View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Archivio della ricerca- Università di Roma La Sapienza

SAPIENZA
UNIVERSITÀ DI ROMA

Analytic Inference in Finite Population Framework Via Resampling

Scuola Di Dottorato In Scienze Statistiche

Dottorato di Ricerca in Statistica Metodologica – XXIX Ciclo

Candidate

Alberto Di Iorio
ID number 1596485

Thesis Advisor

Prof. PierLuigi Conti

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Methodological Statistics

October 2016

Thesis defended on TBA
in front of a Board of Examiners composed by:

TBA (chairman)
Caterina Conigliani
Pietro Coretto
Maria Giovanna Ranalli
REFEREE: Fulvia Mecatti
REFEREE: Pier Francesco Perri

Analytic Inference in Finite Population Framework Via Resampling
Ph.D. thesis. Sapienza – University of Rome

© 2016 Alberto Di Iorio. All rights reserved

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: January 17, 2017

Author's email: alberto.diiorio@uniroma1.it

Abstract

The aim of this dissertation is to provide nonparametric tools for analytic inference on superpopulation models. To pursue the goal we approach to the problem in two different ways. The first one is analytic. Following the classical empirical process theory, we first derive a functional central limit theorem that fully characterizes the asymptotic distribution of the Hàjek estimator of the distribution function of the superpopulation. In addition, assuming some regularity conditions on the (superpopulation) parameters of interest, we extend this analytic characterization to a large class of possible paramaters of interest. The second one is more "practical": our aim is to construct a computer intensive procedure that allows for inferring the superpopulation, also when the (asymptotic) distribution of an interest parameter has an unmanageable analytic form. Clearly, such a procedure is resampling. Unfortunately, the most famous resampling technique, the bootstrap procedure, does not work in our framework. In fact, even if a superpopulation is assumed, the selected units cannot be assumed independent in the presence of a non trivial sampling design when dealing with a finite population. This lack of independence makes the classic bootstrap inadequate for our purposes. Of course, in the survey sampling literature, many resampling procedures have been proposed, but them do not suit our purposes because of two reasons: *i*) a largest part of these resampling techniques have been developed to infer the finite population and not the superpopulation; *ii*) we want to make a parallel between the classical non parametric theory and survey sampling. Almost all of these procedures are justified by mimicking the first two moments of the distribution of the considered estimator, and this is not the argument used to justify Efron's bootstrap in classical nonparametric statistics. Thus, we introduce the " multinomial" scheme as a resampling procedure for the superpopulation and we provide an asymptotic validation of this method, that involves the whole distribution of the considered estimators, exactly as it happens for classic bootstrap. In the last part of this work, the results obtained are applied to different inferential problems and, for each one of the concerned problem, a simulation

study is performed to test the validity of our proposal. For these applications, we especially focused on problems where the interest parameter is not a linear function of the data.

Acknowledgments

Vorrei principalmente ringraziare il mio Supervisore il Prof. Conti per il supporto, i consigli e la passione che mi ha trasmesso in questi anni. Un sentito ringraziamento va anche ai Referee la Prof.ssa Fulvia Mecatti e il Prof. Pier Francesco Perri per gli utili suggerimenti e per la disponibilità dimostrata nel leggere questa tesi. Non posso non ringraziare i miei compagni di "viaggio", senza di loro questo percorso in questa Università per me nuova non sarebbe stato lo stesso; tra questi, un ringraziamento speciale va a Giorgia per essere stata, fin da subito, ben più di una semplice collega. Infine ringrazio la mia famiglia per essere sempre stata al mio fianco, quali che siano state le scelte che ho intrapreso nel corso della mia vita.

Contents

Introduction

The use of superpopulation models in survey sampling has a long history, going back (at least) to Cochran (1939), where the limits of assuming the population characteristics as *fixed*, especially in economic and social studies, are stressed. As clearly appears, for instance, from Särdnal et al. (1992) and Pfeffermann (1993), there are basically two types of inference in the finite populations setting. The first one is *descriptive* or *enumerative* inference, namely inference about finite population parameters. This kind of inference is a static "picture" on the current state of a population, and does not take into account the mechanism generating the characters of interest of the population itself. The second one is *analytic* inference, and consists in inference on superpopulation parameters. This kind of inference is about the process that generates the finite population. In contrast with *enumerative* inference results, *analytic* ones are more general, and still valid for every finite population generated by the same superpopulation model.

The present dissertation essentially focuses on providing nonparametric estimates for superpopulation parameters. To this purpose we first characterize the asymptotic distribution of the Hàjek estimator (or equivalently Horvitz-Thompson esimator) for a large class of parameters. However, depending on the considered parameter of interest, this characterization could be quite unmanageable in practical situation. In fact, the analytic variance of the estimated parameter, may have a really complex form. This drawback is overtaken by resorting to a well known idea of classical nonparametric statistics; approximate the estimator distribution *via* bootstrap (cfr. Efron (1979), Romano (1988), Romano (1989) and references therein). Efron's bootstrap procedure (Efron (1979)) is based on a crucial assumption: data are independent and identically distributed ($i.i.d$). Unfortunately, this is not the case of finite population framework, where the presence of a complex sampling design induces dependences in the data. A deeper discussion on this "unlucky" situation will be faced in Chapter 3, but it is important to say that the

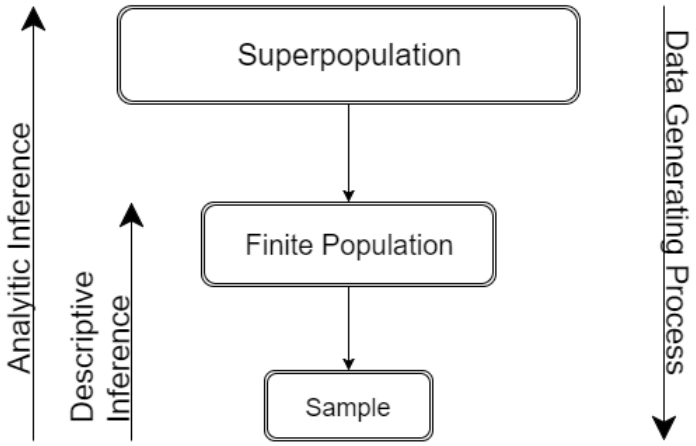


Fig.0.1. Diagram summarizing analytic inference and descriptive inference in superpopulation approach.

main worst consequence of these dependences is the *inconsistency* of the original bootstrap procedure.

Because of the bootstrap failure in survey sampling, several different resampling techniques in finite populations setting have been proposed in the literature. However no one of these procedures suit our purposes. In fact, a large portion of such techniques essentially refers to descriptive inference, and rests on the idea of mimicking the moments of the sampling distributions of statistic of interest. In particular, in case of Horvitz-Thompson estimator of the population mean, this idea reduces to require that the variance of the resampled statistic should be equal (or at least very close) to the variance estimate of the original statistic. This is usually attained by resampling units according to some special sampling design that takes into account the dependence between units: cfr. Antal and Tillé (2011) and references therein.

In addition, the arguments above are considerably different from those commonly used to justify the classical bootstrap, that are based on asymptotic considerations involving the whole sampling distribution of a statistic, not only the first two moments. In particular, in Bickel and Freedman (1981), usual Efron's bootstrap is justified by proving that the asymptotic distribution of a bootstrapped statistic coincides with that of the original statistic. To our knowledge, the only papers that develop resampling methods for finite populations justified *via* asymptotic arguments are Chatterjee (2011), Conti and Marella (2015), Conti et al. (2015). All the

above mentioned papers are based on the fixed population approach, *i.e.* refer to the estimation of finite population parameters (descriptive inference). Furthermore, Chatterjee (2011) is confined to quantile stimation under simple random sampling. The results are then extended to general πps designs in Conti and Marella (2015).

In Conti et al. (2015) a class of resampling procedures based on a predictive approach is defined, and their asymptotic distribution is studied. Such procedures are essentially taylored for the estimation of finite population parameters, in a descriptive inference perspective. For this reason Conti et al. (2015), represents our starting point. In this dissertation, we will generalize the results in Conti et al. (2015) to analytic inference. As it will be seen in the sequel, the analytic-inference perspective dramatically changes the asymptotic distributions to be considered. As a consequence, the resampling procedures defined in Conti et al. (2015) do not work when superpopulation parameters are involved; the only exception is the so-called "multinomial" approach, defined first in Pfeffermann and Sverchkov (2006).

The present dissertation is organized as follows. In Preliminaries Chapter, we cite some well-known definitions and results about different fields (*Probability*, *Empirical Processes*, *Sampling Theory*), that will be used in the whole work. In Chapter 1 we define the framework in which we will develop our theory, specifing how the asymptotic is made, how the superpopulation generates the finite population and deeply analizing the key assumption on how the data are collected from the finite population. In the second part of the chapter, we define the class of parameters on which we are interested and we discuss what kind of regularity conditions are needed on these paramaters in order to derive easily the asymptotic results. The first part of Chapter 2 is devoted to a review of Conti et al. (2015) that is the foundation on which we build our methodology. The second section of Chapter 2 contains the first crucial results, that addresses our first aim. In fact, in Propositions 2.2.2 we fully characterize the asymptotic distribution of the Hàjek estimator of superpopulation distribution function. Then, thanks to the regularity conditions introduced in Chapter 1 we also characterize the asymptotic distribution of Hajek estimator of interest parameters in Proposition 2.2.5. The last Section of Chapter 2 contains a review of Boistard et al. (2015) where the authors obtain a result substantially equivalent to Proposition 2.2.2, with some remarkable differences between our and their methodology. The main goal of Chapter 3 is to address the aim of providing a bootstrap-like procedure to avoid the explicit finding of the limiting distribution of a interest parameter. The first Section is devoted to briefly summarize the most

contributions to the literature of the last thirty years about resampling in survey theory. In the second Section we introduce the "multinomial" resampling scheme as a procedure to recover the asymptotic distribution of the Hajek estimator of a superpopulation parameter. In addition, in the same spirit of Bickel and Freedman (1981) we prove the validity of this procedure on the basis of asymptotic considerations. In the last Section, we introduce the Monte Carlo procedure to practically implement the resampling procedure proposed. In the Application Chapter we show some of the possible applications of our method to common statistical problems, with a special attention on measuring inequality. At the end, in the Appendix, the proofs of all the original results contained in the present work are showed.

Preliminaries

The aim of this chapter is to refresh some well known results that will be useful to a better understanding of this dissertation as well as to fix a large part of the notation that will be employed in the sequel. The first part is devoted to some classical results of probability theory. The second part is devoted to empirical process under the classical assumption of $i.i.d$ data. The last section is about some results of sampling theory in finite population framework that are massively used in this work.

0.1 Probability

In this section we briefly analyze a particular functional space and its basic probabilistic and topological properties. Using the same notation of Billingsley (1968) let $D[-\infty, +\infty]$ be the space of the *càdlàg* (French: "Continue à droite, limite à gauche", that means right continuous with left limits) functions on the extended real line. By definition, of course, every continuous function is in $D[-\infty, +\infty]$, that is $C[-\infty, +\infty] \subset D[-\infty, +\infty]$. It is also evident by construction that Cumulative Distribution Functions are *càdlàg* functions. For the sake of brevity we will use some symbols to indicate different types of convergences.

$\xrightarrow{weak}$ expresses *weak convergence*.

$\xrightarrow{probability}$ expresses *convergence in probability*. In the sequel the superscript might contain explicitly the distribution to which the convergence refers, for example $\xrightarrow{\mathbb{P}-probability}$.

$\xrightarrow{\mathcal{D}}$ expresses *convergence in distribution*. Although the meaning is the same, it is used instead of $\xrightarrow{weak}$ when dealing with succession of random variables.

$\xrightarrow{a.s.}$ expresses *almost sure convergence*. In the sequel the superscript might contain explicitly the distribution to which the convergence refers, for example $\xrightarrow{a.s.-\mathbb{P}}$.

In the whole work we will consider the space $D[-\infty, +\infty]$ endowed with the *Skorohod Topology*. The *Skoroohod topology* is a generalization of the *uniform topology* (the one generated by the sup norm metric) usually used with the space $C[-\infty, +\infty]$. In fact the *Skorohod topology* relativized to C coincides with the *uniform topology* (for more formal definitions see Billingsley (1968) chapters 2-3). The reason why we consider $D[-\infty, +\infty]$ endowed with the *Skorokhod Topology* rather than the *uniform convergence topology* is that the first one makes the space D *separable* and *complete* and, in addition, it can be seen as generated by a metric called the *Skorokhod metric*. In order to make more easily understandable the rest of the present section a few definitions, remarks and results are given in the sequel.

- A topological space $(S, \mathcal{S})$ is *separable* if there exist a dense, countable subset A contained in it. In a more intuitive way, a space is *separable* if you can approximate every element of such a space with a countable sequence of elements of A. For example $\mathbb{R}$ is a separable space because $\mathbb{Q} \subset \mathbb{R}$ and $\mathbb{Q}$ is countable and dense in $\mathbb{R}$. In fact you can describe every real number by a countable sequence of rational numbers. Essentially this property implies that you do not need all the elements of a space to describe some properties of it.

- A metric space (S, d) is *complete* if every Cauchy sequence in S converges in S. In a less formal language, this means that every "well behaved" sequence of points in the space must converge in the same space. In this way the space has no *holes*. For example the open interval $(0, 1)$ with the absolute value metric is not complete. In fact the sequence $a_n = \frac{1}{n}$ is a Cauchy sequence, it converges to 0, but 0 is not in $(0, 1)$

Separability and completeness of the considered (metric) space are necessary assumptions of the Prohorov's theorem, that follows

Theorem 0.1.1 (Prohorov)**.** *Let* Π *be a family of probability measures on a complete, separable metric space* S *(that is* S *is a Polish space). The family* Π *is tight if and only if it is relatively compact.*

With reference to Prohorov's Theorem two remarks are necessary.

- A family of probability measures Π on a generic metric space S is *tight* if for every $\epsilon > 0$ there exist a compact subset K of S such that $\mathbb{P}(K) > 1 - \epsilon$ for every $\mathbb{P} \in \Pi$.

- A family of probability measures Π on a measure space $(S, \mathcal{S})$ ($\mathcal{S}$ is the Borel σ-algebra) is *relatively compact* if every sequence of elements of Π contains a weakly convergent subsequence. Formally for every sequence $\{\mathbb{P}_n\} \subset \Pi$ there exist a subsequence $\{\mathbb{P}_{n'}\} \subset \{\mathbb{P}_n\}$ and a probability measure Q defined on $(S, \mathcal{S})$ such that $\mathbb{P}_{n'} \xrightarrow{weak} Q$.

Usually to show weak convergence of a sequence of random elements of D to some limit process, we have to show that the finite-dimensional distributions of the considered processes converge to the finite-dimensional distribution of the limit process and that the distributions of these processes are tight (see Theorem 15.1 p.124 Billingsley (1968)). Prohorov's theorem provides an operational tool to show the tightness, since it ties the notion of *tightness* to the more operational notion of *relative compactness*. Separability is also a necessity for another important result, known as Skorohod's representation Theorem.

Theorem 0.1.2. *Suppose that $\mathbb{P}_n \xrightarrow{weak} \mathbb{P}$ and that $\mathbb{P}$ is defined on a separable space. Then there exist a sequence of random variable X_n and a variable X defined on a common probability space, such that $\mathbb{P}_n$ is the the distribution of X_n for every n and $\mathbb{P}$ is the distribution of X. Moreover it holds that $X_n \xrightarrow{a.s.} X$.*

Weak convergence allows the probability space to be different for every considered probability, while for a stronger form of convergence, like almost sure convergence, this is not true. What is generally true, in fact, is that almost sure convergence implies weak convergence; the converse is generally false. Skorohod's Representation theorem, makes the converse of the last proposition almost true, ensuring the existence of a *representation* of the weak convergence as an "almost" sure convergence.

The last results that we want to recall here is a Theorem that matches together two types of convergences, the convergence in distribution and the convergence in probability. This Theorem is known as Slutsky Theorem (see Corollary 2. in Billingsley (1968) p. 31)

Theorem 0.1.3. *Given two sequences of random variables Y_n and X_n, suppose that $Y_n \xrightarrow{\mathcal{D}} Y$ where Y is a random variable and $X_n \xrightarrow{probability} c$ where c is a constant.*

Then it holds that:

$$Y_n \times X_n \xrightarrow{\mathcal{D}} Y \times c \tag{0.1}$$

$$\frac{Y_n}{X_n} \xrightarrow{\mathcal{D}} \frac{Y}{c} \text{ if } c \neq 0 \tag{0.2}$$

What this last Theorem states is quite important. Generally we have that if two sequences of random variables converges in distribution to a non degenerate random variable, it is not true that the product (or equivalently the ratio) of the considered sequences converges to the product (or the ratio) of the limit random variable. The Slutsky Theorem ensures that this convergence holds if one of the considered sequence converges in probability to a constant or equivalently to the convergence in distribution to a degenerate random variable.

0.2 Empirical Processes

In this section we will provide some definitions and theorems related to the empirical process, that is a well known topic of classic non-parametric statistics. If not differently specified, we will identify a virtual or infinite population by its distribution function. Consider a sample of independent, real valued random variables $Y_1, \ldots, Y_n$ with common distribution F.

The most known non-parametric estimator of the distribution function is *Empirical Cumulative Distribution Function* (ECDF), which is defined as:

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I_{(Y_i \leq y)}, \tag{0.3}$$

where

$$I_{(Y_i \leq y)} = \begin{cases} 1 & \text{if } Y_i \in (-\infty, y] \\ 0 & \text{otherwhise} \end{cases}$$

are i.i.d. Bernoulli random variables with:
$\mathbb{E}[I_{\{Y_i \leq y\}}] = F(y)$
$\mathbb{V}[I_{\{Y_i \leq y\}}] = F(y)(1 - F(y))$.
The symbols $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denote, respectively, the expected value and the variance with respect to the distribution of the population.

Clearly, for every fixed $y \in \mathbb{R}$, $F_n(y)$ is itself a random variable with

$$\mathbb{E}[F_n(y)] = F(y) \tag{0.4}$$

$$\mathbb{V}[F_n(y)] = \frac{F(y)(1-F(y))}{n}. \tag{0.5}$$

Concerning the pointwise convergence (remember that y is fixed) of the random sequence $F_n(y)$, by the strong law of the large numbers, we have:

$$F_n(y) \xrightarrow{a.s.} F(y). \tag{0.6}$$

Thus, asymptotically $F_n(y)$ converges almost surely to the true value $F(y)$, that is $F_n(y)$ is a (strongly) consistent estimator of $F(y)$. Focusing on the limit distribution of the random sequence $F_n(y)$, using the Central Limit Theorem it is immediate to see that:

$$\sqrt{n}\frac{F_n(y) - F(y)}{F(y)(1-F(y))} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)\,, \ \forall y \in \mathbb{R}. \tag{0.7}$$

We want to highlight that if we do not fix y, the whole F_n is a random function and, by construction, is a random element of the space $D[-\infty, +\infty]$. In particular, the process obtained centering and opportunely scaling the random function F_n is well known in non-parametric statistic and it is referred to as empirical process. In details the empirical process is defined as

$$\sqrt{n}\,(F_n - F) \tag{0.8}$$

where the scaling factor $\sqrt{n}$ and the centering factor F are obvious in virtue of (0.4), (0.5). We now proceed with a characterization of the objects defined above.

We start with a result that is known as the Glivenko-Cantelli Theorem or also as the *Fundamental Theorem of Statistics* (for a proof see Van der Vaart (2000) p.266).

Theorem 0.2.1 (Glivenko-Cantelli)**.** *If $X_1, \ldots, X_n$ are independent random variables with common distribution F, then*

$$\|F_n - F\|_\infty = \sup_{y \in \mathbb{R}} |F_n(y) - F(y)| \xrightarrow{a.s.} 0 \tag{0.9}$$

This theorem strengthens (0.6). In fact it ensures almost surely a *uniform* convergence of the ECDF to the population distribution function, that is the convergence does not happen for a fixed point, as in (0.6), but for all the points of the

real line. Term "fundamental" is now clearer. This theorem, in fact, guarantees that in the presence of a "large" sample using the empirical cumulative distribution function is almost equivalent to use the actual population distribution function F, independently from F. Moving from a pointwise analysis to a distributional one, we cite the Donsker's Theorem (for a proof see Billingsley (1968), Th. 16.4, p. 141)

Theorem 0.2.2 (Donsker)**.** *Let $X_1, \ldots, X_n$ are independent random variables with common distribution F, then*

$$\sqrt{n}(F_n - F) \xrightarrow{\mathcal{D}} B(F) \tag{0.10}$$

where $B(\cdot)$ is a Brownian bridge, that is a Brownian motion tied down to 0 at time 1.

Going a bit deeper we have that the limit in distribution of the Empirical Process is a Gaussian process with zero mean function and a covariance function

$$C_2(s,t) = F(s) \wedge F(t) - F(t)F(s) \tag{0.11}$$

It is clear that we can look at Donsker's Theorem like a functional extension of the Central Limit Theorem.

In Figures 0.2-0.4, below, three plots of some simulations of the Empirical Process for a uniform on $(0,1)$ population are reported. We want to highlight that in case of uniform population, that is $F(x) = x$, the limit process in Donsker's Theorem is exactly a Brownian bridge. From all the situations listed in Figures 0.2-0.4, the symmetry of the Empirical Process and the zero mean function are evident. Of course, we have a better approximation to the Brownian bridge in case illustrated in Fig. 0.4, when the sample size is very large. In fact in this case the actual expectation (the zero line) and the simulated mean (in black) are indistinguishable. Also with a sample size of $n = 100$, that is the case of Fig. 0.3, we have a good approximation, while in case of only $n = 10$ observations the approximation to the Brownian bridge is clearly inadequate.

At the end of this section we want to introduce a result known as Dvoretzky–Kiefer–Wolfowitz (DKW) inequality from the name of the authors of the paper in which it firstly appeared (see Dvoretzky et al. (1956)).

