\pi}_j) = \begin{cases} (\min(1, \pi_i + \pi_j), \max(\pi_i - \pi_j - 1, 0)) & \text{with probability } q \\ (\min(\pi_i + \pi_j, 1), \max(0, \pi_i - \pi_j - 1)) & \text{with probability } 1 - q, \end{cases}$$

where

$$q = \frac{\min(1, \pi_i + \pi_j) - \pi_j}{2\min(1, \pi_i + \pi_j) - \pi_i - \pi_j}.$$

This operation is repeated at each step. Since at each step one unit has its inclusion probability updated to 0 or to 1, in maximum $N - 1$ steps, all the inclusion probabilities are randomly transformed to 0 or 1, which means that the sample is selected. Roughly speaking, the two units fight with strength proportional to their original inclusion probability until one of the two inclusion probabilities is randomly set to 0 or 1.

If the two units are sequentially selected according to their order in the population, the method is called *sequential pivotal method* (or *ordered pivotal sampling* or *Deville's systematic sampling*) (Chauvet, 2012). The sequential pivotal method is also closely related to the sampling designs introduced by Fuller (1970). If the two units are randomly selected at each step, the method is called *random pivotal method*.

Grafström et al. (2012) have proposed using the pivotal method for spatial sampling. This method is called *local pivotal sampling*. At each step of the method, two neighboring units are selected. Next a step of the pivotal method is applied. So, if the probability of one of these two units is increased, the probability of the other is decreased, which in turn induces some repulsion between the units proportionally to their original inclusion probabilities. The probability of selecting two neighboring units is then small and the sample is thus well spread. Several variants exist depending on how the two neighboring units are selected.

2.7.4 Spreading and balancing: local cube method

In the local pivotal, two units compete to be selected. The natural extension of this idea is to let a cluster of units fight. The local pivotal method has been generalized by Grafström and Tillé (2013) to provide a sample that is at the same time well spread in space and balanced on auxiliary variables in the sense of Expression (2.3). This method, called *local cube*, consists in running the flight phase of the cube method on a subset of $p + 1$ neighboring units, where p is the number of auxiliary variables. After this step, the inclusion probabilities are updated such that:

- one of the $p + 1$ units has its inclusion probability updated to 0 or 1,
- the balancing equation is satisfied.

When a unit is selected, it decreases the inclusion of the p other units of the cluster. Hence, it induces a negative correlation in the selection of neighboring units, which in turn spreads the sample.

2.7.5 Spatial sampling for non-spatial problems

Spatial methods can also be used in a non-spatial context. Indeed, assume that a vector $\mathbf{x}_k$ of auxiliary variables is available for each unit of the population. These variables can be, for

instance, turnover, profit or the number of employees in business surveys. Even if these variables are not spatial coordinates, they can be used to compute a distance between the units. For instance, the Mahalanobis distance can be used:

$$d^2(k,\ell) = (\mathbf{x}_k - \mathbf{x}_\ell)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}_k - \mathbf{x}_\ell),$$

where

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{k\in U}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^\top \text{ and } \bar{\mathbf{x}} = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k.$$

Grafström et al. (2012) advocate the use of spreading on the space of the auxiliary variables. Indeed, if the response variable is correlated with the auxiliary variable, then spreading the sample on the space of auxiliary variables also spreads the sampled response variable. It also induces an effect of smooth stratification on any convex set of the space of variables. The sample is thus stratified for any domain, which can be interpreted as a property of robustness.

2.7.6 Beyond the finite population framework

Spatial sampling in the sense discussed here has been motivated by the application of finite population sampling to environmental problems, in particular by forest inventory problems. But there are many settings in which the population is not finite. In this case, the theory of sampling in finite populations does not apply and different methodologies are needed to address such issues.

It should be noticed that a sample in the continuum, i.e. a point pattern, is usually referred to as the realization of a point process which is thus the continuous analogue of a sampling design. It is a well-known mathematical object (Daley and Vere-Jones, 2002, 2008). Macchi (1975) gives a construction of point processes which is similar to the approach of finite population sampling. In particular, the continuous analogue of the common quantities such as the inclusion probabilities are given. The continuous version of survey sampling theory was developed by Deville (1989) and Cordy (1993).

The definition of a total of a function defined on a continuous domain is an integral. Monte-Carlo and quasi Monte-Carlo methods have been developed for estimating integrals from samples. Actually, this is equivalent to choosing at random the quadrature points when using a numerical scheme for the integration. Caflisch (1998) gives an overview of these methods and Dick et al. (2013) establish a review of quasi Monte-Carlo methods. These methods are particularly efficient for high dimensional integration. For low dimensions, deterministic methods such as quadrature rules are very efficient but do not allow us to quantify the uncertainty. Instead, upper bounds of the error are usually provided by the theory under regularity assumptions. The idea of considering the computation of an integral as an estimation rather than a deterministic computation is an active field of research, coined as probabilistic integration. In this case, the Bayesian paradigm is particularly suited. Diaconis (1988) first mentioned the fact that Bayesian methods could be used to both estimate an integral and provide a quantification of uncertainty (see Briol et al., 2016, for a review of the current state of the research) .

In the case where a sample is drawn not only to estimate a total but to estimate some unknown function from pointwise measurements, the paradigm becomes more intricate. The same issue arises in designs for computer experiments. Many different methods for sampling in this context have been developed, most of them relying on the Bayesian paradigm and on Gaussian process assumptions (Santner et al., 2003). They are usually sequential: starting from

an initial sample, other measurements are made sequentially by optimizing some empirical risk measure. The risk measure is usually based on model assumptions such as Gaussianity of the observed random field. The initial sample is usually based on latin hypercube type methods (McKay et al., 1979). More modern ideas include orthogonal-arrays based latin hypercube sampling designs (Owen, 1992; Tang, 1993).

It is worth mentioning that in most environmental surveys, the continuous nature of the space is bypassed using a discretization and transforming the problem in a finite population setting. When a fine grid is needed, the size of the population can be very large. Hence, it can be useful to tackle the problem without resorting to any discretization.

2.8 Illustrative example

A simple example illustrates the advantages of the sampling designs discussed in Section 2.7. Consider a square of $N = 40 \times 40 = 1600$ dots that are the sampling units of the population. A sample of size $n = 50$ is selected from this population by means of different sampling designs. Figure 2.1 contains two samples that are not spatially balanced: SRS and balanced sampling by means of the cube method. For balanced sampling, three variables are used: a constant equal to 1, the x coordinate and the y coordinate. So the sample has a fixed sample size and is balanced on the coordinates. These samples are not well spread or spatially balanced.

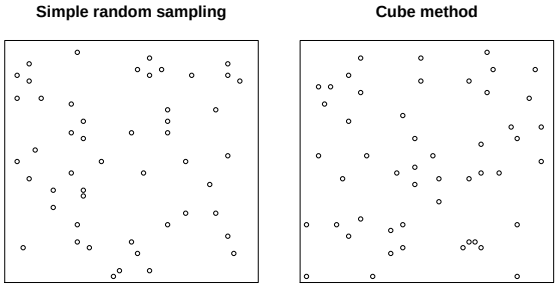

FIGURE 2.1: Sample of size $n = 50$ in a population of 1600 plots by means of SRS and the cube method. These samples are not very well spread.

Figure 2.2 contains the most basic sampling designs used to spread a sample: systematic sampling and stratification. Unfortunately, spatial systematic sampling cannot be generalized to unequal probability sampling. It is also not possible to apply it to a population that is lying on a lattice.

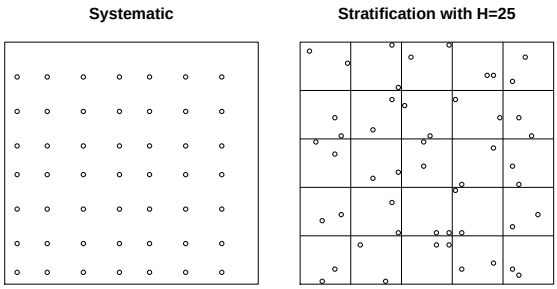

FIGURE 2.2: Systematic sampling and stratification with 2 units per stratum.

Figure 2.3 contains modern well spread sampling methods such as the local pivotal method, GRTS and the local cube method. At first glance, it is difficult to evaluate which design gives the most well spread sample.

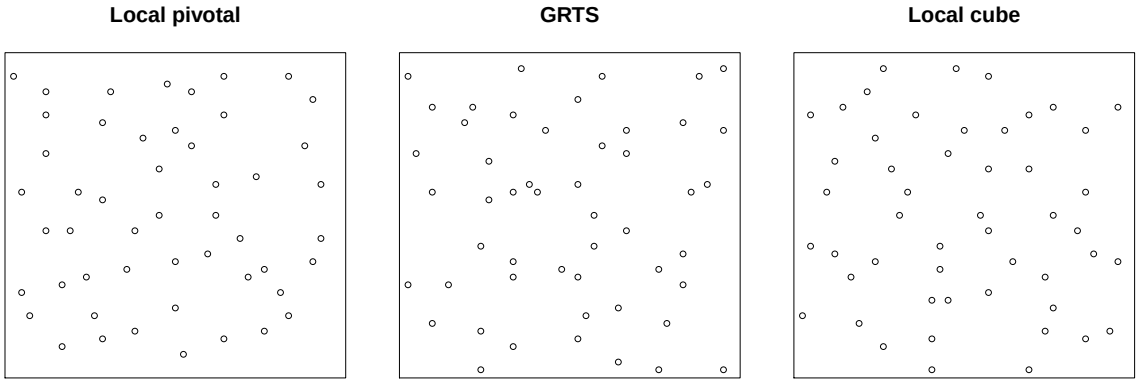

FIGURE 2.3: GRTS, pivotal method, and local cube method. Samples are well spread.

The Voronoï polygons are the set of the elements of the population that are closer to a given point than any other points in the population. Figure 2.4 contains the Voronoï polygons for SRS, stratification and local pivotal method. The variance of the sum of inclusion probabilities of the population units that are included in a polygon is an indicator of the quality of spatial balancing (Stenvens Jr. and Olsen, 2004).

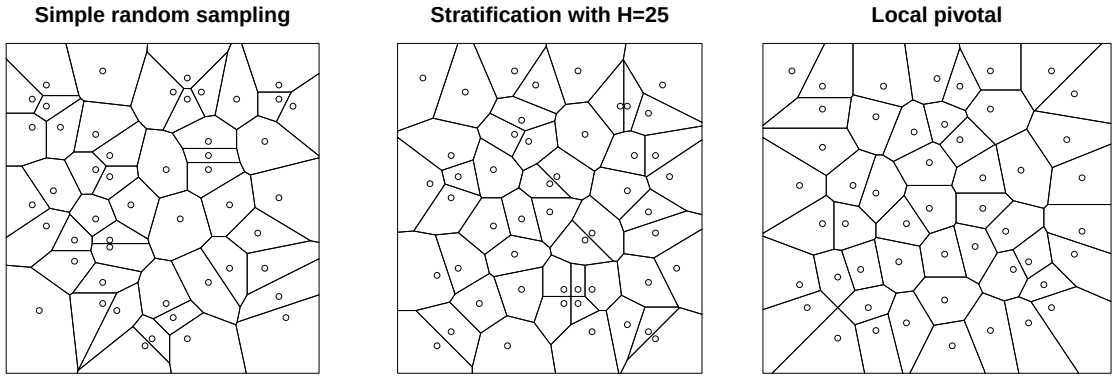

FIGURE 2.4: Example of Voronoï polygons for three sampling designs.

Table 2.2 contains the average of the indicators of spatial balance for 1000 selected (Grafström and Lundström, 2013). For systematic sampling, the index is not null because of the edge effect. The best designs are the local pivotal methods and the local cube method. Grafström and Lundström (2013) extensively discuss the concept of balancing and the implication on the estimation. In particular, they show under some assumptions that a well spread sampling design is an appropriate design under model (2.7).

TABLE 2.2: Indices of spatial balance for the main sampling designs with their standard deviations

Design	Balance indicator
Systematic	0.05 (0.04)
Simple random sampling	0.31 (0.10)
Stratification with H=25	0.10 (0.02)
Local pivotal	0.06 (0.01)
Cube method	0.21 (0.06)
Local Cube method	0.06 (0.01)
GRTS	0.10 (0.02)

2.9 Discussion

Three principles with theoretical appealing properties have been established. Modeling the population can be used as a tool for the implementation of the principle of overrepresentation and of restriction. Indeed, the use of auxiliary variables through a model determines the inclusion probabilities (overrepresentation) and imposes a balancing condition (restriction). Thus, balanced sampling is a crucial tool to implement these principles.

However, some limitations of the scope of this paper must be outlined. First, it is worth noting that beyond the theoretical principles there are also a practical constraints. Practitioners have to take the context into account. A very large number of practical issues affect the direct applications of the suggested theoretical principles. So we recommend to keep those principles in mind when designing a survey even though we acknowledge that it is probably not always possible to apply them because of constraints such as time, inaccurate sampling frame or budget.

In addition to this, a simplicity principle can be predominant. A large number of environmental monitoring surveys are based on a systematic spatial sampling just because this design has the advantage of being simple, spread and easy to implement.

Moreover, in the case of multi-objective surveys, a single model that summarizes the link between the variables of interest and the auxiliary variables is not always available. There is sometimes an interest for regional or local estimations or for complex statistics. The aim can thus not be reduced to the estimation of a simple total. Compromises should then be established between the different objectives of the samples (Falorsi and Righi, 2008, 2016).

Finally, surveys are also repeated in time, which makes the problem much more intricate. Transversal and longitudinal estimations require very different sampling designs. It is always better to select the same sample to estimate evolutions, while for transversal estimations independent samples are more efficient. In business statistics, great attention is given to the survey burden that should be fairly distributed between companies. For these reasons, in a large number of surveys, statisticians foster a partial rotation of the units in the sample. Rotation is sometimes difficult to reconcile with the optimization of the transversal designs.

The three principles formalized and developed in this paper should guide the choice of the sampling design whenever possible. The principle of randomization should always be considered by trying to maximize the entropy, possibly under some constraints. The other two principles can only be applied when the population is explicitly modelled. This modeling may or may not be used as an assumption for the inference, depending on whether a design-based or a model-based approach is adopted. Using a model-assisted approach, we advocate the use

of a model to apply the principles of overrepresentation and of restriction while preserving the design unbiasedness of the Horvitz-Thompson estimator.

Acknowledgements

The authors would like to thank two referees and the Associate Editor for constructive remarks that have helped to greatly improve the paper. We would like to thank Mihaela Anastasiade for a careful reading and Audrey-Anne Vallée for useful comments on an early draft of the present paper.

Chapter 3

Sampling Designs on Finite Populations with Spreading Control

Abstract

We present new sampling methods in finite population that allow one to control the joint inclusion probabilities of units and especially the spreading of sampled units in the population. They are based on the use of renewal chains and multivariate discrete distributions to generate the difference of population ranks between successive selected units. With a Bernoulli sampling design, these differences follow a geometric distribution, and with a simple random sampling design they follow a negative hypergeometric distribution. We propose to use other distributions and introduce a large class of sampling designs with and without fixed sample size. The choice of the rank-difference distribution allows us to control units joint inclusion probabilities with a relatively simple method and closed-form formula. Joint inclusion probabilities of neighboring units can be chosen to be larger, or smaller, compared to those of Bernoulli or simple random sampling, thus allowing more or less spread of the sample in the population. This can be useful when neighboring units have similar characteristics or, on the contrary, are very different. A set of simulations illustrates the qualities of this method. [a]

[a] This chapter is essentially a reprint of Tillé et al. (2017).

3.1 Introduction

In this paper, we propose sampling methods for fixed and random sample sizes. We more particularly focus on the spacings that are the difference of population ranks between two successive selected units. We propose a large set of new methods that allows one to control the spacings and thus the joint inclusion probabilities of population units in the sample. These methods are useful in that they allow one to make less (or more) likely the selection of neighboring units. Indeed, when the variable of interest takes similar values on neighboring units, spreading the sample improves estimation because the selection of similar units is avoided.

A sampling design is a probability distribution on all the finite subsets of a population. It can be implemented by means of sampling algorithms. Several different sampling algorithms can implement the same sampling designs. Examples are given in Tillé (2006) where a large number of algorithms is given for designs like Simple Random Sampling (SRS) with and without replacement or maximum entropy sampling designs. Algorithms such that the decision

of selecting or not a unit into the sample is taken for each population unit successively according to the order of the population sampling frame are called "sequential" or "one-pass" algorithms. These algorithms are particularly useful when the population list is dynamic, like on a production chain or in real time sampling applications.

Systematic sampling is one of the most common sampling designs. It has been studied among others by Madow and Madow (1944), Cochran (1946), Madow (1949), Bellhouse and Rao (1975), Iachan (1982), Iachan (1983), Murthy and Rao (1988), Bellhouse (1988), Bellhouse and Sutradhar (1988), and Pea et al. (2007). One advantage of systematic sampling is that it spreads the sample very well over the population, thus allowing one to get precise estimators for totals and averages in the case of "auto-correlated" interest variables. Indeed, it can be shown to be an optimal design in this case under some conditions (Bondesson, 1986). However, it presents the important drawback that lots of unit couples have null joint inclusion probabilities. This makes impossible an unbiased estimation of the variance.

This drawback has led to a quest for other sampling designs that would retain good estimation properties. Deville (1998) proposed the Deville-systematic method, also called ordered pivotal method by Chauvet (2012) (see also Tillé, 2006, pp. 128-130). Tillé (1996) proposed a moving stratification algorithm that avoids the selection of neighboring units. Bondesson and Thorburn (2008) and Grafström (2010a) also proposed a method that allows one to control joint-inclusion probabilities. Recently, Loonis and Mary (2015) proposed using determinantal point processes that are known for their repulsiveness property (see for example Daley and Vere-Jones, 2002, p. 138). This last method necessitates to work with a huge matrix.

We advocate the use of point processes with simple specifications, motivated by usual sampling designs: the systematic design has deterministic spacings between selected units, the Bernoulli sampling design (see for example Tillé, 2006, pp. 43–44) has geometrically distributed spacings, and circular spacings of the simple random sampling design follow a negative hypergeometric distribution (see Vitter (1984, 1985, 1987)). It is worth mentioning that the most efficient known algorithm to draw a simple random sample is the Z algorithm proposed by Vitter (1985). This algorithm is optimal up to constant factor. In this paper, we will use other distributions to tune the joint selection probability of neighboring units. For each of these methods, we are able to compute positive joint inclusion probabilities and unbiased variance estimators. Special attention to edge effects must be given to ensure correct first order inclusion probabilities. Part of these sampling designs, with independent and identically distributed spacings, were introduced by Bondesson (1986).

The paper is organized as follows: Section 3.2 is devoted to the main definitions of survey sampling theory. Sections 3.3 and 3.4 present renewal chain sampling designs for random size samples. In Sections 3.5 and 3.6, we discuss fixed size sampling obtained through the generation of circular spacings with multivariate discrete distributions. Simulation results are given in Section 3.7. The paper ends with our conclusions in Section 3.8.

3.2 Sampling from a finite population

Consider the finite population of N units, $U = \{1, \ldots, N\}$. A sample without replacement of U is a subset $s \subset U$. A sampling design $P(.)$ is a probability distribution on samples,

$$P(s) \geq 0, \ s \subset U, \text{ such that } \sum_{s \subset U} P(s) = 1.$$

Let S denote the random sample, so that $\Pr(S = s) = P(s)$. The sample size $n = \#S$ can be random or not. The inclusion probability of unit k is its probability of being selected into a sample

$$\pi_k = \Pr(k \in S) = \sum_{s \ni k} P(s).$$

The joint inclusion probability of units k and ℓ is their probability of being selected together into a sample

$$\pi_{k\ell} = \pi_{\ell k} = \Pr(k \text{ and } \ell \in S) = \sum_{s \ni k,\ell} P(s).$$

Let Y be a variable of interest and let y_k be the value of Y associated to unit k of the population. The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) is defined by

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

It is an unbiased estimator of the population total

$$t_Y = \sum_{k \in U} y_k,$$

provided that $\pi_k > 0$, $k \in U$. Let

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell & \text{if } k \neq \ell, \\ \pi_k(1 - \pi_k) & \text{if } k = \ell. \end{cases}$$

The variance of the Horvitz-Thompson estimator is

$$\operatorname{var}\left(\widehat{Y}\right) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.$$

If the sampling design has a fixed size, the variance can also be written as (see Sen, 1953; Yates and Grundy, 1953)

$$\operatorname{var}\left(\widehat{Y}\right) = -\frac{1}{2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell}.$$

Estimators can be derived from these two expressions. For the general case, the variance estimator is given by:

$$\widehat{\operatorname{var}}_{HT}\left(\widehat{Y}\right) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}}, \tag{3.1}$$

where $\pi_{kk} = \pi_k$. When the sample size is fixed, the Sen-Yates-Grundy variance estimator (Sen, 1953; Yates and Grundy, 1953) is given by

$$\widehat{\operatorname{var}}_{SYG}\left(\widehat{Y}\right) = -\frac{1}{2} \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{3.2}$$

These estimators are unbiased provided that $\pi_{k\ell} > 0$, $k \neq \ell \in U$. Estimator (3.2) is non-negative when $\Delta_{k\ell} \leq 0$, $k \neq \ell$ (Sen-Yates-Grundy conditions).

3.3 Renewal chain sampling designs

The idea of selecting samples through the use of renewal processes is not new. It can be traced back at least to Bondesson (1986) (see also Meister, 2004). We give a different presentation in this section in that we focus on the parametrization of the distribution of spacings between selected units whereas Bondesson (1986) and Meister (2004) focus on the parametrization of the so-called *renewal sequence*, the conditional inclusion probabilities given the past. Their aim was to provide solutions for real time sampling, and the proposed methods are intrinsically sequential, allowing one to spread the sample by introducing a negative correlation between the sample inclusion indicators. Bondesson and Thorburn (2008) generalize this idea using a splitting method (see Deville and Tillé, 1998) that allows use of a unequal probability sampling designs for real time sampling.

3.3.1 Definition

In this section, we present a family of sampling algorithms that are parametrized by a discrete probability distribution. By a careful choice of the generating distribution, we obtain sampling designs with desirable properties. Consider a sequence $J_1, \dots, J_N$ of independently and identically distributed (i.i.d.) random variables in $\mathbb{N}^* = \{1, 2, 3, \dots\}$. The partial sums $S_j = \sum_{i=1}^{j} J_i$, $j \geq 1$, form a discrete process that is called a simple renewal chain (see for example Feller, 1971 and Barbu and Limnios, 2008, p. 18), by analogy with renewal processes (see Cox, 1962; Daley and Vere-Jones, 2002; Mitov and Omey, 2014). Using these J_i's as spacings (jumps) between successive units selected into the sample, we obtain the family of sampling designs of Definition 3.3.1.

Definition 3.3.1. *A sampling design is said to be a (simple) renewal chain sampling design if its random sample can be written*

$$\tilde{S} = \{1, \dots, N\} \bigcap \left\{ \sum_{i=1}^{j} J_i,\ 1 \leq j \leq N \right\},$$

where $J_1, \dots, J_N$ are i.i.d. random variables in $\mathbb{N}^*$.

The first order inclusion probability of a renewal chain design can be obtained from the common distribution $f(\cdot)$ of the J_i's:

$$\pi_k = \Pr(k \in \tilde{S}) = \sum_{j=1}^{k} f^{j*}(k), \tag{3.3}$$

where $f^{j*}(\cdot)$ is the distribution of the sum of j i.i.d. variables with distribution $f(\cdot)$. Indeed, unit k is selected if $J_1 = k$, or $J_1 + J_2 = k$, or $\cdots$, or $J_1 + \cdots + J_k = k$. These events are non-overlapping thanks to the J_i's being positive. We obtain that:

$$\pi_k = \sum_{j=1}^{k} \Pr\left(\sum_{i=1}^{j} J_i = k \right),$$

which is exactly Equation (3.3). It is a well-known property of renewal process theory given, for example, in Barbu and Limnios (2008, p. 21), Cox (1962, p. 53) or in Mitov and Omey (2014, pp. 44-47).

Even with i.i.d. spacings, a simple renewal chain sampling design usually has unequal first order inclusion probabilities, as we can see in Example 3.3.1.

Example 3.3.1. *Let J_i, $i \in \mathbb{N}^*$ be a sequence of i.i.d. variables such that $\Pr(J_i = 1) = 1/2$ and $\Pr(J_i = 2) = 1/2$. Then,*

$$\begin{aligned}
\pi_1 &= \Pr(J_1 = 1) = 1/2, \\
\pi_2 &= \Pr(J_1 = 2) + \Pr(J_1 + J_2 = 2) = 1/2 + 1/4 = 3/4, \\
\pi_3 &= \Pr(J_1 + J_2 = 3) + \Pr(J_1 + J_2 + J_3 = 3) = 1/2 + 1/8 = 5/8, \\
\pi_4 &= \Pr(J_1 + J_2 = 4) + \Pr(J_1 + J_2 + J_3 = 4) + \Pr(J_1 + J_2 + J_3 + J_4 = 4) = 11/16, \\
&\vdots
\end{aligned}$$

3.3.2 Equilibrium renewal chains

A delayed renewal chain is a discrete process $(S_j)_{j\in\mathbb{N}}$ with $S_j = \tilde{J}_0 + \sum_{i=1}^{j} J_i$, where the J_i's, $i \geq 1$ are i.i.d. random variables taking values in $\mathbb{N}^*$ and $\tilde{J}_0$ is an independent random variable taking values in $\mathbb{N}$ (see e.g. Barbu and Limnios, 2008, p. 31). Of particular interest is the delayed renewal chain obtained when the distribution of $\tilde{J}_0$ is obtained from the distribution of J_1 using

$$\Pr(\tilde{J}_0 = k) = \frac{\Pr(J_1 \geq k + 1)}{\mathrm{E}(J_1)}, \; k \in \mathbb{N}, \tag{3.4}$$

provided that $\mathrm{E}(J_1)$ exists. The distribution of $\tilde{J}_0$ is called the stationary or equilibrium distribution of the renewal chain and the resulting delayed renewal chain is called an equilibrium renewal chain. As written by Barbu and Limnios (2008, Proposition 2.2), this choice of the initial distribution $\tilde{J}_0$ of the delayed renewal chain is the only one where all $k \in \mathbb{N}$ have the same probability of being in the sample path. Proposition 3.3.1 is a general result of renewal process theory (see for example Mitov and Omey, 2014, p. 46) that we applied to the discrete case. We propose a direct proof of Proposition 3.3.1 in Appendix.

Proposition 3.3.1. *If $f(\cdot)$ is a probability distribution on $\mathbb{N}^*$ with cumulative distribution function $F(\cdot)$, expectation μ, and if $f_0(\cdot)$ is defined by $f_0(k) = f(\{k + 1, \dots\})/\mu$, $k \in \mathbb{N}$, then $f_0(\cdot)$ is a probability distribution and*

$$f_0(k) + \sum_{t=1}^{k} f_0(k - t) \sum_{j=1}^{t} f^{j*}(t) = \frac{1}{\mu}, \text{ for all } k \geq 1. \tag{3.5}$$

Corollary 3.3.1. *Let S_j, $j \in \mathbb{N}$ be a delayed renewal chain with $\mathrm{E}(J_1) = \mu$ and $\tilde{J}_0$ have the distribution of (3.4). For all $k \in \mathbb{N}$, if π_k is the probability that k is in the sample path, then*

$$\pi_k = f_0(k) + \sum_{t=1}^{k} f_0(k - t) \sum_{j=1}^{t} f^{j*}(t), \tag{3.6}$$

and is equal to $1/\mu$.

Proof. The event $\{\exists i \in \mathbb{N} \text{ such that } S_i = k\}$ can be decomposed as

$$\{\exists i \in \mathbb{N} \text{ such that } S_i = k\} = \bigcup_{t=0}^{k} \left\{\tilde{J}_0 = k - t\right\} \bigcap \bigcup_{j=1}^{t} \left\{\sum_{i=1}^{j} J_i = t\right\},$$

where all the events in the union are non-overlapping. It follows that

$$\pi_k = f_0(k) + \sum_{t=1}^{k} f_0(k-t) \sum_{j=1}^{t} f^{j*}(t) = \frac{1}{\mu},$$

by Proposition 3.3.1. □

Definition 3.3.2. *Let X be a random variable with values in $\mathbb{N}$ and finite expectation. A random variable X_F is called a forward transform of X if its distribution is given by*

$$\Pr(X_F = k) = \frac{\Pr(X \geq k)}{\mathrm{E}(X+1)},\ k \in \mathbb{N}.$$

Remark 3.3.1. *Moments of X_F can be derived from those of X using the property, proven in the Appendix, that if X is a random variable on $\mathbb{N}$ with finite moment of order $m+1$, $\mathrm{E}(X^{m+1})$, $m \geq 0$, then its forward transform X_F has a finite moment of order m and*

$$\mathrm{E}(X_F^m) = \frac{\mathrm{E}[F_m(X)]}{\mathrm{E}(X+1)},$$

where $F_m(x)$ is the Faulhaber polynomial integer function of degree $m+1$: $F_m(x) = \sum_{k=0}^{x} k^m$.

The equilibrium distribution $\tilde{J}_0$ is the forward transform of the distribution of $J_1 - 1$, according to Definition 3.3.2. Spacing distributions considered in Section 3.4 are defined as shifted variables $J_1 = 1 + X$ where X follows a classical probability distribution on $\mathbb{N}$. The reader can find in Table A.1 a collection of distributions that are used in Sections 3.4 and 3.6, as well as their forward transforms.

3.3.3 Equilibrium renewal chain sampling designs

By taking the intersection of the sample path of an equilibrium renewal chain with the population $U = \{1, \ldots, N\}$, one obtains a random sampling design. Corollary 3.3.1 ensures that all units of the population have the same inclusion probability. The distribution of the first selected unit index X_1 satisfies Equation (3.7).

$$\Pr(X_1 = k) = \Pr(\tilde{J}_0 = k) + \Pr(\tilde{J}_0 = 0)\Pr(J_1 = k) = \frac{\Pr(J_1 \geq k)}{\mathrm{E}(J_1)},\ k \in U. \tag{3.7}$$

By definition, the following sampled units are obtained by adding independent variables distributed like J_1.

Definition 3.3.3. *An equilibrium renewal chain sampling design is the distribution of a random sample S with*

$$S = \{1, \ldots, N\} \bigcap \left\{\sum_{i=0}^{j} J_i,\ 0 \leq j \leq N-1\right\},$$

where $J_1, \ldots, J_{N-1}$ are i.i.d random variables in $\mathbb{N}^*$ with finite expectation, and J_0 is an independent variable with distribution given by (3.7), $\Pr(J_0 = k) = \Pr(J_1 \geq k)/\mathrm{E}(J_1)$, $k \in U$.