Theorem 0.2.3 (DKW inequality)**.** *Let F_n be the ECDF built on a sample $X_1, \ldots, X_n$*

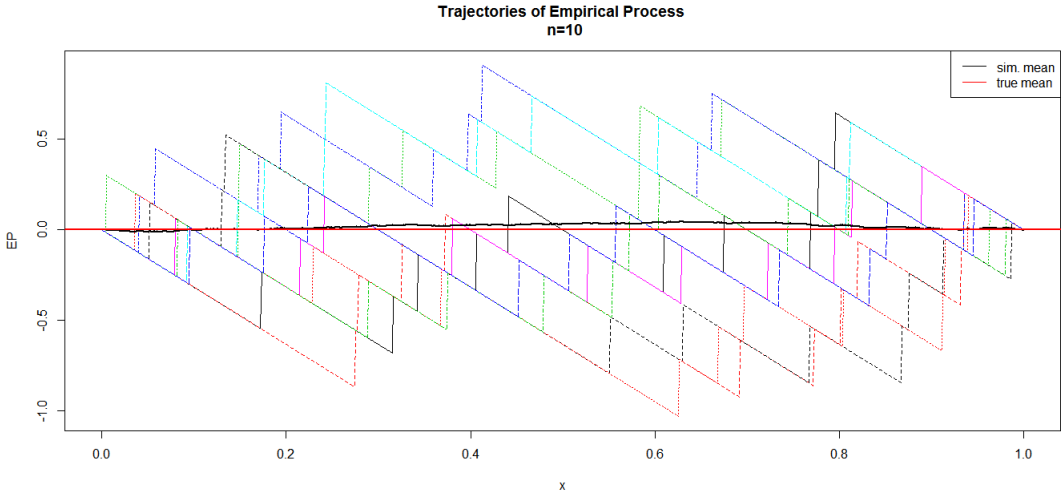


Fig.0.2. Ten trajectories of the empirical process for a uniform $(0,1)$ population with sample size $n = 10$

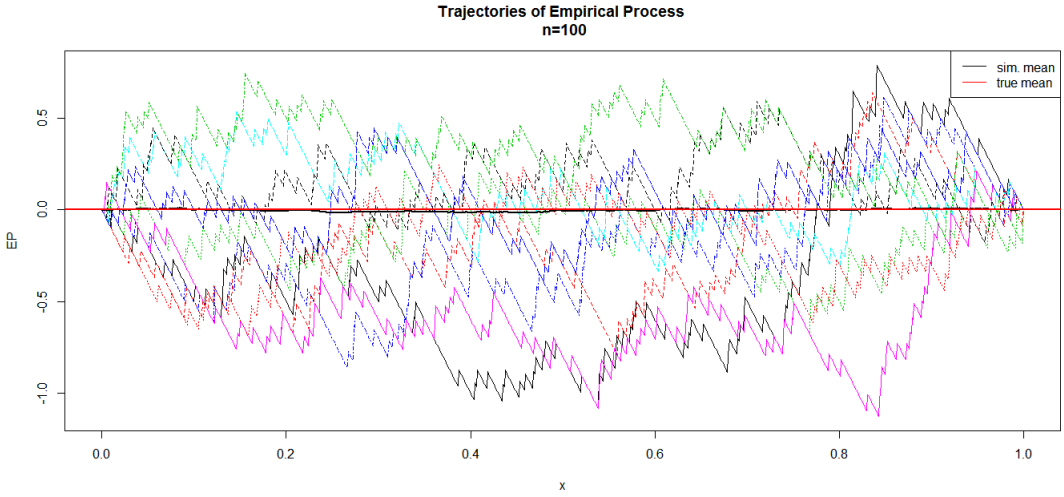


Fig.0.3. Ten trajectories of the empirical process for a uniform $(0,1)$ population with sample size $n = 100$

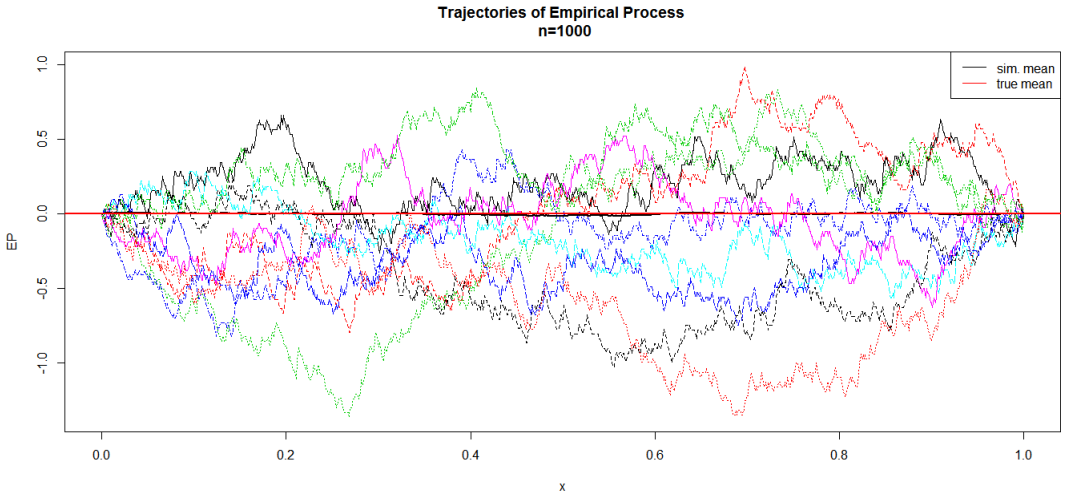


Fig.0.4. Ten trajectories of the empirical process for a uniform $(0,1)$ population with sample size $n = 1000$

of independent variables with common distribution F. It holds that:

$$Pr\left(\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}, \ \forall\epsilon > 0 \tag{0.12}$$

To be more precise, the last result is the two-sided version of the DKW inequality and is due to Massart (1990). In fact, in Dvoretzky et al. (1956), the inequality is given with an unknown constant $D < \infty$, only in 1990 Massart shows that it holds for $D = 2$. This last result generalizes Glivenko-Cantelli Theorem. In particular it, not only shows that the supremum of the absolute difference between the real distribution function F and the ECDF F_n goes, almost surely, to zero when the sample size grows, but it also quantifies the rate of this convergence. Moreover, looking at the quantity

$$\sup_{x\in\mathbb{R}}|F_n(x) - F(x)|$$

as the Kolmogorov-Smirnov (one sample KS) test statistic (for more on these tests see Kolmogorov (1933), Smirnov (1939b) and Smirnov (1939a)), the DKW inequality upper bounds the tails of the distribution of the KS test.

0.3 Sampling Theory

In this last section of this chapter we move from the framework of infinite population to the finite population one. Changing the point of view from infinite to finite

population entails some modification of the approach to inferential problems. First of all, the presence of a sampling design, usually without replacement, and the finiteness of the population implies the inadequacy of the $i.i.d$ assumption for the data. Moreover, in the presence of a sampling design with non trivial sampling weights, the usual estimator like sample means, sample variance and similar that use uniform wights are biased. From this the need of other estimators that take into account the sampling weights. This is the case of the well known Horvitz-Thompson estimator and the Hajék estimator that will be recalled later.

Formally speaking let $\mathcal{U}_N = \{1, 2, \ldots, N\}$ be a finite population of size N and $s \subset \mathcal{U}_N$ a sample of effective size n_s. Define $\mathcal{S}$ the space of all the possible samples (for example if we admit all the possible subsets of $\mathcal{U}_N$ as a sample, $\mathcal{S}$ could be the power set $\mathcal{P}(\mathcal{U}_N)$ of $\mathcal{U}_N$). A (probabilistic) sampling design P is a probability distribution over $\mathcal{S}$. We say that the sampling design has fixed sample size n if $\mathcal{S} = \{s \in \mathcal{P}(\mathcal{U}_N) | n_s = n\}$, that is the sample space contains only samples of n different units. Now let Y be a character of interest defined on the population $\mathcal{U}_N$; clearly there exists a one-to-one relation between $i-$th unit of the population and the values of Y that it takes. Thus, in the sequel the notations $\mathcal{U}_N = \{1, 2, \ldots, N\}$ and $\mathcal{U}_N = \{Y_1, Y_2, \ldots, Y_N\}$ are interchangeable.

For each unit in the population, let denote by

$$D_i = \begin{cases} 1 & \text{if unit } i \in s \\ 0 & \text{otherwise} \end{cases}$$

be the sample inclusion (Bernoulli) random variable (r.v.), and let $\mathbf{D}_N$ be the vector composed by the N random variables $D_1, \ldots, D_N$. The knowledge of s implies the knowledge of (a realization of) $\mathbf{D}_N$ and vice-versa. First and second order inclusion probabilities $(\pi_i,\ i = 1, 2, \ldots, N)$, $(\pi_{i,j},\ i, j = 1, 2, \ldots, N)$ can be defined using the variables D_is respectively as the probability for D_i to take value 1 and the probability for the product $D_i D_j$ to assume value 1. Formally

$$\pi_i = P(D_i = 1),\ \pi_{i,j} = P(D_i D_j = 1).$$

The Horvitz-Thompson estimator of the population mean

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

is defined as the weighted mean of the sample units y_i, with weights equal to the reciprocal of the first order inclusion probabilities, that is:

$$\hat{Y}_{HT} = \frac{1}{N}\sum_{i\in s} \pi_i^{-1} y_i. \tag{0.13}$$

Clearly the estimator (0.13) is unbiased and its variance has form

$$Var(\hat{Y}_{HT}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \tag{0.14}$$

in addition, if the sampling design has a fixed sample size the variance can be written as

$$Var(\hat{Y}_{HT}) = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2(\pi_i\pi_j - \pi_{ij}). \tag{0.15}$$

For the sake of completeness we must add that expression (0.15) for the Horvitz-Thompson estimator variance, has a fundamental role in sampling theory. In fact, its sample version leads to the well known Yates-Grundy variance estimator (see Yates and Grundy (1953)). It possesses a fundamental property: it is a positive quantity provided that the condition $\pi_{ij} < \pi_i\pi_j$ holds.

In Hajek (1971) the following ratio estimator is introduced

$$\hat{Y}_H = \frac{\sum_{i\in s}\pi_i^{-1} y_i}{\sum_{i\in s}\pi_i^{-1}}. \tag{0.16}$$

It is usually known as Hajek's estimator. As suggested in Särdnal et al. (1992) pp. 182-184, thanks to its ratio structure this estimator shows a reduced variance than the Horvitz-Thompson if the sample size is not fixed, if the units are homogeneous (in the sense that the difference between a unit and the population mean is small) or if there is a weak or negative correlation of the interest character Y with the inclusion probabilities (in this case to "huge" value of Y corresponds a small value of the inclusion probability, hence the sum of the reciprocals of the inclusion probabilities at the denominator of the ratio balance the numerator). Hàjek estimator

is asymptotically unbiased and using a Taylor expansion it is possible to show that

$$\hat{Y_H} = \bar{Y} + \sum_{i \in s} \frac{y_i - \bar{Y}}{\pi_i} + O_p(n^{-1}). \tag{0.17}$$

Thus, its variance can be approximated by using $y_i - \bar{Y}$ instead of y_i in (0.14).

In this dissertation we consider only πps[1] sampling designs with fixed sample size. This class of sampling designs takes advantage of the knowledge of an auxiliary character X (size variable) correlated with the interest character Y, to improve the efficiency of the usual Horvitz-Thompson estimator. In fact from (0.15) it easy to see that in case of perfect proportionality of the inclusion probabilities to the character Y, the variance is 0. Thus the inclusion probabilities are usually chosen proportional to the size variable X.

Let us now introduce a measure that quantifies how much a probability distribution is "random". The form of measure of uncertainty that we will consider is *Shannon's entropy* (Shannon, 1948). Given a discrete probability distribution $q = q_1, q_2, \ldots, q_k$ entropy is defined as:

$$H(q) = \mathbb{E}[\log(q)] = -\sum_{j=1}^{k} q_j \log(q_j) \tag{0.18}$$

For the sake of brevity, we will skip all the properties that the quantity (0.18) has, but we want to highlight that entropy is a non-negative quantity and it is zero if the probability mass is degenerate, while it takes its maximum value (in this case $\log(k)$) when all the q_js are equal, that is the case of maximum uncertainty. Although entropy $H(q)$ was first introduced in Information Theory, it has a key role in different branches of statistics like in measuring inequalities (see for instance Theil (1967)) or in econometrics (cfr. Golan et al. (1996)). Entropy has been widely studied also in sampling theory. In particular it has been shown that sampling designs that have a *High Entropy*, like successive design, Poisson design, Sampford design etc.. (cfr. Grafström (2010) and Brewer and Donadio (2003)) have lots of interesting properties. A special role in this framework is played by the Poisson (Po) and the rejective sampling design (R). We remind here that a Poisson sampling

[1] This acronym generally stands for *probability proportional to size*

design of parameter $p_1, p_2, \ldots, p_N$ such that $p_1 + p_2 + \ldots p_N = n$ has mass function

$$Po(\mathbf{D}_N) = \prod_{i=1}^{N} {p_i}^{D_i}(1-p_i)^{1-D_i}. \tag{0.19}$$

In fact it is characterized by the independence of the inclusion random variables D_is and the inclusion probabilities π_i coincide with its parameters. Formally

$$\pi_i^{Po} = Po(D_i = 1) = p_i.$$

The rejective sampling is obtained by a Poisson sampling by conditioning on a fixed sample size. Clearly this condition implies the non-independence of the random variables D_is for the rejective sampling. Moreover for the rejective sampling is not true that the inclusion probabilities coincide with the parameters of the Poisson sampling that originates the rejective design.

$$\pi_i^R = \mathbb{E}_{Po}[D_i | n_s = n] \neq p_i$$

Of course asymptotically we can approximate the inclusion probabilities with the parameters (see Hájek (1964)). In fact as intuition suggests, asymptotically increasing n and N (Hàjek point of view for asymptotics in finite population framework) in the rejective sampling the dependence between the variables D_is vanishes, while for the Poisson sampling due to the law of large numbers the effective sample size n_s tends to its expected value that is $\mathbb{E}_{Po}[n_s] = \mathbb{E}_{Po}[D_1 + D_2 + \ldots D_N] = n$. However, even when the sample size is not very large, it is possible to compute the inclusion probabilities starting from the parameters and vice-versa, (see Tillé (2006) and references therein). In Hájek (1959) it has been shown that the Poisson sampling design within unequal probability sampling designs, is the maximum entropy one. In the same paper it is also shown that the rejective sampling, inherits from the Poisson design the high entropy property. Thus within the sampling design with fixed first order inclusion probabilities and with fixed sample size, R is the maximum entropy one.

In this work we deal only with fixed sample size designs, thus the rejective sampling will play the role of benchmark design while considering the entropy. Among the many properties of high entropy sampling design, we are particularly interested in three of them. The first one is that if a (unequal probability, fixed sample size) sampling design is asymptotically a high entropy sampling (meaning

that entropy of the considered design asymptotically reaches the rejective sampling entropy) it is equivalent in terms of first order inclusion probabilities, to the rejective sampling (for more on this see the interesting work of Berger (1998)). The second result we are interested in is the well known Hàjek approximation (crf. Hájek (1964) p. 1511) for second order inclusion probabilities. This approximation has the following form:

$$\pi_{ij} = \pi_i \pi_j \left\{ 1 - \frac{(1-\pi_i)(1-\pi_j)}{d} [1 + o(1)] \right\}, \tag{0.20}$$

where

$$d = \sum_{i=1}^{N} \pi_i (1 - \pi_i) \tag{0.21}$$

is assumed to diverge. The importance of Hàjek approximation (0.20), as it is evident, is in the fact that it is possible to express the second order inclusion probabilities only in terms of the first order inclusion probabilities. The latter approximation is given by Hàjek only for rejective sampling, but it is clear from what we said above, that if we deal with high entropy sampling designs it is still valid.
After these two results we can claim that high entropy sampling designs are asymptotically equivalent and the variance of the Horvitz-Thompson estimator does not depend on the second order inclusion probabilities, but can be derived considering only the first order inclusion probabilities.

The last result we want to recall here is Theorem 2 in Berger (1998). This Theorem generalizes the result obtained in Hájek (1964). In particular it is a Central Limit Theorem for Horvitz-Thompson estimator in the presence of a generic high entropy sampling design (not only for the rejective sampling design as made by Hàjek in 1964).

Theorem 0.3.1 (Berger (1998))**.** *Consider a sampling design P with first order*

inclusion probabilities $\pi_1, \ldots, \pi_N$ *and define the quantities:*

$$d = \sum_{i=1}^{N} \pi_i(1 - \pi_i), \tag{0.22}$$

$$R = \frac{1}{d} \sum_{i=1}^{N} y_i(1 - \pi_i), \tag{0.23}$$

$$Z_i = y_i - R\pi_i, \tag{0.24}$$

$$S_N^2 = \sum_{i=1}^{N} Z_i^2 \left(\frac{1 - \pi_i}{\pi_i} \right), \tag{0.25}$$

$$L(\epsilon) = \frac{1}{S_N^2} \sum_{i:|Z_i| > \epsilon \pi_i S_N} Z_i^2 \left(\frac{1 - \pi_i}{\pi_i} \right), \tag{0.26}$$

$$e = \inf\{\epsilon : L(\epsilon) \leq \epsilon\} \tag{0.27}$$

then the Horvitz-Thompson estimator has a Normal asymptotic distribution if and only if:

i) $d \to +\infty$,

ii) P *is a high entropy sampling design,*

iii) $e \to 0$ *(Lindeberg-Hàjek condition)*

As it is clear from the latter Theorem the high entropy property is a key condition to obtain the asymptotic normality of the Horvitz-Thompson estimator. Condition *ii*) has been specified in a very generic way. In the following chapters the high entropy condition will be discussed in depth by introducing several conditions to establish the high entropy property for a generic sampling design P.

Chapter 1

Assumptions

In this chapter the assumptions to develop our work are listed and commented. The present chapter is divided in two sections. The first one deals with the basic assumptions that put the basis to develop the asymptotic theory that will be the subject of the next chapter. The second one is about the assumptions on what kind of parameters we are interested in this work and their regularity.

1.1 Basic Assumption

H1. $(\mathcal{U}_N,\ N \geq 1)$ is a sequence of finite population of increasing size N

H2. Let Y be the character of interest, and let $T_1, T_2, \ldots, T_L$ be the design variables. Denote further by $\mathbb{P}$ the superpopulation proability distribution of the r.v.s $(Y_i, T_{i1},\ \ldots,\ T_{iL})$. For each size N, $(y_i, t_{i1}, \ldots, t_{iL})$, $i = 1, 2, \ldots, N$ are realizations of a superpopulation $\{(Y_i, T_{i1},\ \ldots,\ T_{iL}),\ i = 1, \ldots, N\}$ composed by $i.i.d$ $(L+1)$-dimensional random vectors. The symbols $\mathcal{Y}_N$, $\mathcal{T}_N$ are used to denote the vector of N population y_is values and the $N \times L$ matrix of population t_{ij}s values $(j = 1,\ \ldots,\ L)$, respectively. Of course, the values $(t_{i1}, \ldots, t_{iL})$, $i = 1, \ldots, N$ are known prior to draw the sample s.

H3. For each population $\mathcal{U}_N$, sample units are selected according to a fixed size sample design with positive first order inclusion probabilities $\pi_1, \ldots, \pi_N$ and sample size $n = \pi_1 + \ldots + \pi_N$. The first order inclusion probabilities are taken proportional to a variable $x_i = g(t_{i1}, \ldots, t_{iL})$, $i = 1, \ldots, N$, where $g(\cdot)$ is an arbitrary positive function. For the sake of simplicity, we will assume that, for each i, $\pi_i = n x_i / \sum_j x_j$. The quantities n, π_is, D_is obviously depend on N. To avoid complications in the notation we will use the symbols n, π_i, D_i,

omitting the explicit dependence on N. Furthermore is assumed that

$$\mathbb{E}_{\mathbb{P}}[\pi_i(1-\pi_i)] = d, \text{with } 0 < d < \infty \tag{1.1}$$

$$\mathbb{P}\left(\sum_{i=1}^{N} \pi_i(1-\pi_i) = d_N \to \infty\right) = 1. \tag{1.2}$$

H4. The sampling fraction tends to a finite, non-zero limit:

$$\lim_{N\to\infty} \frac{n}{N} = f,\ 0 < f < 1.$$

H5. The actual sampling design P, with inclusion probabilities $\pi_1, \ldots, \pi_N$ satisfies the relationship

$$d_H(P, R) \to 0, \text{ as } N \to \infty,$$

where R is the rejective sampling with the same inclusion probabilities as P and $d_H(\cdot,\cdot)$ is the Hellinger's Distance.

H6. $\mathbb{E}_{\mathbb{P}}[X_1^2] < \infty$.

Let now analyze the assumptions we made.

The first hypothesis is necessary to obtain asymptotic results in "Hàjek's way". In fact asymptotic results here are obtained by pushing the sample size to infinity and having a sequence of increasing size finite populations.

Assumption H2 is particularly relevant. In fact we assume that the finite population is obtained as a realization of a superpopulation. This assumption change the classic point of view of sampling theory in which the finite population is seen as a set of fixed unknown numbers, such that the only source of randomness is due to the sampling design, that is the randomness is introduced by the statistician in order to control the units selection process. By introducing a superpopulation we are introducing a new source of variability in addition to the sampling design. Thus in our work we are in the presence of two sources of variability; this point will be, in more details, discussed later. In addition, with assumption H2 we are allowing the presence of a dependence relationship between the interest character Y and the design variables T_i. We want to highlight that we admit the possibility of a dependence between the variables without specifying any form for this dependence, thus we are in a very general framework.

In assumption H3 we are essentially formalizing what usually happens in πps sampling designs. Inclusion probabilities are taken proportional to a size variable X_i that is a positive function of the design variables. We want to point out that in our framework the inclusion probabilities are random variables and the regularity condition (1.1) is necessary in the sequel of this work in order to explicitly derive some results.

Hypothesis H4 is made to avoid trivial cases in our analysis, while H6 is an integrability condition on the size variable X that brings as main consequence that the quantity in (1.1) can be written as follows

$$d = f\left(1 - \frac{\mathbb{E}[X_1^2]}{E[X_1]^2}\right) + f(1-f)\frac{\mathbb{E}[X_1^2]}{E[X_1]^2}. \tag{1.3}$$

As far as assumption H5 is concerned we will spend a few more words. Firstly we remember here the formal definition of the *Hellinger's Distance* $d_H(\cdot,\cdot)$, the *Total Variation Distance* $d_v(\cdot,\cdot)$ and the *Kullback Leibler Divergence* $D(\cdot\|\cdot)$:

$$d_H(P,R) = \sum_{s\in\mathcal{S}}\left(\sqrt{P(s)} - \sqrt{R(s)}\right)^2 \tag{1.4}$$

$$d_v(P,R) = \frac{1}{2}\sum_{s\in\mathcal{S}}|P(s) - R(s)| \tag{1.5}$$

$$D(P\|R) = \sum_{s\in\mathcal{S}} P(s)\log\left(\frac{P(s)}{R(s)}\right). \tag{1.6}$$

Distances 1.4-1.6 (although the Kullback Liebler's divergence is not a mathematical distance) quantify the similarity between two probability distributions. In our discussion, one distribution corresponds the rejective sampling design, that is the benchmark distribution when dealing with entropy, as discussed in the previous chapter, and the other to a generic sampling design P. The following relationships hold:

$$d_H(P,R)^2 \leq d_v(P,R) \leq 2d_H(P,R) \tag{1.7}$$

$$D(P\|R) \geq \frac{1}{2}d_v(P,R)^2. \tag{1.8}$$

Using the maximum entropy property of the rejective sampling, as showed in Berger (2011), it is possible to see that:

$$D(P\|R) = H(R) - H(P) \tag{1.9}$$

so that the *Kullback Liebler Divergence* quantifies the difference in the entropies. Hence the property of "high entropy" of a sampling design can be characterized by an asymptotic null *Kullback Liebler Divergence*. As one can see from (1.7) requiring that the *Hellinger's Distance* goes to zero, implies that the *Total Variation Distance* goes to zero. If the *Total Variation Distance* between P and R goes to zero it means that $P(s) \approx R(s)$ thus $H(P) \approx H(R)$ and $D(P|R) \to 0$. Thus, at the end we may conclude that hypothesis H5 requires that the sampling design P must be a high entropy sampling design, and as we said in the previous chapter, this condition is essential for several reasons, including the asymptotic normality of the Horvitz-Thompson estimator.

For a finite population $\mathcal{U}_N$ of size N we define the population distribution function (p.d.f.) of the interest character Y as:

$$F_N(y) = \frac{1}{N}\sum_{i=1}^{N} I_{(y_i \leq y)},\ y \in \mathbb{R} \tag{1.10}$$

where I is the same indicator variable defined for the ECDF. It is worth to notice that what we said about the ECDF is still valid for the population distribution function. More specifically under the randomness of the superpopulation, as a consequence of the Strong Law of Large Numbers, for every $y \in \mathbb{R}$ $F_N(y)$ is a strongly consistent estimator of the distribution function $F(y)$ of the superpopulation, namely

$$F_N(y) \to F(y),\ a.s. - \mathbb{P} \tag{1.11}$$

Clearly, the population distribution function contains all the information about the finite population, but is an unknown quantity. The goal is to estimate it on the basis of a sample s of the finite population $\mathcal{U}_N$. We consider here two possible estimators of the population distribution function F_N

$$\hat{F}_H(y) = \frac{\sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(y_i \leq y)}}{\sum_{i=1}^{N} \frac{1}{\pi_i} D_i} \tag{1.12}$$

$$\hat{F}_{HT}(y) = \frac{\sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(y_i \leq y)}}{N} \tag{1.13}$$

The first one is the Hàjek estimator of the p.d.f. and the second is clearly the Horvitz-Thompson one. Both of the above mentioned estimators play the same role in finite population framework, played by the $ECDF$ when dealing with an infinite population. By virtue of their definition, both (1.12) and (1.13) are elements of $D[-\infty, +\infty]$ the *càdlàg* space introduced in the previous chapter. In particular, if we do not fix any specific y, $\widehat{F}_H(y)$ and $\widehat{F}_{HT}(y)$ are random elements of the space D. It is important to highlight that both $\widehat{F}_H$ and $\widehat{F}_{HT}$ are subject to two sources of randomness, the superpopulation variability and the sampling one. Thus even if we neglect the superpopulation approach by considering the finite population as a set of fixed unknown numbers, these estimators are still random elements of the space D. As already done for the ECDF, we want to briefly characterize these quantities. We now list some basic properties of (1.12) (1.13)

The basic properties that identify the cumulative distribution functions are: *i*) *right continuity*, *ii*) they tend to value 1 when the variable approaches huge positive values, *iii*) they tend to 0 when the variable approaches huge negative values. We define a *proper* estimator of a distribution function an estimator that satisfies all of these conditions. The first claim is about that.

Claim 1: *$\hat{F}_H$ is a proper estimator of F while $\hat{F}_{HT}$ is not.*

The right continuity is valid for both $\widehat{F}_H$ $\widehat{F}_{HT}$ by construction. It is trivial to see that when $y \to -\infty$ $\widehat{F}_H \to 0$ and $\widehat{F}_{HT} \to 0$. Let now focus on property *ii*). If $y \to +\infty$ all the indicators variables Is will take value 1, thus $\widehat{F}_H$ will be exactly equal to 1 in this situation.
This is not true for $\widehat{F}_{HT}$ because of the presence of N instead of the sum of the weights at the denominator, thus we have no warranty that the limit is equal to 1. It is worth to notice that the following lemma holds

Lemma 1.1.1. *Under the assumptions $H1$-$H6$, the quantity*

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{\pi_i} \tag{1.14}$$

tends to 1 *as N increases, for a set of (sequences of) y_is, t_{ij}s having $\mathbb{P}$-probability* 1*, and for a set of $\mathbf{D}_N$s of P-probability tending to* 1*.*

Hence, the limit of F_{HT} when $y \to +\infty$ is 1 in P-probability, a.s.-$\mathbb{P}$.

Claim 2: *For every $y \in \mathbb{R}$, both $\hat{F}_H(y)$ and $\hat{F}_{HT}(y)$ are consistent estimators of $F_N(y)$, hence they are also consistent estimator of $F(y)$.*

The consistency as estimators of $F_N(y)$ is obtained with the same reasoning of Lemma 1.1.1 and observing that $0 \leq I(y_i \leq y) \leq 1$. In virtue of (0.6) the last part of the claim is trivial.

1.2 Regularity Assumptions on Parameters

In this section we first define the objects of the inference, i.e. the parameters of interest. We will define a finite population parameter as functional of the population distribution function, formally

$$\theta_{FP} = \theta(F_N) \tag{1.15}$$

where $\theta : D[-\infty, +\infty] \to E$ and the space E is usually the real line or more generally multidimensional euclidean spaces. Clearly, in the same way, we can define a parameter of the superpopulation or hyper-parameter as

$$\theta_{SP} = \theta(F). \tag{1.16}$$

This work focuses on inference for hyper-parameters or *analitycal inference*. Thus, in the sequel, for the sake of brevity and in order to make this work more readable, the symbol θ will be used instead of the symbol θ_{SP} if not explicitly specified.

The consideration of parameters that can be expressed as functionals of the distribution function of a character of interest Y, make us able to impose some regularity condition on the considered functionals so that the asymptotic results can be easily derived for a whole class of estimators.
Usually when dealing with statistical functionals, the sought regularity condition is the *Frechét* differentiability. In some cases this condition is too restrictive. In fact some statistical functionals, like variance and quantiles, do not satisfy the usual *Frechét* differentiability assumption (see Serfling (1980), p. 220, and Osier (2009)).

In view of the above remarks, we resort to the *Hadamard* differentiability condition.

Definition 1.2.1. *Let $\theta(\cdot) : l^\infty(-\infty, \infty) \to E$ be a map having as domain the Banach space (equipped with the sup-norm) of the bounded functions, and taking values on a normed space E with norm $\|\cdot\|_E$. The map $\theta(\cdot)$ is Hadamard-differentiable at*

F if there exist a continuous linear functional $\theta'_F(\cdot) : l^\infty(-\infty,\infty) \to E$ such that

$$\left\| \frac{\theta(F + th_t) - \theta(F)}{t} - \theta'_F(h) \right\|_E \to 0, \ \ as \ \ t \downarrow 0, \ \forall h_t \to h. \tag{1.17}$$

The map $\theta'_F(\cdot)$ is the *Hadamard derivative* of θ at F.
For the sake of completeness we report here the definition of a functional *Frechét* differentiable

Definition 1.2.2. *Let $\theta(\cdot) : l^\infty(-\infty,\infty) \to E$ be a map having as domain the Banach space (equipped with the sup-norm) of the bounded functions, and taking values on a normed space E with norm $\|\cdot\|_E$. The map $\theta(\cdot)$ is Frechét-differentiable at F if there exist a continuous linear functional $\theta'_F(\cdot) : l^\infty(-\infty,\infty) \to E$ such that*

$$\frac{\|\theta(F + h) - \theta(F) - \theta'_F(h)\|_E}{\|h\|} \to 0, \ \ as \ \ \ \|h\| \downarrow 0. \tag{1.18}$$

From the last two definitions it appears that *Frechét*-differentiability implies the *Hadamard*-differentiability. In fact the *Frechét* condition requires the same rate of convergence for each direction h, while the *Hadamard* condition allows different rates for each direction h. The converse generally does not hold. It is easy to see that if the domain of the map θ is the usual euclidean space $\mathbb{R}^d$ (this is not the case of our work), (1.17) and (1.18) coincide.

The *Hadamard*-differentiability assumption is well known in the empirical processes theory, because it shows some good properties related to the weak convergence. In order to make this more clear we report here Theorem 20.8 in Van der Vaart (2000).

Theorem 1.2.1. *Let $\mathbb{D}$ and $\mathbb{E}$ be normed linear spaces. Let $\theta : \mathbb{D} \to \mathbb{E}$ be Hadamard differentiable at ϕ. Let $T_n : \Omega_n \to \mathbb{D}$ be maps such that $r_n(T_n - \phi) \xrightarrow{weak} T$ for some sequence of numbers $r_n \to +\infty$ and a random element T that takes its values in $\mathbb{D}$. Then $r_n(\theta(T_n) - \theta(\phi)) \xrightarrow{weak} \theta'_\phi(T)$.*

One of the most famous consequence of the Theorem 1.2.1 is the well known *delta method*, that is obtained considering $\mathbb{D} = \mathbb{E} = \mathbb{R}$.

We will examine in depth other consequences of this property in the next chapter while studying the large sample distribution of the Hàjek's estimator of the superpopulation distribution function.

Chapter 2

Main Asymptotic Results

In this chapter we study the large sample distribution of the estimator $\widehat{F}_H$. In the first section a brief review of the paper by Conti et al. (2015) is proposed. This work represents the starting point of the present dissertation and some of the results contained there will be used in the following. The second section is the core of this thesis, since it contains the main results about the asymptotic distribution of the Hajék's estimator $\hat{F}_H$ of the superpopulation distribution function F. In particular we fully recover the distribution of the limiting process and we also characterize the limiting distribution when dealing with hyper-parameters.

2.1 Asymptotic Results When Considering A Fixed Finite Population

This thesis takes its inspiration from the work of Conti et al. (2015). Our first contribution, consists in an extension of the results contained there, as well as in new applications to test problems. In Conti et al. (2015) the authors study the asymptotic distribution of the quantity $\hat{F}_H$ considered as an estimator of the finite population distribution function F_N. In particular, assuming hypothesis H1-H6 of the previous chapter, they study the asymptotic distribution of the stochastic process $W_n^H(y)$ given by the deviation of the Hàjek estimator from the finite population distribution function of a interest character Y, assuming that the finite population is fixed[1]. Formally

$$W_n^H(y) = \sqrt{n}\left(\hat{F}_H(y) - F_N(y)\right),\ y \in \mathbb{R}. \tag{2.1}$$

[1]This point is fundamental and it will be discussed in the sequel

It is clear from (2.1) that we are in presence of a variation of the classic empirical process recalled in the Preliminaries chapter.

It is worth to notice that although they assume hypothesis H2 (that is the finite population is produced by a superpopulation model), they consider the finite population as *fixed*. In such an approach the only source of randomness is the one introduced by the statistician, that is the sampling design. In addition, considering a fixed finite population requires a further clarification. In their asymptotic approach, the authors consider the sequences $\mathbf{y}_\infty = (y_1, y_2, \ldots)$, $\mathbf{x}_\infty = (x_1, x_2, \ldots)$ of the interest character Y and the auxiliary variable X when the finite population size N is pushed to the infinity. Hence the actual finite population, composed by the vectors $\mathbf{y}_N = (y_1, y_2, \ldots, y_N)$, $\mathbf{x}_N = (x_1, x_2, \ldots, x_N)$, is viewed as the segments of the first N y_is x_is in the sequences $\mathbf{y}_\infty$, $\mathbf{x}_\infty$. Clearly the infinite sequences live in an appropriate probability space, say $((\mathbb{R}^2)^\infty, \mathcal{B}(\mathbb{R}^2)^\infty, \mathbb{P}^\infty)$ where $\mathcal{B}(\mathbb{R}^2)^\infty$ is the Borel sigma-algebra over $(\mathbb{R}^2)^\infty$) and $\mathbb{P}^\infty$ is the product measure on $((\mathbb{R}^2)^\infty, \mathcal{B}(\mathbb{R}^2)^\infty)$ generated by $\mathbb{P}$.

We now list preliminary Lemmas to prepare to the most interesting result contained in Conti et al. (2015), that is the characterization of the large sample distribution of the process W_n^H when the finite population is considered fixed.

Lemma 2.1.1. *Let* $d_N = \sum_{i=1}^{N} \pi_i(1-\pi_i)$. *Then as* $N \to \infty$,

$$\frac{d_N}{N} \to d = f\left(1 - \frac{E_{\mathbb{P}}[X_1^2]}{E_{\mathbb{P}}[X_1]^2}\right) + f(1-f)\frac{E_{\mathbb{P}}[X_1^2]}{E_{\mathbb{P}}[X_1]^2} \ a.s. - \mathbb{P} \tag{2.2}$$

Lemma 2.1.2. *The following results hold:*

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}\left(I_{(y_i \le y)} - F_N(y)\right) \to \frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}\left(K_{-1}(y) - \mathbb{E}_{\mathbb{P}}[X_1^{-1}]\right)F(y) \ as \ N \to \infty, \ a.s. - \mathbb{P} \tag{2.3}$$

$$\frac{1}{N}\sum_{i=1}^{N}(1-\pi_i)\left(I_{(y_i \le y)} - F_N(y)\right) \to fF(y)\left(1 - \frac{K_{+1}}{\mathbb{E}_{\mathbb{P}}[X_1]}\right) \ as \ N \to \infty, \ a.s. - \mathbb{P}, \tag{2.4}$$

where $K_\alpha(y) = \mathbb{E}_{\mathbb{P}}[X_1^\alpha | Y_1 \le y], \ y \in \mathbb{R}, \ \alpha = \pm 1$

Lemma 2.1.3. *Define the quantities*

$$Z_{i,N}(y) = \left(I_{(y_i \leq y)} - F_N(y)\right) - \pi_i \frac{\sum_{i=1}^{N}(1-\pi_i)\left(I_{(y_i \leq y)} - F_N(y)\right)}{\sum_{i=1}^{N}\pi_i(1-\pi_i)}, \ i = 1, 2, \ldots, N, \tag{2.5}$$

$$S_N^2 = \sum_{i=1}^{N}\left(\frac{1}{\pi_i} - 1\right) Z_{i,N}^2. \tag{2.6}$$

Then, as N approaches to infinity, almost surely w.r.t. $\mathbb{P}$, the following results are true

$$Z_{i,N} - \left(I_{(y_i \leq y)} - F_N(y)\right) \to -\frac{f}{\mathbb{E}_{\mathbb{P}}[X_1]} X_i \frac{f(1 - k_{+1}(y)/\mathbb{E}_{\mathbb{P}}[X_1])}{d} F(y), \tag{2.7}$$

$$\frac{1}{N} S_N^2(y) \to \left(\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f} K_{-1}(y) - 1\right) F(y)(1 - F(y)) - \frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}(K_{-1}(y) - \mathbb{E}_{\mathbb{P}}[X_1^{-1}])F(y)^2$$
$$- \frac{f^2}{d}\left(1 - \frac{k_{+1}(y)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)^2 F(y)^2. \tag{2.8}$$

Lemma 2.1.4. $\forall \epsilon > 0$, *there exists an integer N_ϵ such that*

$$\mathbb{P}\left\{|Z_{i,N}(y)| \leq \epsilon \pi_i S_N \ \forall \ N \geq N_\epsilon\right\} = 1, \ i = 1, 2, , \ldots, N. \tag{2.9}$$

Lemma 2.1.5. *Let ϵ be a positive number and let*

$$A_N(\epsilon) = \{i \in \mathcal{U}_N : |Z_{i,N}(y)| > \epsilon \pi_i S_N\},$$
$$L_N(\epsilon)^2 = \sum_{i \in A_N(\epsilon)}\left(\frac{1}{\pi_i} - 1\right) Z_{i,N}^2.$$

Then:

$$\mathbb{P}\left\{\lim_{N \to \infty} \frac{L_N(\epsilon)^2}{S_N^2} = 0\right\} = 1, \ \forall \epsilon > 0. \tag{2.10}$$

Lemmas 2.1.1-2.1.2 are preparatory and they can be proved very easily using the Laws of Large Numbers. Remembering what we said about Hajek estimator variance it is clear that Lemma 2.1.3, through the same approach used in Hájek (1964), specifies the asymptotic form of the variance of $\hat{F}_H$ (it is sufficient to observe that the arithmetic mean of the indicator variables $I_{(y_i \leq y)}$ is the finite population distribution function $F_N(y)$). The last two Lemmas (Lemma 2.1.4-2.1.5) ensures the fulfillment, with $\mathbb{P}-$probability 1 of condition *iii*) of Theorem 0.3.1 (conditions *i*)

and *ii*) are fulfilled by assumptions H3-H5 in order to obtain the asymptotic normality of the Horvitz-Thompson estimator and equivalently of the Hàjek estimator by the Slutsky's Theorem.

We are now in the position to easily introduce the proposition that fully characterizes the large sample distribution of the process (2.1)

Proposition 2.1.1 (Conti et al. (2015))**.** *If the sampling design P satisfies assumptions $H1 - H6$, the sequence of random functions $(W_n^H(\cdot),\ N \geq 1)$ converges weakly, conditionally on the population $\mathcal{U}_N$, in $D[-\infty, \infty]$ equipped with the Skorokhod topology, to a Gaussian process $\widetilde{W}_1(\cdot) = (\widetilde{W}_1(y),\ y \in \mathbb{R})$ with zero mean function and covariance kernel*

$$\begin{aligned} C_1(y,t) = & f\left\{\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}K_{-1}(y \wedge t) - 1\right\}F(y \wedge t) \\ & - \frac{f^3}{d}\left(1 - \frac{K_{+1}(y)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)\left(1 - \frac{K_{+1}(t)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)F(y)F(t) \\ & - f\left\{\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}(K_{-1}(y) + K_{-1}(t) - \mathbb{E}_{\mathbb{P}}[X_1^{-1}] - 1)\right\}F(y)F(t) \end{aligned} \tag{2.11}$$

almost surely w.r.t. $\mathbb{P}$, with d given by (1.1)*.*

As we can imagine the limiting process is Gaussian, in fact as we have seen the high entropy of the considered sampling design makes asymptotically normal the finite distribution of the process W_n^H. In addition the process is centered and this is a simple consequence of the consistence of $\hat{F}_H$ as an estimator of the finite population distribution function F_N Looking at the covariance Kernel, is immediately visible the difference between the classic empirical process framework and the actual case, with non *i.i.d* data. In fact the limiting process is Gaussian but quite far from being a Brownian bridge, that is the limiting process of the classic non-parametric empirical process.

By the covariance function is clear the role of the dependence between the interest character Y and the auxiliary variable X that acts through the expected values $\mathbb{E}_{\mathbb{P}}[X_1], \mathbb{E}_{\mathbb{P}}[X_1^{-1}]$ and the conditional expected value $K_\alpha(y),\ \alpha = \pm 1$. Clearly this dependence takes into account also the first order inclusion probabilities because of the assumption of proportionality between the π_is and the size variable X_i.

Clearly if we consider the Horvitz-Thompson estimator instead of the Hajek's one, we will have a slight different covariance kernel. We will introduce now a Propo-

sition equivalent to Proposition 2.1.1 but with specific reference to the Horvitz-Thompson estimator:

Proposition 2.1.2. *If the sampling design P satisfies assumptions $H1 - H6$, the sequence of random functions $(\sqrt{n}(\hat{F}_{HT}(y) - F_N(y)),\ N \geq 1,\ y \in \mathbb{R})$ converges weakly, conditionally on the population $\mathcal{U}_N$, in $D[-\infty, \infty]$ equipped with the Skorokhod topology, to a Gaussian process $\widetilde{W}_1^{HT}(\cdot) = (\widetilde{W}_1^{HT}(y),\ y \in \mathbb{R})$ with zero mean function and covariance kernel*

$$C_1^{HT}(y,t) = f\left\{\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}K_{-1}(y \wedge t) - 1\right\}F(y \wedge t) - \frac{f}{d}\left(1 - f\frac{K_{+1}(y)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)\left(1 - f\frac{K_{+1}(t)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)F(y)F(t) \tag{2.12}$$

almost surely w.r.t. $\mathbb{P}$, with d given by (1.1).

As it appears from (2.12), the covariance kernel when considering the Horvitz-Thompson estimator is quite different from the one in (2.12), although the elements that are involved in its form are always the same. In particular, we note the presence of the sampling fraction that takes into account that we are sampling from a finite population and of $K_\alpha(y)$ that takes into account the dependence between the interest variable Y and the size variable X. The special case when there is independence between the interest variable Y and the auxiliary variable X is of some interest. In this situation, observing that $K_\alpha(y) = \mathbb{E}_{\mathbb{P}}[X^\alpha]$, it is easy to see that the covariance kernel (2.11) becomes:

$$C_1(y,t) = f(A-1)(F(y \wedge t) - F(y)F(t)). \tag{2.13}$$

where

$$A = \frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}\mathbb{E}_{\mathbb{P}}[X_1^{-1}]. \tag{2.14}$$

From (2.13) and keeping in mind (0.11) it is evident that in case of independence the limiting process is proportional to a Brownian Bridge scaled by a finite population correction factor $f(A-1)$ that takes into account the sampling fraction (because we are sampling from a finite population) and the presence of the sampling weights (because we are sampling with non trivial inclusion probabilities). It is easy to see that if we consider inclusion probabilities $\pi_i = \dfrac{n}{N}$, that is the case of the Simple Random Sampling, the correction factor $f(A-1)$ becomes the usual and well known finite population correction factor $1 - f$. This situation is well resumed by the following corollary

Corollary 2.1.1 (Conti (2014))**.** *Suppose that the sampling design P satisfies assumptions $H1 - H6$ and that there is independence between the interest variable Y and the size variable X. Then, pushing N to the infinity the sequence of random functions $(W_n^H(\cdot),\ N \geq 1)$ converges weakly, conditionally on the population $\mathcal{U}_N$, in $D[-\infty, \infty]$ equipped with the Skorokhod topology, to a Gaussian process that can be represented in the form*

$$(\sqrt{f(A-1)}B(F(y));\ y \in \mathbb{R}) \tag{2.15}$$

where B is a Brownian Bridge.

Let now see what would happen to the limit of the process W_n^{HT} when we consider the independence between the interest variable and the size variable. Noting that in this case

$$K_\alpha(y) = \mathbb{E}_{\mathbb{P}}[X_1]$$

the covariance kernel (2.12) becomes:

$$C_1^{HT} = f\left\{(A-1)F(y \wedge s) - \frac{(1-f)^2}{d}F(y)F(s)\right\} \tag{2.16}$$

with A already defined in (2.14). As it is possible to see from (2.16) when the Horvitz-Thompson estimator is concerned, the limiting process is no more proportional to a Brownian Bridge. In particular the two quantities

$$C_1^H(y,y) = f(A-1)(F(y) - F(y)^2) \tag{2.17}$$

$$C_1^{HT}(y,y) = f\left\{(A-1)F(y) - \frac{(1-f)^2}{d}F(y)^2\right\} \tag{2.18}$$

are respectively, the asymptotic variances of $\hat{F}_H(y)$ and $\hat{F}_{HT}(y)$. As shown in Conti (2014) the inequality

$$A - 1 \geq \frac{(1-f)^2}{d} \tag{2.19}$$

holds and this implies that

$$C_1^H(y,y) \leq C_1^{HT}(y,y) \tag{2.20}$$

The last inequality tells us a very interesting fact, i.e. the Hajek estimator is more efficient of the Horvitz-Thompson estimator. Clearly, the gain in efficiency is due to the restriction at the end of the *time* of the process. In other words the property of

the Hajek estimator of being a proper estimator of a distribution function implies that the process $W_n^H(y)$ goes to zero when y grows indefinitely. This is not true for the Horvitz-Thompson estimator, where we have no warranty of the behavior when $y \to +\infty$. This excess of variability of the Horvitz-Thompson estimator is the reason of the gap between the Brownian Bridge and the limiting process of W_n^{HT}, (in independence situation).

Remark 2.1.1. As highlighted in the previous chapter, when dealing with πps sampling designs, the presence of a correlation between the size variable and the interest variable is fundamental to gain some efficiency of the considered estimators. Thus, from a practical point of view, considering a dependence between the size variable and the interest variable is the most relevant case, although this implies an augmented complexity in the covariance kernel of the limiting processes W_n^H and W_n^{HT}. On the other hand, considering the independence case reduces the complexity of the large sample distribution of the considered processes, but from a practical point of view this case is not of main interest. In fact, the independence case represents the "worst" scenario where you are not able to improve the efficiency taking advantage of the auxiliary informations.

In order to make more evident these considerations, we have reported set of figures representing the processes W_n^H, W_n^{HT} in different situations.

Trajectories in Figures 2.1a–2.1d are obtained by assuming a finite population generated by a uniform on $(0, 1)$ superpopulation model, while the inclusion probabilities show a negligible correlation with the interest character. The sampling fraction is $f = 1/4$ and the horizontal zero line represents the theoretic mean of the process. Samples are selected according to a Pareto design. The first evident thing is that with a smaller sample size $n = 50$ we have some pathological trajectories. In fact, in both the case of W_n^H and W_n^{HT}, there are some trajectories that depart from the others exhibiting more variability if compared to the others. In the case of Hàjek estimator, that is a proper estimator, the trajectories are constrained to go to the zero line when $y = 1$, while for the Horvitz-Thompson they are allowed to end far from the zero. These figures show why the process W_n^{HT} could not be a Brownian Bridge. Another consequence that is immediately showed by Figures 2.1a–2.1d is that the Hàjek estimator is more efficient than the Horvitz-Thompson one. In fact the process W_n^H shows a smaller variability if compared to the process W_n^{HT}. Both the process are "visually" symmetric, and they fluctuate around the

zero line. The approximation to a Gaussian process is clearly better in the situation of a large sample size, where it is exhibited a more regular behavior.

We now move to analyze the case of dependence between the size variable X and the interest variable Y. Also in this case a finite population from a uniform on $(0, 1)$ distribution is generated and the inclusion probabilities are taken proportional to a size variable X that shows a correlation, in the finite population, of about 0.40 with the interest character and the sampling fraction is 0.25. The horizontal zero line represent the theoretical mean of the process. The first thing to be noticed is that, the dependence between X and Y reduces the variability of the processes as described in Figures 2.2a–2.2d compared to Figures 2.1a–2.1d, in particular for the process W_n^{HT} where we have not trajectories that go away from the others. Also in this case the Hajek estimator shows a better efficiency (Figures 2.2a and 2.2b) with respect to the Horvitz-Thompson estimator that exhibits a more variable behavior being unconstrained at right limit of x-axis. The symmetry of the processes is well visible in every of the situation considered.

Remark 2.1.2. The Pareto sampling design was first introduced in Rosén (1997a) and Rosén (1997b). One of the key concept of this work is considering only sampling designs that asymptotically reach maximum entropy. In our knowledge for Pareto sampling design there is not a proof of its maximum asymptotic entropy. Although this lack of an analytic proof, generally Pareto sampling is regarded as a high entropy sampling design. In Grafström (2010) the entropy of the (adjusted)[2] Pareto design is computed and compared to other designs like (Poisson, Sampford ecc) in two simple cases where the finite population has size of $N = 6, 10$ with a sampling fraction of 1/2. Results of these experimentations shows that the entropy of the (adjusted) Pareto design is substantially equivalent to the entropy of high entropy sampling designs like Poisson and Sampford designs. Clearly, these considerations are not a proof, but they make more plausible the conjecture of the asymptotic high entropy of the Pareto design. Since the ease of its implementation and the last considerations about its entropy, the Pareto design will be largely used in the sequel of this dissertation, especially for the simulations studies reported in Chapter 4.

[2]With adjusted Pareto design, is intended a Pareto design where parameters are modified in order to have the prescribed inclusion probabilities. For some of these modifications see Lundquist (2009)-Bondesson et al. (2006). For our simulations we have used the adjustment proposed by Bondesson et al..