For $J_1 = 1 + X$, the random variable J_0 of (3.7) has the same distribution as $1 + X_F$ where X_F is a forward transform of X.

The equilibrium renewal chain design that corresponds to the renewal distribution of Example 3.3.1 is given in Example 3.3.2. Its first order inclusion probabilities are equal.

Example 3.3.2. *Consider the sequence J_i, $i \in \mathbb{N}^*$ of Example 3.3.1, and define J_0 to be independent of the J_i's, with $P(J_0 = 1) = 2/3$ and $P(J_0 = 2) = 1/3$ according to (3.7). The new inclusion probabilities $\tilde{\pi}_i$ of this equilibrium renewal sampling design are related to those of Example 3.3.1 by:*

$$\begin{aligned}
\tilde{\pi}_1 &= \Pr(J_0 = 1) &&= 2/3,\\
\tilde{\pi}_2 &= \Pr(J_0 = 1)\pi_1 + \Pr(J_0 = 2) &&= 1/3 + 1/3 = 2/3,\\
\tilde{\pi}_3 &= \Pr(J_0 = 1)\pi_2 + \Pr(J_0 = 2)\pi_1 &&= 1/2 + 1/6 = 2/3,\\
&\vdots
\end{aligned}$$

3.3.4 Joint inclusion probabilities

Joint inclusion probabilities of a renewal chain sampling design can be derived from the Probability Mass Function (PMF) $f(.)$ of J_1. Indeed, the selection of unit ℓ given that unit k, $0 < k < \ell$, is selected can be decomposed according to the number of selected units between k and ℓ, and this number does not depend on J_0. We can write that:

$$\Pr(\ell \in S | k \in S) = \sum_{j=1}^{\ell-k} f^{j*}(\ell - k).$$

The joint inclusion probability of units $k \neq \ell$ is thus given by (3.8).

$$\pi_{k\ell} = \pi_k \sum_{j=1}^{\ell-k} f^{j*}(\ell - k),\ k < \ell. \tag{3.8}$$

3.3.5 Bernoulli sampling

The Bernoulli sampling design with inclusion probabilities π is obtained by selecting or not units into the sample through independent Bernoulli trials with parameter π (see for example Tillé, 2006, p. 43). Its probability distribution is given by

$$P(s) = \pi^n (1 - \pi)^{N-n},\ s \subset U,$$

where $n = \#s$ is the size of sample s. The joint inclusion probabilities are equal to $\pi_{k\ell} = \pi^2$, $k \neq \ell$. The usual algorithm used to select a sample according to the Bernoulli sampling design simply consists of generating N independent Bernoulli variables and selecting units according to the observed values.

Bernoulli sampling can also be implemented using Definition 3.3.1. Indeed, it is clear that spacings of a Bernoulli sampling design are i.i.d. distributed variables with shifted geometric distributions $J_i = 1 + X_i$ where $\Pr(X_i = k) = (1 - \pi)^k \pi$, $k \geq 0$. Bernoulli sampling is thus a simple renewal chain sampling design satisfying Definition 3.3.1. On the other hand it is easy

to prove that, if X_i follows a geometric distribution, then X_i has the same distribution as its forward transform (it is the only distributions on $\mathbb{N}$ that enjoy this property). The random variable J_0 of Definition 3.7 has the same distribution as J_1 in this particular case. Consequently, Bernoulli sampling is also an equilibrium renewal chain sampling design according to Definition 3.3.3.

Using (3.8), we find the second order inclusion probabilities $\pi_{k\ell} = \pi^2$, $k \neq \ell$. Indeed, the sum of j i.i.d. geometric random variables with parameter π follows a negative binomial distribution with parameters j and π. The negative binomial distribution with parameters $j \geq 1$ and π in $(0, 1)$ is defined by its PMF:

$$f_{\mathcal{NB}}(x) = \binom{j+x-1}{x}(1-\pi)^x\pi^j,\ x \in \mathbb{N}, \tag{3.9}$$

where $\binom{a}{b} = a!/[b!(a-b)!]$ if $b \leq a$ are non-negative integers, and $\binom{a}{b} = 0$ if a, b or $a-b$ is negative. Considering that

$$\sum_{i=1}^{j} J_i = j + \sum_{i=1}^{j} X_i,\ j \geq 1,$$

we have that

$$f^{j*}(x) = \binom{x-1}{x-j}(1-\pi)^{x-j}\pi^j,\ x \geq j.$$

From (3.8), we get that the joint inclusion probabilities are equal to:

$$\begin{aligned}
\pi_{k\ell} &= \pi \sum_{j=1}^{\ell-k} f^{j*}(\ell-k),\ k < \ell, \\
&= \pi \sum_{j=1}^{\ell-k} \binom{\ell-k-1}{\ell-k-j}(1-\pi)^{\ell-k-j}\pi^j, \\
&= \pi^2(\pi+1-\pi)^{\ell-k-1} = \pi^2.
\end{aligned}$$

3.3.6 Systematic sampling

Systematic sampling with rate $1/r$, $r \in \mathbb{N}^*$, from a population $U = \{1, \ldots, N\}$ is obtained by generating a random start u with a uniform discrete distribution between 1 and r, and selecting units k of U such that $k \equiv u \pmod r$ into the sample (see Madow and Madow, 1944). The first order inclusion probabilities of this sampling design are given by $\pi_k = 1/r$, $k \in U$, and its joint inclusion probabilities by

$$\pi_{k\ell} = 1/r \text{ if } k \equiv \ell \pmod r \text{ and } 0 \text{ otherwise.}$$

If $N = mr$, with $m, r \in \mathbb{N}^*$, the sample size is deterministic and equal to m.

Systematic sampling is an equilibrium renewal chain sampling design, agreeing with Definition 3.3.3 where the J_i's, $i \geq 1$ are deterministic and equal to r. Indeed, the forward transform X_F of $X = r - 1$, $r \in \mathbb{N}^*$ is such that:

$$\Pr(X_F = k) = \frac{\Pr(X \geq k)}{\mathrm{E}(X+1)} = \frac{\mathbf{1}_{\{r-1 \geq k\}}}{r},\ k \in \mathbb{N},$$

and X_F follows a uniform distribution on $\{0, \dots, r-1\}$. Hence J_0 follows a uniform distribution on $\{1, \dots, r\}$.

The joint inclusion probabilities are obtained from (3.8). Indeed, the sum of j spacings J_i, $i \geq 1$ is deterministic, equal to jr, and

$$f^{j*}(x) = \mathbf{1}_{\{jr=x\}}, \ x \geq 1.$$

We then have that, for $k < \ell$,

$$\pi_{k\ell} \ = \ \pi \sum_{j=1}^{\ell-k} f^{j*}(\ell-k) = \frac{1}{r} \sum_{j=1}^{\ell-k} \mathbf{1}_{\{jr=\ell-k\}} = \frac{1}{r} \mathbf{1}_{\{k \equiv \ell \pmod r\}}.$$

We confirm with this expression that most of the joint inclusion probabilities are null, making it impossible to estimate the variance of Horvitz-Thompson estimators without bias.

3.4 Spreading renewal chain sampling designs

We have seen in Sections 3.3.5 and 3.3.6 two examples of renewal chain sampling designs with very different spreading properties. In Bernoulli sampling, the selection of units are independent, even if they are adjacent in the population list. In systematic sampling, the selection of adjacent units is impossible, provided that the sampling rate is smaller than 1. This translates to the variance of the spacings distribution: it is null for systematic sampling, that has perfect spreading properties, and it is quite large, equal to $(1-\pi)/\pi^2$, for Bernoulli sampling.

Using Definition 3.3.3, we can build sampling designs with any given spacing distribution on $\mathbb{N}^*$. The expectation of this distribution is forced by the sampling rate, which is usually itself decided as a function of cost or precision constraints. In Section 3.4.1, we give an application with shifted negative binomial spacings, allowing for a limited control on the variance and spreading properties of the design. As a limiting case, we find the shifted Poisson spacings of Section 3.4.2. To have a variance that is arbitrarily small, in Section 3.4.3 we use shifted binomial distributions that have a variance always smaller than their expectation.

These are only examples, and any distribution or family of distributions on $\mathbb{N}^*$ that offers sufficient control on its shape can be used. Table A.1 in Appendix contains a list of useful discrete probability distributions with their probability mass functions, their supports, means and variances.

3.4.1 Negative binomial spacings

The definition of the negative binomial distribution in (3.9) can be extended to parameters $r > 0$ and p in $(0, 1)$ by considering the PMF:

$$f_{\mathcal{NB}(r,p)}(x) = \frac{\Gamma(r+x)}{x!\Gamma(r)} p^r (1-p)^x, \ x \in \mathbb{N},$$

where $\Gamma(r) = \int_0^{+\infty} t^{r-1} e^{-t} \, dt$, $r > 0$ and $\Gamma(k) = (k-1)!$, $k \in \mathbb{N}^*$. The expectation of this distribution is $r(1-p)/p$ and its variance is $r(1-p)/p^2$.

We consider equilibrium renewal sampling designs with positive spacings J_i, $i \geq 1$ such that $J_i - 1$ follows a negative binomial distribution with parameters r and p, denoted $\mathcal{NB}(r, p)$.

For a given sampling rate $\pi \in (0,1)$, we find that $\mathrm{E}(J_i) = 1/\pi$ implies that

$$p = \frac{r\pi}{r\pi + 1 - \pi}.$$

It follows that

$$\mathrm{var}(J_i) = \frac{1-\pi}{\pi} + \frac{1}{r}\left(\frac{1-\pi}{\pi}\right)^2. \tag{3.10}$$

When $r = 1$, we find, as a special case, the Bernoulli sampling design. From (3.10), we deduce that the variance of spacings is smaller than that of Bernoulli sampling when $r > 1$ and in that case there is a repulsion between selected units: the sample is spread more evenly on the population than if drawings were independent. On the contrary, if $r < 1$, there is an attraction between units and selecting neighbor units together is more likely.

The sum of j independent random variables with negative binomial distribution and parameters r, p, has a negative binomial distribution with parameters jr and p.

Proposition 3.4.1. *The second order inclusion probabilities of an equilibrium renewal chain sampling design with shifted negative binomial spacings, sampling rate π and first parameter r are*

$$\pi_{k\ell} = \pi \sum_{j=1}^{\ell-k} \frac{\Gamma(jr+\ell-k-j)}{(\ell-k-j)!\Gamma(jr)} p^{jr}(1-p)^{\ell-k-j},\ k < \ell,$$

where $p = r\pi/(r\pi + 1 - \pi)$.

Proof. Since $J_i - 1$ has a $\mathcal{NB}(r,p)$ distribution, $\left(\sum_{i=1}^{j} J_i\right) - j$ has a $\mathcal{NB}(jr,p)$ distribution. Using (3.8) one gets the result. □

These joint inclusion probabilities remain positive for any value of r. They are plotted in Figure 3.1 for $\pi = 1/30$ and different values of r.

In order to have an equilibrium renewal chain sampling design and equal first order inclusion probabilities, the first sample unit index has to be generated from a shifted forward negative binomial. We get that $J_0 - 1 \sim \mathcal{ForNB}(r,p)$, where the definition of $\mathcal{ForNB}(r,p)$ can be found in Table A.1.

3.4.2 Poisson spacings

The limit of negative binomial distributions when r tends to infinity and p tends to 0 while keeping a constant expectation $\lambda = r(1-p)/p$, is a Poisson distribution $\mathcal{P}(\lambda)$ with parameter λ which is also its expectation and its variance.

We consider the equilibrium renewal chain sampling design with shifted Poisson spacings: $J_i - 1 \sim \mathcal{P}(\lambda)$, where $\lambda = (1-\pi)/\pi$, $i \in \mathbb{N}^*$. The first spacing J_0 is selected using a shifted forward Poisson distribution: $J_0 - 1 \sim \mathcal{ForP}(\lambda)$, where the definition of distribution $\mathcal{ForP}(\lambda)$ can be found in Table A.1.

Proposition 3.4.2. *The second order inclusion probabilities of an equilibrium renewal chain sampling design with shifted Poisson spacings and sampling rate π are, with $\lambda = (1-\pi)/\pi$*

$$\pi_{k\ell} = \pi \sum_{j=1}^{\ell-k} \frac{e^{-j\lambda}\,(j\lambda)^{\ell-k-j}}{(\ell-k-j)!},\ k < \ell.$$

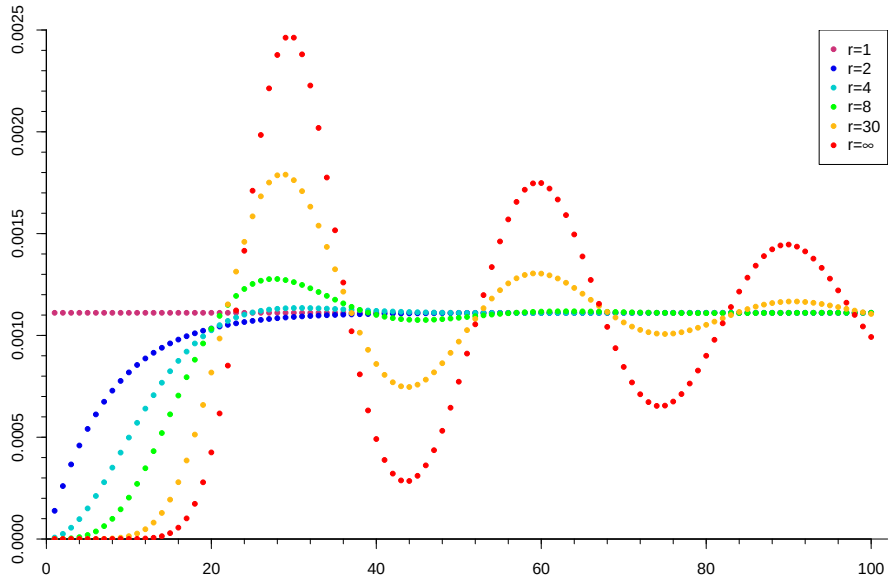

FIGURE 3.1: $\pi_{k,k+i}$ in function of i for negative binomial spacings with $\pi = 1/30$, $r = 1, 2, 4, 8, 30$ and $r = +\infty$ (Poisson spacings). When $r = 1$, we obtain the Bernoulli sampling design and a flat line on the plot. Oscillations are stronger for larger values of r.

3.4.3 Binomial spacings

The variance of spacings in Sections 3.4.1 and 3.4.2 are bounded from below, by $(1-\pi)/\pi$. However, to get a sample spread close to that of systematic sampling, we need to be able to have a variance that is arbitrarily close to 0. For this, we consider the equilibrium renewal chain sampling design that is obtained with shifted binomial spacings: $J_i - 1 \sim \mathcal{B}in(r, p)$, $i \in \mathbb{N}^*$, $r \in \mathbb{N}$, $p \in [0, 1]$. The first spacing J_0 is selected using a shifted forward binomial distribution: $J_0 - 1 \sim \mathcal{F}or\mathcal{B}in(r, p)$, where the definition of distribution $\mathcal{F}or\mathcal{B}in(r, p)$ can be found in Table A.1.

We find that with a sampling rate equal to π, r must necessarily be greater or equal to $(1-\pi)/\pi$, and that

$$p = \frac{1}{r}\left(\frac{1-\pi}{\pi}\right).$$

The variance of spacings is then given by (3.11),

$$\mathrm{var}(J_i) = \frac{1-\pi}{\pi} - \frac{1}{r}\left(\frac{1-\pi}{\pi}\right)^2, \; i \in \mathbb{N}^*. \tag{3.11}$$

Considering the constraints on r and p, this variance is minimal when r is the smallest integer that is greater or equal to $(1-\pi)/\pi$. With this r, the variance of spacings is always smaller than 1, which is really small for an integer valued random variable with a usually very large expectation $1/\pi$. When $1/\pi$ is an integer, the variance of spacings is null when $r = (1-\pi)/\pi$ and $p = 1$. The sampling design obtained then is just the systematic sampling design.

If r tends to infinity and $p = (1-\pi)/r\pi$, the binomial distribution with parameters r and

p converges in distribution toward the Poisson distribution with parameter $(1-\pi)/\pi$. Hence, the sampling design of Section 3.4.2 is also the limiting case of Binomial spacings renewal chain sampling design when r tends to infinity.

Proposition 3.4.3. *The second order inclusion probabilities of an equilibrium renewal chain sampling design with shifted binomial spacings, sampling rate π and first parameter r are equal to*

$$\pi_{k\ell} = \pi \sum_{j=1}^{\ell-k} \binom{jr}{\ell-k-j} p^{\ell-k-j} (1-p)^{j(r+1)-\ell-k}, \ k < \ell,$$

where $p = (1-\pi)/r\pi$.

3.4.4 Summary

The different renewal chain sampling designs we considered are listed in Table 3.1 with the variance of their spacings. If $1/\pi$ is not an integer, the variance of spacings cannot be null and

TABLE 3.1: Renewal chain sampling designs and variance of their spacings.

Distribution of $J_1 - 1$	$\mathrm{var}(J_1)$
Negative binomial, $r > 0$	$\frac{1-\pi}{\pi} + \frac{1}{r}\left(\frac{1-\pi}{\pi}\right)^2$
Poisson	$\frac{1-\pi}{\pi}$
Binomial, $r \geq (1-\pi)/\pi$	$\frac{1-\pi}{\pi} - \frac{1}{r}\left(\frac{1-\pi}{\pi}\right)^2$
Systematic or binomial with $r = (1-\pi)/\pi$	0

is at least $(\lceil 1/\pi \rceil - 1/\pi)(1/\pi - \lfloor 1/\pi \rfloor)$. This lower bound is not reached with shifted binomial spacings but the binomial renewal chain sampling design enjoys the desirable property of having positive joint inclusion probabilities. Other spacing distributions can be used but we retained only common families of distribution that have useful properties such as stability under convolution.

3.5 Fixed size sampling designs with exchangeable circular spacings

Except in very special situations, renewal chain sampling designs do not have fixed sample size. This is due to the independence of spacings. However, in many applications fixed size is required. In this section, we propose to define sampling designs using exchangeable instead of independent spacings. We obtain fixed size designs with equal inclusion probabilities, and we are able to control the sample spread by the choice of the random spacings distribution.

3.5.1 Circular spacings

A sampling design of fixed size n in a population $U = \{1, \ldots, N\}$ is entirely specified by the joint distribution of one of the unit indexes, e.g. X_1, and the "circular" spacings $J_i = X_{i+1} - X_i$, $i = 1, \ldots, n-1$, and $J_n = N + X_1 - X_n$, where X_1 is the smallest sample unit index and X_n the largest. If we represent the population U around a table, as in Figure 3.2, the J_i's are the

difference of units position. Note that considering the population as circular is not new in survey sampling and goes back at least to Fuller (1970).

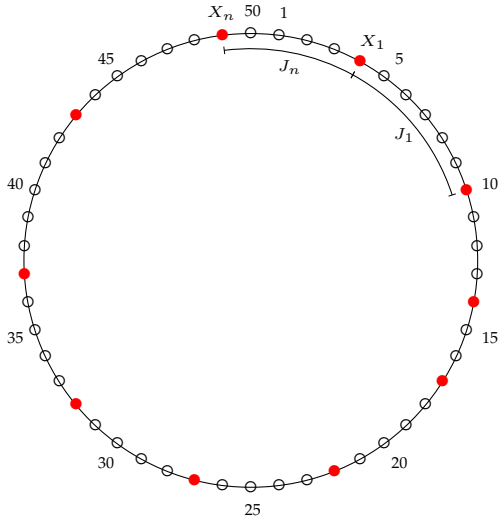

FIGURE 3.2: Illustration of a sample of 10 units, in a population of 50 units. The selected units are in red.

We intend to work with Equation (3.12) that defines without loss of generality the random sample S of a fixed size sampling design,

$$S = \{S_j \pmod N,\ j = 1, \ldots, n\}, \tag{3.12}$$

where $J_1, \ldots, J_n$ are positive integer random variables that sum to N, J_0 is a random variable in U, and

$$S_j = \sum_{i=0}^{j} J_i.$$

3.5.2 First order inclusion probabilities

The first order inclusion probabilities of a sampling design that results from (3.12) depend on the joint distribution of the J_i's, $i = 0, \ldots, n$. Intuitively, one sees that, for a given joint distribution of $J_1, \ldots, J_n$ with $S_n = N$, choosing J_0 to be independent of the other J_i's and uniform on $\{1, \ldots, N\}$ allows to obtain equal first order inclusion probabilities. To prove this assertion, one can compute in the general case the inclusion probability of a unit k. Consider an independent J_0, and let

$$f_0(t) = \Pr(J_0 = t),\ f_j(k) = \Pr(S_j \pmod N = k),\ t,\ k \in U,\ 1 \leq j \leq n.$$

By conditioning on the event $\{J_0 = t\}$ and using the law of total probability on the disjoint events $\{k > t\}$, $\{k = t\}$ and $\{k < t\}$, we get that:

$$\pi_k = \sum_{t=1}^{N} f_0(t) \left[\mathbf{1}_{t=k} + \mathbf{1}_{t<k} \sum_{j=1}^{k-t} f_j(k-t) + \mathbf{1}_{t>k} \sum_{j=1}^{N+k-t} f_j(N+k-t) \right],$$

with the convention that $f_j(k) = 0$ if $j > n$. Hence we can write that vectors $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ and $\mathbf{f}_0 = (f_0(1), \ldots, f_0(N))$ are solutions of the linear equation

$$\boldsymbol{\pi} = \mathbf{A}\mathbf{f}_0, \tag{3.13}$$

where $\mathbf{A}$ is the square matrix of size N with general term

$$a_{kt} = \mathbf{1}_{t=k} + \mathbf{1}_{t<k} \sum_{j=1}^{k-t} f_j(k-t) + \mathbf{1}_{t>k} \sum_{j=1}^{N+k-t} f_j(N+k-t),\ 1 \leq k, t \leq N. \tag{3.14}$$

It is not our purpose to solve the system and find designs with any given inclusion probabilities, especially since solutions depend on the $f_j(k)$'s, but one arrives rapidly to the result that all the lines of $\mathbf{A}$ sum to n (see Proposition A.1.1 in Appendix). Hence, if $f_0(t) = 1/N$ for all t, then the inclusion probabilities are all equal to n/N.

3.5.3 Joint inclusion probabilities

Let $\ell > k$ in $\{1, \ldots, N\}$, and consider $\Pr(\ell \in S | k \in S)$ so that, by definition, $\pi_{k\ell} = \pi_k \Pr(\ell \in S | k \in S)$. Knowing that $k \in S$, the event $\ell \in S$ can be decomposed according to the number of units selected between k and ℓ:

$$\pi_{k\ell} = \pi_k \sum_{j=1}^{\ell-k} \Pr\left(\ell \in S \text{ and } \#S \cap \{k+1, \ldots, \ell\} = j | k \in S\right). \tag{3.15}$$

The term $\Pr\left(\ell \in S \text{ and } \#S \cap \{k+1, \ldots, \ell\} = j | k \in S\right)$ is usually difficult to compute: one must decompose according to which S_i is equal to k. However, if the joint distribution of $J_1, \ldots, J_n$ has some additional properties, as in Proposition 3.5.1, one can obtain a simple expression.

Proposition 3.5.1. *Consider a positive integer random vector $(J_1, \ldots, J_n)$ that sums to N and such that the distributions of any sum of k successive J_i's are equal, this condition also holding for the "circular" sums of J_{n-i} up to J_n and J_1 up to J_{k-i} if $i < k$. Then, the second order inclusion probabilities are given by*

$$\pi_{k\ell} = \pi_k \sum_{j=1}^{\ell-k} f_j(\ell - k),\ k < \ell. \tag{3.16}$$

Proof. Indeed, we then have that:

$$\Pr\left(\ell \in S \text{ and } \#S \cap \{k+1, \ldots, \ell\} = j | k \in S\right) = f_j(\ell - k),\ j = 1, \ldots, \ell - k,$$

and the result follows immediately. □

In the situation of Proposition 3.5.1, we also get that the conditional inclusion probability $\Pr(\ell \in S | k \in S)$ is a function of $\ell - k \pmod N$:

$$\Pr(\ell \in S | k \in S) = \sum_{j=1}^{\ell-k} f_j(\ell - k), \text{ if } k < \ell,$$

$$\Pr(\ell \in S | k \in S) = \sum_{j=1}^{N+\ell-k} f_j(N + \ell - k), \text{ if } \ell < k.$$

It also follows in that case that if the first order inclusion probabilities are all equal, for example when J_0 has a uniform distribution, then the joint inclusion probabilities $\pi_{k\ell}$ depend only on $\ell - k \pmod N$.

Actually, all distributions considered for spacings $J_1, \ldots, J_n$ in this paper enjoy a stronger property, they are exchangeable distributions (Aldous, 1985; Kallenberg, 2005).

Definition 3.5.1. *A family $J_1, \ldots, J_n$ of random variables is said to be exchangeable if, for all $1 \leq k \leq n$ and permutation σ of $\{1, \ldots, n\}$, the joint distribution of $(J_{\sigma(1)}, \ldots J_{\sigma(k)})$ is equal to the joint distribution of $(J_1, \ldots, J_k)$. If the J_i's are discrete distributions, this is equivalent to say that $\Pr(J_1 = a_1, \ldots, J_n = a_n)$ is a symmetric function of $(a_1, \ldots, a_n)$.*

Exchangeable integer distributions $J_1, \ldots, J_n$ that sum to N clearly satisfy the conditions of Proposition 3.5.1. They are the natural equivalents to the i.i.d. spacing distributions used in Section 3.3 when the spacings are constrained, by the fixed sample size, to sum to N.

Definition 3.5.2. *Fixed size sampling designs with exchangeable circular spacings and uniform inclusion probabilities are the sampling designs with random samples $S = \{S_j \pmod N,\ j = 1, \ldots, n\}$ where $S_j = \sum_{i=0}^{j} J_i$, J_0 is a uniform random distribution on $\{1, \ldots, N\}$ independent from $\mathbf{J} = (J_1, \ldots, J_n)$, and the J_i's, $i = 1, \ldots, n$, are exchangeable positive integer distributions that sum to N.*

The PMF of a fixed size sampling design with exchangeable circular spacings and uniform inclusion probabilities $J_1, \ldots, J_n$ is given simply by

$$P(s) = \frac{n}{N} \Pr(J_1 = x_2 - x_1, \ldots, J_{n-1} = x_n - x_{n-1}, J_n = N + x_1 - x_n), \tag{3.17}$$

where $x_1, \ldots, x_n$ are the ordered indexes of units sampled in s. Indeed, $P(s)$ can be decomposed according to the value of J_0 into:

$$\begin{aligned}
&\Pr(S = \{x_1, \ldots, x_n\}) \\
&= \Pr(J_0 = x_1) \Pr(J_1 = x_2 - x_1, \ldots, , J_{n-1} = x_n - x_{n-1}, J_n = N + x_1 - x_n) \\
&+ \Pr(J_0 = x_2) \Pr(J_1 = x_3 - x_2, \ldots, J_{n-1} = N + x_1 - x_n, J_n = x_2 - x_1) \\
&\vdots \\
&+ \Pr(J_0 = x_n) \Pr(J_1 = N + x_1 - x_n, \ldots, J_n = x_n - x_{n-1}),
\end{aligned}$$

and the $\Pr(J_0 = x_i)$ are all equal to $1/N$ while the probabilities involving $J_1, \ldots, J_n$ are all equal due to the exchangeability of the circular spacings.

3.5.4 Simple Random Sampling

The Simple Random Sampling (SRS) without replacement design of fixed sample size n is defined by:

$$P(s) = \binom{N}{n}^{-1} \text{ if } \#s = n, \text{ and } P(s) = 0 \text{ otherwise.}$$

A SRS sample can be selected using the following algorithm (Fan et al. (1962), see also Tillé, 2006, p. 46): define a counter $j = 0$, then, for $k = 1$ to N, select unit k with probability $(N - j)/(N - k - 1)$ and update $j = j + 1$ if k is selected. It is also possible to obtain this design by generating successive jumps according to negative hypergeometric distributions with parameters that depend on the previously selected units (see Vitter (1984, 1985, 1987)).

Proposition 3.5.2 asserts that SRS is a sampling design with exchangeable circular spacings, where the spacings follow a shifted multivariate negative hypergeometric distribution. The (singular) multivariate negative hypergeometric distribution (see for example Johnson et al., 1997, pp. 171-199) of size $n \geq 1$, with parameters $m \in \mathbb{N}$ and $\mathbf{r} = (r_1, \ldots, r_n)$, $r_i > 0$, $i = 1, \ldots, n$ is a probability distribution on integer vectors $(x_1, \ldots, x_n)$ that sum to m. It is denoted here by $\mathcal{MNH}(m, \mathbf{r})$, and has a PMF given by:

$$f_{\mathcal{MNH}(m,\mathbf{r})}(x_1, \ldots, x_n) = \frac{m!\Gamma(R)}{\Gamma(m+R)} \prod_{i=1}^{n} \frac{\Gamma(r_i + x_i)}{\Gamma(r_i)x_i!}, \tag{3.18}$$

where $R = \sum_{i=1}^{n} r_i$.

Proposition 3.5.2. *SRS is the sampling design of (3.12), where J_0 has a uniform distribution on U, is independent of the J_i's, $i \geq 1$, and the integer random vector $\mathbf{J} = (J_1, \ldots, J_n)$ follows a shifted multivariate negative hypergeometric distribution: $\mathbf{J} - \mathbf{1}_n \sim \mathcal{MNH}(N - n, \mathbf{1}_n)$, where $\mathbf{1}_n$ is the $n-$vector of ones.*

Proof. With parameter $N - n$ and $\mathbf{1}_n$, the PMF given in (3.18) reduces to

$$f_{\mathcal{MNH}(N-n,\mathbf{1}_n)}(x_1, \ldots, x_n) = \binom{N-1}{n-1}^{-1},$$

where $x_1, \ldots, x_n$ are non-negative integers that sum to $N - n$. Hence, $\mathbf{J}$ has a uniform distribution on the vectors of positive integer numbers that sum to N,

$$\Pr[\mathbf{J} = (j_1, \ldots, j_n)] = \binom{N-1}{n-1}^{-1},$$

for all positive integers $(j_1, \ldots, j_n)$ that sum to N. Moreover, this PMF is symmetric in its arguments and the J_i's are exchangeable. Applying (3.17), we get that

$$\Pr(S = \{x_1, \ldots, x_n\}) = \frac{n}{N}\binom{N-1}{n-1}^{-1} = \binom{N}{n}^{-1},$$

for all $x_1 < \cdots < x_n$. □

The marginal distributions of the circular spacings are shifted negative hypergeometric distributions. Indeed, the marginal distributions of a $\mathcal{MNH}(m, \mathbf{r})$-distributed vector are negative hypergeometric distributions (see for example Janardan and Patil, 1972) with respective parameters m, r_i, $R = \sum_j r_j$, and PMF

$$f_{\mathcal{NH}(m,r_i,R)}(x) = \frac{m!\Gamma(R)}{\Gamma(m+R)} \frac{\Gamma(r_i + x)}{\Gamma(r_i)x!} \frac{\Gamma(R - r_i + m - x)}{\Gamma(R - r_i)(m - x)!}, \; x \leq m, \; x \in \mathbb{N}.$$

Their expectation and variance are, respectively, mr_i/R and $m(r_i/R)(1-r_i/R)(R+m)/(R+1)$. It follows that $J_k - 1$ has a negative hypergeometric distribution with parameters $N - n$, 1, and n, $k = 1, \ldots, n$. In particular we have that

$$\mathrm{E}(J_k) = \frac{N-n}{n} + 1 = \frac{N}{n},$$

$$\mathrm{var}(J_k) = \frac{N-n}{n}\left(1-\frac{1}{n}\right)\frac{N}{n+1}.$$

The second order inclusion probabilities can be derived from (3.16). Indeed, the sum of components of a multivariate negative hypergeometric distribution follows a negative hypergeometric distribution (see Janardan and Patil, 1972). Its parameters are derived from the parameters m and $\mathbf{r}$ by summing the r_i's that correspond to the components that are in the sum. Hence we have that

$$\sum_{i=1}^{j} J_i = j + K_j,$$

where K_j follows a negative hypergeometric distribution with parameters $N-n$, j and n. We can deduce that

$$\begin{aligned} f_j(\ell-k) &= \frac{(N-n)!\Gamma(n)}{\Gamma(N)}\frac{\Gamma(j+\ell-k-j)}{\Gamma(j)(\ell-k-j)!}\frac{\Gamma(n-j+N-n-\ell+j+k)}{\Gamma(n-j)(N-n-\ell+j+k)!}, \\ &= \frac{(N-n)!(n-1)!(\ell-k-1)!(N-\ell+k-1)!}{(N-1)!(j-1)!(\ell-k-j)!(n-j-1)!(N-n-\ell+k+j)!}, \\ &= \frac{n-1}{N-1}\binom{N-2}{\ell-k-1}^{-1}\binom{n-2}{j-1}\binom{N-n}{\ell-k-j}, \end{aligned}$$

and that

$$\pi_{k\ell} = \frac{n(n-1)}{N(N-1)}\sum_{j=1}^{\ell-k}\binom{N-2}{\ell-k-1}^{-1}\binom{n-2}{j-1}\binom{N-n}{\ell-k-j},\ k<\ell,$$

via (3.16). However, if we rename $u=j-1$, $v=\ell-k-1$, $t=n-2$ and $s=N-2$, this last sum is

$$\binom{s}{v}^{-1}\sum_{u=0}^{v}\binom{t}{u}\binom{s-t}{v-u},$$

and Vandermonde's identity ensures that it is equal to 1. Hence we find the well known result:

$$\pi_{k\ell} = \frac{n(n-1)}{N(N-1)},\ k\neq\ell.$$

3.5.5 Systematic sampling

If $N = rn$ with $r \in \mathbb{N}$, the systematic sampling design presented in Section 3.3.6 is a fixed size sampling design with exchangeable circular spacings. It is trivially obtained by taking J_0 uniform on U and $J_i = r$, $i = 1,\dots,n$. The joint inclusion probabilities can also easily be derived from (3.16) using that $f_j(\ell-k) = \mathbf{1}_{\{\ell=k+jr\}}$.

3.6 Spreading fixed size sampling designs with exchangeable circular spacings

Similar to what we did in Section 3.4, we introduce in Sections 3.6.1, 3.6.2 and 3.6.3 new sampling designs with spreading properties by choosing different circular spacings distributions.

Following the structure of Section 3.4, we work, in Section 3.6.1, on sampling designs with multivariate negative hypergeometric spacings and a spreading control parameter $r > 0$. When

$0 < r < 1$, there is an attraction between the selected units: if a unit is selected, then its neighbors are more likely to be selected. If $r = 1$, the design is SRS and if $r > 1$, the sampling is better spread than SRS. As a limit case when r is large, we obtain the multinomial circular spacings design of Section 3.6.2. The spacings variance of these sampling designs is bounded from below. Smaller variances and better spreading properties are obtained with multivariate hypergeometric circular spacings in Section 3.6.3, furthering the parallel with binomial spacings of Section 3.4.3.

3.6.1 Multivariate negative hypergeometric circular spacings

The multivariate negative hypergeometric distribution $\mathcal{MNH}(m, \mathbf{r})$ has exchangeable marginals exactly when $r_1 = \cdots = r_n$, $\mathbf{r} = r\mathbf{1}_n$ for some positive real number r. If $\mathbf{J} - \mathbf{1}_n \sim \mathcal{MNH}(N - n, r\mathbf{1}_n)$, the sampling design of Definition 3.5.2 has circular spacings with a variance given by:

$$\operatorname{var}(J_k) = \frac{N-n}{n}\left(1 - \frac{1}{n}\right)\frac{rn + N - n}{rn + 1},\ k = 1, \ldots, n.$$

These variances are decreasing functions of r, with SRS corresponding to $r = 1$.

According to (3.17), the sampling design PMF is given by:

$$P(s) = \frac{n}{N}\frac{\Gamma(nr)}{[\Gamma(r)]^n}\frac{(N-n)!}{\Gamma[N + n(r-1)]}\frac{\Gamma(r + N + x_1 - x_n - 1)}{(N + x_1 - x_n - 1)!}\prod_{i=1}^{n-1}\frac{\Gamma(r + x_{i+1} - x_i - 1)}{(x_{i+1} - x_i - 1)!},$$

where $x_1, \ldots, x_n$ are the ordered indexes of units sampled in s. The second order inclusion probabilities are obtained from (3.16):

$$\pi_{k\ell} = \frac{n}{N}\sum_{j=1}^{\ell-k}\binom{N-n}{\ell-k-j}\frac{\mathrm{B}[\ell - k + j(r-1), N + n(r-1) - \ell + k - j(r-1)]}{\mathrm{B}(jr, nr - jr)},\ k < \ell,$$

where $\mathrm{B}(\cdot, \cdot)$ denotes the beta function, defined by:

$$\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1}dt,$$

if a and b are positive real numbers.

These joint inclusion probabilities are plotted in Figure 3.3 for different values of r, including their limit when $r \to \infty$. On this plot, we see strong oscillations of the joint inclusion probabilities when r is large.

3.6.2 Multinomial circular spacings

Let $\mathcal{MNom}(m, \mathbf{p})$ be the multinomial distribution with parameters $m \in \mathbb{N}$ and $\mathbf{p} = (p_1, \ldots, p_n) \in [0, 1]^n$, $\sum_i p_i = 1$. It is the probability distribution on integer vectors $(x_1, \ldots, x_n)$ such that $\sum_i x_i = m$ with PMF:

$$f_{\mathcal{MNom}(m,\mathbf{p})}(x_1, \ldots, x_n) = \binom{m}{x_1 \cdots x_n}\prod_{i=1}^{n} p_i^{x_i}, \tag{3.19}$$

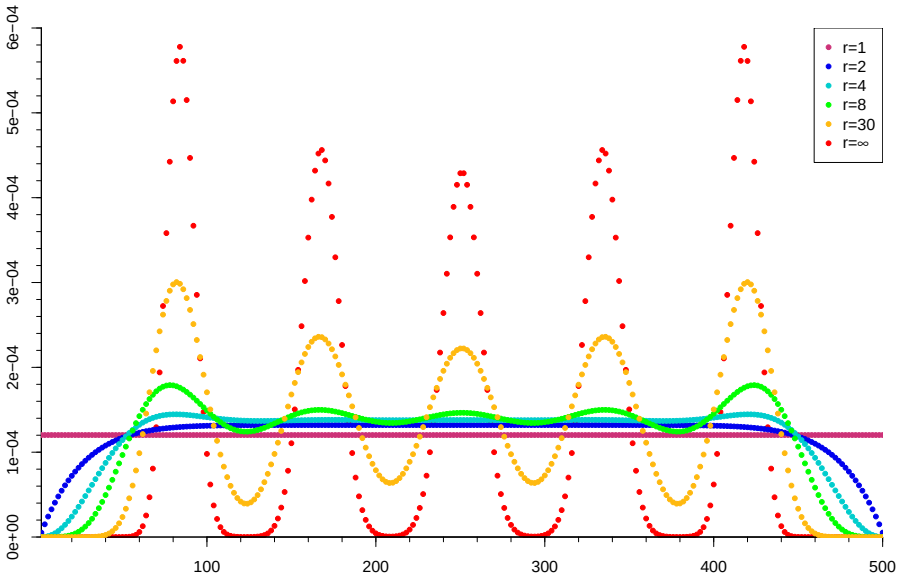

FIGURE 3.3: $\pi_{1,1+i}$ in function of i for a fixed size sampling design with shifted multivariate negative hypergeometric circular spacings, $n = 6$, $N = 500$ and $r = 1,\ 2,\ 4,\ 8,\ 30$ and $r = +\infty$ (multinomial circular spacings). When $r = 1$, we obtain the SRS design and constant joint inclusion probabilities. The larger r is, the more contrasted are the joint inclusion probabilities.

where

$$\binom{m}{x_1 \cdots x_n} = \frac{m!}{x_1! \cdots x_n!}.$$

Its marginal distributions are binomial with respective parameters m and p_i. They are exchangeable exactly when $\mathbf{p} = n^{-1}\mathbf{1}_n$

When r tends to infinity, the multivariate negative hypergeometric distribution with parameters m and $r\mathbf{1}_n$ tends to a multinomial distribution with parameters m and $\mathbf{1}/n$ (see, for instance, Terrell, 1999, p. 182). Thus the fixed size sampling design with shifted multinomial exchangeable circular spacings is the limit case of multivariate negative hypergeometric spacings of Section 3.6.1 when r tends to infinity.

Spacings J_k follow a shifted binomial distribution with parameters $N - n$ and $1/n$. The corresponding sampling design PMF is given by

$$P(s) = \frac{n}{N}\frac{1}{n^{N-n}}\frac{(N-n)!}{(N + x_1 - x_n - 1)!\prod_{j=2}^{n}(x_j - x_{j-1} - 1)!},$$

where $x_1, \ldots, x_n$ are the ordered indexes of units sampled in s. The sum of any j components of a $\mathcal{MN}om(m, \mathbf{p})$ multinomial vector follows a binomial distribution with parameters m and p where p is the sum of corresponding p_i's. Hence we have here that

$$f_i(\ell - k) = \binom{N-n}{\ell - k - j}\left(\frac{j}{n}\right)^{\ell-k-j}\left(1 - \frac{j}{n}\right)^{N-n-\ell+k+j},$$

$$\pi_{k\ell} = \frac{n}{N}\sum_{j=1}^{\ell-k}\binom{N-n}{\ell-k-j}\left(\frac{j}{n}\right)^{\ell-k-j}\left(1-\frac{j}{n}\right)^{N-n-\ell+k+j}, \ k<\ell.$$

3.6.3 Multivariate hypergeometric circular spacings

Variances of circular spacings in Sections 3.6.1 and 3.6.2 are bounded from below. In order to have smaller variances, one can use shifted multivariate hypergeometric spacings.

Let $\mathcal{MH}(m,\mathbf{r})$ be the (singular) multivariate hypergeometric distribution with parameters $m\in\mathbb{N}$ and $\mathbf{r}=(r_1,\ldots,r_n)$, $\mathbf{r}$ is an integer vector that sums to R and $m\leq R$. It is a probability distribution on integer vectors $(x_1,\ldots,x_n)$ that sum to m, $x_i\leq r_i$ and has a PMF given by:

$$f_{\mathcal{MH}(m,\mathbf{r})}(x_1,\ldots,x_n) = \binom{R}{m}^{-1}\prod_{i=1}^{n}\binom{r_i}{x_i}. \tag{3.20}$$

The marginal distributions of $\mathcal{MH}(m,\mathbf{r})$ are hypergeometric variables with respective parameters m, r_i and R, and their variance is $m(r_i/R)(1-r_i/R)(R-m)/(R-1)$.

The multivariate hypergeometric distribution has exchangeable marginals exactly when $\mathbf{r}=r\mathbf{1}_n$ for some integer r larger than m/n. Here, we consider the fixed size sampling design with circular spacings $\mathbf{J}$ such that $\mathbf{J}-\mathbf{1}_n\sim\mathcal{MH}(N-n,r\mathbf{1}_n)$ with $r\geq N/n-1$. We get that the spacing variances are given by

$$\operatorname{var}(J_k) = \frac{N-n}{n}\left(1-\frac{1}{n}\right)\frac{rn-N+n}{rn-1}.$$

The parameter r can be used to tune the variance. If $r=(N-n)/n$ is integer, we have $\operatorname{var}(J_k)=0$ and we obtain the systematic sampling design.

The sampling design PMF is obtained via (3.17):

$$P(s) = \frac{n}{N}\binom{rn}{N-n}^{-1}\binom{r}{x_{i+1}-x_i-1}\prod_{i=1}^{n-1}\binom{r}{N+x_1-x_n-1},$$

where $x_1,\ldots,x_n$ are the ordered indexes of units sampled in s. The sum of components of a $\mathcal{MH}(m,\mathbf{r})$ vector follows a hypergeometric distribution. In the present case we find that

$$f_j(x) = \binom{rn}{N-n}^{-1}\binom{jr}{x-j}\binom{rn-jr}{N-n-x+j}, \ j\leq x\leq N-n,$$

and that the joint inclusion probabilities are

$$\pi_{k\ell} = \frac{n}{N}\binom{rn}{N-n}^{-1}\sum_{j=1}^{\ell-k}\binom{jr}{\ell-k-j}\binom{rn-jr}{N-n-\ell+k+j}, \ k<\ell.$$

Note that some joint inclusion probabilities may be null, even when $r>(N-n)/n$ and the design is not the systematic sampling design. For example, if $N=10$, $n=2$, $r=5$ and $\ell=k+1$, then

$$\pi_{k\ell} = \frac{1}{5}\binom{10}{8}^{-1}\binom{5}{0}\binom{5}{8}=0.$$

3.6.4 Summary

The different fixed size sampling designs with exchangeable circular spacings that we considered are listed in Table 3.2 with the variance of their spacings. Other exchangeable circular

TABLE 3.2: Fixed size sampling designs and variance of their spacings.

Distribution of $\mathbf{J}-\mathbf{1}_n$	$\mathrm{var}(J_i),\ i=1,\ldots,n$
Multivariate negative hypergeometric, $r>0$	$\frac{N-n}{n}\left(1-\frac{1}{n}\right)\frac{rn+N-n}{rn+1}$
Multinomial	$\frac{N-n}{n}\left(1-\frac{1}{n}\right)$
Multivariate hypergeometric, $r\geq N/n-1$	$\frac{N-n}{n}\left(1-\frac{1}{n}\right)\frac{rn-N+n}{rn-1}$
Systematic or hypergeometric with $r=N/n-1$	0

spacings may be used, and are easily obtained as distributions of vectors of i.i.d. random variables conditioned on the sum of the vectors components. The families of distribution of Section 3.6 encompass the SRS and fixed-size systematic designs. They allow one to use designs with low spacings variance, but it is not always possible to avoid having null joint inclusion probabilities with shifted multivariate hypergeometric spacings. Finally, the designs are not strictly sequential as the population list may need to be run over twice in order to finish selecting a sample.

3.7 Simulations

A single artificial population of $N=200$ units was generated with an interest variable Y that had a trend and are autocorrelated: $y_k=k+z_k$, where $z_k=0.6z_{k-1}+\varepsilon_k$ and $\epsilon_k\sim N(0,\sigma_\varepsilon=0.3)$. With this kind of autocorrelation, having well spread samples ought to be an efficient strategy. The "spacing" $N+x_1-x_n$ between the last sampled unit and the first one is treated like any other spacing, so that ideally one would also want to have some similarity between units at the beginning of the population list and those at the end. This feature can easily be obtained in a setting of continuous population sampling (see Wilhelm et al., 2017), but is not common in finite population applications.

For each situation, a set of 100,000 samples was generated. All samples were of fixed size $n=50$ and were selected using the following sampling designs:

- Multivariate negative hypergeometric (MNH) with $r=0.5$, $r=1$ (SRS), $r=5$, $r=10$, $r=50$,
- Multinomial (MULT),
- Multivariate hypergeometric (MH) with $r=50$, $r=10$, $r=6$, $r=4$.

We used different values for the tunings parameter r in all kinds of sampling design in order to show the effect of this tuning parameter.

For each sample an estimate $\widehat{\overline{Y}}$ of the mean and of the variance $\widehat{\mathrm{var}}_{SYG}\left(\widehat{\overline{Y}}\right)$ (with the Sen-Yates-Grundy formula) were produced. Compiling our simulation results, we computed the following values, presented in Table 3.3:

- the Bias Ratio

$$\mathrm{BR} = 100 \frac{\mathrm{E}_{sim}\left(\widehat{\overline{Y}} - \overline{Y}\right)}{\left[\mathrm{var}_{sim}\left(\widehat{\overline{Y}}\right)\right]^{\frac{1}{2}}},$$

  where $\mathrm{E}_{sim}(\cdot)$ and $\mathrm{var}_{sim}(\cdot)$ denote the empirical means and variances of the simulation results.

- the standard error

$$\mathrm{SE} = \left[\mathrm{var}_{sim}\left(\widehat{\overline{Y}}\right)\right]^{\frac{1}{2}};$$

- the square root of the variance estimator average

$$\mathrm{REVAR} = \left\{\mathrm{E}_{sim}\left[\widehat{\mathrm{var}}_{SYG}\left(\widehat{\overline{Y}}\right)\right]\right\}^{\frac{1}{2}};$$

- the coefficient of variation of the variance estimator

$$\mathrm{CV} = \frac{\left\{\mathrm{var}_{sim}\left[\widehat{\mathrm{var}}_{SYG}\left(\widehat{\overline{Y}}\right)\right]\right\}^{\frac{1}{2}}}{\mathrm{var}_{sim}\left(\widehat{\overline{Y}}\right)};$$

- and the coverage rate of the 95% confidence interval.

The simulation results in Table 3.3 confirm, with column BR, that the estimator of the mean is unbiased. The accuracy of the mean estimator improves as the circular spacings variance decreases, from Design MNH $r = 0.5$ to Design NH $r = 4$.

The conclusions are different for the variance estimator. For all situations in our simulations, the joint inclusion probabilities are positive. The variance estimator is unbiased, and this is confirmed by the fact that columns SE and REVAR are mostly equal. However, when the variance of the spacings are close to 0, the variance estimator is unstable. Indeed, with these parameters, some joint inclusion probabilities are very small (less than $1/1000$) compared to others (on average 0.0625). In column CV that the accuracy of the variance estimator improves at first as the circular spacings variance decreases, from Design MNH $r = 0.5$ to Design MNH $r = 5$, and then the coefficient of variation goes up again from Design MNH $r = 0.5$ to Design MH $r = 4$. The coverage rate deviates strongly from its nominal value of 95% in the last couple of designs. Thus the design that performs best for the point estimation of the mean does not allow to properly estimate the precision, and even gives seriously misleading confidence intervals. The same kind of problem arises when a stratified sampling design is used with too many strata.

An arbitration needs to be made between the accuracy of the point estimator and that of its variance estimator. In our simulations, a reasonable solution consists in choosing the sampling design with shifted multinomial distribution (MULT). This method is simple to implement, more so than the MNH or MH. It allows for accurate point estimation while presenting a correct coverage rate of its confidence intervals.

TABLE 3.3: Results of the 100,000 simulations. The designs are ordered in decreasing order of the variance of the spacings.

	BR	SE	REVAR	CV	coverage
MNH $r = 0.5$	-0.25	0.46	0.45	0.48	93.97
SRS	-0.12	0.35	0.35	0.23	94.52
MNH $r = 5$	0.08	0.23	0.23	0.21	94.39
MNH $r = 10$	-0.22	0.21	0.21	0.26	94.08
MNH $r = 50$	0.13	0.19	0.19	0.33	93.90
MULT	0.36	0.19	0.19	0.35	93.64
MH $r = 50$	-0.17	0.18	0.18	0.37	93.58
MH $r = 10$	-0.35	0.16	0.16	0.52	92.05
MH $r = 6$	-0.74	0.14	0.14	0.72	83.97
MH $r = 4$	-0.52	0.11	0.15	1.60	40.55

3.8 Conclusions

In Sections 3.3 and 3.5, we proposed general methods to generate uniform inclusion probabilities sampling designs with i.i.d. or exchangeable spacings. We used them in Sections 3.4 and 3.6 to obtain sample selection methods with controlled spreading properties and gave, in Section 3.7, an example where such methods are useful. If the response variable is similar among units that are close in the population list, the choice of the spreading parameter allows one to make a trade-off between precision of the point estimator and precision of variance estimator.

Some of the designs that we consider have concentrated spacings, but, unlike systematic sampling, they retain positive joint inclusion probabilities and thus allow for an unbiased estimation of variance. These joint inclusion probabilities have computable closed-form expressions and depend only on the "distance" between units in the population list, thus at most $N - 1$ joint inclusion probabilities need to be computed. However, the ranks of sampled units in the population must be known in order to compute a variance estimator.

We do not have a clear solution to extending these results in all generality to unequal first order inclusion probabilities sampling designs. One partial solution is to work on the distribution of J_0. The choice of a different distribution for J_0 allows one to have a limited control on the inclusion probabilities via Equations (3.6) and (3.13), while leaving the spacings untouched. Another solution is the thinning approach. It consists in selecting a large enough first phase sample with a spreading design and uniform inclusion probabilities and selecting a second phase sub-sample with appropriate inclusion probabilities. However, this does not preserve the spreading properties.

Acknowledgments

This work was supported in part by the Swiss Federal Statistical Office. The views expressed in this paper are solely those of the authors. M. W. was partially supported by a Doc.Mobility fellowship of the Swiss National Science Foundation (grant no. P1NEP2_162031). The authors would like to thank three referees and an associate editor for their constructive comments that helped us improve this paper, and Prof. Lennart Bondesson who has kindly sent us copies of his work on a similar topic.

Chapter 4

Quasi-Systematic Sampling From a Continuous Population

Abstract

A specific family of point processes are introduced that allow to select samples for the purpose of estimating the mean or the integral of a function of a real variable. These processes, called quasi-systematic processes, depend on a tuning parameter $r > 0$ that permits to control the likeliness of jointly selecting neighbor units in a same sample. When r is large, units that are close tend to not be selected together and samples are well spread. When r tends to infinity, the sampling design is close to systematic sampling. For all $r > 0$, the first and second-order unit inclusion densities are positive, allowing for unbiased estimators of variance.
Algorithms to generate these sampling processes for any positive real value of r are presented. When r is large, the estimator of variance is unstable. It follows that r must be chosen by the practitioner as a trade-off between an accurate estimation of the target parameter and an accurate estimation of the variance of the parameter estimator. The method's advantages are illustrated with a set of simulations. [a]

[a] This chapter is essentially a reprint of Wilhelm et al. (2017). However, some simulations have been re-run in order to make them reproducible. Hence, some Figures and Tables are not exactly the same as in Wilhelm et al. (2017).

4.1 Introduction

We propose to use a specific family of point processes to select samples for the purpose of estimating the mean or the integral of a function of a real variable. We draw a parallel with sampling designs which are themselves point processes on finite spaces. Systematic sampling is widely used in finite population. It has been introduced by Madow and Madow (1944) and Madow (1949). It is easily implemented and, by spreading the sample over the population, it results in precise mean and total estimators when the variable of interest is similar for neighboring units. The main drawback of systematic sampling is that most of the unit joint inclusion probabilities are null, making it impossible to estimate the variance of the Horvitz-Thompson estimator without bias (see Horvitz and Thompson, 1952).

The aim of this paper is to develop a method that is a compromise between a base point process such as the Poisson process or the binomial process and the systematic process for sample selections in a continuous population. A similar objective is pursued in Breidt (1995) in a finite population setting supported by a superpopulation model. Breidt (1995) considers one-per-stratum sampling designs from a population that is split into strata of a successive

units where a divides the population size. He introduces a class of sampling procedures that encompasses systematic sampling with constant rate $1/a$ and simple random sampling of one unit per stratum.

Point processes, that we refer to as *sampling processes* in the context of sampling, are the subject of a vast literature (see for example Daley and Vere-Jones, 2002, 2008, and references therein). Cordy (1993) and Deville (1989) introduced independently the continuous analogue to the Horvitz-Thompson estimator for infinite population sampling. Different communities have studied point processes: mathematical physicists, probabilists and statisticians. A detailed state of the art in the study and simulation of some complex point processes can be found in Møller and Waagepetersen (2004, 2007). Many simulation methods for point processes are implemented in the **R** package `spatstat` (Baddeley and Turner, 2005b; Baddeley et al., 2015b).

We introduce a new family of sampling methods that enable to continuously tune the distance between units in the sample. These processes allow to obtain small probabilities of jointly selecting neighboring units. These sampling methods are particularly efficient when the function of interest is smooth. Moreover, joint inclusion densities are positive and it is possible to estimate the sampling variance without bias.

The paper is organized as follows: in Section 4.2, we give a definition of sampling processes in continuous populations and we define the Poisson process, the binomial process and the systematic process. Important results of renewal process theory are recalled in Section 4.3. In Section 4.4, we define the systematic-Poisson and the systematic-binomial processes with tuning parameter r, and compute the joint densities. Section 4.5 contains proofs for the asymptotic processes when r tends to infinity. Simulations are presented in Section 4.6 and our ideas on the choice of the tuning parameter in Section 4.7. Finally, we give a brief discussion of the method and its advantages in Section 4.8.

4.2 Sampling from a continuous population

Following Macchi (1975) (see also Moyal, 1962), a finite sample of size n from a bounded and open subset Ω of $\mathbb{R}$ is a collection of units $X = \{x_1, \dots, x_n\}$ without consideration for the order of the x_i's. This definition matches those commonly used in finite population sampling (see for example Cochran, 1977, for an introduction to finite population sampling theory). A sampling process is a probability distribution on the space $\mathcal{S}$ of all such collections, for all $n \in \mathbb{N}$. Note that it is not directly a distribution on $\Omega^{\mathbb{N}}$ equipped with the tensor product of Borel sigma algebras $\mathcal{B}(\Omega)$ as the sample units are not ordered. An extensive discussion on the definition of a sampling point process on Ω and the corresponding symmetric measure on $(\Omega^{\mathbb{N}}, \mathcal{B}^{\otimes\mathbb{N}}(\Omega))$ is given in Macchi (1975). It is sufficient for our purpose to know that a sampling point process is a probability distribution on $(\mathcal{S}, \mathcal{B})$ where $\mathcal{S} = \bigcup_{n\in\mathbb{N}} \Omega^n/\mathcal{R}^n$, with x and y in Ω^n being in the same class for the equivalence relation $\mathcal{R}^n$ if x is a permutation of elements of y, and $\mathcal{B}$ is the sigma algebra generated by the family of counting events:

$$\{s \in \mathcal{S} \text{ such that } N(s, A) = p,\ A \in \mathcal{B}(\Omega),\ p \in \mathbb{N}\},$$

and $N(s, A)$ is the number of elements of s that are in A.

The first and second factorial moment measures of a sampling point process X (Moyal, 1962) are defined respectively as

$$M_1 = \begin{pmatrix} \mathcal{B}(\Omega) & \to & \mathbb{R}_+ \\ A & \mapsto & \mathrm{E}\left[N(X,A)\right] \end{pmatrix},$$

where $N(X,A)$ is the random number of elements of X that are in A, and the second factorial moment measure is the extension to $\mathcal{B}(\Omega)^{\otimes 2}$ of

$$M_2 = \begin{pmatrix} \mathcal{B}(\Omega)\times\mathcal{B}(\Omega) & \to & \mathbb{R}_+ \\ A\times B & \mapsto & \mathrm{E}\left[N_2(X,A\times B)\right] \end{pmatrix},$$

where $N_2(X, A\times B)$ is the random number of pairs (x_i, x_j), $i \neq j$ of elements of X such that $x_i \in A$ and $x_j \in B$.

We call first and joint (second) order inclusion densities the respective densities of M_1 and M_2 with respect to the Lebesgue measure on Ω and Ω^2 when they exist. In that case, the first order inclusion density π is such that $M_1(A) = \int_A \rho(x)dx$, for all $A \in \mathcal{B}(\Omega)$, and the second-order inclusion density $\rho^{(2)}$ satisfies $M_2(A\times B) = \int_A\int_B \rho^{(2)}(x,y)dxdy$ for all $A\times B \in \mathcal{B}(\Omega)\times\mathcal{B}(\Omega)$. Heuristically, the term $\rho(x)dx$ can be viewed as the probability that one unit of the sample lies between x and $x+dx$, and $\pi^{(2)}(x,y)dxdy$ as the probability that one unit of the sample lies between x and $x+dx$ and another between y and $y+dy$, disregarding what happens outside of these sets. Likewise, one can define $k-$th order factorial moments and, when they exist, inclusion densities for $k \geq 3$.