Situation Of Independence Between Y And X

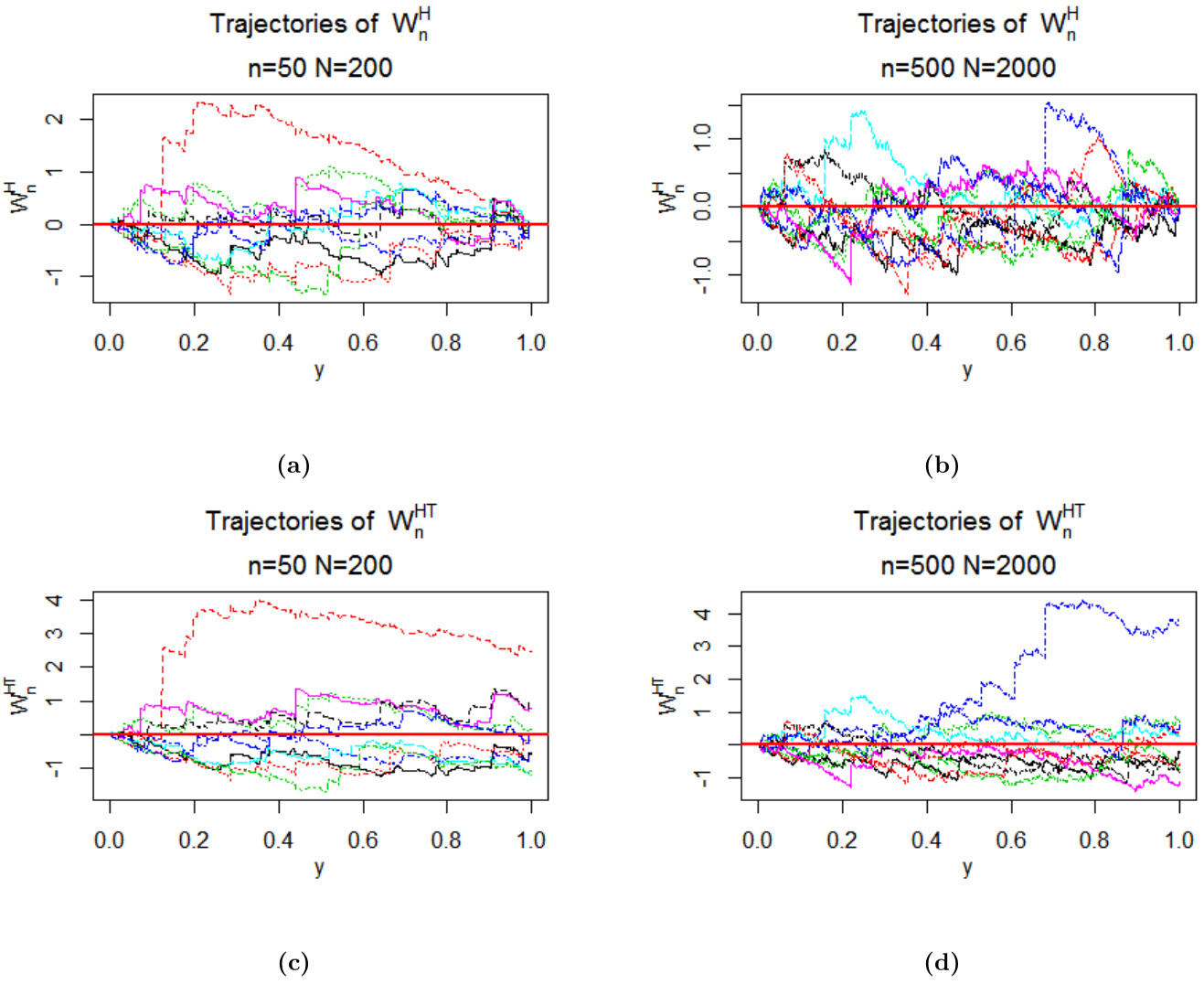


Fig.2.1. Trajectories of the empirical process (for finite population) where the finite population is fixed and the centering factor of the process is the finite population distribution function F_N. Independence between interest and size variables is assumed.

In the next section we will see how these results change when allowing the finite population varying, that is the natural consequence of assuming a superpopulation model. Consequently we will concern the process of the deviations of the Hajek estimator from the superpopulation distribution function, not from the finite population distribution function.

Situation Of Dependence Between Y And X

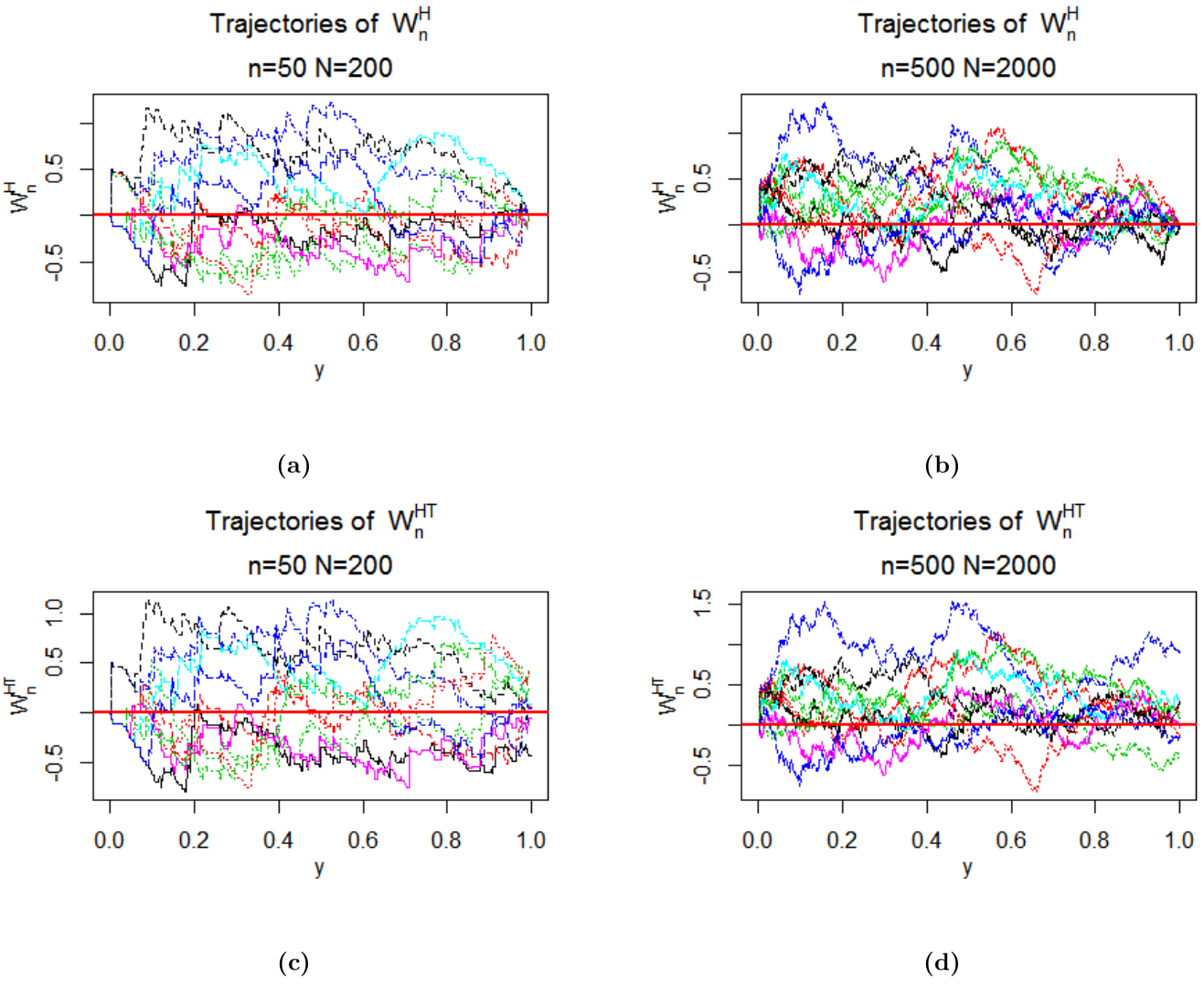


Fig.2.2. Trajectories of the empirical process (for finite population) where the finite population is fixed and the centering factor of the process is the finite population distribution function F_N. Dependence between interest and size variables is assumed.

2.2 Asymptotic Results When Considering A Varying Finite Population

The aim of this section is to study the large sample distribution of the process

$$W^H(y) = \sqrt{n}(\hat{F}_H(y) - F(y)),\ y \in \mathbb{R}. \tag{2.21}$$

The main difference with what discussed in the previous section is that in this case we focus on the superpopulation, and not on the finite population. The main consequence of this changed point of view is that the finite population has to be considered as variable and not fixed in order to make the process of the deviation of the Hàjek estimator from the superpopulation distribution function able to catch the variability of the superpopulation model.

It is easy to see that the process can be decomposed in the sum of two processes. Formally:

$$W^H(y) = \sqrt{n}(\hat{F}_H(y) - F_N(y)) + \frac{\sqrt{n}}{\sqrt{N}}\sqrt{N}(F_N(y) - F(y)) = W_n^H(y) + \sqrt{f}W_N(y) \tag{2.22}$$

We stress here that in this approach, based on considering the finite population a random variable, the process (2.21) depends on two sources of variability, i.e. the sampling design P (variability introduced by the statistician) and the variability of the random mechanism that generates the finite population (superpopulation model variability) $\mathbb{P}$. As far as (2.22) is concerned, a few remarks are necessary. The first one is that the two processes in which W^H is decomposed depend on the two sources of variability mentioned before. The process $W_n^H(y)$ catches the variability due to the sampling design while the process $W_N(y)$ is affected only by the superpopulation variability. Another important observation is that the process W_N is the deviation of the finite population distribution function F_N from the superpopulation distribution function F. Thus, under our assumption of a finite population obtained as *i.i.d.* replications of a superpopulation model, it is a classic empirical process as is introduced in Section 0.2.It is also fascinating to see how formula (2.22) follows our intuition. In fact if we consider a small sample size compared to the finite population size (that is considering a very small sampling fraction f) we are rightly led to believe that we have not enough information to distinctly catch the role of the random mechanism that generates the finite population. For the biggest part we are influenced only by the finite population role. This latter consideration is confirmed

also by formula (2.22) where the process W_N is scaled by a factor $\sqrt{f}$. The smaller is f, the smaller is the contribution of the classic empirical process W_N, thus of the superpopulation variability to the whole process W^H.

From the decomposition showed in (2.22) it is clear that results about the process W_n^H recalled in the previous section (Lemmas 2.1.1-2.1.5, Proposition 2.1.1) of the present chapter are fundamental to study the asymptotic behavior of the process W^H. The first problem that we face in this chapter is the extension of the mentioned results to the case of interest, that is admitting a variable finite population. To this purpose we recall here an interesting Lemma contained in Csörgő and Rosalsky (2003)

Lemma 2.2.1 (Csörgő and Rosalsky (2003))**.** *Let $F_n \subset F$, $n \in \mathbb{N}$ be an arbitrary sequence of σ-algebras. If $V_1, V_2, \ldots,$ and V are real or complex-valued random variables such that $\mathbb{E}[|V_n|] < \infty$ and $\mathbb{E}[V_n|F_n] \leq 1$ a.s. for all $n \geq 1$ and $\mathbb{E}[V_n|F_n] \xrightarrow{\mathcal{D}} V$, then $E[V_n] \to E[V]$. In particular if $\{A_n\}_{n=1}^{\infty}$ is a sequence of events such that $P\{A_n|F_n\} \xrightarrow{P} p$ for some constant p, then $P\{A_n\} \to p$.*

Thanks to the second statement of the latter Lemma, it is seen that the convergence of the finite distributions of the process W_n^H holds also unconditionally. Hence, we are now able to state the main result of this section.

Proposition 2.2.1 (Unconditional Convergence)**.** *If the sampling design P satisfies assumptions $H1-H6$, the sequence of random functions $(W_n^H(\cdot),\ N \geq 1)$ converges weakly, in $D[-\infty, \infty]$ equipped with the Skorokhod topology, to a Gaussian process $\widetilde{W}_1(\cdot) = (\widetilde{W}_1(y),\ y \in \mathbb{R})$ with zero mean function and covariance kernel*

$$\begin{aligned} C_1(y,t) = & f\left\{\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}K_{-1}(y \wedge t) - 1\right\}F(y \wedge t) \\ & - \frac{f^3}{d}\left(1 - \frac{K_{+1}(y)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)\left(1 - \frac{K_{+1}(t)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)F(y)F(t) \\ & - f\left\{\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}(K_{-1}(y) + K_{-1}(t) - \mathbb{E}_{\mathbb{P}}[X_1^{-1}] - 1)\right\}F(y)F(t) \end{aligned} \tag{2.23}$$

almost surely w.r.t. $\mathbb{P}$, with d given by (1.1).

Hence we have that the limiting distribution of the process W_n^H does not change if we consider the finite population as not fixed. For what concern the second process W_N, we have already observed that it is a classic empirical process so that the Donsker's Theorem holds. As a consequence we have that W_N converges weakly to a Gaussian process with zero mean function and covariance kernel $C_2(y,t)$ defined

in (0.11). In order to study the whole process W^H we have to put together the two pieces above and get the limiting distribution. The next Proposition that is the most important result of this chapter, will address this need.

Proposition 2.2.2. *The two sequences* $(W_n^H(y),\ y \in \mathbb{R})$ *and* $(W_N(y),\ y \in \mathbb{R})$ *are asymptotically independent. As a consequence, the whole process* $(W^H(y),\ y \in \mathbb{R})$ *converges weakly in* $D[-\infty, \infty]$ *endowed with the Skorokhod topology, to a Gaussian process* W *with zero mean function and covariance kernel*

$$C(y,t) = C_1(y,t) + fC_2(y,t) \tag{2.24}$$

almost surely w.r.t. $\mathbb{P}$*, where* $C_1(y,t)$ *and* $C_2(y,t)$ *are given by* (2.11) *and* (0.11) *respectively.*

Proposition 2.2.2 tell us some interesting facts. The first one is that in the process W^H the sampling design variability and the superpopulation variability are uncorrelated when considering large samples. This (asymptotic) independence is reflected in the covariance kernel of the process W^H, where we have a larger variability if compared to the process W_n^H analyzed in the previous chapter. The last observation confirms the remark that the contribution to the covariance kernel of the standard empirical process W_N is scaled by a factor f having as a consequence that considering a small sampling fraction, makes the role of the superpopulation negligible with respect to the role of the finite population.

As we did in the previous chapter we are able to state a Proposition equivalent to Proposition 2.2.2 that consider the Horvitz-Thompson estimator of the superpopulation distribution function. Of course, the same reasoning followed to show the unconditional convergence of the process W_n^H is still valid also for the process

$$W_n^{HT} = \sqrt{n}(\hat{F}_{HT} - F). \tag{2.25}$$

Results contained in Proposition 2.2.2 are still valid if we consider the process $W^{HT} = \sqrt{n}(\hat{F}_{HT} - F)$ where the Horvitz-Thompson is concerned, with a few changes only in the covariance kernel.

Proposition 2.2.3. *The two sequences* $(W_n^{HT}(y),\ y \in \mathbb{R})$ *and* $(W_N(y),\ y \in \mathbb{R})$ *are asymptotically independent. As a consequence, it holds true that the whole process* $(W^{HT}(y),\ y \in \mathbb{R})$ *converges weakly in* $D[-\infty, \infty]$ *endowed with the Skorokhod*

topology, to a Gaussian process W' with zero mean function and covariance kernel

$$C^{HT}(y,t) = C_1^{HT}(y,t) + fC_2(y,t) \tag{2.26}$$

almost surely w.r.t. $\mathbb{P}$, where $C_1^{HT}(y,t)$ and $C_2(y,t)$ are given by (2.12) *and* (0.11) *respectively.*

We now investigate closer the case of independence between the interest variable and the size variable, that, as seen in the previous section, is of special interest. Let us first see what happens to the covariance kernel (2.24) when independence is assumed. The covariance kernel becomes:

$$\begin{aligned} C^{indep}(y,t) &= f(A-1)(F(y \wedge t) - F(y)F(t)) + f(F(y \wedge t) - F(y)F(t)) \\ &= fA(F(y \wedge t) - F(y)F(t)) \end{aligned} \tag{2.27}$$

where A is defined in (2.14).
As we can see the limiting process is proportional to a Brownian Bridge on the scale of F, with a factor of proportionality fA that takes into account the fact that we are sampling from a finite population (factor f) and the dependence between units induced by the sampling design P with inclusion probabilities $\pi_1, \ldots, \pi_N$.

Going further, what would happen if we consider a simple random sampling as sampling design? We have already observed that in presence of inclusion probabilities equal to $\pi_1 = \ldots = \pi_N = f$ the factor A becomes equal to $1/f$. Hence the covariance kernel (2.27) becomes

$$C_{SRS}^{indep}(y,t) = F(y \wedge t) - F(y)F(t)^3 \tag{2.28}$$

that is the exactly the covariance kernel of a Brownian Bridge (on the scale of F). This means that the process W^H has the same asymptotic behavior of an empirical process under the standard assumption of *i.i.d.* data. This result is not really surprising, since if we put in the Hajek estimator $\widehat{F}_H$ weights equal to N/n, we have $\widehat{F}_H = \widehat{F}_n$, i.e. under the simple random sampling (S.R.S. in the sequel) the Hàjek estimator of the distribution function is equal to the empirical cumulative distribution function computed on the sample data ($\hat{F}_n$ should not be confused with F_N that is substantially an ECDF but computed on the whole finite population). As

[3] We kept the superscript *indep* because, if considering the Simple Random Sampling design, as a consequence you will have no relationship between Y and the design variable X

suggested by intuition, asymptotically the role of the simple random sampling design is negligible, increasing indefinitely the size of the population makes it virtually infinite, and units can be seen as independently selected from the superpopulation.

With the same reasoning for the process W^{HT} it is easy to see that under the assumption of independence, the covariance kernel becomes:

$$C^{HT,indep}(y,t) = f\left\{AF(y \wedge t) - \left(\frac{(1-f)^2}{d} + 1\right) F(y)F(s)\right\} \tag{2.29}$$

where A is defined in (2.14).
In this case we do not have the proportionality to the classic Brownian Bridge. As we have discussed in the previous chapter, the property of the Horvitz-Thompson estimator of not being a proper estimator of the distribution function F, brings an excess of variability that destroys the Brownian Bridge structure. Let us see now what happens if we consider a simple random sampling design. In this case it is easy to see that the Horvitz-Thompson estimator, exactly as it happen for the Hàjek estimator, becomes equal to the Empirical Cumulative Distribution Function computed on the sample data ($\hat{F}_{HT} = \hat{F}_n$). As a consequence, under the S.R.S., also the Horvitz-Thompson estimator becomes a proper estimator of the distribution function F. In addition we have that under the S.R.S. design the divergence d defined in (1.1) becomes

$$d = f - f^2 = f(1-f).$$

Hence, it is easy to see that the limiting covariance kernel is equal to

$$C_{SRS}^{HT,indep}(y,t) = F(y \wedge t) - F(y)F(s) \tag{2.30}$$

Hence, due to the property of normalizing the Horvitz-Thompson estimator of the Simple Random Sampling design we have that also the process W^{HT} asymptotically behaves as a Brownian Bridge.

The next corollary summarizes the latest observations

Corollary 2.2.1. *The following claims hold:*

Claim 1 *If the interest character Y and the size variable X are independent the sequence W^H defined in* (2.21) *converges weakly to the process*

$$\sqrt{fA}B(F(y)),\ y \in \mathbb{R} \tag{2.31}$$

where A is defined in (2.14) *and B is a Brownian Bridge.*

Claim 2 *If the interest character Y and the size variable X are independent the sequence W^{HT} defined in* (2.25) *converges weakly to a Gaussian process with zero mean function and covariance kernel given by* (2.29)

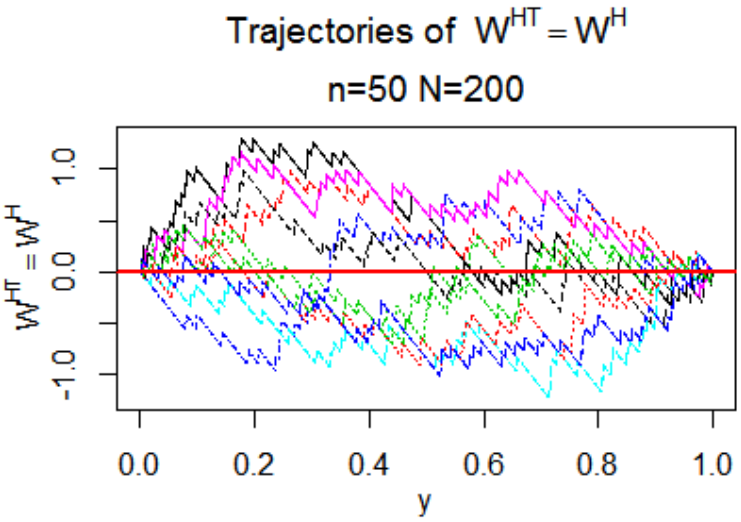
Claim 3 *If the actual sampling design P is a Simple Random Sampling, both the sequences W^H and W^{HT} converges to the process*

$$B(F(y)),\ y \in \mathbb{R} \tag{2.32}$$

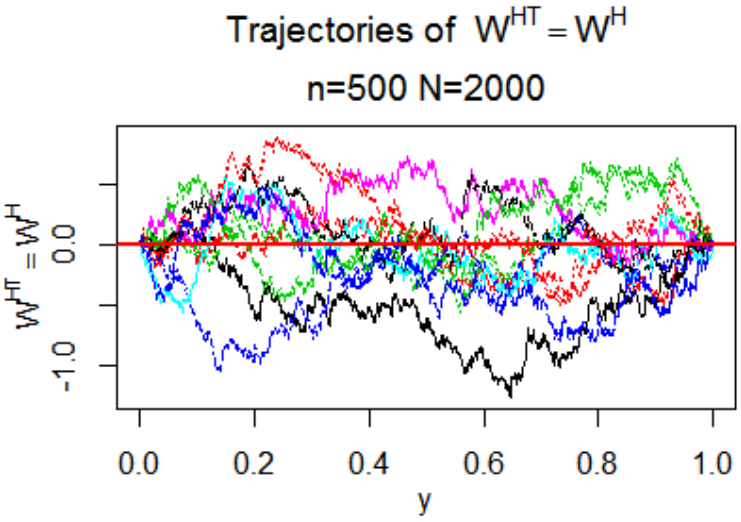
where B is a Brownian Bridge.

Since in our opinion a graphic vision of very abstract concepts can sometimes help the understanding, as done in the previous section we report graphs where some simulated trajectories of the processes examined in the various situations considered in this section are represented.

Situation of S.R.S. Design



(a)



(b)

Fig.2.3. Trajectories of the empirical process (for finite population) where the finite population is variable and the centering factor of the process is the superpopulation distribution function F. The SRS sampling design is assumed.

The trajectories in Figures 2.3, 2.4 and 2.5 are obtained by assuming a uniform $(0, 1)$ model for the superpopulation. In particular, Figure 2.3 shows the case where a Simple Random Sampling design is assumed; as we have seen, in this case the processes W^{HT} and W^{H} coincides and their limiting process is a Brownian Bridge subordinated to F. It is worth to highlight that in the case under examination $F(y) = y$, thus the limiting process is a standard Brownian Bridge on a linear scale. Looking at both Figures 2.3a and 2.3b is evident the symmetry and the larger variability around $y = 0.5$. Clearly the approximation to the Brownian Bridge is better when considering a bigger sample size (2.3b) with smoother trajectories, although we have no pathological situation that evidently deviates from the Brownian Bridge, also when the sample size is lower (2.3a).

Let consider now a situation of independence between the interest variable and the size variable, but assuming a generic πps sampling design (Pareto design in this case, see Remark 2.1.2 below for some clarifications) that is the case of Figure 2.4. In this case two things are quite important to see, when considering a lower sample size (Figures 2.4a and 2.4c) independently from considering the Hajek estimator or the Horvitz-Thompson one, we have some not well-behaved trajectories. This pathological situation vanishes when the sample size (Figures 2.4b and 2.4d) increases providing a better approximation to the limiting processes. The second remark is that, as seen in the previous section for a fixed finite population, and as suggested by theoretical results, when considering the Horvitz-Thompson estimator, we have an extra variability with respect to the Hàjek estimator.

Looking at the most general case (Figure 2.5), that is assuming a correlation between the interest variable Y and the auxiliary variable X (for this simulations the correlation between these two variables is about 0.40) we first see that the variability of the processes is smaller than the independence case. Thus, assuming more information brings us a reduction of the variability. In addition, incorporating more information entails well behaved trajectories also with a sample size of $n = 50$, as it is seen from Figures 2.5a and 2.5c, while this does not happen in the independence case. When the sample size increases the processes are more regular (Figures 2.5b and 2.5d) and quite similar, and the most important difference is due to the *properness* property of the Hàjek estimator.