We now turn to the problem of estimating the mean of a Lebesgue integrable function z defined on Ω:

$$\overline{z} = \frac{1}{|\Omega|}\int_\Omega z(x)dx,$$

where $|\Omega|$ denotes the Lebesgue measure of Ω, using a finite random sample $X = \{x_1,\ldots,x_n\}$ of points in Ω. Assuming that Ω is bounded, $|\Omega|$ is known and X is a sampling process with inclusion density ρ, Cordy (1993) defines the continuous analogue of the Horvitz-Thompson estimator as:

$$\widehat{\overline{z}} = \frac{1}{|\Omega|}\sum_{i=1}^{n}\frac{z(x_i)}{\rho(x_i)},$$

and gives its properties. Under the assumption that $\rho(x) > 0$ on Ω and that z is bounded or non-negative, this estimator is unbiased (Cordy, 1993, Theorem 1). If, moreover, $\int_\Omega 1/\rho(x)dx < +\infty$, the variance of this estimator is given by:

$$\mathrm{var}\left(\widehat{\overline{z}}\right) = \frac{1}{|\Omega|^2}\left[\int_\Omega \frac{[z(x)]^2}{\rho(x)}dx + \int_\Omega\int_\Omega z(x)z(y)\left[\frac{\rho^{(2)}(x,y)-\rho(x)\rho(y)}{\rho(x)\rho(y)}\right]dxdy\right],$$

and if the joint inclusion density exists with $\rho^{(2)}(x,y) > 0$ for all x, y in Ω then:

$$\widehat{\mathrm{var}}\left(\widehat{\overline{z}}\right) = \frac{1}{|\Omega|^2}\left[\sum_{x_i\in X}\left[\frac{z(x_i)}{\rho(x_i)}\right]^2 + \sum_{x_i\in X}\sum_{\substack{x_j\in X\\ i\neq j}} z(x_i)z(x_j)\left[\frac{\rho^{(2)}(x_i,x_j)-\rho(x_i)\rho(x_j)}{\rho(x_i)\rho(x_j)\rho^{(2)}(x_i,x_j)}\right]\right], \tag{4.1}$$

is an unbiased estimator of the variance of $\widehat{\overline{z}}$ (Cordy, 1993, Theorem 2). As pointed out in Cordy (1993) the Horvitz-Thompson variance and variance estimator for a continuous population are slightly different from the finite population case. Conditions to ensure that these estimators are unbiased are, however, similar.

In the case of fixed size sampling process, the continuous analogue of the Sen-Yates-Grundy (Sen, 1953; Yates and Grundy, 1953) variance formula and estimator are:

$$\operatorname{var}\left(\widehat{\overline{z}}\right) = \frac{1}{2|\Omega|^2}\int_\Omega\int_\Omega\left[\frac{z(x)}{\rho(x)} - \frac{z(y)}{\rho(y)}\right]^2 [\rho(x)\rho(y) - \rho^{(2)}(x,y)]dxdy, \tag{4.2}$$

and

$$\widehat{\operatorname{var}}\left(\widehat{\overline{z}}\right) = \frac{1}{2|\Omega|^2}\sum_{x_i\in X}\sum_{\substack{x_j\in X\\ i\neq j}}\left[\frac{z(x_i)}{\rho(x_i)} - \frac{z(x_j)}{\rho(x_j)}\right]^2\left[\frac{\rho(x_i)\rho(x_j) - \rho^{(2)}(x_i,x_j)}{\rho^{(2)}(x_i,x_j)}\right], \tag{4.3}$$

(see Cordy, 1993, pp. 358-359).

Throughout this paper, we assume that $\Omega \subset \mathbb{R}$ but the construction we used up to here also allows to work with other spaces. Indeed, Macchi (1975) and Cordy (1993) consider finite dimensional real vector spaces, and Daley and Vere-Jones (2002) work on complete separable metric spaces (polish spaces). Our purpose is to define sampling processes that have good properties regarding the estimation of $\overline{z}$.

In the following, we assume that $\Omega = (0,1)$. For an ordered set $\{x_1,\ldots,x_n\}$, we define the corresponding *spacings* [1] $\{j_1,\ldots,j_{n-1}\}$ as the differences between two successive units, namely $j_i = x_{i+1} - x_i$, for $i = 1,\ldots,n-1$. If X is a point process, the corresponding spacings (also called waiting times) are random variables. A special class of point processes, called renewal processes, are obtained when the spacings are independent and identically distributed (see for example Mitov and Omey, 2014). In this paper, except when explicitly stated, the random spacings of our sampling processes are neither assumed to be identically distributed nor independent.

The binomial process (see Møller and Waagepetersen, 2004, pp. 23-28) is one of the most basic point processes and has a fixed sample size.

Definition 4.2.1 (Binomial process)**.** *Let f be a PDF on $\Omega = (0,1)$ and let $n \in \mathbb{N}$ be a natural number. The* binomial point process *of n points in Ω with PDF f is the point process whose realizations consist of n points generated from i.i.d distributions with common PDF f.*

When the sample space Ω is bounded, the spacings of the binomial process are not independent. Indeed, the sum of these spacings is necessarily no larger than the diameter of Ω. In the following, we only use binomial processes in $(0,1)$ with i.i.d. points selected according to a uniform distribution on $(0,1)$.

The $k-$th order joint inclusion density of a binomial process of size n at $x_1 < \cdots < x_k$ is given by:

$$\rho^{(k)}(x_1,\ldots,x_k) = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!},\ k = 1,\ldots,n.$$

In particular, $\rho(x_i) = n$ if $x_i \in (0,1)$, and $\rho^{(2)}(x_i,x_j) = n(n-1)$ if $x_i,\ x_j \in (0,1)$. The $n-$th order joint inclusion density is equal to $n!$ on samples $x_1,\ldots,x_n$ with $0 < x_1 < x_2 < \cdots < x_n < 1$.

With $\Omega = (0,1)$ and a fixed size n, we can define the circular spacings as $J_i = (x_{i+1} - x_i) \mod 1$, $i = 1,\ldots,n-1$ and $J_n = (x_1 - x_n) \mod 1$. As we see in Proposition 4.2.1, the binomial

[1] In Wilhelm et al. (2017), we used the term *inter-arrival time* instead of *spacing*.

process can be obtained by generating the circular spacings according a Dirichlet distribution. The Dirichlet distribution with parameter $\boldsymbol{\alpha}$, denoted $\mathcal{Dir}(\boldsymbol{\alpha})$ is a multivariate distribution with PDF given by

$$f(x_1, \ldots, x_n) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{n} x_i^{\alpha_i - 1}, \tag{4.4}$$

where $x_i > 0$, for $i, 1, \ldots, n \sum_{i=1}^{n} x_i = 1$, $\alpha_i > 0$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ and $\mathrm{B}(\boldsymbol{\alpha})$ is the multinomial Beta function. Properties of the Dirichlet distribution are given in (Kotz et al., 2000, pp. 485-528).

Proposition 4.2.1. *Let $\mathbf{J}^c = (J_1^c, \ldots, J_n^c) \sim \mathcal{Dir}(\mathbf{1}_n)$, where $\mathbf{1}_n$ is a vector of n ones and $u \sim \mathcal{U}(0,1)$, uniformly distributed on $(0,1)$, is independent from $\mathbf{J}^c$. The sorted values in*

$$\left\{ \left(\sum_{j=1}^{i} J_1^c + u \right) \mod 1,\ i = 1, \ldots, n \right\}, \tag{4.5}$$

follow a binomial process on $(0,1)$ with uniform density.

Proof. With parameter $\mathbf{1}_n$, the PDF in (4.4) simplifies to

$$f_{\mathbf{J}^c}(j_1, \ldots, j_n) = (n-1)!,$$

with $\sum_{i=1}^{n} j_i = 1$. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be the sorted values (4.5). Since the sum of all j_i's is equal to 1, we see that a given set of numbers $x_1 < \cdots < x_n$ in $(0,1)$ is obtained exactly when $u = x_i$ for some i and the spacings allow to obtain $(x_1, \ldots, x_n)$. These events are almost surely non overlapping and u is independent from $\mathbf{J}^c$. It follows, if f_u is the density of u and $f_{\mathbf{J}^c}$ the density of $\mathbf{J}^c$, that

$$\begin{aligned}
& f_{\mathbf{X}}(x_1, \ldots, x_n) \\
&= f_u(x_1) f_{\mathbf{J}^c}(x_2 - x_1, \ldots, x_n - x_{n-1}, x_1 - x_n + 1) \\
&+ f_u(x_2) f_{\mathbf{J}^c}(x_3 - x_2, \ldots, x_1 - x_n + 1, x_2 - x_1) \\
&\vdots \\
&+ f_u(x_n) f_{\mathbf{J}^c}(x_1 - x_n + 1, \ldots, x_{n-1} - x_{n-2} + 1, x_n - x_{n-1}) \\
&= n(n-1)! = n!.
\end{aligned}$$

□

The Poisson process (see for example Daley and Vere-Jones, 2002) is one of the basic and most studied point processes. It is particularly useful for the construction of more complex processes.

Definition 4.2.2 (Poisson process)**.** *A point process X on Ω is a Poisson process with intensity $\lambda > 0$ if the following properties are satisfied:*

1. *For any $A \in \mathcal{B}(\Omega)$, $N(X, A)$ follows a Poisson distribution with parameter $\lambda|A|$, where $|A|$ denotes the Lebesgue measure of A. If $|A| = 0$, then $N(X, A) = 0$ almost surely.*
2. *For any $n \in \mathbb{N}$, conditional on $N(X, A) = n$, the distribution of X_A (the trace of the random set X on A) is that of a binomial process on A with size n and constant PDF on A.*

There exist several equivalent definitions of the Poisson process, but this one highlights the link with the binomial process. There is a similar link in finite population sampling, where conditioning a Bernoulli sampling design on its size yields a simple random sampling design (see Tillé, 2006, pp. 43-50). Bernoulli sampling can thus be considered as the discrete analogue to the Poisson sampling process. Spacings of the Poisson process with intensity λ are i.i.d. and follow an exponential distribution with parameter λ (Daley and Vere-Jones, 2002).

It follows from the definition that the first order inclusion density of the Poisson sampling process on $\Omega = (0, 1)$ is equal to λ, and using the independence property, that the $k-$th order joint inclusion density is equal to $\rho^{(k)}(x_1, \ldots, x_k) = \lambda^k$ if $x_1 < \cdots < x_k$.

The systematic process, or deterministic renewal process in the interval $(0, 1)$ is defined as follows:

Definition 4.2.3 (Systematic process). *Let $0 < c < 1$ and $u \sim \mathcal{U}(0, c)$. A systematic sampling process with sampling interval $0 < c < 1$ is defined as the distribution of $\{x_1, \ldots, x_n\}$ where*

$$x_k = u + k \cdot c, \quad k = 0, \ldots, n - 1,$$

and n is such that $u + n \cdot c < 1 \leq u + (n + 1) \cdot c$.

4.3 Renewal processes

A renewal process, or renewal sequence, is a stochastic process defined on the positive real line. It is completely characterized by the distribution of its independent and identically distributed spacings. For example, the Poisson process is a renewal process with exponentially distributed spacings when its intensity λ is constant. The following definition can be found in Mitov and Omey (2014).

Definition 4.3.1 (Renewal process). *A renewal process is any process $X = \{X_k,\ k = 0,\ 1,\ 2, \ldots\}$ with*

$$X_k = X_0 + \sum_{i=1}^{k} J_i, \quad k = \ 1,\ 2, \ldots$$

where X_0 is a given non-negative random variable and $J_1,\ J_2, \ldots$ is a sequence of i.i.d non-negative random variables with common Cumulative Distribution Function (CDF) F. If $X_0 = 0$ a.s., the process is called a pure renewal process (or simply a renewal process). If $P(X_0 > 0) > 0$ then the process is called a delayed renewal process (see Resnick, 1992).

The counting measure $N(t)$ (or renewal counting process) of a pure renewal process X is defined in Mitov and Omey (2014) as:

$$N(t) = \sup\{k \geq 0 : X_k \leq t\} = \sum_{i=1}^{\infty} \mathbf{1}_{\{X_i \leq t\}},$$

where $\mathbf{1}_{\{X_i \leq t\}}$ denotes the indicator function. Daley and Vere-Jones (2002) then define the *forward recurrence time* of a renewal process as:

$$B(t) = X_{N(t)+1} - t, \quad t \geq 0.$$

It is the random time between an arbitrarily chosen instant t and the following occurrence of the process (see Figure 4.1).

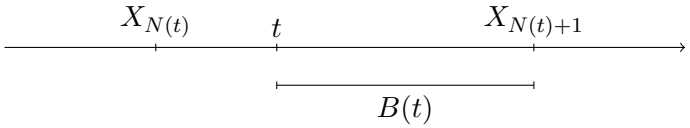

FIGURE 4.1: forward recurrence time $B(t)$

An important result of renewal theory concerns the limiting distribution of the forward recurrence time $B(t)$ when $t \to \infty$. Under some mild conditions (see Mitov and Omey, 2014, Theorem 1.18), if the spacings have CDF F and finite expectation $\mu > 0$, $B(t)$ converges in distribution when $t \to \infty$ to a random variable with CDF F_0 defined as:

$$F_0(x) = \lim_{t\to\infty} P(B(t) \leq x) = \frac{1}{\mu}\int_0^x [1 - F(t)]\, dt,\ x \geq 0. \tag{4.6}$$

The PDF of this limiting distribution is equal to:

$$f_0(x) = \frac{1}{\mu}\left[1 - F(x)\right],\ x \geq 0.$$

For example, if the spacings follow a Gamma distribution with shape parameter r and rate parameter λ, denoted $\mathcal{G}amma(r, \lambda)$, their distribution function is given by $F(x) = \gamma(r, \lambda x)/\Gamma(r)$, where $\Gamma(r) = \int_0^{+\infty} t^{r-1}e^{-t}dt$ and $\gamma(r, x) = \int_0^{\lambda x} t^{r-1}e^{-t}dt$. The corresponding limiting forward recurrence time distribution follows a forward Gamma distribution $\mathcal{F}or\mathcal{G}(r, \lambda)$ with PDF

$$f_0(x) = \frac{\lambda\Gamma(r, \lambda x)}{\Gamma(r+1)},\ x \geq 0,$$

with $\Gamma(r, \lambda x) = \int_{\lambda x}^{+\infty} t^{r-1}e^{-t}dt$.

Another property of renewal processes that will be essential in the following is given in Proposition 4.3.1.

Proposition 4.3.1. *Let $(J_i)_{i\geq 1}$ be a sequence of i.i.d non-negative continuous random variables, with expectation $\mathrm{E}(J_i) = \mu$, CDF $F(x)$, and PDF $f(x)$. Let also f_0 be the function defined by:*

$$f_0(x) = \frac{1}{\mu}\left[1 - F(x)\right] \text{ if } x \geq 0 \text{ and } 0 \text{ if } x < 0.$$

Then, equation 4.7 holds

$$f_0(x) + \int_0^x f_0(x-t)\sum_{k=1}^{\infty} f^{k*}(t)dt = \frac{1}{\mu},\ \textit{for all } x \geq 0, \tag{4.7}$$

where f^{k*} denotes the $k-$fold convolution of the function $f(x)$ with itself, i.e. the PDF of $\sum_{i=1}^k J_i$.

Proposition 4.3.1 is a classical result of renewal process theory. We give a simple proof of it in appendix. Different proofs can be found for instance in Mitov and Omey (2014, p. 47) or in Daley and Vere-Jones (2002, p. 75). Proposition 4.3.1 implies that the delayed renewal process, obtained by generating X_0 with CDF F_0 and the J_i's independently with CDF F, has, among other properties, a constant first-order inclusion density equal to $1/\mu$ on $\mathbb{R}_+$. Such a delayed renewal process has stationary increments and is called a stationary renewal process (Mitov and Omey, 2014).

A special case is that of the Poisson process with intensity λ. It is a renewal process whose spacings follow an exponential distribution $\mathcal{E}xp(\lambda) = \mathcal{G}amma(1, \lambda)$. It turns out that its limiting forward recurrence time distribution is also an exponential distribution with parameter λ, so that $F_0 = F$. This is a consequence of the memory-less property of the exponential distribution.

4.4 Quasi-systematic sampling

Our aim is to propose new sampling processes that allow to control the selection probability of neighboring units by adjusting the joint inclusion density. Spreading the sample units over Ω has some advantages when units close together are similar (e.g. when the function z has small variations).

The systematic sampling process allows to select samples that are very well spread. However, it does not possess a positive second order inclusion density and, as a consequence, the Horvitz-Thompson variance estimator given in Cordy (1993) may not be used. We are thus interested in sampling processes with spacings that have a positive variance smaller than that of Poisson or binomial processes. Without auxiliary information that would encourage us to do otherwise, we focus on sampling processes with constant first order inclusion density on Ω.

The family of sampling processes that we consider can be seen as a compromise between basic sampling processes (Poisson and binomial processes) and systematic sampling. The rough idea is the following: in a first phase sampling procedure, a sample of expected size $n \cdot r$, with $n, r \geq 0$, is selected using an elementary sampling process. In the second selection phase, we use a systematic sampling to draw one unit every r units of the first phase sample. We call these processes quasi-systematic sampling processes. We consider the "systematic-Poisson" and "systematic-binomial" processes obtained when the first phase processes are respectively the Poisson and the binomial process. The first and second order inclusion densities of these sampling processes have a closed-form.

Consider the following two-phases sampling process: a first phase sample is generated from a Poisson sampling process with constant intensity λ. Then, a systematic sample is drawn inside this first phase sample with rate $1/r$ (i.e. a starting unit is randomly chosen among the r first units of the first phase sample and is kept in the second phase sample along with every other r unit). In an interval of length 1, the expected number of units selected by the Poisson process is λ. Thus, by setting $\lambda = n \cdot r$, where n is the targeted final average sample size and r is freely chosen, we ensure that the expected final sample size is n.

The spacings of the first sample are, by definition, realizations of an exponential random variable with parameter λ. After the systematic sampling phase, spacings are realizations of sums of r independent exponential random variables i.e. of non-negative random variables $\mathcal{G}amma(r, \lambda)$ with PDF $f(x) = x^{r-1}e^{-\lambda x}\lambda^r/\Gamma(r)$. Thus, except for the first spacing, this process is a renewal process with $\mathcal{G}amma(r, \lambda)$ renewal distribution.

As we are set on having a constant first order inclusion density, and thanks to Proposition 4.3.1, we choose to generate the first spacing with a $\mathcal{F}or\mathcal{G}(r, \lambda)$ distribution and the following ones with independent $\mathcal{G}amma(r, \lambda)$ distributions. The first and second order densities of the obtained systematic-Poisson sampling process are given in Proposition 4.4.1. Note that parameters n and r do not in fact need to be integer numbers. Algorithm 1 can be used to select a systematic-Poisson sample in $(0, 1)$.

Proposition 4.4.1. *Let us consider a systematic-Poisson process on $(0, 1)$ with positive parameters r and λ. Then*

Algorithm 1 Generates a systematic-Poisson sample with parameters λ and r.

Require: $\lambda > 0, r > 0$;
  Generate $x_1 \sim \mathit{ForG}(r, \lambda)$;
  i=2;
  while $x_i < 1$ **do**
    Generate $J_i \sim \mathit{Gamma}(r, \lambda)$
    $x_i = x_{i-1} + J_i; i = i + 1$;
    if $x_i > 1$ **then**
      $n = i - 1$
    end if
  end while
  return $\{x_1, x_2, \ldots, x_n\}$ ordered systematic-Poisson sample with parameters r and λ.

1. *the first order inclusion density is given by:* $\rho(x) = \lambda/r$, *for any* $x \in (0, 1)$,

2. *the second order inclusion density is given by*

$$\rho^{(2)}(x, y) = \frac{\lambda}{r} e^{-\lambda|x-y|} \sum_{m=1}^{\infty} \frac{\lambda^{mr}}{\Gamma(mr)} |x - y|^{mr-1}, \tag{4.8}$$

   for any $x, y \in (0, 1)$.

Proof.

1. is a direct application of Proposition 4.3.1, since the expectation of a $\mathit{Gamma}(r, \lambda)$ distribution is equal to r/λ.

2. From Daley and Vere-Jones (2002, Example 5.4(b), p. 139), we have that, for $0 \leq x < y$,

$$\rho^{(2)}(x, y) = \frac{\lambda}{r} u(y - x),$$

   where u is the first order density of the renewal process $X = (X_i)_{i \geq 2}$, $X_i \sim \mathit{Gamma}(r, \lambda)$. $u(x)$ is equal to $\sum_{k=1}^{\infty} f^{k*}(x)$, where $x \geq 0$ and f is the PDF of X_i. As the sum of m independent $\mathit{Gamma}(r, \lambda)$ variables is a $\mathit{Gamma}(mr, \lambda)$ and has PDF:

$$f(h; m) = \frac{\lambda^{mr}}{\Gamma(mr)} e^{-\lambda h} h^{mr-1}, \ h \geq 0,$$

   we can infer that the counting measure of the renewal process X has renewal density

$$u(h) = \sum_{m=1}^{\infty} f(h; m) = \sum_{m=1}^{\infty} \frac{\lambda^{mr}}{\Gamma(mr)} e^{-\lambda h} h^{mr-1}, \ h \geq 0,$$

   and the result follows.

□

The joint inclusion density equation simplifies for some values of r. Set $\lambda = n \cdot r$ with n the expected the sample size. With $r = 1$ we get the usual Poisson process and thus $\rho^{(2)}(x, y) = \lambda^2 = n^2$.

The plot of $\rho^{(2)}(x, y)$ as a function of $|x - y|$ is given in Figure 4.2 for different values of r. Except for $r = 1$, $\rho^{(2)}(x, y) = 0$ if $x = y$. The larger r is, the flatter the plot is near the origin: the sampling design avoids selecting neighboring units. We see that, when r is very large, the function concentrates on the inverse of the sampling rate and its multiples. It illustrates that the systematic-Poisson sampling design is close to a systematic sampling when r is large.

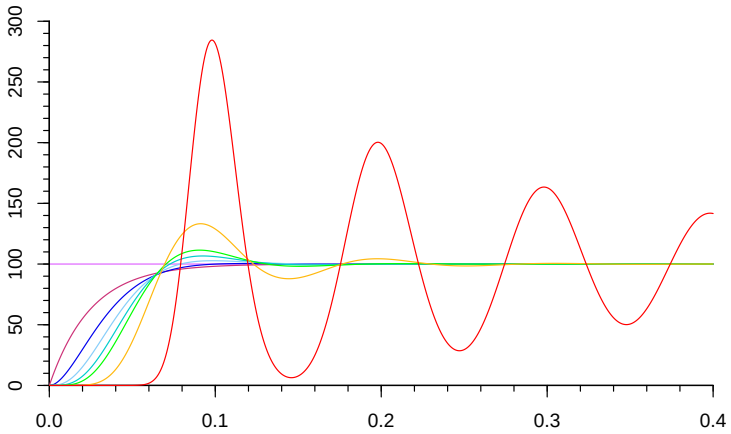

FIGURE 4.2: Joint inclusion density $\rho^{(2)}(x, y)$ as a function of $|x-y|$ for systematic-Poisson sampling, for $n = 10$, $r = 1, 2, 3, 4, 5, 6, 10, 50$ and $\lambda = n \cdot r$. The range of oscillations increases with r. When $r = 1$, $\rho^{(2)}(x, y)$ is constant.

The systematic-binomial process is a fixed size sampling process with constant inclusion density on $(0, 1)$. It is obtained, for example, by taking a realization of a binomial process of size $n \cdot r$, selecting a systematic sub-sample with rate $1/r$ inside the first phase units and finally adding, modulo 1, a random number u generated from a $\mathcal{U}(0, 1)$ distribution. This last step ensures that the circular spacing $x_1+(1-x_n)$ has the same distribution as the other spacings. An illustration of the sampling procedure is given in Figure 4.3. An implementation is proposed

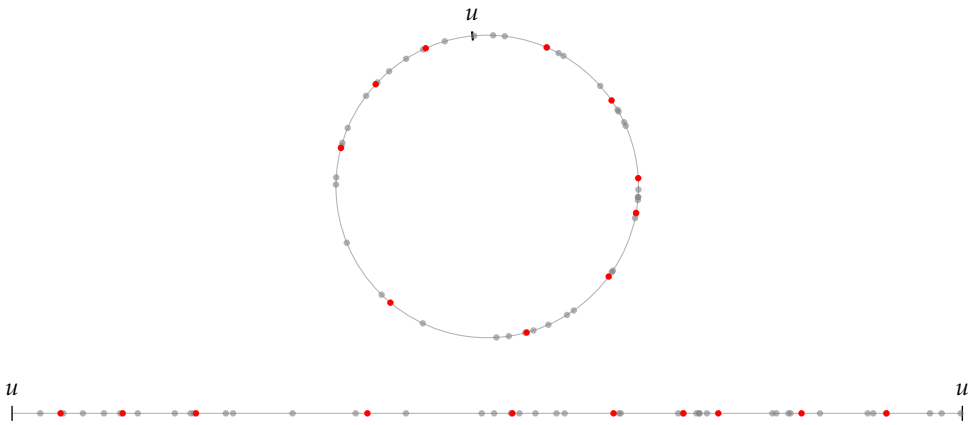

FIGURE 4.3: Systematic-binomial sampling procedure with fixed size $n = 10$ and $r = 5$. In gray, the units sampled at the first phase, and in red the units in the final selection. On the top, we see the random shift u plotted on a circle. On the bottom, we see the final sample on the interval $[0, 1]$.

in Algorithm 2.

Algorithm 2 Systematic-binomial sample with size n and integer parameter r.

Require: $n, r \in \mathbb{N}_*$.
  Generate $\tilde{y}_1, \ldots, \tilde{y}_{nr}$ the sequence of order statistics of $n \cdot r$ i.i.d. variables $\mathcal{U}(0,1)$.
  for $i = 1, \ldots, n$, **do**
    $\tilde{x}_i = \tilde{y}_{ir}$
  end for
  Generate $u \sim \mathcal{U}(0,1)$
  for $i = 1, \ldots, n$ **do**
    $x_i = (\tilde{x}_i + u) \mod 1$
  end for
  return $\{x_1, x_2, \ldots, x_n\}$ ordered systematic-binomial sample with parameter r and size n.

Another way to obtain a realization of a systematic-binomial process is to work with circular spacings. The first phase binomial sample is selected by generating $(\widetilde{J}_i^c)_{i=1,\ldots,nr}$, realization of a $\mathcal{D}ir(\mathbf{1}_{nr})$ distribution, then these spacings are aggregated in packets of r to form the circular spacings of the final sample,

$$J_i^c = \sum_{k=1}^{r} \widetilde{J}_{(i-1)r+k}^c, \ i = 1, \ldots, n, \tag{4.9}$$

and finally a random uniform shift u is used to set the origin. The selected units are

$$\left(\sum_{j=1}^{i} J_j^c + u \right) \mod 1, \text{ for } i = 1, \ldots, n. \tag{4.10}$$

However, the aggregation properties of the Dirichlet distribution ensure that the vector $\mathbf{J}^c = (J_1^c, \ldots, J_n^c)$ of Equation (4.9) follows a $\mathcal{D}ir(r\mathbf{1}_n)$ distribution. We also get that

$$J_i^c \sim \mathcal{B}eta(r, r(n-1)), \tag{4.11}$$

and

$$\sum_{j=1}^{m} J_{i+j}^c \sim \mathcal{B}eta(mr, mr(n-1)), \ 1 \leq m \leq n-i-1, \tag{4.12}$$

where $\mathcal{B}eta(\cdot,\cdot)$ denotes the beta distribution. Taking advantage of this consideration, we can use Algorithm 3 to select samples from a systematic-binomial process. This method is not restricted to integer values of r.

Algorithm 3 Generate a systematic-binomial sample with size n and real parameter $r > 0$.

Require: $n \in \mathbb{N}_*, r \in \mathbb{R}_+^*$.
  Generate $\mathbf{J}^c = (J_1^c, \ldots, J_n^c) \sim \mathcal{D}ir(r\mathbf{1}_n)$.
  Generate $u \sim \mathcal{U}(0,1)$
  for $i = 1, \ldots, n$ **do**
    $x_i = \left(\sum_{j=1}^{i} J_j^c + u \right) \mod 1$
  end for
  return $\{x_1, x_2, \ldots, x_n\}$ ordered systematic-binomial sample with parameter r and size n.

Inclusion densities of the systematic-binomial process are given in Proposition 4.4.2.

Proposition 4.4.2. *Consider a systematic-binomial process of size n with parameter r. Its inclusion densities are given below.*

1. *The first order inclusion density is given by:*

$$\rho(x) = n, \text{ for } x \in (0,1).$$

2. *The second order inclusion density is given by*

$$\rho^{(2)}(x,y) = n \sum_{m=1}^{n-1} \frac{\Gamma(nr)}{\Gamma(mr)\Gamma[(n-m)r]} |x-y|^{mr-1}(1-|x-y|)^{(n-m)r-1}, \tag{4.13}$$

   for $x \neq y \in (0,1)$.

3. *The $n-$th order inclusion density is given by:*

$$\rho^{(n)}(x_1,\ldots,x_n) = n\frac{\Gamma(nr)}{[\Gamma(r)]^n}(1+x_1-x_n)^{r-1}(x_2-x_1)^{r-1}\cdots(x_n-x_{n-1})^{r-1},$$

   for $x_1 < \cdots < x_n \in (0,1)$.

Proof.

1. Due to the random uniform shift used to set the origin, the point process canonically induced on the unit circle is clearly stationary (i.e. rotation invariant). Its first moment measure is thus a Haar measure and proportional to the Lebesgue measure. It follows that the first moment measure of the considered systematic-binomial process is proportional to the Lebesgue measure on $(0,1)$, and the proportionality coefficient is the total mass n.

2. The point process being stationary, its second order inclusion density reduces to

$$\rho^{(2)}(x,y) = n \cdot u(y-x), \text{ if for example } 0 \leq x < y < 1,$$

   where u is the first order density of the point process $J_2^c, \ldots, J_n^c$ on $[0,1]$. However we have that the corresponding counting function $U(h) = N(0,h)$ is given by:

$$U(h) = \sum_{m=1}^{n-1} F^{m*}(h),$$

   where F^{m*} is the CDF of $\sum_{i=1}^{m} J_{2+i-1}^c$ and is thus the CDF of a $\mathcal{B}eta(mr, mr(n-1))$ distribution. Hence

$$u(h) = \sum_{m=1}^{n-1} \frac{\Gamma(nr)}{\Gamma(mr)\Gamma[(n-m)r]} h^{mr-1}(1-h)^{(n-m)r-1}, \ 0 < h < 1,$$

   and the result follows.

3. As with ordinary binomial sampling, a given sample is obtained exactly when u is equal to one of the units and the spacings agree with the sample. Moreover the Dirichlet distribution with parameter $r\mathbf{1}_n$ is symmetric and u is independent from $\mathbf{J}^c$. We get that:

$$\begin{aligned} & f_{\mathbf{X}}(x_1,\ldots,x_n) \\ &= f_u(x_1) f_{\mathbf{J}^c}(x_2 - x_1, \ldots, x_n - x_{n-1}, x_1 - x_n + 1) \\ &+ f_u(x_2) f_{\mathbf{J}^c}(x_3 - x_2, \ldots, x_1 - x_n + 1, x_2 - x_1) \\ &\vdots \\ &+ f_u(x_n) f_{\mathbf{J}^c}(x_1 - x_n + 1, \ldots, x_{n-1} - x_{n-2} + 1, x_n - x_{n-1}) \\ &= n\frac{\Gamma(nr)}{[\Gamma(r)]^n}(1 + x_1 - x_n)^{r-1}(x_2 - x_1)^{r-1}\cdots(x_n - x_{n-1})^{r-1}. \end{aligned}$$

□

Some straightforward computations lead to Equation 4.14

$$\int_0^1 \int_0^1 \rho^{(2)}(x,y)dxdy = n(n-1). \tag{4.14}$$

A plot of $\rho^{(2)}(x,y)$ as a function of y is given in Figure 4.4, for $x = 0.4$ and different values of r. Except for $r = 1$, $\rho^{(2)}(x,y) = 0$ if $x = y$. The larger r is, the flatter the joint inclusion density is around $x = y$. The selection of neighboring units is thus very unlikely with such a sampling design and a large r. When r is very large the function concentrates on regularly spaces pikes as in the systematic-Poisson case.

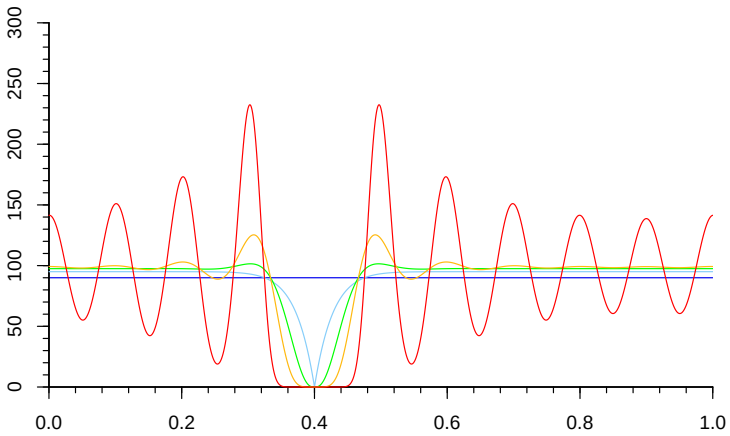

FIGURE 4.4: Joint inclusion density $\rho^{(2)}(x,y)$ as a function of y for $x = 0.4$ in systematic-binomial sampling with $n = 10$ and $r = 1, 2, 4, 8, 30$. The range of oscillations increases with r. When $r = 1$, $\rho^{(2)}(x,y)$ is constant.

4.5 Asymptotic results

The sampling processes introduced in Section 4.4 depend on a parameter r. When r gets large, they look more and more like systematic sampling processes. Indeed, we will see that these processes converge in distribution to the systematic sampling process when n is fixed and r goes to infinity. We first need Lemma 4.5.1.

Lemma 4.5.1. *A forward gamma random variable $\mathcal{ForG}(r, rn)$ converges in distribution to a continuous uniform variable $\mathcal{U}(0, 1/n)$ when r tends to infinity and n is fixed.*

Proof. It is easy to prove that, if ϕ_f is the characteristic function of a positive probability distribution with expectation $\mu > 0$, PDF f and CDF F, then the characteristic function ϕ_{f_0} of the probability distribution with density $f_0 = (1 - F)/\mu$ is such that:

$$\phi_{f_0}(t) = \frac{1}{i\mu}\left[\frac{\phi_f(t) - 1}{t}\right], \ t \in \mathbb{R},$$

where $i^2 = -1$. However, the characteristic function of a $\mathcal{Gamma}(r, \lambda)$ is given by $\phi_\Gamma(t) = (1 - it/\lambda)^{-r}$. It follows that the characteristic function of a $\mathcal{ForG}(r, \lambda)$ is given by

$$\phi(t; r, \lambda) = \lambda \frac{\left(\frac{\lambda}{\lambda - it}\right)^r - 1}{irt},$$

Replacing λ by rn and letting r tend to infinity, we obtain that the characteristic function has a limit:

$$\lim_{r\to\infty} \phi(t; r, rn) = \frac{e^{it/n} - 1}{it/n},$$

which is the characteristic function of a continuous uniform random variable $\mathcal{U}(0, 1/n)$. Lévy's continuity theorem applies and gives the result. □

We can now prove the announced result. We start with the systematic-Poisson process in Proposition 4.5.1.

Proposition 4.5.1. *Let us consider a systematic-Poisson process on $(0, 1)$ with parameters $r > 0$ and $\lambda = rn$. Then, the process weakly converges to a systematic process of size n when r tends to infinity.*

Proof. In systematic-Poisson process with parameter r and $\lambda = rn$, the first spacing time follows a forward Gamma distribution $\mathcal{ForG}(r, rn)$ and the next ones follow a Gamma distribution $\mathcal{Gamma}(r, rn)$. We have seen in Proposition 4.5.1 that $\mathcal{ForG}(r, rn)$ converges to a $\mathcal{U}(0, 1/n)$ when r tends to infinity. We also have that the $\mathcal{Gamma}(r, rn)$ distribution converges to a degenerate distribution $\mathcal{Dirac}(1/n)$. Indeed, the expectation of a $\mathcal{Gamma}(r, rn)$ is equal to $1/n$ and its variance to $1/(rn^2)$. As the spacings are independent, we get that any finite family of them jointly converges to the matching distributions of spacings of a systematic process, as defined in Section 4.2. However, in the case of point processes the weak convergence of finite distributions is equivalent to the weak convergence of the process (see, e.g., Theorem 11.1.VII of Daley and Vere-Jones, 2008, p. 137). □

The case of the systematic-binomial process is dealt with in Proposition 4.5.2.

Proposition 4.5.2. *Consider a systematic-binomial process of size n on $(0, 1)$ and with parameter $r > 0$. Then the process converges in distribution to a systematic sampling process when r tends to infinity.*

Proof. It is sufficient to show that the circular spacings converge in distribution to a $\mathcal{D}irac(1/n)$. Indeed, the random start is already accounted for in the procedure. However, the spacings follow a Beta distribution with mean $1/n$ and variance $r^2(n-1)/[(rn)^2(rn+1)]$, and indeed, their variance tends to 0 when r tends to infinity. As in the proof of Proposition 4.5.1, Theorem 11.1.VII in Daley and Vere-Jones (2008) allows to finish the proof. □

4.6 Simulations

Some simulations are useful to illustrate the properties of the systematic-binomial sampling process. We also ran simulations with the systematic-Poisson process and found that it behaves similarly but gives results that are less accurate than the systematic-binomial process with our test function. We considered the following test function:

$$h(x) = 100 \sin\left(\frac{3x^2}{2x^2+1}\right) \exp\left\{-\left[\sin(4\pi x)^2\right]\right\},$$

plotted in Figure 4.5 (left). We aim at estimating its mean using the Horvitz-Thompson estimator on a sample selected with a systematic-binomial process. A set of $10,000$ samples was generated using a systematic-binomial process with fixed size $n = 30$ and for each value of the parameter $r = 1,\ 2,\ 5,\ 10,\ 30,\ 50$ and 100. Figure 4.5 (right) shows that the accuracy of the Horvitz-Thompson estimator increases with r. As expected, the systematic process performs better than any quasi-systematic process. Corresponding simulation Root Mean Square

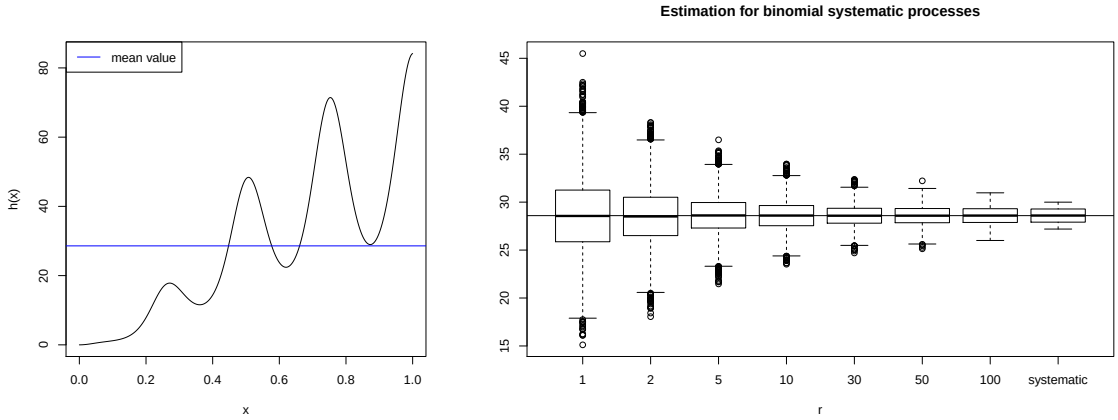

FIGURE 4.5: Test function (left) and boxplots of the estimated totals over all the simulations (right). The parameter r varies between $r = 1$ and $r = 100$ and we included the systematic sampling estimation. The horizontal line represent the true value of the total.

Errors (RMSE) are given in Table 4.1. We see in this table that the RMSE decreases rapidly with moderate values of r.

TABLE 4.1: RMSE of simulation results with a systematic-binomial process of size $n = 30$ and different values of r.

1	2	5	10	30	50	100	Systematic
3.98	2.92	1.95	1.52	1.09	1.00	0.91	0.81

Estimating the variance of the Horvitz-Thompson estimator is a different issue. As previously stated, the variance estimator becomes unstable as r increases, due to the fact that the second order inclusion density tends to 0 almost everywhere when r goes to infinity. The estimated variance can also be negative in some cases. To alleviate these problems, the sample size n should be increased when using large values of r. We give, in Table 4.2 the mean over $10,000$ simulation samples of the variance estimator, their standard deviation as well as the true variance, for different combinations of n and r. Since the systematic-binomial process has a fixed size, we use the Sen-Yates-Grundy variance estimator. The estimator is theoretically unbiased and decreases on average as the sample size n increases when r is fixed. We see that the standard deviation of the variance estimator values obtained in the simulations is consistently smaller for $r = 2$ than for $r = 1$ but gets a lot worse for larger values of r. Note that the simulation RMSEs of Table 4.1 mostly agree with the true variances in Table 4.2.

TABLE 4.2: Estimated variance of a systematic-binomial process for different sample sizes n and parameter r values. In each cell, the simulation mean of the variance estimator with the corresponding Standard Deviations (SD) within parentheses, and the true target variance on the right.

	$n = 30$		$n = 50$		$n = 70$		$n = 100$	
		var	avg.$\widehat{\text{var}}$ (SD)	var	avg.$\widehat{\text{var}}$ (SD)	var	avg.$\widehat{\text{var}}$ (SD)	var
$r = 1$	15.87 (3.62)	15.88	9.51 (1.66)	9.53	6.81 (1.00)	6.81	4.76 (0.58)	4.76
$r = 2$	8.49 (2.14)	8.52	4.96 (0.97)	4.96	3.50 (0.52)	3.50	2.43 (0.31)	2.43
$r = 4$	4.58 (3.61)	4.65	2.59 (1.58)	2.62	1.81 (0.81)	1.82	1.25 (0.56)	1.25
$r = 8$	2.94 (55.36)	2.67	1.40 (4.95)	1.43	0.95 (1.71)	0.97	0.66 (1.89)	0.66
$r = 30$	0.30 (20.22)	1.20	0.31 (7.02)	0.56	0.25 (4.46)	0.35	0.16 (1.85)	0.22

We also see in table 4.2 that the variance estimator gets very unstable for large values of r. One reason for this instability of the variance estimator is the joint inclusion density function getting close to 0 for large values of r as can be seen on Figures 4.2 and 4.4. Actually, this function, with y in a neighborhood of a fixed x in $(0, 1)$, is driven by the first term in Equation (4.13), and behaves like $|x - y|^{r-1}$. This is considered in Section 4.7 where we discuss the choice of the tuning parameter r.

Another cause of instability in this example is that the test function has different values in 0 and 1 whereas the probability of jointly selecting $x > 0$ but close to 0 and $y < 1$ but close to 1 is small. When a sample is selected that contains such units, the variance estimator (4.3) takes a very large value. In our simulations, this case was responsible for most of the observed atypical very large values of the variance estimator.

To solve this problem, if the test function f is such that $f(0) \neq f(1)$, we define a new function g by

$$g(x) = \begin{cases} f(2x) & \text{if } 0 \leq x \leq 1/2, \\ f(2x - 2) & \text{if } 1/2 < x \leq 1. \end{cases}$$

The function g is such that $\int g(x)\, dx = \int f(x)\, dx$ and satisfies $g(0) = g(1)$. As we see in Table 4.3, replacing f with g does not increase the simulated RMSEs, nor the true variances found in Table 4.4.

The variance estimator is much more stable with the transformed function g than with the interest function f, as can be seen in Table 4.4, compared with Table 4.2. The variance itself is slightly lower, meaning that the loss in spreading efficiency due to the transformation of the

TABLE 4.3: RMSE using the transformed function with a systematic-binomial process of size $n = 30$ and different values of r.

$r = 1$	$r = 2$	$r = 4$	$r = 8$	$r = 30$
3.97	2.85	2.06	1.46	0.76

interest function is more than compensated by the absence of extreme values that were caused by $f(1)$ being different from $f(0)$.

TABLE 4.4: Estimated variance of a systematic-binomial process using the transformed function for different sample sizes n and parameter r values. In each cell, the simulation mean of the variance estimator with the corresponding Standard Deviations (SD) within parentheses, and the true target variance on the right.

	$n = 30$		$n = 50$		$n = 70$		$n = 100$	
	avg.$\widehat{\text{var}}$ (SD)	var	avg.$\widehat{\text{var}}$ (SD)	var	avg.$\widehat{\text{var}}$ (SD)	var	avg.$\widehat{\text{var}}$ (SD)	var
$r = 2$	8.30 (1.39)	8.31	4.86 (0.61)	4.85	3.44 (0.36)	3.44	2.39 (0.21)	2.39
$r = 4$	4.26 (0.57)	4.26	2.44 (0.23)	2.44	1.72 (0.13)	1.72	1.20 (0.07)	1.20
$r = 8$	2.17 (1.17)	2.15	1.22 (0.23)	1.22	0.86 (0.08)	0.86	0.60 (0.06)	0.60
$r = 30$	0.32 (9.26)	0.58	0.25 (2.91)	0.33	0.24 (1.46)	0.23	0.15 (0.48)	0.16

When r is not too large, confidence intervals exhibit coverage rates very close to the nominal rate of 95% as shown in Table 4.5. These confidence intervals are computed assuming a normal approximation which seems compatible with our simulation results. However, for large values of r, $r \geq 30$ in our simulations, the estimation of the variance is very unstable, and the coverage rate of estimated confidence intervals deviates strongly. Indeed, for $r = 30$ the low coverage rates in our simulations are explained by the variance estimator often taking negative values. In this case it would certainly be preferable to use a plain systematic process as the systematic-binomial process does not allow to get good confidence interval estimates.

TABLE 4.5: Empirical coverage rates with a systematic-binomial sampling process and a transformed interest function, for different values of n and r.

	$n = 30$	$n = 50$	$n = 70$	$n = 100$
$r = 2$	0.9426	0.9472	0.9480	0.9492
$r = 4$	0.9447	0.9476	0.9469	0.9452
$r = 8$	0.9333	0.9463	0.9469	0.9475
$r = 30$	0.4789	0.5054	0.5522	0.5993

4.7 Choice of the tuning parameter

By choosing the tuning parameter r one can make a compromise between an accurate estimation of the target parameter with a poor estimation of the precision and a less accurate estimation of the target parameter but with a reliable estimation of the estimator variance. Ideally one would have at its disposal a proxy interest function and could run simulations to select a suitable r, that is to say a r that corresponds to one's preferred compromise.

When no useful proxy function is available, some general remarks apply. Judging from our simulations, it seems that a small value of r already helps reducing variance considerably compared to plain binomial process sampling. It is to be noted that, with values of r between

1 and 2, the joint inclusion probability function $\rho^{(2)}(x, y)$ takes small values only when x and y are extremely close, as can be seen on Figures 4.2 and 4.4. This is not the case anymore when r is larger than 2. In our simulations of Section 4.6, using the transformed function, we observed large values of the variance estimator only with r larger than 2.

A second point that could be inferred from our simulations is that larger sample sizes can accommodate for larger values of r. However, we do not have solid arguments to support that and we may just be lacking more simulation results here. It is to be noted though that, for fixed size processes such as the systematic-binomial process, one can check in advance which values of r and sample size n, allow to satisfy the Sen (1953), Yates and Grundy (1953) conditions: $\rho^{(2)}(x, y) \leq \rho(x)\rho(y)$ for all x, y. When these conditions hold, the variance estimator (4.3) is non-negative. Based on a numerical exploration, our conjecture is that this condition holds for $r = 2$ and any sample size, but not for $r = 3$. We also conjecture that, for fixed $r \geq 3$, increasing the sample size does not help reducing the maximal value of $\rho^{(2)}(x, y)/\rho(x)\rho(y)$. However, for a large enough n, and a given x, values of y such that $\rho^{(2)}(x, y)/\rho(x)\rho(y)$ is greater than 1 are concentrated around x, and thus these couples do not contribute much to the variance estimator (4.3). Based on these considerations, it seems that $r = 2$ could be a good compromise between stability of the variance estimator and stability of the target parameter estimator when no other information is available. The associated estimator true variance is however clearly greater than that obtained with larger values of r.

Finally, the regularity of the interest function has its importance. We can observe that having a function that satisfies a Hölder condition with exponent $\alpha \geq 0$ implies that the variance estimator (4.3) is bounded for all $r \leq 2\alpha + 1$ (n.b.: we need to take the restriction of the function to $[0, 1)$ and transport its source to the unit circle first in order to account for what happens near 0 and 1).

4.8 Conclusion and discussion

In this paper, we only worked on sampling processes with constant first order inclusion density. It is however common in finite population survey sampling to choose different inclusion probabilities for different population units using auxiliary information available (e.g. the size of businesses or the approximate dispersion of the interest variable in a sub-population). Suppose we want to have a sampling process with first order inclusion density proportional to a non-negative continuous function ϕ, and note $\Phi(x) = \int_0^x \phi(t)dt$. Assume that the set of zeroes of ϕ have no interior, so that Φ is increasing. We just need to select a sample $x_1, \ldots, x_n$ with a constant inclusion density process, and retain $\Phi^{-1}(x_1), \ldots, \Phi^{-1}(x_n)$ as our sample. Indeed, if $\tilde{U}(x) = \mathrm{E}\{\tilde{N}([0, x])\}$ is the counting function of the new process and $U(x) = \mathrm{E}\{N([0, x])\}$ is the counting function of the constant density process, we have that

$$\tilde{U}(x) = U[\Phi(x)] = \lambda\Phi(x) \text{ for some } \lambda > 0.$$

It follows that $\tilde{U}(x) = \int_0^x \lambda\phi(t)dt$ and that the first order inclusion density of the new process is given by $\tilde{\rho}(x) = \lambda\phi(x)$. The second inclusion density $\tilde{\rho}^{(2)}$ of this new process can also be derived from that, denoted by $\rho^{(2)}$, of the process used to select $x_1, \ldots, x_n$. We find that $\tilde{\rho}^{(2)}(x, y) = \rho^{(2)}[\Phi(x), \Phi(y)]\phi(x)\phi(y)$.

Both algorithms proposed in Section 4.4 work with any positive value of r. The use of a parameter $0 < r < 1$ results in an attractive or clustering process where units tend to be selected

in grouped clusters. This can be useful in some modelization problems. However, the interest of sampling with such clustering processes is probably limited to very specific objectives.

In future work, we intend to explore the possibility of developing similar sampling tools in spaces with more than one dimension. The generalization is far from being obvious as we only worked here on $\mathbb{R}$ equipped with its field ordering and some notions strongly depend on it.

Quasi-systematic sampling processes are useful to the practitioner who wants to make his own compromise between a more accurate estimation of a functions mean and a good estimation of the uncertainty of his estimator. Our simulations illustrate this trade-off between precision in the estimation of the mean and accuracy of the variance estimator. The former is better with a systematic sampling process while the latter is better with small values of r. We argue that quasi-systematic sampling processes could be used in place of plain binomial or Poisson processes for the purpose of estimating a mean in a continuous universe. A possible application is the estimation of the total or the mean of a variable of interest over time.

Acknowledgements

The authors are grateful to one associate editor and three reviewers for their insightful comments that helped considerably improve the quality of this paper. This work was supported in part by the Swiss Federal Statistical Office. The views expressed in this paper are solely those of the authors. M. W. was partially supported by a Doc.Mobility fellowship of the Swiss National Science Foundation (grant no. P1NEP2_162031).

Chapter 5

Point Processes

In order to make this document self-contained, we state some basic facts about point processes and we give some definitions. We also review some results about the Horvitz-Thompson estimator in a continuous universe and give proofs. Those results are well-known and there is some overlap with Chapter 4. The proofs presented here are probably known but we give them for the sake of completeness.

5.1 Point processes

Let $\Omega \subseteq \mathbb{R}^d, d = 2, 3$, be an open and bounded set equipped with the Lebesgue measure. We follow the construction of point processes given by Macchi (1975) (see also Moyal, 1962). A realization of a point process is a set of units $X = \{x_1, \ldots, x_n\}$ without consideration for the order of the x_i's. This definition is analogue to the one used in finite population sampling (see for example Cochran, 1977, for an introduction to finite population sampling theory). In the following, we assume that the x_i's are all distinct and such a process is called *simple*. A point process is thus a probability distribution on the space $\mathcal{S}$ of all such collections. It is not a distribution on $\Omega^{\mathbb{N}}$ with the corresponding $\sigma-$algebra being the tensor product of the univariate $\sigma-$algebra as we are considering unordered sets. A careful construction of point processes on Ω is given in Macchi (1975). A point process is thus a probability distribution on $(\mathcal{S}, \mathcal{B})$ where $\mathcal{S} = \bigcup_{n\in\mathbb{N}} \Omega^n/\mathcal{R}^n$, with x and y in Ω^n being in the same class for the equivalence relation $\mathcal{R}^n$ if x is a permutation of elements of y, and $\mathcal{B}$ is the $\sigma-$algebra generated by the family of counting events:

$$\{s \in \mathcal{S} \text{ such that } N(s, A) = p,\ A \in \mathcal{B}(\Omega),\ p \in \mathbb{N}\},$$

and $N(s, A)$ is the number of elements of s that are in A.

Definition 5.1.1. *A point process X defined on Ω can be defined by specifying*

1. *A sequence $\{p_k\}_{k\geq 0}$, such that $p_k = P(\#X = k)$ is the distribution of the number of points of the process;*
2. *A family of symmetric probability density functions $\{\pi^{(k)}\}_{k\geq 0}$ such that $\pi^{(k)}$ is defined on Ω^k, for all $k \geq 0$.*

The definition 5.1.1 is similar to the one given by Daley and Vere-Jones (2002, Conditions 5.3.I).

Example 5.1.1. *1. In the case of the Poisson process of constant intensity $\lambda > 0$, the distribution of the number of points follows a Poisson distribution proportional to $|\Omega|$, the Lebesgue measure of*

Ω, *that is*

$$p_k = e^{-\lambda|\Omega|}\frac{(\lambda\,|\Omega|)^k}{k!}.$$

We know that, conditional to n, locations of a realization of a Poisson process are independent and uniformly distributed over Ω. Thus $\pi^{(k)}$ is constant and must sum to 1 when integrated on Ω^k. Thus

$$\pi^{(k)}(x_1,\dots,x_k) = \frac{1}{|\Omega|^k}, \quad k = 1,\dots.$$

2. *For the binomial process of size $n \in \mathbb{N}$ and with density f (see Definition 4.2.1), we have $p_k = \delta_{kn}$ and*

$$\pi^{(n)}(x_1,\dots,x_n) = f(x_1)\cdots f(x_n).$$

Definition 5.1.2 (Product densities)**.** *Let X be a point process as in Definition 5.1.1. We define the product densities as*

$$\rho^{(k)}(x_1,\dots,x_k) = \sum_{m=0}^{\infty}\frac{1}{m!}\int_{\Omega^m}(k+m)!\;p_{k+m}\;\pi^{(k+m)}(x_1,\dots,x_k,y_1,\dots,y_m)\;dy_1\dots dy_m. \quad (5.1)$$

Remark 5.1.1. *Equation (5.1) has a straightforward interpretation. Indeed, the terms*

$$p_k\;k!\;\pi^{(k)}(x_1,\dots,x_k)\;dx_1\dots dx_k$$

can be interpreted as the probability that there are exactly n points in the process, one in each of the n distinct balls $B_{x_i}(dx_i)$. These are referred to as the Janossy densities *(Daley and Vere-Jones, 2002, p. 125). Hence*

$$\rho^{(k)}(x_1,\dots,x_k)\;dx_1\dots dx_k$$

is the probability that the process has a point in each balls $B_{x_i}(dx)$, disregarding the rest of the process. In Chapter 4, they are referred to as inclusion densities*, but in the following, we call them product densities, since it is a more usual denomination in the literature of point processes.*

Remark 5.1.2. *Here, we have considered that $\Omega \subset \mathbb{R}^d$. If instead we had considered that Ω was a finite set, then a point process is exactly the same mathematical object as a sampling design. Indeed, the product densities play the same role as the inclusion probabilities in survey sampling. See the heuristic discussion about this fact in Ben Hough et al. (2010, pp. 9-10)*

If they exist, and for any non negative measurable function $h : \Omega^k \to \mathbb{R}$, the product densities $\rho^{(k)} : \Omega^k \to \mathbb{R}_+,\ k = 1,\dots,$ satisfy

$$E\left[\sum_{x_1,\dots,x_k\in X}^{\neq} h(x_1,\dots,x_k)\right] = \int_{\Omega^k} h(x_1,\dots,x_k)\rho^{(k)}(x_1,\dots,x_k)\;dx_1\dots dx_k. \quad (5.2)$$

This expression is a generalization of the *Campbell's formula* (Campbell, 1909). By analogy with a Poisson process, we usually refer to $\rho^{(1)}(x) = \rho(x)$ as the *intensity*.

Roughly speaking, repulsion means that the probability of occurrence of two points tends to zero if the distance between the points tend to zero. A way of quantifying the repulsion is to compare this probability with the scenario where the points occur independently. We thus

define the *pair correlation function* as:

$$g(x,y) = \frac{\rho^{(2)}(x,y)}{\rho(x)\rho(y)}.$$

If the points occur independently, then $g = 1$. If $g > 1$, we say that the process is clustering and if $g < 1$, we say that the process is repulsive. We see that the pair correlation function is a local measure of repulsiveness. A global measure of repulsiveness has been introduced by Lavancier et al. (2015), is studied in details by Biscio and Lavancier (2016) and is related to the Ripley's K function (Ripley, 1976, 1977).

5.2 Some results about Point processes

We will show how we can derive the Campbell's formula from Definition 5.1.1. These results can certainly be found in different reference books. However, we have not found them in this very same form, and thus, we give some proofs. Note that the assumptions can be considerably weakened but they are reasonable in the framework we consider.

Proposition 5.2.1 (Campbell's formula)**.** *Let X a point process defined on a bounded and open set Ω, with distribution of the number of points $\{p_k\}_{k\geq 0}$ and a family of symmetric probability density functions of the locations of points $\{\pi^{(k)}\}_{k\geq 0}$. Then, we have:*

$$\mathrm{E}\left[\sum_{x_i \in X} f(x_i)\right] = \int_\Omega f(x)\rho(x)\, dx,$$

for any continuous function $f : \overline{\Omega} \to \mathbb{R}$.

Proof. We define $\tilde{f} : \mathcal{S} \to \mathbb{R}$ as $\tilde{f}(X) = \sum_{x_i \in X} f(x_i)$. Note that since f is continuous, $\tilde{f}$ is measurable. Note also that if $\#X = 0$, then $\tilde{f}(\emptyset) = 0$. We thus have

$$\begin{aligned}
\mathrm{E}[\tilde{f}(X)] &= \sum_{k=0}^{\infty} p_k \int_{\Omega^k} \tilde{f}(x_1,\ldots,x_k)\pi^{(k)}(x_1,\ldots,x_k)\, dx_1\cdots dx_k \\
&= \sum_{k=1}^{\infty} p_k \int_{\Omega^k} \sum_{i=1}^{k} f(x_i)\pi^{(k)}(x_1,\ldots,x_k)\, dx_1\cdots dx_k \\
&= \sum_{k=1}^{\infty} p_k \sum_{i=1}^{k} \int_{\Omega} f(x_i) \left[\int_{\Omega^{k-1}} \pi^{(k)}(x_1,\ldots,x_k)\, dx_1\cdots dx_k\right] dx_i \\
&= \sum_{k=0}^{\infty} p_k k \int_{\Omega} f(x_1) \left[\int_{\Omega^{k-1}} \pi^{(k)}(x_1,x_2,\ldots,x_k)\, dx_2\cdots dx_k\right] dx_1 \qquad (5.3) \\
&= \sum_{k=0}^{\infty} \int_{\Omega} f(x) \left[\int_{\Omega^k} p_{k+1}(k+1)\pi^{(k+1)}(x,y_1,\ldots,y_k)\, dx_2\cdots dx_k\right] dx \\
&= \int_{\Omega} f(x) \left[\sum_{k=0}^{\infty} \frac{1}{k!} \int_{\Omega^k} p_{k+1}(k+1)!\pi^{(k+1)}(x,y_1,\ldots,y_k)\, dx_2\cdots dx_k\right] dx. \\
&= \int_{\Omega} f(x)\rho(x)\, dx. \qquad (5.4)
\end{aligned}$$

By symmetry of the densities $\{\pi^{(k)}\}$, the integrals

$$\int_\Omega f(x_i) \left[\int_{\Omega^{k-1}} \pi^{(k)}(x_1, \dots, x_k)\, dx_1 \cdots dx_k \right] dx_i$$

are the same for all $i = 1, \dots, k.$, which yields the equality (5.3). Note also that by convention, we have used that

$$\int_{\Omega^k} \pi^{(k+1)}(x, y_1, \dots, y_k)\, dy_1 \dots y_k = \pi^{(1)}(x), \text{ for } k = 0.$$

□

This result is stated in terms of the Janossy measures in Daley and Vere-Jones (2002, exercise 5.3.8, p. 132). The following result is a generalization of the previous one, and the proof is exactly the same.

Proposition 5.2.2 (Generalized Campbell's formula)**.** *Let X a point process defined on a bounded and open set Ω, with distribution of the number of points $\{p_k\}_{k\geq 0}$ and a family of symmetric probability density functions of the locations of points $\{\pi^{(k)}\}_{k\geq 0}$. Then, we have:*

$$E\left[\sum_{x_1,\dots,x_k \in X}^{\neq} f(x_1, \dots, x_k) \right] = \int_{\Omega^k} f(x_1, \dots, x_k) \rho^{(k)}(x_1, \dots, x_k)\, dx_1 \dots dx_k.$$

for any continuous function $f : \overline{\Omega^k} \to \mathbb{R}$.