Making a comparison with the situations reported in Figures 2.1 and 2.2, it is evident the difference in the variability of the processes. Processes analyzed in Figures 2.4 and 2.5 show a higher variability due to the presence of the superpopulation

Situation Of Independence Between Y And X

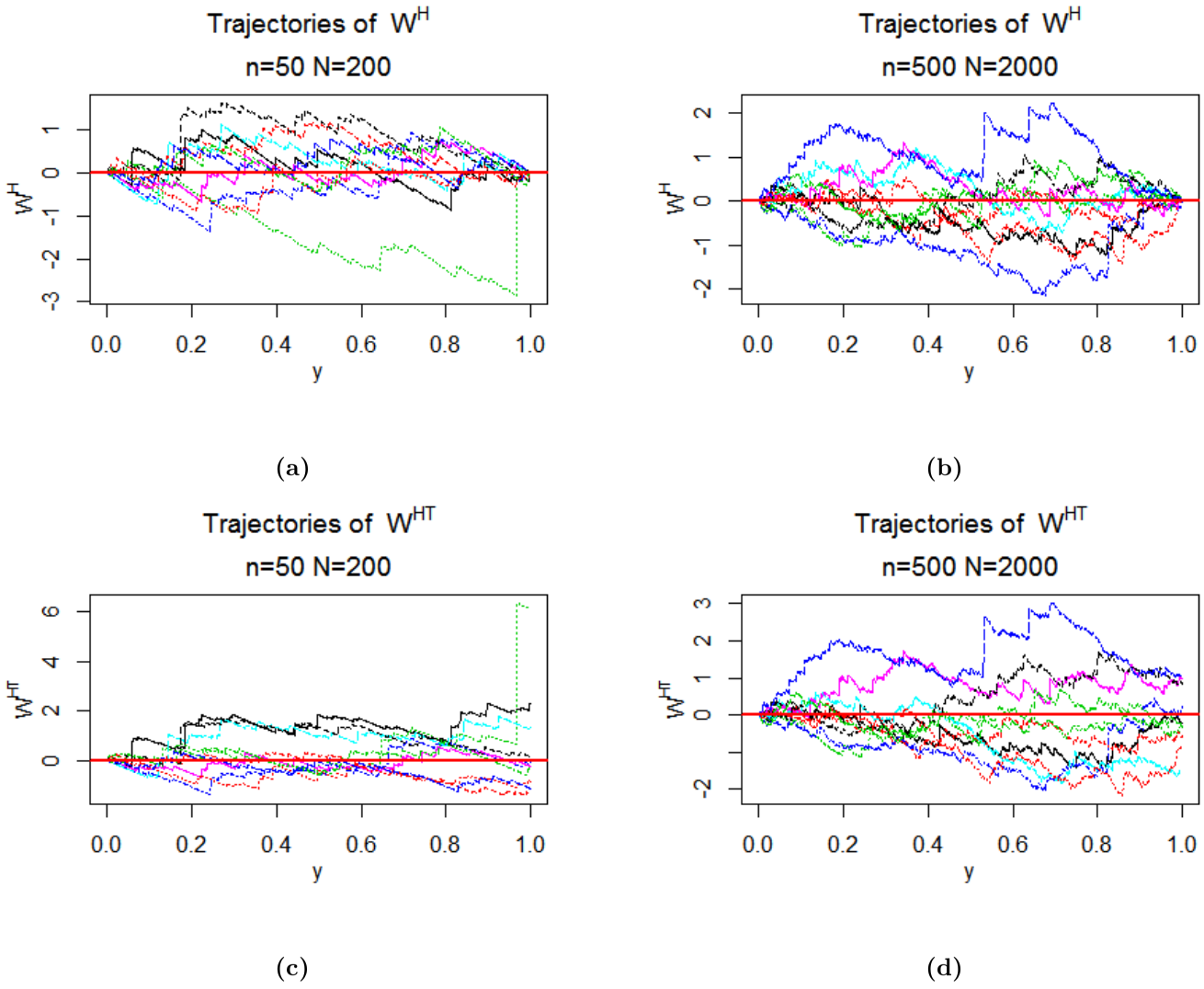


Fig.2.4. Trajectories of the empirical process (for finite population) where the finite population is variable and the centering factor of the process is the superpopulation distribution function F. Independence between interest and size variables is assumed.

Situation Of Dependence Between Y And X

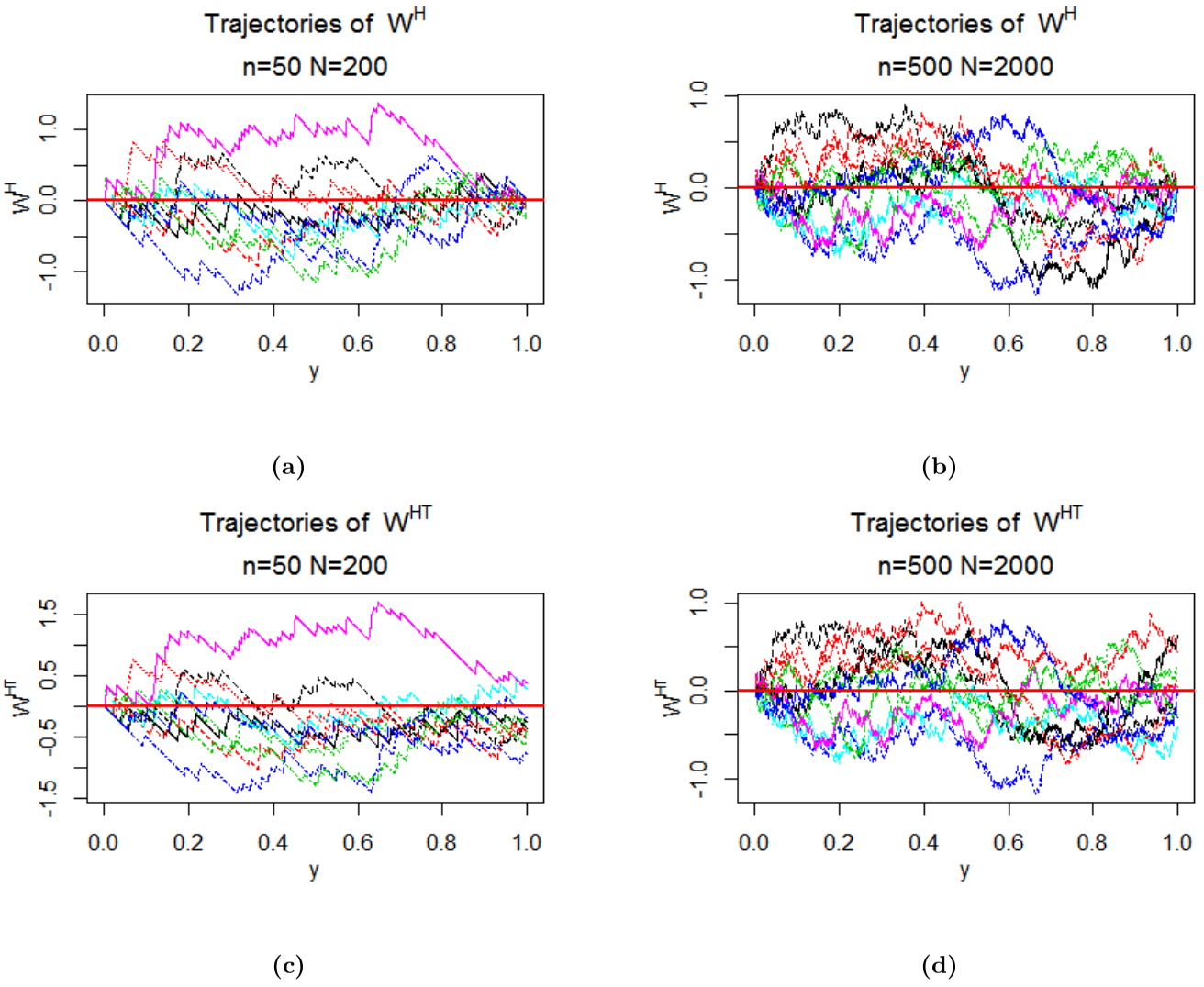


Fig.2.5. Trajectories of the empirical process (for finite population) where the finite population is variable and the centering factor of the process is the superpopulation distribution function F. Dependence between interest and size variables is assumed.

randomness that is not present in the cases examined in the previous section where the attention were focused on the finite population distribution function.

We state now a generalization of the Glivenko-Cantelli Theorem that will be useful in the next chapter

Proposition 2.2.4. *Under the hypotheses $H1 - H6$, we have:*

$$\sup_y \left|\widehat{F}_H(y) - F(y)\right| \to 0 \text{ as } N \to \infty \tag{2.33}$$

for a set of (sequences of) Y_is, T_{ij}s having $\mathbb{P}$-probability 1, and for a set of $\mathbf{D}_N$s of P-probability tending to 1 as N increases.

Proposition 2.2.4 states the uniform convergence of the Hajék estimator to the real superpopulation distribution function F. What is different from the Glivenko-Cantelli is the form of convergence. While for the Theorem 0.2.1 the convergence is *almost sure* (as a consequence of the strong Law of Large Numbers), here we have to consider two types of convergence because of the presence of two source of randomness. Thus the convergence w.r.t. the sampling design is *in probability* (as a consequence of the weak Law of Large Numbers and the dependence between units) and the convergence w.r.t. the superopopulation model is almost sure (as a consequence of the Strong Law of Large Numbers and independence of the finite population units).

Remark 2.2.1. The usual Glivenko-Cantelli Theorem does not hold in this framework even if the Y_is are *i.i.d* in the superpopulation. In fact the usual empirical cumulative distribution function based on the sample data defined as

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{N} D_i I_{(y_i \leq y)}. \tag{2.34}$$

is not a consistent estimator of the superpopulation distribution function F.[4] Computing the expected value of the ECDF we have

$$\begin{aligned} \mathbb{E}_{\mathbb{P},P}\left[\widehat{F}_n(y)\right] &= \frac{1}{n}\sum_{i=1}^{N} \mathbb{E}_{\mathbb{P}}\left[\pi_i I_{(y_i \leq y)}\right] \\ &\to \mathbb{E}_{\mathbb{P}}\left[X I_{(Y \leq y)}\right] / \mathbb{E}_{\mathbb{P}}\left[X\right] \neq F(y) \end{aligned} \tag{2.35}$$

[4] If a Simple Random Sampling is considered the Hàjek estimator is equal to the ECDF based on the sample data. Hence $\hat{F}_n$ is consistent.

as N increases. Relationship (2.35) shows that the ECDF (2.34) is asymptotically biased, and hence inconsistent.

The above result can be slightly refined. Using the same approach as in Lemma 2.1.1, it is not difficult to show that by the Laws of Large Numbers, as N increases,

$$\widehat{F}_n(y) \to \mathbb{E}_{\mathbb{P}}\left[XI_{(Y\leq y)}\right]/\mathbb{E}_{\mathbb{P}}\left[X\right] \neq F(y)$$

for a set of (sequences of) y_is, t_{ij}s having $\mathbb{P}$-probability 1, and for a set of $\mathbf{D}_N$s of P-probability tending to 1. This makes it stronger the assertion about the inconsistency of $\widehat{F}_n(y)$, because it shows that such an inconsistency is due to the inability of the ECDF to take into account a complex sampling design.

Moving to parameter estimation we remember here that we are interested in parameters that can be expressed as functional of the superpopulation distribution function. Thanks to Propositions 2.2.2 and 1.2.1 we have that the following result holds true.

Proposition 2.2.5. *Suppose that $\theta(\cdot)$ is (continuously) Hadamard-differentiable at F, with Hadamard derivative $\theta'_F(\cdot)$. Assuming $H1 - H6$, it holds that*

i) The sequence $(\sqrt{n}(\theta(\hat{F}_H) - \theta(F)),\ y \in \mathbb{R})$ converges weakly to $\theta'_F(W)$, almost surely w.r.t. $\mathbb{P}$, as N increases.

ii) The sequence $(\sqrt{n}(\theta(\hat{F}_{HT}) - \theta(F)),\ y \in \mathbb{R})$ converges weakly to $\theta'_F(W')$, almost surely w.r.t. $\mathbb{P}$, as N increases.

The Hadamard-differentiability assumption essentially allows us to characterize the large sample distribution of the interest parameter without any additional effort. In addition it is worth to notice that if $\theta(\cdot)$ takes value on the real line, which is of a primary interest in dealing with statistical parameters, the limiting random variable $\theta'_F(W)$ is Gaussian and centered. In fact, the linearity of the Hadamard derivative preserve both normality and the zero mean. Thus, the variance of $\theta'_F(W)$ is equal to

$$\sigma^2_\theta = \mathbb{E}[\theta'_F(W)^2]. \tag{2.36}$$

Clearly with the right modifications this latter observation holds when the Horvitz-Thompson estimator is considered.

2.3 Parallel Results

As a support to the modernity of the results shown in the present work, in this section we will discuss the recent paper Boistard et al. (2015), contemporary to the present dissertation, that reaches the same type of results shown in this chapter.

What Boistard et al. make in their work is to establish a (design-based) Functional Central Limit Theorem valid in survey sampling framework. This is translated into recovering the large sample distribution of the Hajek empirical process and Horvitz-Thompson empirical process by taking in exam both the situation of considering only the sampling design variability (i.e. in our notation considering the process W_n^H and W^H) and also taking into account the superpopulation randomness (i.e. in our notation considering the process W_n^{HT} and W^{HT}), that is exactly the aim of the present chapter. In particular looking at results established by Propositions 2.1.1, 2.1.2, 2.2.2 and 2.2.3 of the present work, it is easy to see that they are very similar to those established by Theorems 3.1, 3.2 and Propositions 4.1, 4.2, but with some notable differences.

The first important difference is in the assumptions. In their paper Boistard et al., in order to recover the large sample distribution of the considered empirical processes, they impose some bounds on high order correlation between the inclusion variables D_is. In particular these assumptions is used to prove the *tightness* of the finite dimensional distribution of the empirical processes considered. As they notice, this approach is well supported by sampling theory literature when dealing with asymptotic approach. Another assumption they make is that the Horvitz-Thompson estimator $\hat{F}_{HT}(y)$ of the distribution function F is normally distributed for every fixed y in $\mathbb{R}$ (such an assumption is needed to prove that the finite dimensional distribution of considered processes are Gaussian). Those assumptions together make one able to establish the large sample behavior of the Horvitz-Thompson empirical process or equivalently the Hàjek one. This approach is quite different from the one we pursued. In our work the key assumption is made on the type of sampling designs that we consider, that is considering only high entropy sampling design. Concerning only sampling design that show a high entropy ensures: *i*) the normality of the Horvitz-Thompson estimator $\hat{F}_{HT}(y)$, and equivalently of the Hàjek estimator $\hat{F}_H$ trough the Theorem 0.3.1 (where the Lindeberg-Hajek conditions are verified by Lemma 2.1.5), with no need of assuming it and *ii*) because of the asymptotic equivalence between the high entropy sampling designs, discussed in Preliminaries, makes

it easy proving the tightness of the finite dimensional distributions, allowing the possibility of limiting the proof to the case were the rejective sampling is concerned (for more on this see the appendix of Conti (2014)). On the other side however we have to admit that although their approach can be judged quite intricate with respect to the one that we propose, it is more general than ours.

Another important difference is in the approach pursued, that is more *mathematical* and less *statistical*. Their asymptotic results are given in terms of the first and second order inclusion probabilities, specifying that random inclusion probabilities are allowed, but they do not develop this situation. They focused on the simplest case where the inclusion probabilities of the first and second order are deterministic. In our opinion this approach does not take into account what usually happens in statistical practice when dealing with survey sampling. In fact, it is a common procedure, when dealing with finite populations that if additional information about some auxiliary variables is available, this extra knowledge is used to produce better estimates, taking the inclusion probabilities proportional to these auxiliary variables (we have discussed this point when we recalled the Horvitz-Thompson estimator in Preliminaries). Clearly, in this framework, if the presence of a superpopulation model is assumed, it is more reasonable to see the inclusion probabilities as random quantities defined through superpopulation variables, than deterministic. Moreover, assuming a link between the inclusion probabilities and some design variables in the superpopulation, allow to explicitly show the role of the dependence between the interest character and the design variables in the covariance kernel. As a consequence, it is possible to fully analyze the behavior of the limiting process of the empirical processes in finite population framework, in different interesting scenarios that are common in sampling theory, as we have done in the previous section of this chapter. The last point we aim at discussing is that assuming a multivariate superpopulation model where the variables exhibits some form of dependencies, that is our approach, is more general instead of considering only a univariate superpopulation.

Let's go back for a while to the key assumption of dealing only with High Entropy sampling designs. As we have seen this assumption considerably simplifies results, when asymptotic normality and tightness are concerned, but its consequences are not ended. In fact looking at the covariance kernels of Theorems 3.1, 3.2 and Propositions 4.1, 4.2 in Boistard et al. (2015), we see that these kernels explicitly depend on the second order inclusion probabilities. In our results second order inclusion

probabilities do not appear in the covariances of the limiting processes. Also this simplification is a consequence of the hypothesis of considering only High Entropy sampling design. In fact, because of approximation given by (0.20), for sampling designs that asymptotically maximize the entropy, the second order inclusion probabilities can be expressed as function of only the first order inclusion probabilities, making the results more easily readable and more easy applicable.

Moving to parameter estimation, also Boistard et al. exactly as we do, see in the Hadamard-differentiability the regularity condition that allows an easy recovering of the limiting law of interest parameter that can be seen as a functional of the finite population distribution function F_N or equivalently of the superpopulation distribution function F. They focus on the poverty rate, while we will introduce in the Applications Chapter some other Hadamard-differentiable functionals that can be useful in very common statistical problems.

Although there are some little differences in our approach and their, our main additional contribution is that in the next chapter we will provide a resampling scheme that is able to recover the limit distribution of the considered processes and also of interest parameters, while their work lacks of such a procedure. In particular, we want to highlight that in order to use the Functional Central Limit Theorem introduced by Boistard et al. we must explicitly derive the analytic form of the Hadamard derivative of the parameter considered. Using a resampling technique, computational power is needed, but it is possible to avoid the need for finding the analytic form of the derivative of the parameter of interest, that can represent a hard problem especially for *statistics practitioners*.

Chapter 3

Resampling

In this section we will face the problem of resampling techniques in survey sampling. Firstly a general overview about resampling methods in finite population framework is proposed (the discussion is focused on fixed-population design-based framework). Then, we will introduce our original generalization to the superpopulation framework of the resampling procedure proposed in Conti et al. (2015). At the end, the properties and the methodological validation of such a procedure are provided.

3.1 State Of The Art

Nowadays resampling methods such as *Bootstrap*, are a standard tool in statistics. Since the innovative paper by Efron (1979) has been published, because of its applicability, and with the exponential growth of computational power and its availability, resampling became a trend topic studied by the scientific community. In the literature, several variations to the usual bootstrap technique have been proposed in order to extend the validity of this useful and simple method to different branches of statistics. The main goal of this section is to discuss the main contributions to resampling in survey sampling.

Original Efron's bootstrap, was proposed for the classical setting of $i.i.d.$ observations. The idea behind the usual bootstrap is simple. Using the data to mimic the process that generates it. The mimicking process can be represented in a visual way, by the following diagram:

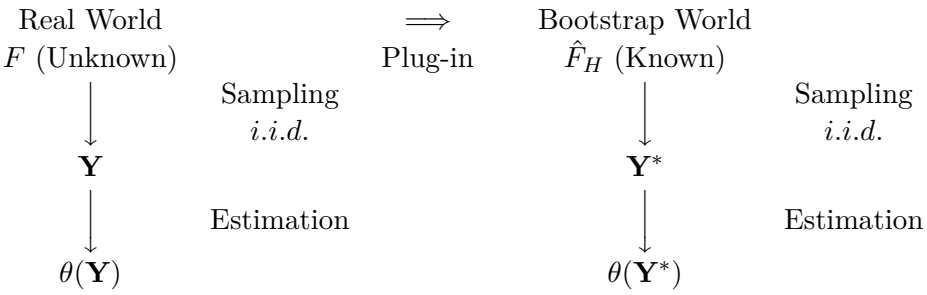


Fig.3.1. Classic Bootstrap mimicking scheme

A few words about Figure 3.1 are necessary. Under *i.i.d.* assumption, what we are assuming to be the truth, is that our sample of observation $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is obtained as independent replications of a variable Y with distribution function F that is partially or totally unknown (maybe we lack information about its parameters but we know its functional form, or maybe we don't know its form but we have some information about some parameters or maybe we ignore everything about it). After collecting the data, we proceed to the estimation step with the need of assessing the accuracy of the estimate produced in such a way. This can be a difficult problem. What "usual" bootstrap does, is to replace the unknown distribution F with a known (consistent) estimator, usually the ECDF $\hat{F}_N$. Then, it follows the same process that generates the original sample $\mathbf{Y}$. In this case this means sampling independently from $\hat{F}_N$ or equivalently, resampling with replacements (in order to have independence) from the sample $\mathbf{Y}$ obtaining a bootstrap sample $\mathbf{Y}^*$. Quoting Efron et al. (2003), the advantage of bootstrap procedure, is that the accuracy of inference can be assessed by using the observed variability of the bootstrap replications $\theta(\mathbf{Y}^*)$. Moving from the left hand side of Figure 3.1 to the right hand side is what Efron defines as *plug in principle*. You move from the real world to the bootstrap world by only plugging in an estimate of the distribution function and this is the only hard part, the rest is just mimicking the data generating process.

Although the idea behind bootstrap is very intuitive, its theoretical justification is not so easy. The validation of usual bootstrap is provided in Bickel and Freedman (1981) and it is based on asymptotic considerations. The key of the bootstrap consistency is that asymptotically the resampling distribution (known, or at least approximable using computational resources) of the statistics of interest is the same of the true distribution (unknown, partially or totally) of the statistics themselves. Hence, asymptotically the resampling distribution and original distribution are interchangeable, in the sense that using one or the other brings to the same inferential

conclusion. A very informal way to figure that is imagining bootstrap as a set of chinese boxes. In the same way a sample brings you information about a population, re-samples of a sample bring you information about the sample. If a "large" sample is available (asymptotic approach), the difference between the population and the sample is negligible hence the information that sub-samples bring about the sample are also information about the population. We have also all the tools to give a more formal sketch on why bootstrap works. Consider a $n-$sized sample $S = (Y_1, \ldots, Y_n)$ of *i.i.d.* replications of a variable $Y \sim F$. We know that the ECDF $\hat{F}_n$ is a consistent estimator of F and also that for every fixed y its variance takes value $\dfrac{F(y)(1-F(y))}{n}$ as showed in (0.5). Consider now a bootstrap sample $S^* = (Y_1^*, \ldots, Y_n^*)$ with the same size of S. Conditionally on the sample S, the ECDF computed on the sample S^* is equal to

$$\hat{F}_n^* = \frac{1}{n}\sum_{i=1}^{n} I_{(Y_i^* \leq y)} \tag{3.1}$$

where

$$Y_i^* = \begin{cases} Y_1 \text{ with probability } 1/n \\ \vdots \\ Y_n \text{ with probability } 1/n \end{cases}. \tag{3.2}$$

Hence,

$$I_{(Y_i^* \leq y)} = \begin{cases} I_{(Y_1 \leq y)} \text{ with probability } 1/n \\ \vdots \\ I_{(Y_n \leq y)} \text{ with probability } 1/n \end{cases}. \tag{3.3}$$

By the law of large numbers, conditionally on $Y_1^*, \ldots, Y_n^*$, we have that

$$\hat{F}_n^* \xrightarrow{a.s.} \mathbb{E}[\hat{F}_n^*],$$

but

$$\mathbb{E}[\hat{F}_n^*] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[I_{(Y_i^* \leq y)}] = \mathbb{E}[I_{(Y_i^* \leq y)}] = \frac{1}{n}\sum_{i=1}^{n} I_{(Y_i \leq y)} = \hat{F}_n \tag{3.4}$$

Thus it holds

$$\hat{F}_n^* \xrightarrow{a.s.} \hat{F}_n \xrightarrow{a.s.} F. \tag{3.5}$$

Let have a look to the variance of $\hat{F}_n^*$

$$\mathbb{V}[\hat{F}_n^*] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[I_{(Y_i^*\leq y)}] = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}I_{(Y_i\leq y)} - \left(\frac{1}{n}\sum_{i=1}^{n}I_{(Y_i\leq y)}\right)^2\right) = \frac{\hat{F}_n(1-\hat{F}_n)}{n}. \tag{3.6}$$

By the consistency of $\hat{F}_n$ and what we have just observed, it easy to prove that

$$\sqrt{n}(\hat{F}_n^* - F_n) \xrightarrow{weak} B(F(y)) \tag{3.7}$$

that is exactly what happens for the classic empirical process.

When dealing with finite populations and complex sampling designs, the usual bootstrap fails. This is quite intuitive, in fact as we have seen to apply usual bootstrap we have to resample from the ECDF of the sample data. Hovewer as we have seen in Remark 2.2.1, the usual ECDF generally is not a consistent estimator of the finite population distribution F_N or also of the superpopulation ditribution F, because it is not able to take into account the presence of the sampling weights. This inconsistency implies that also if we are dealing with very large sample, $\hat{F}_n$ can be very different from the real distribution function, hence the bootstrap is unable to recover information about the real distribution function. Under the same assumptions and using the same notation introduced until now, we are able to show in a more formal way the inconsistency of bootstrap procedure when dealing with finite populations and complex designs.

We have already shown that the ECDF $\hat{F}_n$ (based on the sample s, selected from a finite population with a complex sampling design P with first order inclusion probabilities proportional to a size variable) is not consistent as an estimator of finite population distribution function if the sampling weights are not trivial. The main consequence of this observation is that the stochastic process

$$\sqrt{n}(\hat{F}_n - F_N) \tag{3.8}$$

does not have a zero mean function neither asymptotically. If we want to use the classic bootstrap procedure, we have to sample independently from $\hat{F}_n$ for obtaining a bootstrap sample s^*. The ECDF $\hat{F}_n^*$ based on the bootstrap sample is defined as in 3.1 and it is a consistent estimator of $\hat{F}_n$ as shown in 3.4. Thus the process

$$\sqrt{n}(\hat{F}_n^* - \hat{F}_n) \tag{3.9}$$

has a zero mean function asymptotically, thus it is different from process 3.8. This is well represented by Figure 3.2. Here are represented some trajectories of the process $\sqrt{n}(\widehat{F}_n - F_N)$ and its resampled version $\sqrt{n}(\widehat{F}_n^* - \widehat{F}_n)$ where the finite population is generated as *i.i.d* replication of a uniform distribution on the unit interval and samples from the finite population are drawn using a Pareto design, using a size variable for the inclusion probabilities that shows an empirical correlation of about 0.40 with the interest variable. It is clear that the classical bootstrap does not recover the original process (Figs. 3.2a 3.2b) neither when a "large" sample size of $n = 500$ is considered (Figs. 3.2c 3.2d). The resampled process shows a Brownian Bridge behavior while the original process if far away from being such a process.

Even if we consider $\hat{F}_H$ that takes into account the presence of the sampling weights and it is a consistent estimator of the distribution function F_N the original bootstrap procedure is not consistent. In fact from the previous chapter we know analytically the limit distribution of the stochastic process W_n^H. Let us see what would happen if we resample from $\hat{F}_H$. If we resample independently from $\hat{F}_H$ we will have a bootstrap sample $s^* = (y_1^*, \ldots, y_n^*)$ where

$$y_i^* = \begin{cases} y_1 \text{ with probability } \dfrac{D_1\pi_1^{-1}}{\sum_{i=1}^{N} D_i\pi_i^{-1}} \\ \vdots \\ y_N \text{ with probability } \dfrac{D_N\pi_N^{-1}}{\sum_{i=1}^{N} D_i\pi_i^{-1}} \end{cases} \tag{3.10}$$

and, as a consequence,

$$I_{(y_i^* \leq y)} = \begin{cases} I_{(y_1 \leq y)} \text{ with probability } \dfrac{D_1\pi_1^{-1}}{\sum_{i=1}^{N} D_i\pi_i^{-1}} \\ \vdots \\ I_{(y_N \leq y)} \text{ with probability } \dfrac{D_N\pi_N^{-1}}{\sum_{i=1}^{N} D_i\pi_i^{-1}} \end{cases} \tag{3.11}$$

By the Law of Large Numbers, the ECDF of the sample s^* is a consistent estimator

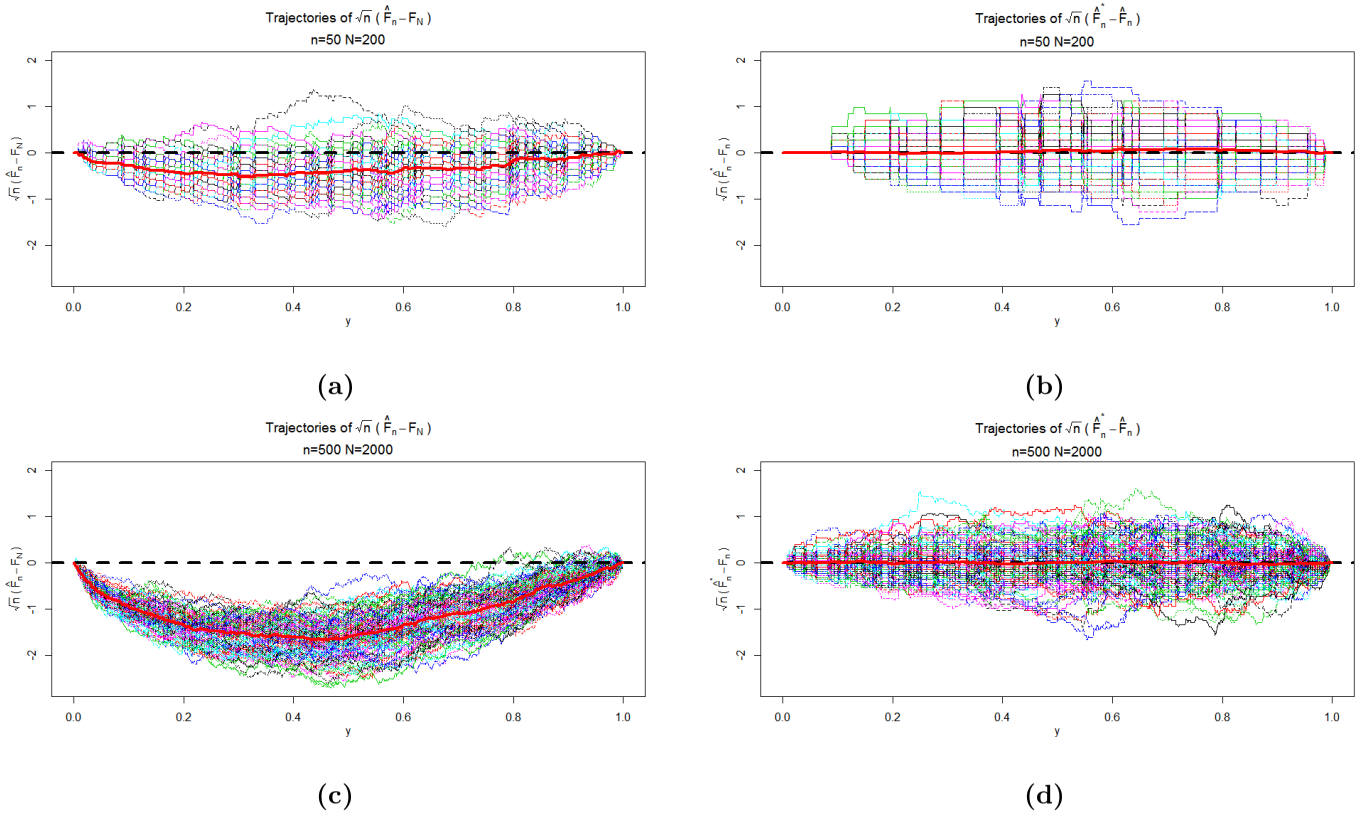


Fig.3.2. Representation of the process $\sqrt{n}(\widehat{F}_n - F_N)$ and its resampled version $\sqrt{n}(\widehat{F}_n^* - \widehat{F}_n)$

of $\hat{F}_H$ (and then of F_N) in fact, conditionally on the original sample s

$$\mathbb{E}[\hat{F}_n^*] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I_{(y_i^* \leq y)}\right] = \mathbb{E}[I_{(y_i^* \leq y)}] = \frac{\sum_{i=1}^{N} I_{(y_i \leq y)} D_i \pi_i^{-1}}{\sum_{i=1}^{N} \frac{D_i}{\pi_i}} = \hat{F}_H \tag{3.12}$$

However the variance of $\hat{F}_n^*$ is not equal, not even asymptotically, to the variance of $\hat{F}_H$. By Lemma 2.1.3, we have that conditionally on $Y_1^*, \ldots, Y_n^*$

$$\mathbb{V}[\hat{F}_n^*(y)] = \frac{1}{n^2}\sum_{i \in s^*} I_{(y_i^* \leq y)} = \frac{1}{n}\left(\mathbb{E}[I_{(y_i \leq y)}^2] - \mathbb{E}[I_{(y_i \leq y)}]^2\right) = \frac{\hat{F}_H(1-\hat{F}_H)}{n} \tag{3.13}$$

and by the Law of Large Numbers

$$\frac{n\mathbb{V}[\hat{F}_n^*(y)]}{F(y)(1-F(y))} \to 1$$

that is different from the variance of $\hat{F}_H(y)$ given by Lemma 2.1.3. Thus, *i.i.d.* resampling fails to recover the distribution of the original sample process. With reference to the last examined case, it is worth to quote the paper Conti and Marella (2015), where assuming independence between the interest variable Y and the size variable X, or equivalently assuming the inclusion probabilities π_is as deterministic, it is proposed a *rescaled bootstrap* procedure, because it holds that

$$\sqrt{n}(\hat{F}_H - F_N) \xrightarrow{weak} \sqrt{f(A-1)}B(F(y)) \tag{3.14}$$

$$\sqrt{n}(\hat{F}_n^* - \hat{F}_H) \xrightarrow{weak} B(F(y)) \tag{3.15}$$

that is, the original limiting process and the resampled limiting process coincide up to a scaling factor, and the considered limiting process is exactly a Brownian Bridge on the scale of F as it happens for the usual bootstrap procedure. We have shown that classic bootstrap fails when trying to recover the distribution of the finite population empirical process $\sqrt{n}(\hat{F}_N - F_N)$ (where $\hat{F}_N$ is an estimator of F_N). Clearly, this happens also for the superpopulation empirical process $\sqrt{n}(\hat{F} - F)$ and it is clear by the decomposition introduced in (2.22). In addition it is easy to understand that for a procedure (bootstrap) that takes into account only a sampling variability (resampling from the original sample) is quite difficult to recover the global variability of a process that takes into account two sources of randomness, that is the case of interest of this dissertation.

Because of the inconsistency of the classic bootstrap in survey sampling, a huge amount of literature focused on resampling methods available for finite populations. The most part of the resampling methods available in literature for survey sampling pursues the aim of mimicking only the first two moments of the distribution of linear functional of totals or means. In other words the goodness of these resampling methods is evaluated on how close the variance is of the resampled mean to the usual unbiased variance estimator of the original sample mean, without any further asymptotic analysis. As well discussed in Ranalli and Mecatti (2012) and Chauvet (2007) we can distinguish between two type of resampling approaches: the *ad hoc* approach and the *plug in* approach.

In *ad hoc* approaches, in order to obtain a resampling variance that is close to the Yates and Grundy unbiased variance estimator, units are resampled according to a specific procedure. We now list some famous resampling *ad hoc* methods. The first one is the *rescaled bootstrap* proposed in Rao and Wu (1988), where a S.R.S. is assumed for the original sample and then the classic bootstrap is applied to this sample. The main difference between the Rao and Wu bootstrap and Efron bootstrap is that the resampled units are scaled by a specific factor that depends on the size of the bootstrap samples, in order to have the variance of the resampled mean equal to the unbiased variance estimator of the original mean estimator. Also the resampling procedure proposed in McCarthy and Snowden (1985) is a simple application of the classic bootstrap with a finite population and a sample selected according to a simple random sampling. The variance of the mean of the resampled units is equal to the unbiased variance estimator of the original sample mean, only if the square of the sample size is equal to the finite population size (that is a quite artificial condition) and if the bootstrap samples have the same size of the original sample. In the procedure analyzed until now, the units in bootstrap samples are selected independently, as in classic bootstrap. A quite different procedure is the *Mirror-Match Bootstrap* proposed in Sitter (1992). In this case the bootstrap samples are selected of a size $n^* \lneq n$ without replacements (thus the units in the bootstrap samples are dependent), and defining $f^* = n^*/n$ bootstrap samples are selected in number of $M = \dfrac{n(1-f^*)}{n^*(1-f)}$. Then defining by S^* the union of the M bootstrap samples, it is shown that the mean computed on S^* is an unbiased estimator of the original sample mean, and also the variance of the bootstrapped mean perfectly recovers the unbiased estimator of the variance of the original mean. A bootstrap method that fully pursues the aim of mimicking the first two moments

of the distribution of a linear statistics is the one of Antal and Tillé (2011). In this paper the authors focus on the random variables that indicate how many times a unit of the original population is resampled. They provide some sufficient conditions on the expectation and covariances of these random variables, in order to have the perfect matching between the resampled moments in the linear case with their sample equivalent quantities. This bootstrap procedure is more general than the *ad hoc* approaches seen before, it allows also unequal probability sampling designs and has also the advantage of not scaling the data.

On the other side the *plug in* approach is based on the mimicking principle of the classic bootstrap. In these resampling procedures, an artificial population that plays the role of the original one is first generated based on the sample data. Then, bootstrap samples are drawn from such a population. This idea first appears in Gross (1980), where the finite population size N is assumed to be a multiple of the sample size n, and where the sample is selected according to a simple random sampling. Then a pseudo-population is generated by expanding every observation in the sample N/n times. The bootstrap samples are obtained selecting samples of size n with a simple design without replacements from the generated pseudo-population. In such a way the bootstrapped mean is an unbiased estimator of the original sample mean and the variance of the bootstrapped mean recovers the unbiased estimator of the original sample mean variance up to a factor $(n-1)/n$. Several modifications to this procedure have been proposed in literature, focusing on the pseudo population generation. For instance, in Chao and Lo (1985) it is suggested to replicate each unit $\lfloor N/n - 1/2 \rfloor$ times with a probability α or $\lfloor N/n - 1/2 \rfloor + 1$ with a probability $1 - \alpha$ (where $\lfloor \cdot \rfloor$ is the integer part operator). Then, a simple random sample without replacements is drawn from the pseudo population generated in such a way. This procedure, on the average, has the same properties of the Gross bootstrap, and avoids the restrictive assumption of N/n to be integer (if this happens this procedure coincides with the Gross one). In Booth et al. (1994) it is suggested to replicate each unit $\lfloor N/n - 1/2 \rfloor$ times, a pseudo population is obtained by adding a sample of size $N - n \times \lfloor N/n \rfloor$ selected from the original sample. Clearly if the inverse of the sampling fraction is an integer, this procedure coincides with the Gross one. A very simple idea is proposed in Presnell and Booth (1994) where a pseudo-population of the same size of the original finite population, is generated by sampling with replacement N times from the original sample. In this framework a particular role is played by the Holmberg scheme. The resampling procedure

proposed in Holmberg (1998), allows more general sampling designs (πps) and it is a generalization of the previous methods because it uses random weights to generate the pseudo population. In particular, it works as follows. For each sampled unit in the sample the sampling weights are decomposed in $1/\pi_i = c_i + R_i$ where $c_i = \lfloor 1/\pi_i \rfloor$ and clearly $0 \leq R_i < 1$. Let ϵ_i be the realization of a Bernoulli random variable of parameter R_i and define $N_i^* = c_i + \epsilon_i$. The pseudo population is defined replicating each unit i, N_i^* times. Thus, the size of the pseudo population is $N^* = \sum_{i \in s} N_i^*$. It is clear that if the size of the finite population N is a multiple of the sample size n, and the sampling design is a simple random sampling, this resampling procedure is equivalent to the Gross one. We will come back on the Holmberg scheme in the next chapter.

It is worth to notice that no one of the mentioned methods is justified by asymptotic considerations as it happens for the classic bootstrap. A procedure that takes into account the whole distribution function of a statistics and not only its first two moments is the rescaled bootstrap proposed in Conti and Marella (2015) that we quoted before. In the next sections we will introduce the resampling scheme, based on the mimicking principle, proposed in Conti et al. (2015). This resampling scheme is used by Conti et al. to make inference about the finite population. We will adapt it in order to be a valid tool to infer the superpopulation. We also provide an asymptotic validation in view of a unification of the resampling procedures in the different framework of the survey sampling and of the classic inference setting.

3.2 Resampling Procedure: Theoretical Properties

On the basis of the asymptotic results exposed in Chapter 1, making inference about hyper-parameters requires the knowledge of the analytic form of the Hadamard-derivative of the functional that defines the parameter. This can be generally quite a hard problem. In addition we want to provide a simple tool addressed also to *statistics practitioners* that could be not so strong in analytic math computation.

A resampling procedure allows us to recover the distribution of an hyper-parameter of interest by avoiding the explicit computation of the Hadamard-derivative. As already said in the previous section, our goal is to provide a resampling procedure and give a theoretical justification looking at the whole distribution function. Using asymptotic considerations is not the common practice in survey sampling bootstrap, where the usual procedure is recovering only the first two moments for linear statis-

tics. This way of validating a resampling procedure is followed in Conti and Marella (2015), where asymptotic considerations end into the mentioned *rescaled bootstrap*, but this procedure is not thought to allow a relationship between an interest variable and the design variables that is common in practice. This last point in developed in Conti et al. (2015), but this work is focused on descriptive inference, i.e. they infer the finite population and this make their class of resampling procedures not suitable for our purpose except for the "multinomial scheme" described below.

As highlighted several times, when the object of inference is an hyper-parameter, two sources of randomness have to be taken into account. Thus, for direct bootstrap methods, like the *ad hoc* procedures described in the previous sections, it is quite difficult to recover the variability of both random mechanisms. For this reason, it is intuitive that the resampling procedure considered in this work is composed by two phases. In the first phase, a prediction of the finite population is generated starting from the sample data, in order to get information about the superpopulation variability. In the second phase, a new sample, of the same size of the original one, is selected according to a sampling design P^* that fulfills the high entropy requirement. The inclusion probabilities of the resampling design are chosen proportional to the size variable X of the predicted population constructed in Phase 1. Intuitively, in such a way the sampling design randomness is also taken into account.

Phase 1.
1. Draw N units independently from the distribution $\hat{F}_H$, such that each unit $i \in s$ is selected with probability $\pi_i^{-1} / \sum_{j \in s} \pi_j^{-1} = \pi_i^{-1} / \sum_{j=1}^{N} D_j \pi_j^{-1}$
2. For $k = 1, 2, \ldots, N$, if the $k - th$ sampled unit is unit $i \in s$, take $y_k^* = y_i$ and $x_k^* = x_i$.
3. Define a predicted population of N units $\mathcal{U}_N^*$, such that unit k possesses y-value y_k^* and x-value x_k^*, $k = 1, 2, \ldots, N$.

Phase 2. Draw a re-sample s^* of size n from the population $\mathcal{U}_N^*$ defined in phase 1, using a high entropy sampling design P^* with first order inclusion probabilities $\pi_k^* = n x_k^* / \sum_{j=1}^{N} x_k^*$.

Although the sampling design P^* used in Phase 2 has to be an high entropy sampling design, it does not necessarily coincide with the sampling design P used to select the sample s from $\mathcal{U}_N$, but the resampling inclusion probabilities have a similar structure to the original ones.

This resampling scheme was first considered in Pfeffermann and Sverchkov (2006), in a different framework. In principle, it is based on a simple idea: Phase

1 mimics the generation process of the finite population from the superpopulation, and Phase 2 mimics the selection of the sample from the finite population. This is sketched in the scheme below.

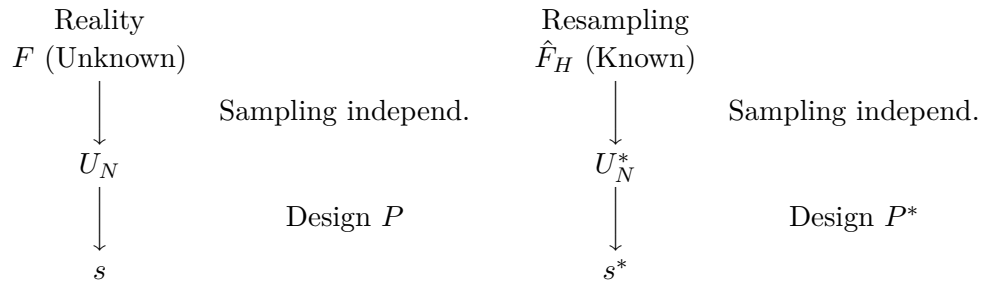


Fig.3.3. Multinomial resampling mimicking scheme

Define now N_i^* as the number of the predicted population units equal to unit i of the sample s, and let $\mathbb{P}^*$ be the probability distribution of the predicted population generating process. It is easy to see that, given s, $\mathcal{Y}_N$, $\mathcal{T}_N$, the r.v.s $(N_i^*,\ i \in s)$ possesses a multinomial distribution with:

$$\mathbb{E}_{\mathbb{P}^*}[N_i^*|\mathbf{D}_N, \mathcal{Y}_N, \mathcal{T}_N] = N\left(D_i\pi_i^{-1}/\sum_{j=1}^{N} D_j\pi_j^{-1}\right) \tag{3.16}$$

$$\mathbb{V}_{\mathbb{P}^*}[N_i^*|\mathbf{D}_N, \mathcal{Y}_N, \mathcal{T}_N] = N\left(D_i\pi_i^{-1}/\sum_{j=1}^{N} D_j\pi_j^{-1}\right)\left(1 - D_i\pi_i^{-1}/\sum_{j=1}^{N} D_j\pi_j^{-1}\right) \tag{3.17}$$

$$\mathbb{C}_{\mathbb{P}^*}[N_i^*, N_j^*|\mathbf{D}_N, \mathcal{Y}_N, \mathcal{T}_N] = -ND_iD_j\pi_i^{-1}\pi_j^{-1}/\left(\sum_{k=1}^{N} D_k\pi_k^{-1}\right)^2,\ j \neq i \tag{3.18}$$

The d.f. of the predicted population can be written as:

$$F_N^*(y) = \frac{1}{N}\sum_{i=1}^{N} I_{(y_i^* \leq y)} = \sum_{i=1}^{N} D_i\frac{N_i^*}{N}I_{(y_i \leq y)}. \tag{3.19}$$

Consider next the bootstrap replicate of the Hajék estimator of F_N^*, which is equal to

$$F_H^*(y) = \frac{\sum_{i=1}^{N}\frac{D_i^*}{\pi_i^*}I_{(y_i^* \leq y)}}{\sum_{i=1}^{N}\frac{D_i^*}{\pi_i^*}}. \tag{3.20}$$

Looking at the proposed resampling procedure, it is clear that pushing the sample size and the finite population size to the infinity makes it negligible the difference

between the real and the predictive population (and the difference on how they are generated, as well). In the sequel we aim at formalizing this remark by showing that the asymptotic distribution of the resampled process

$$W^{H*}(y) = \sqrt{n}(\widehat{F}^*_H(y) - \widehat{F}_H(y)),\ y \in \mathbb{R},\ N \geq 1. \tag{3.21}$$

coincides with the asymptotic distribution of W^H given in Proposition 2.2.2.

Exactly as it happens to the resampling procedure that mimics the data generation process, the theory will retrace what we have done in Chapter 1. We start by stating the validity of the preparatory Lemmas 2.1.1-2.1.5 also for the resampled process (3.21)

Lemma 3.2.1. *Under hypothesis $H1-H6$, conditionally on $\mathcal{Y}_N, \mathcal{T}_N, \mathbf{D}_N$, as N approaches to infinity the statements of Lemmas 2.1.1-2.1.5 hold true for the predicted population $\mathcal{U}^*_N$, in probability w.r.t. $\mathbb{P}^*$, for a set of samples $\mathcal{S}$ of $P-$probability tending to 1, and for a set of y_is and t_{ij}s of $\mathbb{P}$-probability 1.*

We are now ready to state the result that establishes the same limiting behavior of W^H for the resampled process W^{H*}

Proposition 3.2.1. *Suppose the sampling design P and the resampling design P^* both satisfy assumptions $H1-H6$. The following claims hold.*
Claim 1 *Conditionally on $\mathcal{Y}_N, \mathcal{T}_N, \mathbf{D}_N$, N_i^*s, the sequence $(W_n^{H*}(y) = \sqrt{n}(\hat{F}^*_H(y) - F^*_N(y)),\ y \in \mathbb{R},\ N \geq 1)$ converges weakly, in $D[-\infty, \infty]$ equipped with the Skorokhod topology, to a Gaussian Process $\widetilde{W}_1^*$ with zero mean function and covariance function given by* (2.11)*. The convergence holds for almost all y_is, t_{ij}s, for a set of $\mathbf{D}_N$s of P-probability tending to* 1*, and for a set of N_i^*s of $\mathbb{P}^*$-probability tending to* 1*.*
Claim 2 *Conditionally on $\mathcal{Y}_N, \mathcal{T}_N, \mathbf{D}_N$, the sequence of random functions $(W_n^{H*}(y) = \sqrt{n}(\hat{F}^*_H(y) - F^*_N(y)),\ y \in \mathbb{R},\ N \geq 1)$ converges weakly, in $D[-\infty, \infty]$ equipped with the Skorokhod topology, to a Gaussian Process W_1^* with zero mean function and covariance function given by (2.11). The convergence holds for almost all y_is and t_{ij}s, and for a set of $\mathbf{D}_N$s of P-probability tending to* 1*.*
Claim 3 *The two sequences $(W_n^{H*}(y),\ y \in \mathbb{R})$ and $(W_N^*(y),\ y \in \mathbb{R})$ are asymptotically independent. Moreover, the following further statements hold true.*

- *R1 The whole process $(W^{H*}(y),\ y \in \mathbb{R})$ converges weakly in $D[-\infty, \infty]$ endowed with the Skorokhod topology, to a Gaussian process W^* with zero mean function and covariance kernel given by* (2.24)*.*

R2 If $\theta(\cdot)$ is continuously Hadamard differentiable at F, then $(\sqrt{n}(\theta(\widehat{F}_H^*)-\theta(\hat{F}_H)),\ N \geq 1)$ converges weakly to $\theta'_F(W^*)$, as N increases.

In both R1, R2 the convergence hold for almost all y_is and t_{ij}s, and for a set of $\mathbf{D}_N$s of P-probability tending to 1 as N increases.

Claim 1 is about the convergence of the resampled process W_n^{H*}. This approach is the one pursued by Conti et al. (2015) to show that the "multinomial scheme" works in their setting, where the object of the inference is the finite population that is assumed fixed even if generated by a superpopulation model. In fact, the convergence holds conditionally on the variables N_i^*s, that in a less formal language means that the predictive population $\mathcal{U}_N^*$ is fixed, the only assumed variability is the one of the resampling design P^*. Claim 2 extends the convergence of the resampled process W_n^{H*} considering a varying predictive population. Clearly this extension is obtained by resorting to Lemma 2.2.1, that is exactly what we have done in order to generalize the convergence of the process W_n^H to the unconditional case. Claim 3 summarizes the limiting behavior of the whole resampled process W^{H*} and of the process obatined concerning a parameter of the superpopulation stating the equivalence between the limiting distributions of the resampled processes and the distributions of original processes.
Clearly in the spirit of Bickel and Freedman (1981), Proposition 3.2.1 provides a full asymptotic justification of the resampling procedure considered in the present section.