Proof. We define $\tilde{f} : \mathcal{S} \to \mathbb{R}$ as $\tilde{f}(X) = \sum_{x_1,\dots,x_k \in X}^{\neq} f(x_1, \dots, x_k)$. Note that since f is continuous, $\tilde{f}$ is measurable. Note also that if $\#X < k$, then $\tilde{f}(\emptyset) = 0$. We introduce some notations for the purpose of this proof. Let $\mathbf{n} = \{1, \dots, n\}$, $\mathbf{i} = \{i_1, \dots, i_k\} \subset \mathbf{n}$ a subset of *ordered* distinct indices and $-\mathbf{i} = \mathbf{n} \setminus \mathbf{i}$. Then we denote by $f(\mathbf{x_i}) = f(x_{i_1}, \dots, x_{i_k})$, $d\mathbf{x_i} = dx_{i_1} \cdots dx_{i_k}$, and similarly, we will use $d\mathbf{x_n}$ and $d\mathbf{x_{-i}}$ in the following. Moreover, we denote the sum over all the ordered subsets of size k of n indices by

$$\sum_{x_{i_1},\dots,x_{i_k}}^{\neq} = \sum_{\mathbf{i}\subset\mathbf{n}}.$$

Note that this sum has $n!/(n-k)!$ different terms, the number of ordered permutation of k out of n elements. We thus have

$$\begin{aligned}
\mathrm{E}\left[\tilde{f}(X)\right] &= \sum_{n=0}^{\infty} p_n \int_{\Omega^n} \tilde{f}(x_1, \dots, x_n) \pi^{(n)}(x_1, \dots, x_n)\, dx_1 \dots dx_n \\
&= \sum_{n=k}^{\infty} p_n \int_{\Omega^n} \sum_{x_{i_1},\dots,x_{i_k}}^{\neq} f(x_{i_1}, \dots, x_{i_k}) \pi^{(n)}(x_1, \dots, x_n)\, dx_1 \dots dx_n \\
&= \sum_{n=k}^{\infty} p_n \int_{\Omega^n} \sum_{\mathbf{i}\subset\mathbf{n}} f(\mathbf{x_i}) \pi^{(n)}(\mathbf{x_n})\, d\mathbf{x_n} \\
&= \sum_{n=k}^{\infty} p_n \sum_{\mathbf{i}\subset\mathbf{n}} \int_{\Omega^k} f(\mathbf{x_i}) \left[\int_{\Omega^{n-k}} \pi^{(n)}(\mathbf{x_n})\, d\mathbf{x_{-i}} \right] d\mathbf{x_i}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=k}^{\infty} p_n \frac{n!}{(n-k)!} \int_{\Omega^k} f(\mathbf{x_i}) \left[\int_{\Omega^{n-k}} \pi^{(n)}(\mathbf{x_n})\, d\mathbf{x_{-i}} \right] d\mathbf{x_i} \\
&= \sum_{m=0}^{\infty} p_{k+m} \frac{(m+k)!}{m!} \int_{\Omega^k} f(\mathbf{x_i}) \left[\int_{\Omega^m} \pi^{(k+m)}(\mathbf{x_i}, \mathbf{x_{-i}})\, d\mathbf{x_{-i}} \right] d\mathbf{x_i} \qquad (5.5) \\
&= \int_{\Omega^k} f(\mathbf{x_i}) \left[\sum_{m=0}^{\infty} \frac{1}{m!} \int_{\Omega^m} p_{k+m}(m+k)! \pi^{(k+m)}(\mathbf{x_i}, \mathbf{x_{-i}})\, d\mathbf{x_{-i}} \right] d\mathbf{x_i} \\
&= \int_{\Omega^k} f(\mathbf{x_i}) \rho^{(k)}(\mathbf{x_i}) d\mathbf{x_i}. \qquad (5.6)
\end{aligned}
$$

By symmetry of the densities $\{\pi^{(n)}\}$, the integrals

$$\int_{\Omega^k} f(\mathbf{x_i}) \left[\int_{\Omega^{n-k}} \pi^{(n)}(\mathbf{x_n})\, d\mathbf{x_{-i}} \right] d\mathbf{x_i}$$

are the same for all $\mathbf{i} \subset \mathbf{n}$, which yields the equality (5.5). Note also that by convention, we have used that

$$\int_{\Omega^m} \pi^{(k+m)}(\mathbf{x_i}, \mathbf{x_{-i}})\, d\mathbf{x_{-i}} = \pi^{(k)}(\mathbf{x_i}), \text{ for } m = 0$$

□

In the following, we provide proofs to results that are only stated by Cordy (1993).

Proposition 5.2.3 (Cordy, 1993)**.** *Let $z : \overline{\Omega} \to \mathbb{R}$ a continuous function of interest and X a point process with continuous and strictly positive first and second product densities $\rho(x)$ and $\rho^{(2)}(x, y)$. We define $\hat{z}$ as*

$$\hat{z} = \sum_{x_i \in X} \frac{z(x_i)}{\rho(x_i)}.$$

Then, we have:

1. $$\mathrm{E}[\hat{z}] = \int_{\Omega} z(x).$$

2. $$
\begin{aligned}
\mathrm{var}(\hat{z}) &= \int_{\Omega} \frac{z^2(x)}{\rho(x)}\, dx + \int_{\Omega}\int_{\Omega} \frac{z(x)z(y)}{\rho(x)\rho(y)} \rho^{(2)}(x,y)\, dxdy - \left(\int_{\Omega} z(x)\, dx \right)^2 \\
&= \int_{\Omega} \frac{z^2(x)}{\rho(x)}\, dx + \int_{\Omega}\int_{\Omega} \frac{z(x)z(y)}{\rho(x)\rho(y)} [\rho^{(2)}(x,y) - \rho(x)\rho(y)]\, dxdy.
\end{aligned}
$$

3. $$\widehat{\mathrm{var}}\,(\hat{z}) = \sum_{x_i \in X} \left[\frac{z(x_i)}{\rho(x_i)} \right]^2 + \sum_{x_i \in X} \sum_{\substack{x_j \in X \\ i \neq j}} z(x_i)z(x_j) \left[\frac{\rho^{(2)}(x_i, x_j) - \rho(x_i)\rho(x_j)}{\rho(x_i)\rho(x_j)\rho^{(2)}(x_i, x_j)} \right], \qquad (5.7)$$

   is an unbiased estimator of $\mathrm{var}(\hat{z})$.

Proof. First, note that the hypothesis ensures that all the integrals are finite.

1. By applying the Campbell's Formula (5.2) to $\hat{z}$, we get:

$$\mathrm{E}\left[\sum_{x_i \in X} \frac{z(x_i)}{\rho(x_i)}\right] = \int_\Omega \frac{z(x)}{\rho(x)} \rho(x)\, dx = \int_\Omega z(x)\, dx.$$

2. We have $\mathrm{var}(\hat{z}) = \mathrm{E}[\hat{z}^2] - (\mathrm{E}[\hat{z}])^2$. We thus first compute the first term and by a simple application of the Campbell's formula, we get:

$$\begin{aligned} \mathrm{E}[\hat{z}^2] &= \mathrm{E}\left[\sum_{x_i, x_j \in X} \frac{z(x_i) z(x_j)}{\rho(x_i)\rho(x_j)}\right] = \mathrm{E}\left[\sum_{x_i \in X} \frac{z^2(x_i)}{\rho^2(x_i)}\right] + \mathrm{E}\left[\sum_{x_i \in X} \sum_{\substack{x_j \in X \\ i \neq j}} \frac{z(x_i) z(x_j)}{\rho(x_i)\rho(x_j)}\right] \quad (5.8) \\ &= \int_\Omega \left(\frac{z(x)}{\rho(x)}\right)^2 \rho(x)\, dx + \int_\Omega \int_\Omega \frac{z(x) z(y)}{\rho(x)\rho(y)} \rho^{(2)}(x, y)\, dxdy \quad (5.9) \\ &= \int_\Omega \frac{z^2(x)}{\rho(x)}\, dx + \int_\Omega \int_\Omega \frac{z(x) z(y)}{\rho(x)\rho(y)} \rho^{(2)}(x, y)\, dxdy. \end{aligned}$$

   By noting that

$$\left(\int_\Omega z(x)\, dx\right)^2 = \int_\Omega \int_\Omega z(x) z(y)\, dxdy,$$

   we get the result.

3. The result is immediate from the equality between Equations (5.8) and (5.9) above.

□

We have more precise results in the case of fixed size processes. The variance of the Horvitz-Thompson estimator and its estimator can be expressed in a more concise form. The result is stated without proof in Cordy (1993).

Proposition 5.2.4 (Cordy, 1993)**.** *Let $z : \overline{\Omega} \to \mathbb{R}$ a continuous function of interest and X a point process with continuous and strictly positive first and second product densities $\rho(x)$ and $\rho^{(2)}(x, y)$. Moreover, we assume that the cardinality of X is fixed, that is $\#X = n$. Then we have*

1. $$\mathrm{var}(\hat{z}) = \frac{1}{2} \int_\Omega \int_\Omega \left(\frac{z(x)}{\rho(x)} - \frac{z(y)}{\rho(y)}\right)^2 [\rho(x)\rho(y) - \rho^{(2)}(x, y)]\, dxdy,$$

2. $$\widehat{\mathrm{var}}(\hat{z}) = \frac{1}{2} \sum_{x_i \in X} \sum_{\substack{x_j \in X \\ i \neq j}} \left[\frac{z(x_i)}{\rho(x_i)} - \frac{z(x_j)}{\rho(x_j)}\right]^2 \left[\frac{\rho(x_i)\rho(x_j) - \rho^{(2)}(x_i, x_j)}{\rho^{(2)}(x_i, x_j)}\right].$$

Lemma 5.2.1. *Let X a point process such that $\#X = n$. We assume that the product densities $\rho, \ldots, \rho^{(k)}$ exist for $k = 1, \ldots, n$. Then, if we denote by $\pi^{(n)}(x_1, \ldots, x_n)$, the (symmetric) joint probability density function of the locations of X, we have*

$$\rho^{(k)}(x_1, \ldots, x_k) = \frac{n!}{(n-k)!} \int_{\Omega^{(n-k)}} \pi^{(n)}(x_1, \ldots, x_k, y_1, \ldots, y_{n-k})\, dy_1 \ldots y_{n-k}.$$

Proof. Since we have $\#X = n$, we have that $p_k = \delta_{kn}$ and thus, the only non null term in (5.1) is for $l = n - k$ and the product density $\rho^{(k)}(x_1, \dots, x_k)$ reduces to

$$\frac{n!}{(n-k)!} \int_{\Omega^{n-k}} \pi^{(n)}(x_1, \dots, x_k, y_1, \dots, y_{n-k})\, dy_1 \dots y_{n-k}.$$

□

Corollary 5.2.1. *Let us assume that X is a finite point process such that $\#X = n$ and with product densities $\rho^{(k)}$, for $k = 1, \dots, n$. Then, we have*

$$\int_\Omega \rho^{(k+1)}(x_1, \dots, x_{k+1})\, dx_{k+1} = (n-k)\rho^{(k)}(x_1, \dots, x_k).$$

Proof. By Lemma 5.2.1, we have that

$$\rho^{(k+1)}(x_1, \dots, x_{k+1}) = \frac{n!}{(n-k-1)!} \int_{\Omega^{n-k-1}} \pi^{(n)}(x_1, \dots, x_{k+1}, y_1, \dots, y_{n-k-1})\, dy_1 \dots y_{n-k-1}.$$

By integrating, we get

$$\begin{aligned}
&\int_\Omega \rho^{(k+1)}(x_1, \dots, x_{k+1})\, dx_{k+1} \\
&= \int_\Omega \left[\frac{n!}{(n-k-1)!} \int_{\Omega^{n-k-1}} \pi^{(n)}(x_1, \dots, x_k, x_{k+1}, y_1, \dots, y_{n-k-1})\, dy_1 \dots y_{n-k-1} \right] dx_{k+1} \\
&= (n-k)\frac{n!}{(n-k)!} \int_{\Omega^{n-k}} \pi^{(n)}(x_1, \dots, x_k, x_{k+1}, y_1, \dots, y_{n-k-1})\, dx_{k+1} dy_1 \dots y_{n-k-1} \\
&= (n-k)\rho^{(k)}(x_1, \dots, x_k).
\end{aligned}$$

□

Corollary 5.2.2. *Let us assume that X is a finite point process such that $\#X = n$ and with intensity ρ and with product densities $\rho^{(k)}$, for $k = 2, \dots, n$. Then, we have*

$$\int_\Omega \rho(x)\, dx = n.$$

Proof. By trivially applying the Lemma 5.2.1 to $k = 1$ and since $\pi^{(n)}(x_1, \dots, x_n)$ is a probability density function, we get the result. □

Proof of Proposition 5.2.4. 1. It is sufficient to show that, in the case where the cardinality of X is fixed, the following equality holds:

$$\begin{aligned}
&\frac{1}{2} \int_\Omega \int_\Omega \left(\frac{z(x)}{\rho(x)} - \frac{z(x)}{\rho(x)} \right)^2 [\rho(x)\rho(y) - \rho^{(2)}(x, y)]\, dxdy \\
&= \int_\Omega \frac{z^2(x)}{\rho(x)}\, dx + \int_\Omega \int_\Omega \frac{z(x)z(y)}{\rho(x)\rho(y)} [\rho^{(2)}(x, y) - \rho(x)\rho(y)].
\end{aligned}$$

Indeed, we have:

$$\frac{1}{2} \int_\Omega \int_\Omega \left(\frac{z(x)}{\rho(x)} - \frac{z(x)}{\rho(x)} \right)^2 [\rho(x)\rho(y) - \rho^{(2)}(x, y)]\, dxdy$$

$$= \frac{1}{2}\int_\Omega\int_\Omega \left[\left(\frac{z(x)}{\rho(x)}\right)^2 - 2\frac{z(x)z(y)}{\rho(x)\rho(y)} + \left(\frac{z(y)}{\rho(y)}\right)^2\right][\rho(x)\rho(y) - \rho^{(2)}(x,y)]\,dxdy$$

$$= \frac{1}{2}\Bigg[\underbrace{\int_\Omega\int_\Omega \left(\frac{z(x)}{\rho(x)}\right)^2 \rho(x)\rho(y)\,dxdy}_{:=A} - \underbrace{\int_\Omega\int_\Omega \left(\frac{z(x)}{\rho(x)}\right)^2 \rho^{(2)}(x,y)\,dxdy}_{:=B}$$

$$+ 2\underbrace{\int_\Omega\int_\Omega \frac{z(x)z(y)}{\rho(x)\rho(y)}\left(\rho^{(2)}(x,y) - \rho(x)\rho(y)\right)\,dxdy}_{:=C}$$

$$+ \underbrace{\int_\Omega\int_\Omega \left(\frac{z(y)}{\rho(y)}\right)^2 \rho(y)\rho(x)\,dxdy}_{:=A'} - \underbrace{\int_\Omega\int_\Omega \left(\frac{z(y)}{\rho(y)}\right)^2 \rho^{(2)}(x,y)\,dxdy}_{:=B'}\Bigg].$$

Trivially, we have $A = A'$ and $B = B'$. For the term A, and by the Corollary 5.2.2, we have:

$$\int_\Omega\int_\Omega \left(\frac{z(x)}{\rho(x)}\right)^2 \rho(x)\rho(y)\,dxdy = \int_\Omega \frac{z^2(x)}{\rho(x)}\rho(x)\left[\int_\Omega \rho(y)\,dy\right]\,dx = n\int_\Omega \frac{z^2(x)}{\rho(x)}\,dx.$$

The term B can be computed by applying the Corollary 5.2.1. Indeed, we have

$$\int_\Omega\int_\Omega \left(\frac{z(x)}{\rho(x)}\right)^2 \rho^{(2)}(x,y)\,dxdy = \int_\Omega \left(\frac{z(x)}{\rho(x)}\right)^2\left[\int_\Omega \rho^{(2)}(x,y)dy\right]\,dx = (n-1)\int_\Omega \frac{z^2(x)}{\rho(x)}\,dx.$$

By putting this together, we have

$$\frac{1}{2}(A-B+2C+A-B) = A-B+C = \int_\Omega \frac{z^2(x)}{\rho(x)}\,dx + \int_\Omega\int_\Omega \frac{z(x)z(y)}{\rho(x)\rho(y)}[\rho^{(2)}(x,y)-\rho(x)\rho(y)],$$

which concludes the proof.

2. The result follows from an application of the Campbell's formula.

□

Remark 5.2.1. *The Corollary 5.2.2 can be shown by applying the Campbell's formula to the identity. Indeed, we have*

$$n = \sum_{x_i\in X} \mathbf{1}_{\{x_i\in\Omega\}} \Rightarrow \mathrm{E}\left[\sum_{x_i\in X} \mathbf{1}_{\{x_i\in\Omega\}}\right] = \int_\Omega \mathbf{1}_{\{x\in\Omega\}}\rho(x)\,dx = \int_\Omega \rho(x)\,dx.$$

Remark 5.2.2. *The estimator $\hat{z}$ is the most basic estimator for Monte-Carlo integration. In most Monte-Carlo integration schemes, the integration points are chosen independently. This does not change the estimator of the integral, since it only depends on the intensity. However, the variance and the estimator of variance depends on the repulsion.*

Chapter 6

Thinned Determinantal Point Processes for Spatial Survey Sampling

Abstract

We suggest to use a thinned Determinantal Point Processes (DPP) to generate samples which are well spread and with a possibly varying intensity. We are interested in such methodology in order to sample points on a wind turbine blade. The final aim is to estimate the aerodynamic force acting on the blade. We resort to an independent thinning to achieve a prescribed intensity while preserving the repulsion property of the original determinantal point process. The estimator of the force is unbiased and its variance can be unbiasedly estimated. [a]

[a] Part of this chapter is a joint work with Lionel Wilhelm.

6.1 Introduction

Sampling in a domain of $\mathbb{R}^d$ is an ubiquitous task arising in many different fields of statistics including Computer Experiments, Monte Carlo integration, or environmental studies just to name a few. When a function of interest is to be estimated, one usually resorts to sampling, for the obvious reason that the function cannot be measured everywhere. From a finite set of measurements, the function can be approximated, by using interpolation techniques. Sometimes only a functional of the function, such as an optimum, a mean or an integral is of interest. In these cases, a careful design of the experiment can be constructed in order to improve the accuracy. In this chapter, we address the issue of estimating the integral or the mean of a function of interest over a two dimensional domain. We address the cases where the domain of interest is planar and when it is a surface embedded in $\mathbb{R}^3$.

When the function of interest is smooth, heuristically, two very close points will give the same information. It is therefore convenient to spread the sample. This basic fact has been the starting point for the development of techniques that try to maximize the spreading of the points of measurements. Strong repulsion leads to the idea of systematic sampling in survey sampling. This can be generalized and this has been explored in the quasi-Monte Carlo (QMC) literature (Dick et al., 2013). In this case, any variance estimator of the estimate is biased, as we will see later. Instead, one often try to find a trade-off between complete randomness and strong repulsion. Complete randomness is the plain Monte Carlo method: the integral is estimated by taking the average of the function evaluated at some independently chosen random points in the domain.

We adopt a Monte Carlo prospective in that we generate our sample from a random process, while imposing repulsion in the sample. It is different from plain Monte Carlo since the pattern of the sample is not generated by independent draws from a given distribution but a repulsion is introduced. We aim at developing random processes that can efficiently handle the task of estimating integrals over complex domains but in moderate dimension, typically 2 or 3. This is in contrast with most QMC methods which aim at estimating integrals over the hypercube of a large dimension. Specifically, an ideal process must satisfy the following properties:

1. It must be easy to sample from on complex domains and to control the effect of the boundaries;
2. It must cope with a prescribed varying intensity;
3. It must be repulsive but with strictly positive second order product density;
4. The first and second order product densities must have analytical forms;

The first condition is related to the aim of conducting surveys over complex domains of interest. In particular, it is important that one can control the effect of the boundaries on the generated process. Imposing a constant intensity is a strong constraint. It is therefore natural to allow the process to be of varying intensity. We will see that this is also essential to get a stationary process on a surface defined through a non-linear mapping. Using a Monte Carlo method with a repulsive point pattern is particularly useful to improve the accuracy of the estimates and it is thus a reasonable requirement. If one uses an Horvitz–Thompson type estimator, its estimate requires the intensity of the process on the sampled points and the estimator of its variance requires the second order product density over all pairs of sampled points. We will see that stationary DPPs satisfy most of these properties. However, it is not straightforward to build kernel with a varying intensity (non-stationary kernels). A simple solution consists in applying an independent thinning to a stationary process. In this case, it results in a non-stationary DPP (Lavancier et al., 2015). In the literature of point processes, it is referred to as second order reweighted stationary process and was first introduced by (Baddeley et al., 2000). It is a common assumption for the inference of non-stationary point processes (Waagepetersen and Guan, 2009).

This work is related to a vast literature. In particular, quasi-Monte Carlo have first exploited the fact that to obtain faster convergence rates (or smaller variance) compared to usual Monte Carlo, it is necessary to spread the sample. Low-discrepancy sequences are a key ingredient and they are essentially deterministic. See for instance the comprehensive review paper of Dick et al. (2013). Randomized quasi-Monte Carlo is a randomization of a quasi-Monte Carlo sample. This ensures an unbiasedness property of the estimated integral but fails to give an analytical form of the second order product densities.

It is worth mentioning the impressive recent work of Bardenet and Hardy (2016), where DPPs are used for estimating integrals over the unit hypercube in the spirit of quasi Monte-Carlo. However, while in randomized quasi Monte-Carlo, the randomness is added after having first sampled a deterministic sample, DPPs provide a way to introduce the repulsion directly in the sample.

6.1.1 Kernels and integral operator

Let us introduce a continuous complex covariance function $K : \Omega \times \Omega \to \mathbb{C}$, that we will refer to as the kernel. It induces an integral operator $\mathcal{K} : L^2(\Omega) \to L^2(\Omega)$ defined as:

$$\mathcal{K}(f)(x) = \int_\Omega K(x,z) f(z)\, dz,$$

which is self-adjoint, semi-positive definite and trace-class (Simon, 2005, Theorem 2.12, p. 24). Using the Mercer's theorem (see for instance Gohberg et al., 2003, Theorem 3.1, p.197), we have

$$K(x,y) = \sum_{k=1}^{\infty} \lambda_k\, \phi_k(x)\overline{\phi_k(y)}, \tag{6.1}$$

where $\lambda_i \geq 0$ are the (real) eigenvalues of the operator $\mathcal{K}$. Here $\overline{\phi_k(y)}$ denotes the complex conjugate of the eigenfunction $\phi_k(y)$. This is an infinite dimensional version of the spectral theorem.

The assumption of being trace class means that the set of eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ satisfies $\sum_{k=1}^{\infty} \lambda_k < \infty$. The trace of the operator $\mathcal{K}$ is defined by

$$\text{tr}(\mathcal{K}) = \sum_{k=1}^{\infty} \langle \phi_k, \mathcal{K}\phi_k \rangle,$$

for any orthonormal family $\{\phi\}_{k=1}^{\infty}$ of $L^2(\Omega)$ and where $\langle \cdot, \cdot \rangle$ denotes the inner product on $L^2(\Omega)$ (Simon, 2005, p. 31). Under the hypothesis we consider, we have

$$\text{tr}(\mathcal{K}) = \sum_{k=1}^{\infty} \lambda_k = \int_\Omega K(x,x)\, dx,$$

where the first equality is referred to as the Lidskii theorem (Lidskii, 1959). See also Simon (2005, Theorems 3.8 and 3.9).

6.1.2 Determinantal Point Processes

A determinantal point process is defined through its product densities.

Definition 6.1.1. *A determinantal process Z defined on Ω and with kernel function K, denoted by $Z \sim DPP(K, \Omega)$, is a point process whose product densities satisfy*

$$\rho^{(n)}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = \det[\mathbf{K}(x_1, \ldots, x_n)]\, dx_1 \ldots dx_n,$$

for $x_1, \ldots, x_n \in \Omega$, $n = 1, \ldots$, and where $\mathbf{K}(x_1, \ldots, x_n)$ denotes the Gram matrix associated to the kernel K, that is the matrix whose entry ij is given by $K(x_i, x_j)$.

Theorem 6.1.1 (Macchi, 1975)**.** *Let K induces a trace class, selfadjoint positive semi-definite integral operator $\mathcal{K}$ with eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$. The kernel K defines a determinantal process on Ω if and only if $\lambda_k \leq 1$, for all $k \in \mathbb{Z}_+$.*

Remark 6.1.1. *In the case of a DPP, the product densities are given in terms of the kernel. In particular, we have:*

$$\rho^{(1)}(x) = \rho(x) = K(x,x), \text{ and } \rho^{(2)}(x,y) = K(x,x)K(y,y) - K(x,y)^2.$$

This highlights the repulsive nature of DPPs since we have:

$$g(x,y) = \frac{\rho^{(2)}(x,y)}{\rho(x)\rho(y)} = \frac{K(x,x)K(y,y) - K(x,y)^2}{K(x,x)K(y,y)} = 1 - \frac{K(x,y)^2}{K(x,x)K(y,y)} < 1.$$

Remark 6.1.2. *If we consider the restriction of the kernel K to $D \times D$, where $D \subset \Omega$ is measurable, the restriction to D of any realization of a DPP with kernel K is the realization of DPP with kernel K restricted to $D \times D$ (see, e.g. Ben Hough et al., 2010, Remark 4.2.5, p.52). Thus, if one wants to sample from a DPP in a complex region, the territory of a country for instance, it is sufficient to sample the DPP on a rectangle that encompasses the whole region and then to consider the points lying inside to region of interest. Another consequence of this fact is that DPPs do not suffer from any edge effect.*

6.2 Sampling from a DPP

Ben Hough et al. (2006) suggested the first algorithm for sampling from a DPP. It relies on the following theorem.

Theorem 6.2.1 (Ben Hough et al., 2006)**.** *Let us consider a DPP X with continuous and trace class kernel K satisfying*

$$K(x,y) = \sum_{k=1}^{\infty} \lambda_k \, \phi_k(x)\overline{\phi_k(y)},$$

and the hypotheses of theorem 6.1.1. For all $k \geq 1$, let I_k be an independent realization of a Bernoulli random variable of parameter λ_k and define the kernel $\tilde{K}$ by

$$\tilde{K}(x,y) = \sum_{k=1}^{\infty} I_k \, \phi_k(x)\overline{\phi_k(y)}.$$

If $\tilde{X}$ is a DPP defined on Ω with kernel $\tilde{K}$, then

$$X \overset{d}{=} \tilde{X}.$$

Remark 6.2.1. *The theory of DPPs is profound and particularly elegant. In this presentation, we try to avoid technicalities and restrict our attention to some very particular aspects. We refer to Soshnikov (2000), to Ben Hough et al. (2006) and to Ben Hough et al. (2010) for details. Note that most of the results stated here are valid under milder conditions.*

From this theorem, we can immediately derive the distribution of the number of points of a realization of a DPP, denoted by N hereafter. Indeed, we have

$$N \sim \sum_{k=1}^{\infty} I_k, \quad \mathrm{E}[N] = \sum_{k=1}^{\infty} \lambda_k < \infty \quad \mathrm{var}(N) = \sum_{k=1}^{\infty} \lambda_k(1-\lambda_k),$$

where the trace class assumption ensures that N is a.s. finite. Let $M = \max\{k \in \mathbb{N} : I_k = 1\}$. It is possible to simulate M by a procedure described in Lavancier et al. (2015). Then, given the spectral decomposition of K, we can simulate the independent Bernoulli variables $I_1, \ldots, I_{M-1}$ of which, say N, are equal to 1, and we relabel the corresponding eigenfunctions $\phi_1, \ldots, \phi_N$.

The projected kernel $\tilde{K}$ is given by

$$\tilde{K}(x,y) = \sum_{k=1}^{N} \phi_k(x)\overline{\phi_k(y)} = \boldsymbol{\phi}(y)^* \boldsymbol{\phi}(x),$$

where $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_N(x))^t$ and $*$ denotes the conjugate transpose.

The sampling algorithm described here relies on the idea of using a Gram-Schmidt orthogonalization procedure to be able to sample each point of the process independently. The algorithm 4 is reproduced from Lavancier et al. (2015) and has been first introduced by Ben Hough et al. (2006).

Algorithm 4 Sample from a DPP of with kernel K.

Require: $\boldsymbol{\phi}(x)$, vector of N eigenfunctions of K.;
Sample x_N from the density $f_N = \frac{\|\boldsymbol{\phi}(x)\|^2}{N}$;
Set $\mathbf{e}_1 = \boldsymbol{\phi}(x_N)/\|\boldsymbol{\phi}(x_N)\|$;
for $k = N-1, \ldots, 1$ **do**
Sample x_k from the density

$$\pi_k(x) = \frac{1}{k}\left(\|\boldsymbol{\phi}(x)\|^2 - \sum_{j=1}^{N-k} \left|\mathbf{e}_j^* \boldsymbol{\phi}(x)\right|^2\right)$$

$\mathbf{w}_k = \boldsymbol{\phi}(x_k) - \sum_{j=1}^{N-k} \left(\mathbf{e}_j^* \boldsymbol{\phi}(x_k)\right) \mathbf{e}_j$;
$\mathbf{e}_{N+1-k} = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|}$;
end for
return $\{x_1, x_2, \ldots, x_N\}$, sample from a DPP with kernel K.

6.3 Thinned Determinantal Point Process

We have seen that DPPs satisfy most of the desired properties that we have mentioned in the introduction. In order to satisfy the varying intensity without resorting to a separable intensity, we use to independent thinning. Thinning is a random procedure that deletes or retains points of an observed point process. It is a natural way to modify the intensity of a point process. The first example is the algorithm of Lewis and Shedler (1979) for simulating inhomogeneous Poisson Processes. It is also a well-known fact that independent thinning preserves the pair correlation function (Møller and Waagepetersen, 2004, Proposition 4.2).

Definition 6.3.1. *We consider a positive and continuous function $\rho : \overline{\Omega} \to \mathbb{R}$, such that*

$$\rho_{max} := \max_{x \in \overline{\Omega}} \rho(x),$$

and a stationary process $X \sim DPP(K, \Omega)$ with intensity ρ_{max}, that is satisfying

$$K(x,x) = \rho_{max}, \quad \forall x \in \Omega.$$

Let $p(x) = \rho(x)/\rho_{max} \in [0,1]$. We call p the retention probability. We define a thinned determinantal point process X_{thin} *with intensity $\rho(x)$ as*

$$X_{thin} = X \cap \{x \in \Omega : R(x) \geq p(x)\},$$

where $R(x) \overset{i.i.d}{\sim} \mathcal{U}(0,1)$, for all $x \in X$ and are independent from X too.

Proposition 6.3.1 (See for instance Møller and Waagepetersen (2004), Proposition 4.3)**.** *Let X_{thin} be a thinned process with intensity $\rho(x)$ whose maximum is $\rho_{max} := \max_{x \in \overline{\Omega}} \rho(x)$. Then, we have,*

$$\rho_{X_{thin}}(x) = \rho(x), \quad \rho^{(2)}_{X_{thin}}(x,y) = \frac{\rho(x)\rho(y)}{\rho^2_{max}} \left[K(x,x)K(y,y) - |K(x,y)|^2\right], \quad g_{X_{thin}}(x,y) = g(x,y).$$

Proposition 6.3.2 (Lavancier et al., 2015)**.** *Let $X \sim DPP(K,\Omega)$ a stationary DPP. Let X_{thin} be a thinned DPP with retention probability $p : \Omega \to [0,1]$. Then, $X_{thin} \sim DPP(\tilde{K},\Omega)$, where*

$$\tilde{K}(x,y) = p(x)K(x,y)p(y).$$

Remark 6.3.1. *The Proposition 6.3.1 gives a recipe to obtain processes that comply to the conditions 1 - 4 given in Section 6.1. Moreover, we see that the pair correlation function which is a local measure of repulsion is invariant under independent thinning. Thus, the repulsion of the initial process is preserved, which shows that the resulting process can have a varying intensity while preserving some repulsion at a smaller scale. The Proposition 6.3.2 completely characterize the thinned DPP by giving an explicit formula of its kernel. Note also that DPPs are stable by independent thinning, that is the thinned process is still determinantal.*

6.4 Simulations

In order to illustrate the effect of the repulsion on the estimation of an integral, we show some simulations. We first consider a toy example where we compute the integral over a square of a test function.

6.4.1 Square domain

In this setting, we only consider the performance of stationary DPPs in computing the integral of a given test function over a square . The test function is a weighted mixture of the distribution of two dimensional normal random variables restricted to the domain $\Omega := [0,1]^2$. Let $\boldsymbol{\mu}_1 = (0.27, 0.33)^t$, $\boldsymbol{\mu}_2 = (0.69, 0.83)^t$, $\boldsymbol{\mu}_3 = (0.85, 0.21)^t$ and $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ the following covariance matrices:

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.3 & 0.01 \\ 0.01 & 0.3 \end{pmatrix}, \boldsymbol{\Sigma}_3 = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}.$$

Then $z : \Omega \to \mathbb{R}$ is defined as:

$$z(x) = 50 \sum_{i=1}^{3} \boldsymbol{\phi}_{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}(x),$$

where $\boldsymbol{\phi}_{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$ is the density of a normal random variable $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Hence the integral of z over the square can be easily computed using the cumulative distribution function. We find

numerically that

$$\int_\Omega z(x)\, dx = 112.7816.$$

Note that since there is no thinning so that the intensity ρ is constant over the square. This is obviously a toy example but it is meant to be illustrative. The graph of the test function is shown on the left panel of the Figure 6.1.

We have used the power exponential spectral kernel (Lavancier et al., 2015). It is given implicitly by specifying its Fourier transform, which is:

$$\varphi(x) = \rho \frac{\alpha^2}{\pi \Gamma(2/\nu + 1)} \exp(-\|\alpha x\|^\nu),$$

where $\alpha = \sqrt{\Gamma(2/\nu + 1)}\frac{\pi}{\rho}$ and Γ denotes the Gamma function. The repulsiveness of this process increases, for fixed ρ, as ν increases (Lavancier et al., 2014, Appendix J) and for $\nu = 2$, it corresponds to the Gaussian covariance function. We thus sampled from a DPP with a power exponential spectral kernel with increasing values of ν to assess the effect of the repulsion on the accuracy of the estimates. We compare it with a Binomial process with a uniform density and such that the number of points of the process follows the same distribution as for a DPP with Gaussian kernel, corresponding to the case $\nu = 2$ in our simulations.

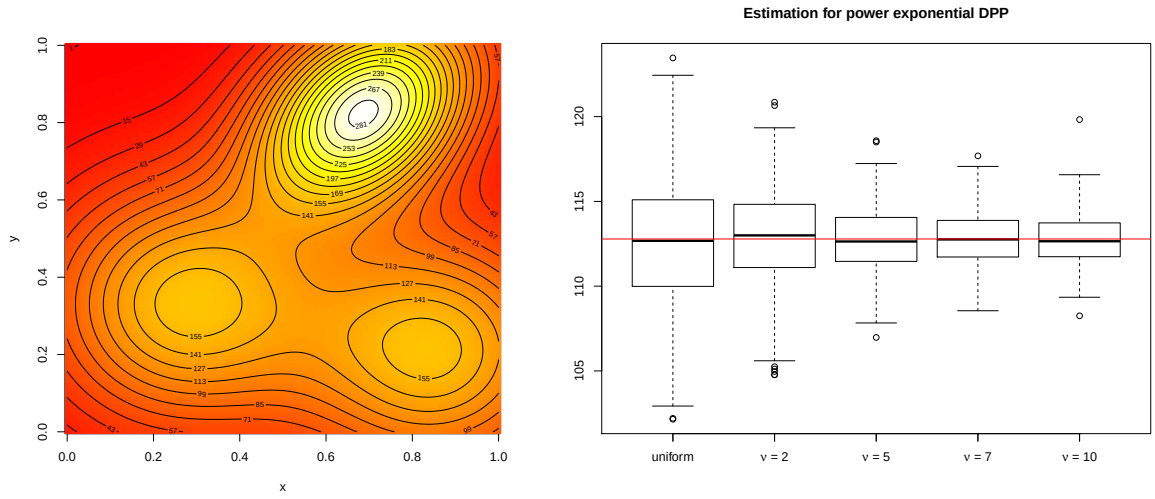

FIGURE 6.1: Test function (left) and boxplots of the estimated totals over all the simulations (right). The parameter ν varies between $\nu = 2$ and $\nu = 10$ and the expected sample size is $\rho = 250$. The horizontal line represent the true value of the total.

On the right panel of the Figure 6.1, we see the effect of the repulsion on the precision of the estimates. The corresponding simulation Rooted Mean Square Errors (RMSE) are given in Table 6.1. We see in this table that the RMSE decrease rapidly with moderate values of ν. The "uniform" estimates are obtained by computing the estimate based on a binomial process with a number of points being the same as for $\nu = 2$. In this way, we highlight the role of the repulsion on the precision of the estimates.

In the Table 6.2, the variance estimations, computed using (5.7) are shown. Even though it seems to be unbiased, the variance estimator is unstable when the repulsion is strong. In order to compute the true variances, we have to compute the integrals given in Proposition 5.2.3, for

TABLE 6.1: RMSE of simulation results with a DPP process and different values of ν.

	$\rho = 100$	$\rho = 150$	$\rho = 200$	$\rho = 250$
Uniform	5.66	5.00	4.38	3.64
$\nu = 2$	4.08	3.46	2.78	2.76
$\nu = 5$	3.18	2.50	2.22	1.87
$\nu = 7$	2.80	2.27	1.92	1.62
$\nu = 10$	2.77	2.00	1.69	1.50

different values of the parameter ν. These integrals are in 4 dimensions and have no analytical form, so we compute them using a rough Monte-Carlo integration with 10^7 random points in the domain Ω^2.

TABLE 6.2: Estimated variance of the integral over the square domain. The process is a DPP defined by the power exponential spectral kernel. We varied the expected sample size ρ and the parameter ν, which tunes the repulsion. In each cell, the simulation mean of the variance estimator with the corresponding Standard Deviations (SD) within parentheses, and the true target variance on the right.

	$\rho = 100$		$\rho = 150$		$\rho = 200$		$\rho = 250$	
	avg.$\widehat{\text{var}}$ (SD)	var	avg.$\widehat{\text{var}}$ (SD)	var	avg.$\widehat{\text{var}}$ (SD)	var	avg.$\widehat{\text{var}}$ (SD)	var
$\nu = 2$	77.78 (55.3)	81.98	56.38 (17.75)	54.63	41.94 (12.45)	40.68	32.91 (8.99)	32.7
$\nu = 5$	44.7 (52.37)	42.72	27.61 (33.37)	28.22	22.14 (17.89)	20.48	17.16 (14.5)	16.72
$\nu = 7$	34.85 (50.77)	33.69	24.39 (31.77)	22.06	15.64 (21.31)	15.94	13.23 (14.51)	12.68
$\nu = 10$	30.75 (60.35)	26.8	16.9 (34.57)	17.3	10.75 (26.92)	12.28	8.75 (22.33)	9.96

Finally, we give the 95% coverage rates of the confidence intervals based on a normal approximation in the Table 6.3.

TABLE 6.3: Coverage rates of simulation results with a DPP process and different values of ν.

	$\rho = 100$	$\rho = 150$	$\rho = 200$	$\rho = 250$
$\nu = 2$	0.94	0.98	0.99	0.99
$\nu = 5$	0.88	0.87	0.91	0.92
$\nu = 7$	0.85	0.88	0.85	0.89
$\nu = 10$	0.84	0.81	0.81	0.82

We face a trade-off between the accuracy of the estimates and the accuracy of the variance estimator. Although it is unbiased, the estimator of the variance of the Hovitz-Thompson estimator is not guaranteed to be positive. Indeed, this is a well known fact in the context of finite population sampling and the same phenomena appears in the context of a one dimensional continuous population (Wilhelm et al., 2017). As it is suggested by Wilhelm et al. (2017), the choice of the repulsion has to be made by the practitioner. The results of these simulations are in line with those of Wilhelm et al. (2017).

The methods for sampling DDPs are implemented in the `spatstat` package (Baddeley and Turner, 2005a; Baddeley et al., 2015a).

6.5 Parametrized surfaces and sampling

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain and $\Sigma \subset \mathbb{R}^3$ be a compact, connected, and oriented surface, defined by a diffeomorphism $\varphi : \Omega \to \Sigma$ such that:

$$\varphi : \Omega \subset \mathbb{R}^2 \to \Sigma \subset \mathbb{R}^3, \quad x = (x_1, x_2) \mapsto s = (s_1, s_2, s_3). \tag{6.2}$$

The Jacobian of the mapping φ, denoted by $\nabla\varphi$, is

$$\nabla\varphi : \Omega \to \mathbb{R}^{3\times 2}, \quad x \mapsto \nabla\varphi(x), \quad (\nabla\varphi)_{i,j}(x) = \frac{\partial\varphi_i}{\partial x_j}(x), \quad i = 1, 2, 3,\ j = 1, 2,$$

and the corresponding metric tensor $\mathbf{G}$ is

$$\mathbf{G} : \Omega \to \mathbb{R}^{2\times 2}, \quad x \mapsto \mathbf{G}(x), \quad \mathbf{G}(x) = [\nabla\varphi(x)]^T \nabla\varphi(x).$$

We denote by $g(x)$ the square root of the determinant of $\mathbf{G}$

$$g : \mathbf{\Omega} \to \mathbb{R}, \quad x \mapsto g(x), \quad g(x) = \sqrt{\det[\mathbf{G}(x)]}.$$

For any measurable function $f : \Sigma \to \mathbb{R}$ and any Borel set $A \subset \Omega$, the formula of change of variable yields

$$\int_{\varphi(A)} f(s)\, ds = \int_A (f \circ \varphi)(x) g(x)\, d(x).$$

Proposition 6.5.1. *Let a process X' with intensity $\tilde{\rho}(s)$ be defined on Σ. Then, the function $\rho : \Omega \to \mathbb{R}$, $\rho(x) = (\tilde{\rho} \circ \varphi)(x) g(x)$ satisfies*

$$\int_A \rho(x)\, dx = \int_{\varphi(A)} \tilde{\rho}(s)\, ds,$$

for any Borel set $A \subset \Omega$. Hence, any process X with intensity ρ defined on Ω is such that the process $\varphi(X)$ has intensity $\tilde{\rho}$ defined on Σ.

Proof. By a simple application of the formula of the change of variable, we have

$$\int_{\varphi(A)} \tilde{\rho}(s)\, ds = \int_A (\tilde{\rho} \circ \varphi)(x) g(x)\, dx.$$

□

Corollary 6.5.1. *Let X be a process defined on Ω, with intensity $\rho(x) = \varrho\, g(x)$. Then, the intensity of $\varphi(X)$ is constant and equal to ϱ.*

A similar discussion about the mapping of Poisson processes can be found in Kingman (1992, pp. 17-21). Again, these results hold under considerably weaker conditions, and this can be found in different reference books. The aim of this section is only to justify the construction of non-stationary processes on a planar domain in order to ensure that the mapping on a given surface of this process is constant.

6.6 Wind turbine blade

In this section, we develop a strategy for sampling on a wind turbine blade. The final goal is to design a wind tunnel experiment and the sample would be the locations of the measurement points of the pressure coefficient. From the measurements of the pressure coefficient, we can deduce the aerodynamic force acting on the blade. Indeed, the pressure coefficient is the relative pressure over the surface and is given by:

$$C_p(\mathbf{x}) = \frac{p(\mathbf{x}) - p_\infty}{\frac{\rho_\infty}{2}\, v_\infty^2},$$

where $p(\mathbf{x})$ is the pressure at point $\mathbf{x}$, v_∞ and p_∞ are the far field wind speed and pressure and ρ_∞ is the air density. The force $\mathbf{F}$ acting on the winglet and due to the contribution of the pressure (see, for instance, Anderson, 2010) is a multiple of the the integral over the surface of the pressure coefficient field

$$\mathbf{F} = \int_\Sigma (p - p_\infty)\ \mathbf{n}_\Sigma\, d\Sigma = \frac{\rho_\infty}{2}\, v_\infty^2 \int_\Sigma C_p\, \mathbf{n}_\Sigma\, d\Sigma,$$

where $\mathbf{n}_\Sigma$ is the unit normal vector to the surface Σ. Hence, the force is a multiple of an integral over the surface, which we aim at estimating.

The wind turbine is designed with 8 different patches of NURBS (Non Uniform Rational B-Splines) which is a generalization of B-Splines. For a very short and gentle introduction to NURBS surfaces in a similar context, see Wilhelm et al. (2016). For a comprehensive review of NURBS, we refer to Piegl and Tiller (1997). A patch is a rectangle that is included in the unit square. For each patch, a mapping is defined in order to design the surface in the space. Even though the starting computational domain Ω is a set of rectangles, the physical domain can be topologically of a different structure. Indeed, the mapping is required to be regular up to a set of null measure, which allows distortions. The design of the blade is such that some edges in the computational domain collapse to a single point in the physical domain. We report in the Figure 6.3 the adjacencies.

To design the sampling process, we have drawn a sample in the computational domain Ω which is a stratified thinned DPP. Indeed, we have sampled four samples independently, the first on the patches 1, the second on the patches 2 and 5, the third on the patches 3 and 4 and the last on the patches 7 and 8. The expected number of points of the samples before thinning are 3, 100, 100 and 10 respectively. We have sampled points on the different patches of the blade independently. The "cylinder" of the blade is composed by the patches 1 to 6, while the end of the blade is composed by the patches 7 and 8. On the Figure 6.2, the blade is displayed with the different patches. We see that the patches 2 and 5 on one hand and 3 and 4 on the other are the two faces of the blade. The patches 7 and 8 are on the top and close the blade. The patches 1 and 6 are not distinguishable and represent 0.122% and 0.002 % of the total area of the blade and this is the reason why we have not sampled any point on the patch 6. They compose front edge of the blade. Note that since the samples are independent on both faces there is repulsion within the same face but not on different faces. This is a reasonable restriction since the pressure coefficient field are almost discontinuous on the edges between the faces.

On the Figure 6.4, different views of the sample on the blade are shown. The sample is locally stationary and exhibit repulsion, as expected. For the estimation of the force, we use an

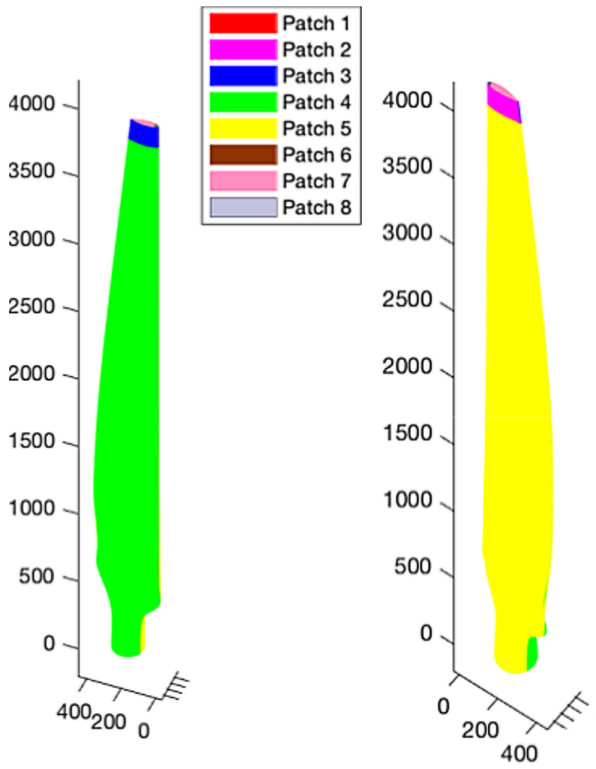

FIGURE 6.2: Wind turbine blade. The different patches are highlighted.

3.4		3.3	6.4		6.3	2.4		2.3	3.4		3.3
3.1	3	3.2	6.1	6	6.2	2.1	2	2.2	3.1	3	3.2
4.4		4.3	1.4		1.3	5.4		5.3	4.4		4.3
	4			1			5			4	
4.1		4.2	1.1		1.2	5.1		5.2	4.1		4.2

FIGURE 6.3: Scheme of the adjacencies between the patches of NURBS in the physical domain Σ.

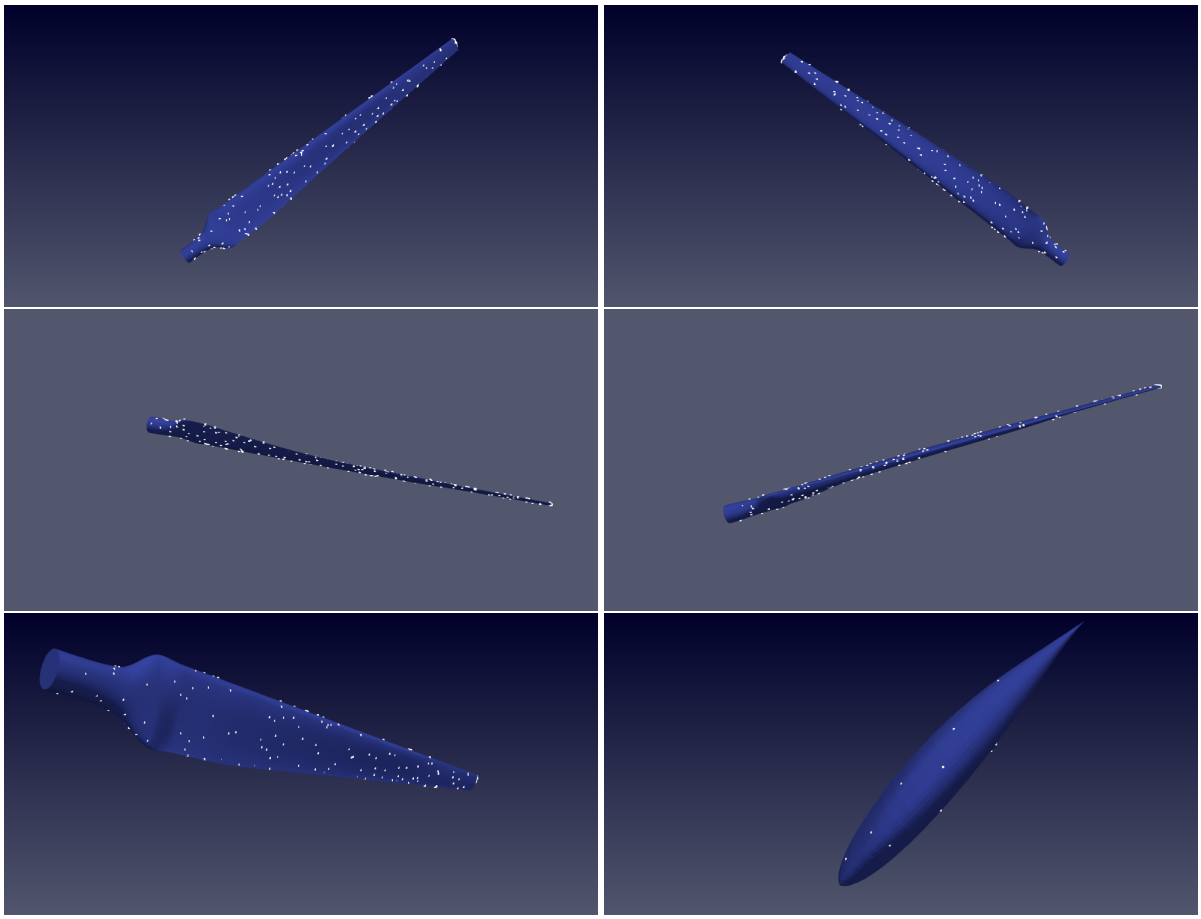

FIGURE 6.4: Multiple views of a sample on the wind turbine blade.

Horvitz-Thompson estimator componentwise. This allows us to get an unbiased estimate of the force.

The subsequent part of the project will be to assess the experimental validity of the suggested method by using Computational Fluid Dynamics (CFD) simulations. We will further explore the possibility of drawing an anisotropic sample on the blade in order to decrease the variance.

6.7 Conclusion

We have shown an application of DPPs for the purpose of survey sampling. This last Chapter is a collection of different and known methodologies, which satisfy the conditions 1 - 4 given in Section 6.1.

The conclusions about the results of the simulations are very similar to the ones of Wilhelm et al. (2017). Indeed, we face the same trade-off between the precision of the estimates and the stability of the variance estimation. A statistician has coined the term "Heisenberg Principle of the Statistics" for this phenomenon [1].

However, the great advantage of such a repulsive process is to provide unbiased and presumably highly accurate estimator. But this is similar to randomized Monte Carlo methods, which are also unbiased (see for instance Dick et al., 2013). In the literature of quasi Monte Carlo methods, the focus is on the theoretical variance. But, the variance is rarely estimated from the data to estimate the uncertainty.

We outline the different directions in which this work could be extended. Establishing a Central Limit Theorem as well as deriving convergence rates is a difficult task but would be of great interest. Based on a general result of Soshnikov (2002), Bardenet and Hardy (2016) prove a central limit theorem for special class of DPPs defined on $[0, 1]^d$. In this case, the CLT holds when the intensity of the process increases. Bardenet and Hardy (2016) also review previous works about CLT in the context of DPPs. A different approach based on the concept $\alpha-$mixing is adopted by Poinas et al. (2017), and a CLT is established for a wide class of DPP, including all the models considered here. However, the setting is a bit different since the observation window is assumed to grow, while the intensity of the process does not vary. It would be of interest to adapt this result to our setting.

One could also explore the way of deriving (hopefully sharp) bounds for the variance that could be used without resorting to second order product densities, which are problematic. This could yield a rigorous construction of conservative bounds on the estimates.

Acknowledgments

M. W. was partially supported by a Doc.Mobility fellowship of the Swiss National Science Foundation (grant no. P1NEP2_162031).

[1] Ray L. Chambers in a private conversation with Y. Tillé about stratification where a similar phenomenon arises.

Chapter 7

Conclusion

In this thesis, we have developed and studied some sampling designs and, in particular, we have considered the concept of repulsion. We have explored different situations: in finite population in Chapter 3, in one and two dimensions in chapters 4 and 6 respectively. The aim of the various sampling designs and point processes used is to improve the accuracy of the Horvitz-Thompson estimator by introducing a repulsion mechanism while maintaining positive (and known) second order inclusion probabilities (or product densities) to unbiasedly estimate the variance of the Horvitz-Thompson estimator. This allows us to make a rigorous estimation of the uncertainty without making any modelling assumption about the variable of interest. However, we saw that the estimator of the variance of the Horvitz-Thompson estimator is unstable when the repulsion is strong. We therefore recommend to introduce only little repulsion, which can greatly improve the estimate while still allowing for a reasonable quantification of the uncertainty.

In the future, the research could be extended in several directions. First, it will be very interesting to find different methods that would allow a practical and rigorous quantification of the uncertainty while being more accurate than simple random sampling, as has been suggested in the conclusion of Chapter 6.

Another question which recently arose is the use of space filling curves for spatial sampling. This is not a new topic since it had been already introduced by Stenvens Jr. and Olsen (2004). The recent paper of He and Owen (2016) highlights the potential of such a method for computing an integral on $[0,1]^d$, for $d > 1$. He and Owen (2016) first sample in the interval $[0,1]$ and then use a Hilbert space filling curve to map the interval $[0,1]$ to $[0,1]^d$. We could use the same metholodogy and apply a quasi systematic process on the interval $[0,1]$ as introduced in Chapter 4. Although in a very different context, the recent discussion paper of Gerber and Chopin (2015) also adapts quasi Monte-Carlo concepts to particle filtering, and shows improvement over the state-of-the-art methods when repulsion is added. Thus, including repulsion when sampling is both an old idea and a fruitful field of research.

In the field of finite population sampling, one of the very crucial question which is not discussed in the present work, is the problem of developing concentration inequalities for unequal probability sampling. This questions has been raised by Philip Stark and is important to be able to quantify the uncertainty without making any distributional assumption about the variable of interest.

Another interesting point is to develop balanced point processes. A continuous analogue to the cube method of Deville and Tillé (2004) would be stimulating. It is quite common to have access to spatial covariates and drawing a balanced sample could reduce the variance of the Horvitz-Thompson estimator if the field of interest is correlated with the spatial covariate.

Finally, a very promising field of research would be to consider sampling as a random measure. Indeed, in all the cases we consider, a sample is a random atomic measure, giving mass to the selected units. In the case of finite population, the state space is finite and is infinite when we consider continuous (spatial) sampling. An extension would be to develop sampling as a possibly diffuse random measure. In this case, the sample could result in some selected areas. To draw such samples, different methods could be used. One would be to consider excursion sets of well-known random fields. However, even in the most trivial case of Gaussian random fields, the excursion sets are only approximately characterized. See for instance Adler and Taylor (2007) or Adler and Taylor (2011) for a (not so) gentle introduction. A different though promising approach is to consider elliptic curves defined over some finite dimensional fields. It is a seed of an idea that I have discussed with my colleague Reda Boumasmoud. It is also related to stochastic geometry and some of the definitions have already been introduced in Chiu et al. (2013).

A personal conclusion

Doctoral studies are not only a document called a thesis. In my case, it gave me the opportunity to do many different things. First, it is a great chance and privilege to be paid for conducting research. It is a very exiting activity that necessitates and stimulates creativity. It is also extraordinary to have the possibility to spend so much time to learn. I am very grateful to the taxpayers and to the Society in general and I hope to be worthy of the money I got for this job.

I met wonderful people from all around the world, in particular when I spent some months at the University of Oxford. Science is a universal language and I really enjoyed talking, sharing ideas and debating about statistics and more generally about science.

I had also the opportunity to serve as a Phd representative for western Switzerland doctoral school in statistics. We spent some time to try to improve the way the doctoral schools were organized and to improve the participation of the Phd students. We organized in Neuchâtel along with Mihaela Anastasiade and Audrey-Anne Vallé the Young Researcher Conference, which is a one day workshop with Phd students from the western Switzerland universities.

I have been an active member of the committee of the Swiss Statistical Society for official statistics (SSS-O). I have organized a workshop geared to a professional audience. This course was organized at Swiss Federal Office of Statistics and was intended to give very broad introductions to some recent topics of statistics. The speakers were Isabel Molina Peralta (Universidad Carlos III, Madrid), Thibaut Lienart (University of Oxford), Mateo Rojas-Carulla (University of Cambridge and Max-Planck Institut, Tübingen) and Mihaela Anastasiade (Université de Neuchâtel). Around 40 people attended the course.

I had to teach two different courses. During my first two years, I taught Mathématiques Appliquées I & II, given by Anne Massiani. It was an introductory course for first year students. Then, I taught twice Time Series Analysis, given by Clément Chevalier, a master level course. I had the great chance of being a teaching assistant for great teachers, and I am very grateful for that.