For the sake of completeness we want to remark that this resampling scheme works also when the Horvitz-Thompson empirical process W^{HT} is considered. Hence the following proposition holds:

Proposition 3.2.2. *Suppose the sampling design P and the resampling design P^* both satisfy assumptions $H1 - H6$. The following claims hold.*
Claim 1 *Conditionally on $\mathcal{Y}_N, \mathcal{T}_N, \mathbf{D}_N$, N_i^*s, the sequence $(W_n^{HT*}(y) = \sqrt{n}(\hat{F}_{HT}^*(y)-F_N^*(y)),\ y \in \mathbb{R},\ N \geq 1)$ converges weakly, in $D[-\infty,\infty]$ equipped with the Skorokhod topology, to a Gaussian Process $\widetilde{W}_1^*$ with zero mean function and covariance function given by (2.12). The convergence holds for almost all y_is, t_{ij}s, for a set of $\mathbf{D}_N$s of P-probability tending to 1, and for a set of N_i^*s of $\mathbb{P}^*$-probability tending to 1.*
Claim 2 *Conditionally on $\mathcal{Y}_N, \mathcal{T}_N, \mathbf{D}_N$, the sequence of random functions $(W_n^{HT*}(y) = \sqrt{n}(\hat{F}_{HT}^*(y) - F_N^*(y)),\ y \in \mathbb{R},\ N \geq 1)$ converges weakly, in $D[-\infty,\infty]$ equipped with the Skorokhod topology, to a Gaussian Process W_1^* with zero mean function and co-*

variance function given by (2.12). The convergence holds for almost all y_is and t_{ij}s, and for a set of $\mathbf{D}_N$s of P-probability tending to 1.

Claim 3 *The two sequences $(W_n^{HT*}(y),\ y \in \mathbb{R})$ and $(W_N^*(y),\ y \in \mathbb{R})$ are asymptotically independent. Moreover, the following statements hold true.*

- *S1 The whole process $(W^{HT*}(y),\ y \in \mathbb{R})$ converges weakly in $D[-\infty, \infty]$ endowed with the Skorokhod topology, to a Gaussian process W'^* with zero mean function and covariance kernel given by* (2.26).
- *S2 If $\theta(\cdot)$ is continuously Hadamard differentiable at F, then $(\sqrt{n}(\theta(\widehat{F}^*_{HT}) - \theta(\hat{F}_{HT})),\ N \geq 1)$ converges weakly to $\theta'_F(W'^*)$, as N increases.*

In both S1, S2 the convergence hold for almost all y_is and t_{ij}s, and for a set of $\mathbf{D}_N$s of P-probability tending to 1 *and N increases.*

In Figures 3.4-3.5 below some trajectories of the original processes and the resampled processes are represented in order to graphically convey the statements of Propositions 2.2.2 and 3.2.2, and to make some comparisons between the use of the Horvitz-Thompson and the Hàjek estimator, as well. In particular Figures 3.4 (where the Hàjek estimator is considered) and 3.5 (where the Horvitz-Thompson estimator is considered) are obtained assuming a uniform distribution on $(0, 1)$ as superpopulation model. The correlation between the interest character Y and the size variable X is about 0.40 in the population. For both sampling and resampling procedures, a Pareto design is assumed. with a sampling fraction of 1/4. On the left side of the considered figures (Figs. 3.4a, 3.4c, 3.5a and 3.5c), the trajectories of the original processes are depicted. On the right side (Figs. 3.4b, 3.4d, 3.5b and 3.5d) the resampled process is represented. The dashed line is the zero line (the theoretical mean of the process) the solid line is the empirical mean of the process. As far as the Hàjek empirical process is concerned, the similarity between the original process and the resampled one is evident also with a sample size of $n = 50$. The variability of the resampled process is very similar to the variability of the original one (cfr. Figs. 3.4a 3.4b where also some spikes in the trajectories are recovered by the resampled process) and the resampled process shows a zero mean function. Clearly, enlarging the size of the sample to $n = 500$ in our simulation, makes it nearly indistinguishable the two processes (Figs. 3.4c 3.4d). When considering the Horvitz-Thompson process W^{HT}, is well visible in the resampled process a bit of bias in the mean function. The empirical mean of the resampled process depart from the zero line in both the situation of a sample size of $n = 50$

and $n = 500$, this is probably a consequence of the lack of restriction at the end of the "time" for the Horvitz-Thompson empirical process. Thus the extra variability makes the convergence of the resampled process a bit slower than the case of the Hàjek one. Although the resampled process W^{HT*} seems to be quite biased in our simulated scenario, it is clear that the whole behavior of this process mimics the original process, especially when the sample size grows (Figs 3.5c 3.5d).

Remark 3.2.1. For a better understanding of why the resampling scheme introduced so far works, reconsider the Holmberg (1998) scheme mentioned before, which is a popular resampling scheme used in finite populations sampling. For each unit i in the sample s, let $R_i = \pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor$, and consider independent Bernoulli r.v.s ϵ_is with $Pr(\epsilon_i = 1 | \mathbf{D}_N, \mathcal{Y}_N, \mathcal{T}_N) = R_i$. Let further $N_i^* = \lfloor \pi_i^{-1} \rfloor + \epsilon_i$. Even if

$$\sum_{i=1}^{N} N_i^* \neq N$$

it is shown in Conti et al. (2015) that a result similar to Proposition 3.2.1 still holds. In other words the Holmberg scheme is able to recover the limit distribution of the process W_n^H. Clearly this is not enough. In our situation we have to take into account the superpopulation randomness (the process W_N that converges to a Brownian Bridge), but the resampled version of W_N under the Holmberg scheme, that is $\sqrt{n}(F_{N^*}^*(y) - \widehat{F}_H(y))$, does not converge to a Brownian bridge. To show this, it is enough to observe first that adding and removing the quantity

$$\frac{\sum_{i=1}^{N} \pi_i^{-1} D_i I_{(y_i \leq y)}}{\sum_{i=1}^{N} \left(\lfloor \pi_i^{-1} \rfloor + \epsilon_i \right) D_i}$$

we have that

$$\sqrt{n}(F_{N^*}^*(y) - \widehat{F}_H(y)) = A(y) + B(y) \tag{3.22}$$

Comparisons Between The Original Process And The Resampled Process/1

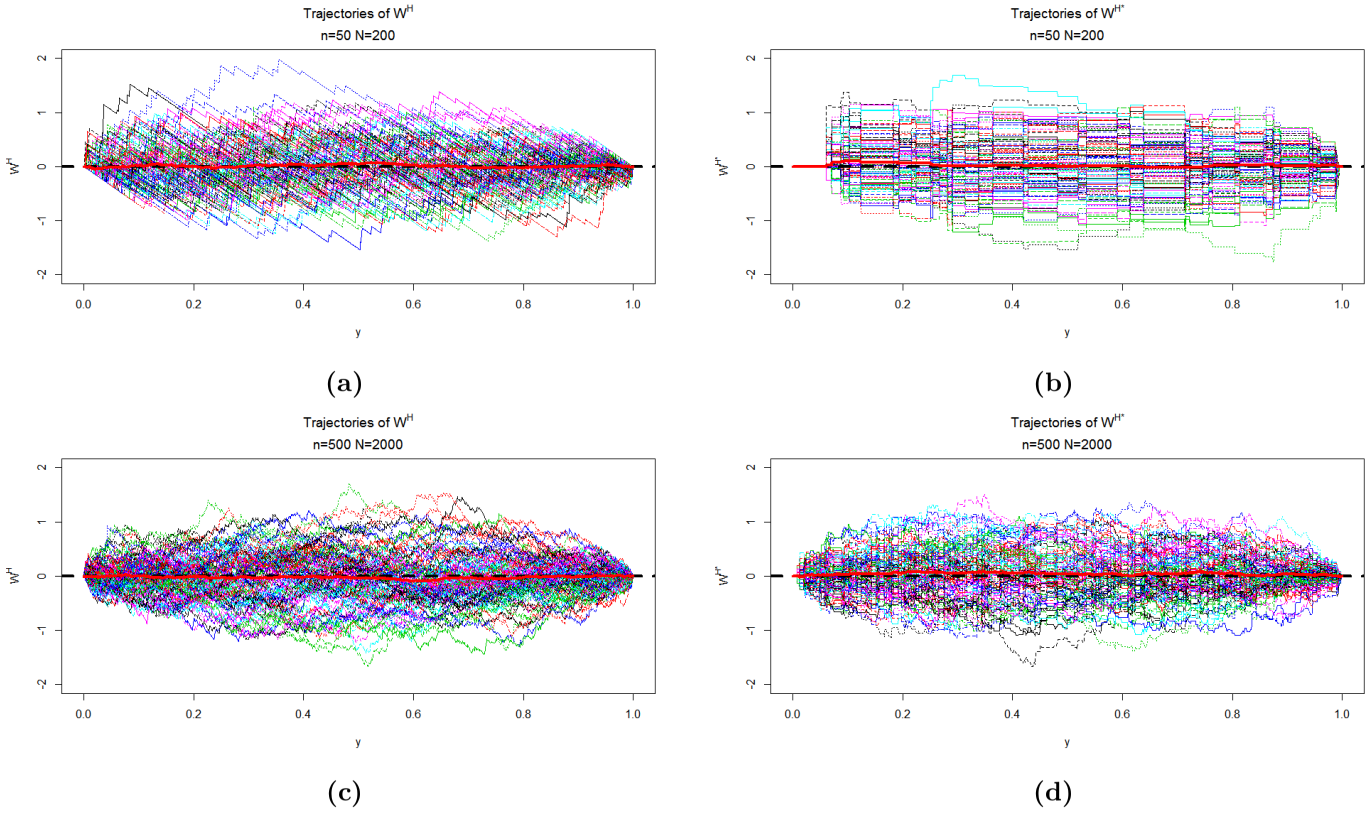


Fig.3.4. Trajectories of the original Hàjek empirical process W^H and the resampled Hàjek process W^{H*}

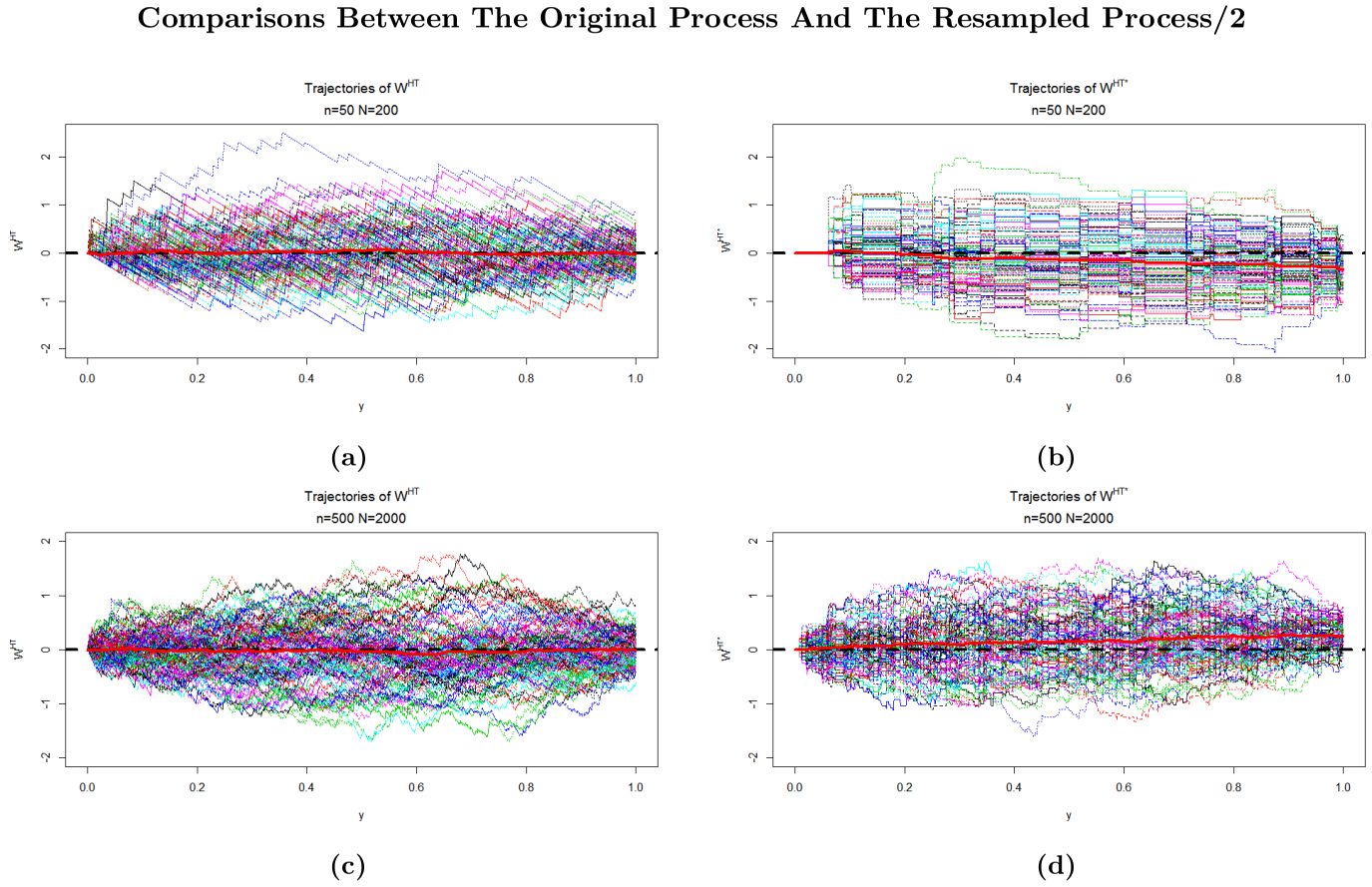


Fig.3.5. Trajectories of the original Horvitz-Thompson empirical process W^{HT} and the resampled Horvitz-Thompson process W^{HT}

where

$$A(y) = \sqrt{n}\left(\frac{\sum_{i=1}^{N}\left(\lfloor\pi_i^{-1}\rfloor+\epsilon_i\right)D_iI_{(y_i\le y)}}{\sum_{i=1}^{N}\left(\lfloor\pi_i^{-1}\rfloor+\epsilon_i\right)D_i} - \frac{\sum_{i=1}^{N}\pi_i^{-1}D_iI_{(y_i\le y)}}{\sum_{i=1}^{N}\left(\lfloor\pi_i^{-1}\rfloor+\epsilon_i\right)D_i}\right)$$

$$B(y) = \sqrt{n}\left(\frac{\sum_{i=1}^{N}\pi_i^{-1}D_iI_{(y_i\le y)}}{\sum_{i=1}^{N}\left(\lfloor\pi_i^{-1}\rfloor+\epsilon_i\right)D_i} - \frac{\sum_{i=1}^{N}\pi_i^{-1}D_iI_{(y_i\le y)}}{\sum_{i=1}^{N}\pi_i^{-1}D_i}\right).$$

Conditionally on $\mathbf{D}_N, \mathcal{Y}_N, \mathcal{T}_N$, the variance of ϵ_i is $R_i(1-R_i) \le 1/4$. Taking into account Lemma 1.1.1, and observing that

$$\mathbb{E}\left[\sum_{i=1}^{N}\left(\lfloor\pi_i^{-1}\rfloor+\epsilon_i\right)D_i\right] = \mathbb{E}_P\left[\mathbb{E}_{\mathbb{P}^*}\left[\sum_{i=1}^{N}\left(\lfloor\pi_i^{-1}\rfloor+\epsilon_i\right)D_i\right]\right] = \tag{3.23}$$

$$\mathbb{E}_P\left[\sum_{i=1}^{N}\left(\lfloor\pi_i^{-1}\rfloor+R_i\right)D_i\right] = \mathbb{E}_P\left[\sum_{i=1}^{N}\left(\pi_i^{-1}\right)D_i\right] = N^1 \tag{3.24}$$

this shows that the limiting distribution of $A(y)$ coincides with the limiting distribution of

$$\frac{\sqrt{n}}{N}\sum_{i=1}^{N}(\epsilon_i - R_i)D_iI_{(y_i\le y)}.$$

In a similar way, it can be shown that the limiting distribution of $B(y)$ coincides with the limiting distribution of

$$-\frac{\sqrt{n}}{N}\sum_{i=1}^{N}(\epsilon_i - R_i)D_iF_N(y)$$

and hence the limiting distribution of (3.22) coincides with the limiting distribution of

$$C(y) = \frac{\sqrt{n}}{N}\sum_{i=1}^{N}(\epsilon_i - R_i)D_i\left(I_{(y_i\le y)} - F_N(y)\right).$$

The arguments of Proposition 2.2.2 can be used to show that, conditionally on

[1] The symbol $\mathbb{E}_{\mathbb{P}^*}[\cdot]$ defines the expected value where the only variability is due to the pseudo-population randomness

$\mathcal{Y}_N$, $\mathcal{T}_N$, $C(y)$ converges to a Gaussian process for almost all y_is, t_{ij}s, and for a set of $\mathbf{D}_N$s of P-probability tending to 1. To show that $C(y)$ does not tend to a Brownian bridge, it is sufficient to show that the asymptotic variance of $C(y)$ is not $F(y)(1-F(y))$. Since the conditional expectation of $\epsilon_1 - R_i$ is zero, we have

$$\begin{aligned}
\mathbb{V}(C(y)|\mathcal{Y}_N, \mathcal{T}_N) &= \frac{n}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[R_i(1-R_i)D_i|\mathcal{Y}_N,\mathcal{T}_N\right]\left(I_{(y_i\leq y)} - F_N(y)\right)^2 \\
&= \frac{n}{N^2}\sum_{i=1}^{N}R_i(1-R_i)\pi_i^{-1}\left(I_{(y_i\leq y)} - F_N(y)\right)^2 \\
&\to \mathbb{E}_{\mathbb{P}}[X_1]\,\mathbb{E}_{\mathbb{P}}\left[R_1(1-R_1)X_1^{-1}\left(I_{(Y\leq y)} - F(y)\right)^2\right] \\
&\neq F(y)(1-F(y)) = \mathbb{E}_{\mathbb{P}}\left[\left(I_{(Y\leq y)} - F(y)\right)^2\right].
\end{aligned}$$

The failure of Holmberg scheme is a consequence of a simple fact: the scheme itself cannot recover the generation process of the finite population from the superpopulation.

3.3 Resampling procedure: Monte Carlo algorithm

Clearly, resampling is performed by resorting to Monte Carlo simulations and this is a computer-intensive procedure. Thus, due to the factorial growth of the cardinality of the space of the boostrap samples, recovering the true asymptotic (resampling) distribution is practically infeasible. To overcome this problem, it necessary to approximate the actual asymptotic distribution of the Hajék estimator (or equivalently of the Horvitz-Thompson estimator) with a simulated resampling distribution. This procedure will be now clarified. For the sake o simplicity we assume $\theta(\cdot)$ to be real-valued, that is considering scalar parameters of the superpopulation.

1. Generate M independent bootstrap samples of size n on the basis of the two-phase resampling procedure described above.

2. For each bootstrap sample, compute the corresponding Hajék estimator (3.20), denoted by $\hat{F}^*_{H,m},\ m = 1, 2, \ldots, M$.

3. Compute the corresponding estimates of $\theta(\cdot)$:

$$\hat{\theta}^*_m = \theta(\hat{F}^*_{H,m}),\ m = 1, 2, \ldots, M.$$

4. Compute the M quantities

$$Z^*_{n,m} = \sqrt{n}(\hat{\theta}^*_m - \theta(\hat{F}^*_H)),\ m = 1, 2, \ldots, M. \tag{3.25}$$

5. Compute the variance of (3.25):

$$\hat{S}^{2*} = \frac{1}{M-1}\sum_{m=1}^{M}(Z^*_{n,m} - \bar{Z}^*_M)^2 = \frac{n}{M-1}\sum_{m=1}^{M}(\hat{\theta}^*_m - \bar{\theta}^*_M) \tag{3.26}$$

where

$$\bar{Z}^*_M = \frac{1}{M}\sum_{m=1}^{M} Z^*_{n,m},\ \bar{\theta}^*_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^*_m.$$

Denote further by

$$\hat{R}^*_{n,M}(z) = \frac{1}{N}\sum_{m=1}^{M} I_{(Z^*_{n,m}\leq z)},\ z \in \mathbb{R} \tag{3.27}$$

the empirical distribution function of $Z^*_{n,m}$, and by

$$\widehat{R}^{*-1}_{n,M}(p) = \inf\{z \in \mathbb{R} : \widehat{R}^*_{n,M}(z) \geq p\},\ 0 < p < 1 \tag{3.28}$$

the corresponding pth quantile.

The empirical distribution (3.27) is the Monte Carlo approximation of the true resampling distribution of $\sqrt{n}(\theta(\hat{F}^*_H(y)) - \theta(\hat{F}_H(y)))$. Next proposition establishes convergence of the empirical distribution (3.27) to the actual asymptotic distribution of the resampled process and the convergence of the quantiles (3.28).

Proposition 3.3.1. *Suppose the assumptions $H1 - H6$ are fulfilled, let $\sigma^2_\theta = \mathbb{V}_{\mathbb{P}}(\theta(F))$, and let Φ_{0,σ^2_θ} be a normal distribution function with expectation 0 and variance σ^2_θ and let $\Phi^{-1}_{\{0,\sigma^2\}}(p)$ be the p-quantile of Φ_{0,σ^2_θ}. Conditionally on $\mathcal{Y}_N, \mathcal{T}_N, \mathbf{D}_N$, the following results hold:*

$$\sup_z |\hat{R}^*_{n,M}(z) - \Phi_{0,\sigma^2_\theta}(z)| \xrightarrow{a.s.-\mathbb{P}^*} 0; \tag{3.29}$$

$$\widehat{R}^{*-1}_{n,M}(p) \xrightarrow{a.s.-\mathbb{P}^*} \Phi^{-1}_{\{0,\sigma^2\}}(p) \tag{3.30}$$

as M, N go to infinity. The convergence holds for almost all y_is and t_{ij}s, for a set of $\mathbf{D}_N$s of $P-$probability tending to 1, and is in probability w.r.t. P^*.
If, in addition, $\sup_{n,m}\mathbb{E}_{\mathbb{P}^*}[Z^{2*}_{n,m}] < \infty$, the sample variance $\widehat{S}^{2*}$ of $(Z^*_{n,m},\ m = 1, 2, \ldots, M)$ is a consistent estimator of σ^2_θ. Formally, for a set of y_is and t_{ij}s

of $\mathbb{P}-$*probability equal to* 1 *and for a set of* $\mathbf{D}_N s$ *of* $P-$*probability tending to* 1*, conditionally on* $\mathcal{Y}_N, \mathcal{T}_N, \mathbf{D}_N$ *it holds that*

$$\widehat{S}^{2*} \to \sigma_\theta^2, \text{ as } M, N \to \infty \tag{3.31}$$

where the convergence in (3.31) *holds in probability w.r.t. the resampling replications and the pseudo-population generation.*

.

Chapter 4

Applications

In this chapter we face the problem of bringing theoretical results obtained in the previous chapters, into statistical practice. Together with this purpose, we want to test the goodness of the proposed resampling procedure. The goodness of the resampling procedure is evaluated via a simulation study, for each one of the applications proposed. All of the simulations produced in this chapter have the main aim of simulating both the superpopulation variability and the sampling design variability.

4.1 Confidence Intervals For Quantiles

The aim of this section is to provide asymptotic confidence intervals, for quantiles of an interest character. Quantiles are very important when measuring poverty or inequality. In fact, poverty and inequality measures are usually expressed as a function of the estimated quantiles of the income or wealth distributions. In practice the confidence intervals are obtained resorting to our resampling procedure.

Let assume a superpopulation as in H2 and let F be the distribution function of the interest variable Y. We define the (superpopulation) *quantile function* as

$$Q(p) = \inf\{y \in \mathbb{R} : F(y) \geq p\} = F^{-1}(p), \text{with } 0 < p < 1. \tag{4.1}$$

Thus, the (superpopulation) quantile function $Q(\cdot)$ can be seen as the inverse of the (superpopulation) distribution function F. This implies that the quantile function can be seen as a functional of the distribution function F. Let now focus on such a

functional. Consider the real valued functional $\theta_p(\cdot) : D[-\infty, +\infty] \to \mathbb{R}$ such that

$$\theta_p(F) = F^{-1}(p) = Q(p), \tag{4.2}$$

that is $\theta_p(\cdot)$ is the functional that brings a distribution function in its quantile function. Hence we define

$$Q_N(p) = \theta_p(F_N), \qquad 0 < p < 1 \tag{4.3}$$

$$\widehat{Q}_H(p) = \theta_p(\widehat{F}_H), \qquad 0 < p < 1 \tag{4.4}$$

In order to use the methodology proposed in this dissertation, we have to show the Hadamard-differentiability of the functional $\theta_p(\cdot)$ defined in (4.2). To this purpose we resort to Lemma 21.3 in Van der Vaart (2000) that gives sufficient conditions in order to have the Hadamard-differentiability of the concerned functional $\theta_p(\cdot)$. These conditions are:

i) The superpopulation distribution function F has to be differentiable at the point $q_p \in \mathbb{R}$ such that $F(q_p) = p$;

ii) The superpopulation distribution have to give a non-zero density to the quantile in exam, i.e. $F'(q_p) \gneq 0$.