Appendix A

Complement to Chapter 3

A.1 Proofs of Proposition 3.3.1 and Remark 3.3.1

Lemma A.1.1. *If $f(\cdot)$ is a probability distribution on $\{1, 2, \dots, \}$ with cumulative distribution function $F(\cdot)$, and $k, j \geq 1$, then*

$$\sum_{t=1}^{k} f^{(j+1)*}(t) = \sum_{t=1}^{k} f^{j*}(t) F(k-t).$$

Proof. Indeed, if $\mathbf{1}_A$ is the indicator function of set A,

$$\begin{aligned}
\sum_{t=1}^{k} f^{(j+1)*}(t) &= \sum_{t=1}^{k} \sum_{u=1}^{t} f^{j*}(u) f(t-u), \\
&= \sum_{t} \sum_{u} f^{j*}(u) f(t-u) \mathbf{1}_{\{1 \leq u \leq t\}} \, \mathbf{1}_{\{1 \leq t \leq k\}}, \\
&= \sum_{u} \sum_{t} f^{j*}(u) f(t-u) \mathbf{1}_{\{1 \leq u \leq k\}} \, \mathbf{1}_{\{1 \leq u \leq t\}} \, \mathbf{1}_{\{1 \leq t \leq k\}}, \\
&= \sum_{u} f^{j*}(u) \mathbf{1}_{\{1 \leq u \leq k\}} \left[\sum_{t} f(t-u) \, \mathbf{1}_{\{1 \leq u \leq t\}} \mathbf{1}_{\{1 \leq t \leq k\}} \right], \\
&= \sum_{u=1}^{k} f^{j*}(u) \left[F(k-u) - \underbrace{F(0)}_{=0} \right], \\
&= \sum_{t=1}^{k} f^{j*}(t) F(k-t).
\end{aligned}$$

□

Proof of Proposition 3.3.1. $f_0(\cdot)$ is a well-defined non-negative function on $\mathbb{N}$. It is sufficient to prove that $\sum_{k \geq 0} f(\{k+1, \dots\}) = \mu$, but

$$\begin{aligned}
\sum_{k \geq 0} f(\{k+1, \dots\}) &= \sum_{k \geq 0} \sum_{j \geq k+1} f(j), \\
&= \sum_{j \geq 0} \sum_{k \geq 0} f(j) \mathbf{1}_{k+1 \leq j}, \\
&= \sum_{j \geq 0} j \cdot f(j), \\
&= \mu.
\end{aligned}$$

As $f_0(k-t) = [1 - F(k-t)]/\mu$, to prove (3.5), it is sufficient to note that

$$\begin{aligned}
&\sum_{t=1}^{k} [1 - F(k-t)] \sum_{j=1}^{t} f^{j*}(t) \\
&= \sum_{t} \sum_{j} [1 - F(k-t)] f^{j*}(t) \mathbf{1}_{1\le t\le k} \mathbf{1}_{1\le j\le t}, \\
&= \sum_{j} \sum_{t} [1 - F(k-t)] f^{j*}(t) \mathbf{1}_{1\le t\le k} \mathbf{1}_{1\le j\le t}, \\
&= \sum_{j} \left[\sum_{t} f^{j*}(t) \mathbf{1}_{1\le t\le k} \mathbf{1}_{1\le j\le t} - \sum_{t} F(k-t) f^{j*}(t) \mathbf{1}_{1\le t\le k} \mathbf{1}_{1\le j\le t} \right], \\
&= \sum_{t=1}^{k} [1 - F(k-t)] \sum_{j=1}^{t} f^{j*}(t) \\
&= \sum_{j=1}^{k} \left[\sum_{t=j}^{k} f^{j*}(t) - \sum_{t=j}^{k} F(k-t) f^{j*}(t) \right], \\
&= \sum_{j=1}^{k} \left[\sum_{t=1}^{k} f^{j*}(t) - \sum_{t=1}^{k} F(k-t) f^{j*}(t) \right] \text{ (indeed, } f^{j*}(t) = 0 \text{ if } t < j), \\
&= \sum_{t=1}^{k} f^{1*}(t) - \sum_{t=1}^{k} F(k-t) f^{k*}(t) \text{ via lemma A.1.1}, \\
&= F(k) - \sum_{t=1}^{k} f^{(k+1)*}(t) = F(k),
\end{aligned}$$

since $f^{(k+1)*}(t) = 0$ if $t \le k$, and the result follows immediately. □

Proof of Remark 3.3.1. Consider X a random variable on $\mathbb{N}$ with finite moment of order $m+1$, $\mathrm{E}(X^{m+1})$, $m \ge 0$, and its forward transform X_F according to Definition 3.3.2. Then we can write:

$$\begin{aligned}
\sum_{k\ge 0} k^m \Pr(X_F = k) &= \sum_{k\ge 0} k^m \frac{\Pr(X \ge k)}{\mathrm{E}(X+1)} \\
&= = \sum_{k\ge 0} \sum_{i\ge k} \frac{k^m \Pr(X=i)}{\mathrm{E}(X+1)} = \frac{1}{\mathrm{E}(X+1)} \sum_{i\ge 0} \sum_{k\ge 0} \mathbf{1}_{k\le i} k^m \Pr(X=i) \\
&= \frac{1}{\mathrm{E}(X+1)} \sum_{i\ge 0} \left(\sum_{k=0}^{i} k^m \right) \Pr(X=i) = \frac{\mathrm{E}[F_m(X)]}{\mathrm{E}(X+1)},
\end{aligned}$$

where $F_m(x) = \sum_{k=0}^{x} k^m$. □

Proposition A.1.1. *The lines of matrix* $\mathbf{A}$ *with general term* a_{kt} *given at (3.14) all sum to* n.

Proof. We have

$$a_{kt} = \mathbf{1}_{t=k} + \mathbf{1}_{t<k} \sum_{j=1}^{k-t} f_j(k-t) + \mathbf{1}_{t>k} \sum_{j=1}^{N+k-t} f_j(N+k-t),$$

with $f_j(t) = 0$ if $j < t$, $t \leq 1$, $t > N$ or $j > n$. We also have that $f_n(N) = 1$ and $f_j(N) = 0$ if $j < n$. The conclusion follows from

$$\begin{aligned}
\sum_{t=1}^{N} \mathbf{1}_{t<k} \sum_{j=1}^{k-t} f_j(k-t) &= \sum_{t=1}^{N}\sum_{j=1}^{N} f_j(k-t)\mathbf{1}_{j\leq k-t}\mathbf{1}_{j\leq n}, \\
&= \sum_{j=1}^{n}\sum_{t=1}^{N} f_j(k-t) = \sum_{j=1}^{n} \Pr(S_j \leq k-1), \text{ and} \\
\sum_{t=1}^{N} \mathbf{1}_{t>k} \sum_{j=1}^{N+k-t} f_j(N+k-t) &= \sum_{t=1}^{N}\sum_{j=1}^{N} f_j(N+k-t)\mathbf{1}_{j\leq N+k-t}\mathbf{1}_{j\leq n}\mathbf{1}_{t>k}, \\
&= \sum_{j=1}^{n}\sum_{t=1}^{N} f_j(N+k-t)\mathbf{1}_{t>k} = \sum_{j=1}^{n} \left[\Pr(S_j \geq k) - f_j(N)\right].
\end{aligned}$$

□

A.2 Discrete probability distributions

Let $\mathbb{R}_+$ denote the set of positive real numbers,

$$\Gamma(r,x) = \int_x^{+\infty} t^{r-1}e^{-t}\,dt, \quad \gamma(r,x) = \int_0^x t^{r-1}e^{-t}\,dt,$$

where $r > 0, x > 0$ and

$$\mathrm{B}_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt, \qquad \mathrm{I}_x(a,b) = \frac{\mathrm{B}_x(a,b)}{\mathrm{B}(a,b)},$$

with $a > 0, b > 0, 0 < x < 1$.

TABLE A.1: Discrete distributions of probability

Name	Notation	PMF	Support	Parameters	Mean	Variance
Bernoulli	$\mathcal{Bern}(p)$	$p^x(1-p)^{1-x}$	$\{0,1\}$	$p \in [0,1]$	p	$p(1-p)$
Forward Bernoulli	$\mathcal{ForBern}(p)$	$\frac{p^x}{p+1}$	$\{0,1\}$	$p \in [0,1], n \in \mathbb{N}$	(see below the table)	
Binomial	$\mathcal{Bin}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,\dots,n\}$	$p \in [0,1], n \in \mathbb{N}$	np	$np(1-p)$
Forward Binomial	$\mathcal{ForBin}(n,p)$	$\frac{\mathrm{I}_p(x,n-x+1)}{np+1}$	$\{0,\dots,n\}$	$p \in [0,1], n \in \mathbb{N}$	(see below the table)	
Geometric	$\mathcal{G}(1-p)$	$p(1-p)^x$	$\mathbb{N}$	$p \in [0,1]$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Negative Binomial	$\mathcal{NB}(r,p)$	$\frac{\Gamma(r+x)}{x!\Gamma(r)}p^r(1-p)^x$	$\mathbb{N}$	$p \in [0,1], r \in \mathbb{N}^*$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Forward Negative Binomial	$\mathcal{ForNB}(r,p)$	$\frac{p\mathrm{I}_{(1-p)}(x,r)}{r(1-p)+p}$	$\mathbb{N}$	$p \in [0,1], r \in \mathbb{N}^*$	(see below the table)	
Poisson	$\mathcal{P}(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	$\mathbb{N}$	$\lambda \in \mathbb{R}_+$	λ	λ
Forward Poisson	$\mathcal{ForP}(\lambda)$	$\frac{1}{\lambda+1}\left[\mathbf{1}_{x=0} + \frac{\gamma(x,\lambda)}{(x-1)!}\mathbf{1}_{x\geq 1}\right]$	$\mathbb{N}$	$\lambda \in \mathbb{R}_+$	(see below the table)	
Hypergeometric	$\mathcal{H}(m,r,R)$	$\frac{\binom{r}{x}\binom{R-r}{m-x}}{\binom{R}{m}}$	$\{0,\dots,m\}\bigcap$ $\{r+m-R,\dots,r\}$	$m,r,R \in \mathbb{N}^*$, $m,r \leq R$	$\frac{mr}{R}$	$\frac{mr(R-r)}{R^2}\frac{R-m}{R-1}$
Negative Hypergeometric	$\mathcal{NH}(m,r,R)$	$\frac{\frac{\Gamma(r+x)}{\Gamma(r)x!}\frac{\Gamma(R-r+m-x)}{\Gamma(R-r)(m-x)!}}{\frac{\Gamma(m+R)}{\Gamma(R)m!}}$	$\{0,\dots,m\}$	$m,r,R \in \mathbb{N}^*$ $1 \leq R-r$	$\frac{mr}{R}$	$\frac{mr(R-r)}{R^2}\frac{R+m}{R+1}$
Uniform	$\mathcal{U}(0,a)$	$\frac{1}{a+1}$	$\{0,\dots,a\}$	$a \in \mathbb{N}$	$\frac{a}{2}$	$\frac{(a+1)^2-1}{12}$

Expectations and variances of forward distributions are easily computed in function of the first three moments of the original distribution (see Remark 3.3.1).

Bibliography

R. J. Adler and J. Taylor. *Random Fields and Geometry*. Springer, New York, 2007.

R. J. Adler and J. Taylor. *Topological Complexity of Smooth Random Functions*. Springer, New York, 2011.

D. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198. Springer, New York, 1985.

J. Anderson. *Fundamentals of Aerodynamics*. McGraw-Hill Education, New York, 2010.

E. Antal and Y. Tillé. A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106:534–543, 2011.

A. Baddeley and R. Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42, 2005a.

A. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC, London, 2015a.

A. J. Baddeley and R. Turner. `spatstat`: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42, 2005b.

A. J. Baddeley, J. Møller, and R. Waagepetersen. Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54:329–350, 2000.

A. J. Baddeley, E. Rubak, and R. P. Waagepetersen. *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC, London, 2015b.

V. Barbu and N. Limnios. *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis*. Springer, New York, 2008.

R. Bardenet and A. Hardy. Monte Carlo with Determinantal Point Processes. Preprint, 2016. URL https://arxiv.org/abs/1605.00361.

D. R. Bellhouse. Systematic sampling. In P. R. Krishnaiah and C. R. Rao, editors, *Handbook of Statistics Volume 6: Sampling*, pages 125–145, Amsterdam, 1988. Elsevier/North-Holland.

D. R. Bellhouse and J. N. K. Rao. Systematic sampling in the presence of a trend. *Biometrika*, 62: 694–697, 1975.

D. R. Bellhouse and B. C. Sutradhar. Variance estimation for systematic sampling when autocorrelation is present. *The Statistician*, 37:327–332, 1988.

J. Ben Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. *Probability Surveys*, 3:206–229, 2006.

J. Ben Hough, M. Krishnapur, Y. Peres, and B. Virág. *Zeros of Gaussian Analytic Functions and Determinantal Point Processes*. American Mathematical Society, Providence, 2010.

Y. G. Berger. Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74:149–168, 1998a.

Y. G. Berger. Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 67:209–226, 1998b.

Y. G. Berger and O. De La Riva Torres. Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78: 319–341, 2016.

C. A. N. Biscio and F. Lavancier. Quantifying repulsiveness of determinantal point processes. *Bernoulli*, 22:2001–2028, 2016.

L. Bondesson. Sampling of linearly ordered population by selection of units at successice random distances. Technical Report 25, Swedish University of agricultural sciences, Section of forest biometry, Umeå, 1986.

L. Bondesson and D. Thorburn. A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, 35:466–483, 2008.

F. J. Breidt. Markov chain designs for one-per-stratum sampling. *Survey Methodology*, 21:63–70, 1995.

F. J. Breidt and G. Chauvet. Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141:479–487, 2011.

F. J. Breidt and G. Chauvet. Penalized balanced sampling. *Biometrika*, 99:945–958, 2012.

F. J. Breidt and J. D. Opsomer. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 2017.

K. R. W. Brewer. Ratio estimation in finite populations: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5:93–105, 1963.

K. R. W. Brewer and M. E. Donadio. The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29:189–196, 2003.

K. R. W. Brewer and M. Hanif. *Sampling with Unequal Probabilities*. Springer, New York, 1983.

F.-X. Briol, C. J. Oates, M. Girolami, M A. Osborne, and D. Sejdinovic. Probabilistic integration: A role for statisticians in numerical analysis? *ArXiv preprint*. submitted, 2016.

R. E. Caflisch. Monte carlo and quasi-monte carlo methods. *Acta Numerica*, 7:1–49, 1998.

N. R. Campbell. The study of discontinuous phenomena. *Proceedings of the Cambridge Philosophical Society, Mathematical and physical sciences*, 15:117 – 136, 1909.

H. Cardot and E. Josserand. Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98:107–118, 2011.

G. Chauvet. On a characterization of ordered pivotal sampling. *Bernoulli*, 18:1320–1340, 2012.

G. Chauvet and Y. Tillé. *Fast SAS macros for balancing samples: user's guide*. Software Manual, University of Neuchâtel, 2005.

G. Chauvet, D. Bonnéry, and J.-C. Deville. Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141:984–994, 2011.

S. X. Chen, A. P. Dempster, and J. S. Liu. Weighted finite population sampling to maximize entropy. *Biometrika*, 81:457–469, 1994.

X. Chen. Poisson-binomial distribution, conditional bernoulli distribution and maximum entropy. Technical report, Department of Statistics, Harvard University, 1993.

S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic Geometry and its Applications*. Wiley, Chichester, third edition, 2013.

W. G. Cochran. Relative accuracy of systematic and stratified random samples for a certain class of population. *Annals of Mathematical Statistics*, 17:164–177, 1946.

W. G. Cochran. *Sampling Techniques*. Wiley, New York, 1977.

C. B. Cordy. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics and Probability Letters*, 18:353 – 362, 1993.

D. R. Cox. *Renewal Theory*. Methuen, London, 1962.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer, New York, 2002.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Probability and Its Applications. Springer, New York, 2008.

J.-C. Deville. Une théorie simplifiée des sondages. In P. L'Hardy and C. Thélot, editors, *Les ménages : mélanges en l'honneur de Jacques Desabie*, pages 191–214, Paris, 1989. Insee.

J.-C. Deville. Une nouvelle (encore une!) méthode de tirage à probabilités inégales. Technical Report 9804, Méthodologie Statistique, INSEE, Paris, 1998.

J.-C. Deville. Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, Rennes, 2000.

J.-C. Deville. Échantillonnage équilibré exact poisonnien. In *Actes du 8^{e} Colloque francophone sur les sondages*, pages 1–6, Dijon, 2014. Société Française de Statistique.

J.-C. Deville and C.-E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.

J.-C. Deville and Y. Tillé. Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85:89–101, 1998.

J.-C. Deville and Y. Tillé. Selection of several unequal probability samples from the same population. *Journal of Statistical Planning and Inference*, 86:215–227, 2000.

J.-C. Deville and Y. Tillé. Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912, 2004.

J.-C. Deville and Y. Tillé. Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:569–591, 2005.

P. Diaconis. Bayesian numerical analysis. In J. Berger and S. Gupta, editors, *Statistical decision theory and related topics IV*, volume 1, pages 163–175, New York, 1988. Springer.

J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The quasi-monte carlo way. *Acta Numerica*, 22:133–288, 2013.

P. D. Falorsi and P. Righi. A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology*, 34:223–234, 2008.

P. D. Falorsi and P. Righi. A unified approach for defining optimal multivariate and multi-domains sampling designs. In *Topics in Theoretical and Applied Statistics*, pages 145–152. Springer, New York, 2016.

C. T. Fan, M. E. Muller, and I. Rezucha. Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *Journal of the American Statistical Association*, 57:387–402, 1962.

W. Feller. *An introduction to Probability Theory and its applications*. Wiley, New-York, 1971.

W. A. Fuller. Sampling with random stratum boundaries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 32:209–226, 1970.

M. Gerber and N. Chopin. Sequential quasi monte carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:509–579, 2015.

V. P. Godambe and V. M. Joshi. Admissibility and bayes estimation in sampling finite populations I. *Annals of Mathematical Statistics*, 36:1707–1722, 1965.

I. Gohberg, S. Goldberg, and M. A. Kaashoek. *Basic Classes of Linear Operators, vol. I*. Birkhäuser, Basel, 2003.

A. Grafström. On a generalization of poisson sampling. *Journal of Statistical Planning and Inference*, 140:982 – 991, 2010a.

A. Grafström. Entropy of unequal probability sampling designs. *Statistical Methodology*, 7(2): 84–97, 2010b.

A. Grafström and J. Lisic. *BalancedSampling: Balanced and spatially balanced sampling*, 2016. R package version 1.5.2.

A. Grafström and N. L. P. Lundström. Why well spread probability samples are balanced? *Open Journal of Statistics*, 3:36–41, 2013.

A. Grafström and Y. Tillé. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 14:120–131, 2013.

A. Grafström, N. L. P. Lundström, and L. Schelin. Spatially balanced sampling through the pivotal method. *Biometrics*, 68:514–520, 2012.

J. Hájek. Optimum strategy and other problems in probability sampling. *Casopis pro Pestování Matematiky*, 84:387–423, 1959.

J. Hájek. *Sampling from a Finite Population*. Marcel Dekker, New York, 1981.

Z. He and A. B. Owen. Extensible grids: uniform sampling on a space filling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:917–931, 2016.

J. L. Jr. Hodges and L. Le Cam. The Poisson approximation to the Poisson binomial distribution. *Annals of Mathematical Statistics*, 31:737–740, 1960.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

R. Iachan. Systematic sampling: A critical review. *International Statistical Review*, 50:293–303, 1982.

R. Iachan. Asymptotic theory of systematic sampling. *Annals of Statistics*, 11:959–969, 1983.

K. G. Janardan and G. P. Patil. A unified approach for a class of multivariate hypergeometric models. *Sankhyā: The Indian Journal of Statistics, Series A*, 34:363–376, 1972.

R. J. Jessen. *Statistical Survey Techniques*. Wiley, New York, 1978.

J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York, 2007.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Wiley, New York, 1997.

O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, New York, 2005.

T. M. Kincaid and A. R. Olsen. *spsurvey: Spatial Survey Design and Analysis*, 2015. R package version 3.1.

J. F. C. Kingman. *Poisson Processes*. Oxford University Press, Oxford, 1992.

S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions*. Wiley, New York, second edition, 2000.

W. Kruskal and F. Mosteller. Representative sampling, I: nonscientific literature. *International Statistical Review*, 47:13–24, 1979a.

W. Kruskal and F. Mosteller. Representative sampling, II: scientific literature, excluding statistics. *International Statistical Review*, 47:111–127, 1979b.

W. Kruskal and F. Mosteller. Representative sampling, III: The current statistical literature. *International Statistical Review*, 47:245–265, 1979c.

W. Kruskal and F. Mosteller. Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review*, 48:169–195, 1980.

P. S. Laplace. *Théorie analytique des probabilités*. Imprimerie royale, Paris, 1847.

F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference: Extended version. Supplementary Materials, 2014. URL https://arxiv.org/abs/1205.4818.

F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:853–877, 2015.

J. C. Legg and C. L. Yu. Comparison of sample set restriction procedures. *Survey Methodology*, 36:69–79, 2010.

P. A. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26:403–413, 1979.

V. B. Lidskii. Non selfadjoint operators with a trace. *Doklady Akademii Nauk SSSR*, 125:488 – 587, 1959.

S. L. Lohr. *Sampling: design and analysis*. Brooks/Cole, Boston, 2009.

V. Loonis and X. Mary. Determinantal Sampling Designs. *ArXiv e-prints*, 2015.

O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7:83–122, 1975.

L. H. Madow and W. G. Madow. On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15:1–24, 1944.

W. G. Madow. On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, 20: 333–354, 1949.

D. A. Marker and D. L. Stevens Jr. Sampling and inference in environmental surveys. In *Sample surveys: design, methods and applications*, volume 29, pages 487–512. Elsevier/North-Holland, Amsterdam, 2009.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21: 239–245, 1979.

K. Meister. *On methods for real time sampling and distributions in sampling*. PhD thesis, Department of Mathematical Statistics, UmeåUniversity, 2004.

K. V. Mitov and E. Omey. *Renewal Processes*. Springer, New York, 2014.

J. Møller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC, London, 2004.

J. Møller and R. P. Waagepetersen. Modern Statistics for Spatial Point Processes. *Scandinavian Journal of Statistics*, 34:643–684, 2007.

J. E. Moyal. The general theory of stochastic population processes. *Acta Mathematica*, 108:1–31, 1962.

M. N. Murthy and T. J. Rao. Systematic sampling with illustrative examples. In P. R. Krishnaiah and C. R. Rao, editors, *Handbook of Statistics Volume 6: Sampling*, pages 147–185, Amsterdam, 1988. Elsevier/North-Holland.

R. D. Narain. On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–174, 1951.

D. Nedyalkova and Y. Tillé. Optimal sampling and estimation strategies under linear model. *Biometrika*, 95:521–537, 2008.

J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97: 558–606, 1934.

J. Neyman. Contribution to the theory of sampling human population. *Journal of the American Statistical Association*, 33:101–116, 1938.

A. B. Owen. Orthogonal arrays for computer experiment integration and visualization. *Statistica Sinica*, 2:439–452, 1992.

J. Pea, L. Qualité, and Y. Tillé. Systematic sampling is a minimal support design. *Computational Statistics and Data Analysis*, 51:5591–5602, 2007.

L. Piegl and W. Tiller. *The NURBS Book*. Springer-Verlag, New York, 1997.

A. Poinas, B. Delyon, and F. Lavancier. Mixing properties and central limit theorem for associated point processes. *ArXiv preprint*, pages 1 – 22, 2017.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

S. I. Resnick. *Adventure in Stochastic Processes*. Birkhäuser, Boston, 1992.

B. D. Ripley. The second order analysis of spatial point patterns. *Journal of Applied Probability*, 13:255–266, 1976.

B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:172–212, 1977.

S. Rousseau and F. Tardieu. La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Technical report, Insee, Paris, 2004.

R. M. Royall. On finite population sampling theory under certain linear regression models. *Biometrika*, 57:377–387, 1970a.

R. M. Royall. Finite population sampling - On labels in estimation. *Annals of Mathematical Statistics*, 41:1774–1779, 1970b.

R. M. Royall. The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71:657–664, 1976a.

R. M. Royall. Likelihood functions in finite population sampling theory. *Biometrika*, 63:605–614, 1976b.

R. M. Royall. The model based (prediction) approach to finite population sampling theory. In M. Ghosh and Pramod K. P., editors, *Current issues in statistical inference: Essays in honor of D. Basu*, volume 17 of *Lecture Notes-Monograph Series*, pages 225–240. Institute of Mathematical Statistics, 1992.

R. M. Royall and J. Herson. Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68:880–889, 1973a.

R. M. Royall and J. Herson. Robust estimation in finite populations II: Stratification on a size variable. *Journal of the American Statistical Association*, 68:890–893, 1973b.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

T. J. Santner, B. J. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer, New York, 2003.

C.-E. Särndal, B. Swensson, and J. H. Wretman. *Model Assisted Survey Sampling*. Springer, New York, 1992.

A. R. Sen. On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127, 1953.

B. Simon. *Trace Ideals and their Applications*. American Mathematical Society, Providence, second edition, 2005.

A. Soshnikov. Determinantal random point fields. *Russian Mathematical Surveys*, 55:923 – 975, 2000.

A. Soshnikov. Gaussian limit for determinantal random point fields. *The Annals of Probability*, 30:171–187, 2002.

C. Stein. Application of Newton's identities to a generalized birthday problem and to the Poisson-Binomial distribution. Technical report, Department of Statistics, Stanford University, 1990.

D. L. Stenvens Jr. and A. R. Olsen. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99:262–278, 2004.

D. L. Stevens Jr. and A. R. Olsen. Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics*, 4:415–428, 1999.

D. L. Stevens Jr. and A. R. Olsen. Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14:593–610, 2003.

P. V. Sukhatme. *Sampling Theory of Surveys, with Applications*. Indian Society of Agricultural Statistics, New Delhi, 1954.

B. Tang. Orthogonal array-based latin hypercubes. *Journal of the American Statistical Association*, 88:1392–1397, 1993.

G. R. Terrell. *Mathematical Statistics: A Unified Introduction*. Springer, New York, 1999.

S. K. Thompson. *Sampling*. Wiley, New York, 2012.

Y. Tillé. A moving stratification algorithm. *Survey Methodology*, 22:85–94, 1996.

Y. Tillé. *Sampling Algorithms*. Springer, New York, 2006.

Y. Tillé and A.-C. Favre. Coordination, combination and extension of optimal balanced samples. *Biometrika*, 91:913–927, 2004.

Y. Tillé and A.-C. Favre. Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74:31–37, 2005.

Y. Tillé and A. Matei. *sampling: Survey Sampling*, 2015. R package version 2.7.

Y. Tillé and M. Wilhelm. Probability sampling designs: Principles for choice of design and balancing. *Statistical Science*, 32:176–189, 2017.

Y. Tillé, L. Qualité, and M. Wilhelm. Sampling designs on finite populations with spreading control. *Statistica Sinica*, pages 1–25, 2017.

R. Valliant, A. H. Dorfman, and R. M. Royall. *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York, 2000.

R. Valliant, J. A. Dever, and F. Kreuter. *Practical tools for designing and weighting survey samples*. Springer, New York, 2013.

J. S. Vitter. Faster methods for random sampling. *Communications of the ACM*, 27:703–718, 1984.

J. S. Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11: 37–57, 1985.

J. S. Vitter. An efficient algorithm for sequential random sampling. *ACM Transactions on Mathematical Software*, 13:58–67, 1987.

R. Waagepetersen and Y. Guan. Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:685–702, 2009.

M. Wilhelm, L. Dedè, L. M. Sangalli, and P. Wilhelm. IGS: an IsoGeometric approach for Smoothing on surfaces. *Computer Methods in Applied Mechanics and Engineering*, 302:70 – 89, 2016.

M. Wilhelm, L. Qualité, and Y. Tillé. Quasi-systematic sampling from a continuous population. *Computational Statistics and Data Analysis*, 105:11–23, 2017.

F. Yates and P. M. Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 15: 235–261, 1953.

Remerciements

Les remerciements sont l'occasion de m'offrir un dernier regard sur ce qui a fait ces quelques années, mais aussi ce qui m'a permis d'y parvenir. Je commencerai donc par exprimer mon infinie gratitude à mes parents.

Mes grands-parents ont toujours été une référence pour moi. Je les remercie d'avoir partagé avec nous ce qu'étaient leurs vies et de m'avoir fait connaître d'où je viens.

Mes chers frères ont été un soutien et une motivation pendant toutes ces années. De la première fois où je suis venu à l'EPFL grâce à Pierre à ces dernières semaines où l'on a collaboré avec Lionel, la présence de mes frères aura littéralement jalonné ma vie d'étudiant et de chercheur. Je leur remercie de m'avoir permis d'agréablement conjuguer travail, conseils et bons moments.

Mon directeur de thèse, Yves Tillé est un homme formidable. Il a toujours été un soutien encourageant, une source intarissable d'idées, un optimiste invétéré et une âme d'une sincère bienveillance. J'ai eu un plaisir fou à faire de la recherche pendant ces années et c'est à lui que je le dois.

Lionel Qualité est un collègue extraordinaire. Son esprit brillant a contribué à rendre ma recherche très agréable. Il m'a aussi patiemment supporté dans son bureau. Qu'il soit remercié pour ce qu'il m'a appris.

Ho tanto imparato lavorando con Luca Dedè sul mio primo articolo. Lo ringrazio anche di avere suscitato il mio entusiasmo per la ricerca e di avermi preso sul serio.

Finalement, je remercie Arnaud Doucet, qui m'a permis de passer quelques mois extraordinaires à Oxford. Je le remercie aussi pour ces longues discussions qui auront formé ma vision que j'ai des statistiques. Finalement, c'est un honneur pour moi qu'il ait accepté de faire partie de mon jury.

Je remercie mes amis de longue date, sans qui les mathématiques n'auraient pas eu la même saveur: Adrien, Denis et Thomas. Je remercie aussi mes camarades d'études, devenus des amis et avec qui j'ai partagé tant de bons moments ces dernières années: David, Pauline et Reda.

I would like to thank my friends I met while visiting in Oxford, Niko and Thibaut, from whom I learned so much in many passionate debates. Among many other things, you made my stay an amazing experience. I also would like to thank Marco who was a great office mate there.

Je suis reconnaissant à Marina, amie de très longue date qui m'a soutenu tout au long de ces années.

Je remercie chaleureusement mes collègues Audrey-Anne et Mihaela de leur amitié et de leur bienveillance. Je remercie Clément et Jean-Marc dont les arrivées parmi nous furent source de dynamisme et de bonne humeur ainsi qu'Odile, qui m'a permis de partir en Angleterre l'esprit serein et qui est une amie au long cours. Merci à Monique de son agréable présence. Je remercie aussi Vittoria dont la présence lumineuse dans notre groupe a toujours été appréciée. Merci à Caren de m'avoir initié à la fête des vendanges, sortie de boîte officieuse de notre groupe. Ti sono molto grato, Maria Michela per queste settimane purtroppo fredde ma così divertenti che hai trascorso da noi. I giorni straordinari passati a Roma insieme rimarranno un bellissimo riccordo. La tua tazzina di Procida è il testimone dei numerosi caffè presi durante questi anni. Finalement, un grand merci à Corine, dont la présence chaleureuse anime régulièrement nos matinées et qui est toujours disponible pour organiser les barbecues estivaux.

Les écoles doctorales auront été une facette marquante de ces années. Je remercie les collègues doctorants pour tous ces fructueux échanges. Je pense notamment à Anjeza, Claudio, Daniel, Hélène, Marie-Hélène, Raphaël, Yoav. Finalement, je tiens à remercier tout particulièrement mes amis Thomas, Sebastian, Peiman et Sophie qui ont eu la gentillesse de m'offrir un asile dans leur bureau si souvent lorsque je venais travailler à l'EPFL. Ces journées passées auprès d'eux auront été autant agréables que productives et je leur en suis reconnaissant. Je ne saurais oublier Yan-Kim, seul capable de prendre un café à 7h30 à Sat.

Finalement, je te remercie, ma chère Marie, d'être cette présence qui rend mon quotidien si joyeux et mon existence si heureuse.