If both conditions *i*) and *ii*) are satisfied, then the functional $\theta_p(\cdot)$ is Hadamard-differentiable at F, and its Hadamard-derivative has the form

$$\theta'_p(h) = -\frac{h(q_p)}{F'(q_p)}, \ h \in D[-\infty, +\infty], \ h \text{ continuous at } q_p. \tag{4.5}$$

Equation (4.5) shows why a resampling procedure, sometimes, could be a better alternative to the analytic computation. In fact, we have that

$$\sqrt{n}(\theta_p(\widehat{F}_H) - \theta_p(F)) \xrightarrow{weak} \theta'_p(W) = -\frac{W(q_p)}{F'(q_p)}, \ q_p \in \mathbb{R}, \tag{4.6}$$

or equivalently

$$\sqrt{n}(\theta_p(\widehat{F}_H) - \theta_p(F)) \xrightarrow{weak} \theta'_p(W) = -\frac{W(Q(p))}{F'(Q(p))}, \ 0 < p < 1. \tag{4.7}$$

Hence, the covariance kernel of the limiting process $\theta'_p(W)$ has the form

$$C^{\theta'}(y,t) = \frac{C(y,t)}{F'(y)F'(t)},\ (y,t) \in \mathbb{R}^2 \tag{4.8}$$

where $C(y,t)$ is defined in (2.24). If you want to look at the process $\sqrt{n}(\theta_p(\widehat{F}_H) - \theta_p(F))$ as the quantile process $\sqrt{n}(\widehat{Q}_H(p) - Q(p)),\ 0 < p < 1$, the covariance kernel can be explicitly written in terms of the quantile function $Q(\cdot)$ as follows:

$$\begin{aligned} C^{\theta'}(u,v) = \frac{1}{F'(Q(u))F'(Q(v))} \Bigg\{ & f\left[\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f} K_{-1}(Q(u) \wedge Q(v))\right] u \wedge v \\ & - \frac{f^3}{d}\left(1 - \frac{K_{+1}(Q(u))}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)\left(1 - \frac{K_{+1}(Q(v))}{\mathbb{E}_{\mathbb{P}}[X_1]}\right) uv \\ & - f\left[\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}(K_{-1}(Q(u)) + K_{-1}(Q(v)) - \mathbb{E}_{\mathbb{P}}[X_1^{-1}] - 1) + 1\right] uv \Bigg\},\ (u,v) \in (0,1)^2. \end{aligned} \tag{4.9}$$

As (4.9) shows, if you want to proceed with an analytic (non parametric) approach, in order to recover the asymptotic variance of the estimated quantile, you need to provide estimates of $\mathbb{E}_{\mathbb{P}}[X], K_\alpha, F'$. This way might be more complex and less efficient. Using resampling procedures allows you to avoid these estimation steps making the inference procedure easier.

Define now

$$\widehat{d}_\alpha = \widehat{R}^{-1*}_{n,M}(\alpha), \tag{4.10}$$

$$z_\alpha = \Phi^{-1}_{\{0,1\}}(\alpha). \tag{4.11}$$

Expression (4.10) is the $\alpha-$quantile of the Monte Carlo approximation of resampling distribution of $\sqrt{n}(\theta_p(\widehat{F}^*_H) - \theta_p(\widehat{F}_H))$ as defined in (3.28), while (4.11) is the $\alpha-$quantile of a standard normal distribution. Due to Proposition 3.3.1 (that ensures the consistency of both the empirical distribution function $\widehat{R}^*_{n,M}$ of the resampled parameter and also of the empirical quantile function $\widehat{R}^{-1*}_{n,M}$), we have that:

$$[\widehat{L}_P, \widehat{U}_P] = \left[\theta_p(\hat{F}_H) + z_{\frac{\alpha}{2}}\frac{\widehat{S}^*}{\sqrt{n}}, \theta_p(\hat{F}_H) + z_{1-\frac{\alpha}{2}}\frac{\widehat{S}^*}{\sqrt{n}}\right]^1 \tag{4.12}$$

$$[\widehat{L}_{NP}, \widehat{U}_{NP}] = \left[\theta_p(\hat{F}_H) + \frac{\widehat{d}_{\frac{\alpha}{2}}}{\sqrt{n}}, \theta_p(\hat{F}_H) + \frac{\widehat{d}_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right] \tag{4.13}$$

are confidence intervals of asymptotic size $1-\alpha$. The confidence interval in (4.12) is a *parametric confidence interval* in the sense that we use the parametric approximation of the quantile distribution to a Gaussian distribution. The quantity in (4.13) is a *non-parametric confidence interval*, because of using the empirical quantiles based on bootstrap procedure, without assuming any additional information about the form of the limiting distribution.

In order to test our resampling procedure applied to the estimation of confidence intervals for quantiles, we conduct a small simulation study. For our simulations we assume the same superpopulation model as in Antal and Tillé (2011), *i.e.*

$$Y = (\beta_0 + \beta_1 X^{1.2} + \sigma\epsilon)^2 + c \tag{4.14}$$

where $X \sim |N(0,7)|$, $\epsilon \sim N(0,1)$, $\beta_0 = 12.5$, $\beta_1 = 3$, $\sigma = 15$ and $c = 4000$. Parameters in (4.14) are chosen in order to make the interest variable distribution similar to an income distribution. In order to stress our resampling procedure from different points of view, we have simulated under some different scenarios. We assumed three different sample sizes $n \in \{50, 150, 500\}$ with two different sampling fraction; a lower one $f = 1/10$ and an higher one $f = 1/3$. In addition we have considered the situation in which in both sampling and resampling a successive sampling design is considered (this will be indicated as $SU - SU$ scenario) and the situation in which in the sampling stage a successive sampling design is used and in resampling stage a Pareto sampling design (We refer to this scenario with $SU - PA$) is concerned. The inclusion probabilities are taken proportional to a size variable, that is equal to $Y^{(0.5)}H$ where $H \sim LogN(0, 0.4)$ when the sampling fraction is $f = 1/10$. In the situation that concerns a larger sampling fraction of $f = 1/3$, the size variable is equal to $Y^{(0.21)}H'$ where $H' \sim LogN(0, 0.175)$. With these choices the inclusion probabilities exhibit a wide range of variation (in this way we have inclusion probabilities that can be very different from the sampling fraction f) and a correlation (in the finite population) between the interest variable and the size variable that is around 0.40. For each sample size, sampling fraction and couple of sampling designs ($SU - SU$, $SU - PA$) we have generated from the model (4.14) $J = 1000$ finite populations and for each sample drawn from these finite populations, $M = 1000$ bootstrap samples are selected.

Using our resampling scheme we have computed confidence intervals for quan-

[1] $\widehat{S}^*$ is the square root of the quantity defined in (3.26)

tiles of order $p = 0.10, 0.25, 0.50, 0.75, 0.9$ according to formulas (4.12)-(4.13) for a nominal confidence level of $1 - \alpha = 0.95$. In order to test the goodness of the proposed "multinomial resampling scheme" we introduce some indicators:

1. Estimated Coverage Probability

$$CP = \frac{1}{J}\sum_{j=1}^{J} I(\widehat{L}^j_{N/NP} \leq \hat{q}_p \leq \widehat{U}^j_{P/NP}). \tag{4.15}$$

2. Estimated Left and Right Errors

$$LE = \frac{1}{J}\sum_{j=1}^{J} I(\widehat{L}^j_{P/NP} > \hat{q}_p); \tag{4.16}$$

$$RE = \frac{1}{J}\sum_{j=1}^{J} I(\widehat{U}^j_{P/NP} < \hat{q}_p). \tag{4.17}$$

3. Average Length

$$AL = \frac{1}{J}\sum_{j=1}^{J} (\widehat{U}^j_{P/NP} - \widehat{L}^j_{P/NP}). \tag{4.18}$$

4. Estimated Relative Bias (for the standard deviation)

$$RB = \frac{1}{J}\sum_{j=1}^{J} \frac{\widehat{\sigma}_{MC} - \widehat{S}^{*j}}{\widehat{\sigma}_{MC}} \tag{4.19}$$

Some clarification are needed about the notation used in the previous indicators. In (4.15) - (4.17) the quantity $\hat{q}_p$ is the $p-$quantile obtained inverting the ECDF computed on 10^7 simulated values from the model (4.14) in order to have a Monte Carlo approximation of the true quantiles. Clearly with the notation $\widehat{L}^j_{P/NP}$ ($\widehat{U}^j_{P/NP}$) we indicate the lower (upper) extreme of the parametric (sub-script P) (or non-parametric with sub script NP) confidence interval for the $j-$th, $j = 1, \ldots, J$ simulated finite population. The function $I(a)$ takes value 1 if a is true and zero otherwise. In (4.19) with the symbol $\widehat{\sigma}_{MC}$ we indicate the estimated Monte Carlo standard deviation of the estimated quantile, that is

$$\widehat{\sigma}_{MC} = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J} \left(\theta_p(\widehat{F}^j_H) - \overline{\theta}_p\right)^2} \tag{4.20}$$

where

$$\overline{\theta}_p = \frac{1}{J}\sum_{j=1}^{J}\theta_p(\widehat{F}_H^j). \tag{4.21}$$

where $\widehat{S}^{*j}$ indicates the bootstrap estimate of the variance (formula (3.26)) for the $j-$th, $j = 1, \ldots, J$ simulated finite population.

Next tables show the estimated quantities (4.15) - (4.19) in scenarios of different sample sizes, sampling fractions and couple of sampling designs.

Looking at Tables 4.1-4.2 that summarize the result of our simulation for the smallest sample size ($n = 50$), it seems that in this case the non-parametric approach has quite better estimated coverage probabilities respect to the parametric case. Only the case of $p = 0.10$ shows a really bad behavior if compared with the parametric approach. It is worth to highlight that the cases of $p = 0.10, 0.90$ are the harder ones. In fact, in this case the parameter falls near the border of the parametric space and this affects the inference procedure. The estimated right and left errors are quite unbalanced, but this is probably due to both the small sample size (our approach is an asymptotic one) and the skewness of the considered population (that is build to resemble an income distribution that in general is highly positively skewed). If we look at the two different sampling fractions considered, we do not see any big difference. As you expect there is a quite better performance when a higher sampling fraction is considered. In fact we have that the coverage probabilities are closer to the nominal level and the average length is shorter if compared to the case of a sampling fraction $f = 1/10$. In addition, also the case of the extreme value for quantile of order $p = 0.10$, shows a better performance when a higher sampling fraction is concerned. Moving to consider the same sampling design in both sampling and resampling stage (SU-SU scenario) or two different designs (SU-PA scenario) produces no relevant differences. Considering a sample size of $n = 150$, Tables 4.3–4.4 implies generally better estimated coverage probabilities, also for the extreme case of $p = 0.10$ in the non-parametric approach. In general the considerations we made above for the sample size of $n = 50$ still hold in this case. We have a better performance of the non-parametric approach and when the higher sampling fraction is considered. Clearly we have more informative confidence intervals, in the sense that the average length is shorter respect to the case of $n = 50$. When the larger sample size of $n = 500$ is considered, Tables 4.5–4.6, the differences between the parametric and non-parametric approach are not so evident. This might be a consequence of a better convergence to the normal distribution of the quantiles distribution. Also the cases of a higher and lower sampling fraction show now, very

		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
CP	P	0.938	0.962	0.935	0.928	0.942	0.94	0.936	0.922	0.93	0.939
	NP	0.896	0.938	0.938	0.949	0.951	0.943	0.939	0.925	0.92	0.915
LE	P	0.042	0.019	0.031	0.021	0.023	0.011	0.016	0.021	0.009	0.011
	NP	0.093	0.051	0.046	0.033	0.029	0.02	0.018	0.02	0.01	0.007
RE	P	0.02	0.019	0.034	0.051	0.035	0.049	0.048	0.057	0.061	0.05
	NP	0.011	0.011	0.016	0.018	0.02	0.037	0.043	0.055	0.07	0.078
AL	P	347.0974	297.3671	721.9078	640.7656	1374.166	1264.24	2623.733	2548.632	5073.915	5071.576
	NP	308.3896	267.1817	669.5726	601.6181	1302.186	1205.231	2480.08	2421.874	4647.269	4674.864

Table 4.1. Indicators (4.15)-(4.18) for SU-SU scenario with $n = 50$ and $\alpha = 0.10$.

		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
CP	P	0.94	0.964	0.933	0.929	0.943	0.938	0.933	0.921	0.935	0.938
	NP	0.894	0.942	0.938	0.940	0.949	0.949	0.943	0.929	0.925	0.918
LE	P	0.044	0.020	0.033	0.021	0.025	0.014	0.016	0.021	0.008	0.01
	NP	0.097	0.05	0.046	0.038	0.029	0.021	0.018	0.020	0.010	0.007
RE	P	0.016	0.016	0.034	0.050	0.032	0.048	0.051	0.058	0.057	0.052
	NP	0.009	0.008	0.016	0.022	0.022	0.03	0.039	0.051	0.065	0.075
AL	P	347.4699	298.7859	723.3885	643.5553	1375.6085	1206.8517	2626.2389	2435.3315	5076.6133	4723.2665
	NP	307.7687	266.7873	669.2072	603.9870	1302.6400	1206.8517	2488.9902	2435.3315	4675.7012	4723.2665

Table 4.2. Indicators (4.15)-(4.18) for SU-PA scenario with $n = 50$ and $\alpha = 0.10$.

		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
CP	P	0.939	0.945	0.939	0.933	0.929	0.927	0.931	0.932	0.921	0.926
	NP	0.939	0.932	0.944	0.946	0.942	0.936	0.946	0.944	0.934	0.95
LE	P	0.026	0.021	0.019	0.023	0.022	0.026	0.016	0.014	0.015	0.020
	NP	0.042	0.052	0.038	0.032	0.023	0.028	0.016	0.012	0.017	0.012
RE	P	0.035	0.034	0.042	0.044	0.049	0.047	0.053	0.054	0.064	0.054
	NP	0.019	0.016	0.018	0.022	0.035	0.036	0.038	0.044	0.049	0.038
AL	P	171.0155	154.1751	395.9150	355.4143	773.2304	703.0053	1472.5921	1418.7555	2851.6720	2862.5691
	NP	160.7140	144.6981	380.0059	342.1344	748.1385	682.9018	1431.9197	1374.0733	2702.4694	27504.9453

Table 4.3. Indicators (4.15)-(4.18) for SU-SU scenario with $n = 150$ and $\alpha = 0.10$.

		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
CP	P	0.935	0.942	0.938	0.933	0.931	0.930	0.930	0.935	0.924	0.936
	NP	0.932	0.928	0.950	0.947	0.940	0.933	0.947	0.944	0.936	0.954
LE	P	0.029	0.025	0.020	0.024	0.021	0.024	0.018	0.011	0.015	0.019
	NP	0.049	0.054	0.035	0.033	0.027	0.028	0.016	0.016	0.014	0.015
RE	P	0.036	0.033	0.042	0.043	0.048	0.046	0.052	0.054	0.061	0.045
	NP	0.019	0.018	0.015	0.020	0.033	0.039	0.037	0.040	0.050	0.031
AL	P	171.2529	155.2792	395.8784	358.0014	774.1934	706.8665	1476.0178	1424.6131	2845.6005	2878.3872
	NP	160.6369	145.3093	380.4871	345.3777	747.6593	688.1376	1434.9387	1377.5727	2700.7302	2767.6330

Table 4.4. Indicators (4.15)-(4.18) for SU-PA scenario with $n = 150$ and $\alpha = 0.05$.

		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
CP	P	0.934	0.931	0.930	0.954	0.942	0.939	0.939	0.929	0.943	0.930
	NP	0.948	0.945	0.938	0.962	0.944	0.944	0.938	0.929	0.947	0.937
LE	P	0.031	0.028	0.032	0.014	0.020	0.022	0.020	0.024	0.020	0.014
	NP	0.035	0.028	0.035	0.012	0.023	0.019	0.020	0.024	0.016	0.008
RE	P	0.035	0.041	0.038	0.032	0.038	0.039	0.041	0.047	0.037	0.056
	NP	0.017	0.027	0.027	0.026	0.033	0.037	0.042	0.047	0.037	0.055
AL	P	88.1538	76.5926	215.2548	192.6489	414.3868	382.6410	801.8705	764.2873	1516.1891	1527.9683
	NP	84.6976	74.1118	211.3069	188.6408	407.7924	375.3835	788.7037	750.6316	1471.9787	1498.6540

Table 4.5. Indicators (4.15)-(4.18) for SU-SU scenario with $n = 500$ and $\alpha = 0.05$.

		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
CP	P	0.928	0.933	0.930	0.951	0.940	0.944	0.941	0.926	0.936	0.930
	NP	0.95	0.941	0.938	0.962	0.947	0.948	0.941	0.929	0.949	0.940
LE	P	0.033	0.027	0.032	0.018	0.021	0.023	0.022	0.025	0.021	0.013
	NP	0.035	0.038	0.034	0.015	0.025	0.025	0.020	0.030	0.017	0.012
RE	P	0.039	0.040	0.038	0.031	0.039	0.033	0.037	0.049	0.043	0.057
	NP	0.015	0.021	0.028	0.023	0.028	0.027	0.039	0.041	0.034	0.048
AL	P	88.2084	77.0434	215.4598	193.8665	414.9369	385.7579	803.3144	768.6136	1517.3526	1533.1298
	NP	85.1046	74.4103	211.3208	189.8237	408.3777	378.0225	789.0756	752.6317	1475.0457	1503.4937

Table 4.6. Indicators (4.15)-(4.18) for SU-PA scenario with $n = 500$ and $\alpha = 0.05$.

RB		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
SU-SU		−0.11	−0.148	−0.031	−0.032	−0.061	−0.045	−0.07	−0.042	−0.098	−0.091
SU-PA		−0.111	−0.154	−0.033	−0.036	−0.062	−0.05	−0.071	−0.048	−0.098	−0.097

Table 4.7. Relative bias (4.19) for the standard deviation of the estimated quantile for $n = 50$.

RB		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
SU-SU		−0.052	−0.012	−0.027	−0.016	−0.013	−0.0003	−0.016	−0.016	−0.044	−0.025
SU-PA		−0.053	−0.02	−0.027	−0.024	−0.014	−0.006	−0.018	−0.02	−0.041	−0.03

Table 4.8. Relative bias (4.19) for the standard deviation of the estimated quantile for $n = 150$.

RB		$p = 0.10$		$p = 0.25$		$p = 0.50$		$p = 0.75$		$p = 0.90$	
		$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
SU-SU		0.003	−0.023	0.027	−0.037	−0.03	0.002	−0.01	0.024	−0.042	−0.003
SU-PA		0.002	−0.03	0.026	−0.044	−0.03	−0.006	−0.013	0.019	−0.043	−0.006

Table 4.9. Relative bias (4.19) for the standard deviation of the estimated quantile for $n = 500$.

similar behaviors. It is worth to notice that some values of the estimated coverage probabilities show a lower performance ($p = 0.25, 0.75$) if compared to the case of $n = 150$. These fluctuations are probably consequence of considering inclusion probabilities more variable when the sample size is increased.

As far as the estimated relative bias for the bootstrap estimate standard deviation (4.19) is concerned, we have that when the smallest sample size is considered, a high relative bias (estimated about 11%) affects the bootstrap estimate of the variance of the quantile of order $p = 0.10$. When the considered quantile is the one of order $p = 0.90$, the estimated relative bias is about 9%. These are the extreme cases and thus they are the more pathological ones. The situation for the extreme values of quantiles is worst if the higher sampling fraction is considered. For the median we have only an absolute relative bias of about 6% if $f = 1/10$ and 4.5% if $f = 1/3$. The SU-SU scenario and the SU-PA scenario have essentially the same behavior. Moving to the case of $n = 500$, the resampling based variance estimator is substantially unbiased in all of the considered cases, showing a maximum absolute relative bias of about 4% when the order $p = 0.90$ is concerned.

4.2 Testing For Conditional Independence

Dependence tests are one of the widely investigated problem of statistical literature. The goal of this paragraph is to perform an independence test for two interest character, conditionally on discrete design variables $T_j s$. For the sake of simplicity we will consider a single design variable T.

To achieve this purpose, the general measure of monotone dependence, proposed in ((Cifarelli et al., 1996)) is extended to the present case. Given two continuous variables X, Y, let $F(x)$ and $G(y)$ be the marginal distributions of the bivariate variable (X, Y) and $H(x, y)$ the joint distribution. A general measure of the monotone dependence γ_g between X and Y, is a real-valued functional γ_g of the bivariate distribution $H(x, y|T)$ defined as follows

$$\theta_\gamma(H(x,y|T)) = \gamma_g = \int_{\mathbb{R}^2} g(|F(x|T)+G(y|T)-1|)-g(|F(x|T)-G(y|T)|)\, dH(x,y|T), \tag{4.22}$$

where $g : [0, 1] \to \mathbb{R}$ is a strictly increasing, continuous and convex function, such that $g(0) = 0$ with continuous first derivative. Under the hypothesis of independence, that is $H(x, y) = F(x)G(y)$, the latter quantity is equal to zero. Thus, we

are considering a test of the form

$$\begin{cases} H_0: & H(x,y|T) = F(x|T)G(y|T) \\ H_1: & H(x,y|T) \neq F(x|T)G(y|T) \end{cases}$$

Using sample data we estimate the joint distribution function $H(x,y)$ with its Hàjek estimator

$$\widehat{H}_H(x,y) = \frac{\sum_{i=1}^{N} \frac{D_i}{\pi_i} I_{(x_i \leq x, y_i \leq y)}}{\sum_{i=1}^{N} \frac{D_i}{\pi_i}}. \tag{4.23}$$

In addition we define the bivariate finite population distribution function as

$$H_N(x,y) = \frac{1}{N} \sum_{i=1}^{N} I_{(x_i \leq x, y_i \leq y)}. \tag{4.24}$$

The basic idea is to estimate the quantity γ_g with a plug-in approach, substituting the distributions functions in (4.22) with their Hajék estimators $\widehat{F}_H$, $\widehat{G}_H$, $\widehat{H}_H$, obtaining

$$\hat{\gamma}_{g,H|T} = \frac{\sum_{i \in s} \frac{1}{\pi_i} \left(g(|F(x_i|T_i) + G(y_i|T_i) - 1|) - g(|F(x_i|T_i) - G(y_i|T_i)|) \right)}{\sum_{i \in s} \frac{1}{\pi_i}}. \tag{4.25}$$

Before analyzing the Hadamard differentiability of $\theta_\gamma(\cdot) : D[-\infty, \infty]^2 \to \mathbb{R}$, we stress that our results are given for the univariate case, but they can be simply generalized to the multivariate case. When considering a bivariate character of interest, in fact, following the same approach of the univariate case, you have that the bivariate version of the process W_n^H is defined as

$$W_n^H(x,y) = \sqrt{n}(\widehat{H}_H(x,y) - H_N(x,y)) \tag{4.26}$$

and you have that

$$W_n^H(x,y) \xrightarrow{weak} W_1(x,y) \tag{4.27}$$

where $W_1(x,y)$ is a Gaussian process and with a covariance kernel of the form

$$C_1((x,y),(s,t)) = f\left\{\frac{\mathbb{E}_{\mathbb{P}}[Z_1]}{f}K_{-1}(y\wedge x, s\wedge t) - 1\right\}H(x\wedge y, s\wedge t) \tag{4.28}$$

$$- \frac{f^3}{d}\left(1 - \frac{K_{+1}(x\wedge y)}{\mathbb{E}_{\mathbb{P}}[Z_1]}\right)\left(1 - \frac{K_{+1}(s\wedge t)}{\mathbb{E}_{\mathbb{P}}[Z_1]}\right)H(x,y)H(s,t)$$

$$- f\left\{\frac{\mathbb{E}_{\mathbb{P}}[Z_1]}{f}(K_{-1}(x,y) + K_{-1}(s,t) - \mathbb{E}_{\mathbb{P}}[Z_1^{-1}] - 1)\right\}H(x,y)H(s,t) \tag{4.29}$$

where $K_\alpha(x,y) = \mathbb{E}_{\mathbb{P}}[Z^\alpha | X \leq x, Y \leq y]$, with $\alpha = \pm 1$ and Z is the size variable. A multivariate extension of the Donsker's Theorem ensures that under the *i.i.d* assumption, it holds that

$$\sqrt{N}(H_N(x,y) - H(x,y)) \xrightarrow{weak} W_2(x,y) \tag{4.30}$$

where $W_2(x,y)$ is a Brownian sheet on the scale of $H(x,y)$, that is a Gaussian process with covariance kernel

$$C_2((x,y),(s,t)) = H(x\wedge y, s\wedge t) - H(x,y)H(s,t) \tag{4.31}$$

Thus, proving the asymptotic independence of the two considered processes as in the univariate case, the whole process $W^H(x,y) = \sqrt{n}(\widehat{H}_H(x,y) - H(x,y))$ converges weakly to a Gaussian process $W(x,y)$ with covariance kernel

$$C((x,y),(s,t)) = C_1((x,y),(s,t)) + f(C_2((x,y),(s,t)). \tag{4.32}$$

To show the Hadamard-differentiabilty of the considered functional, it is sufficient to use the same arguments as the proof of Theorem 4.1. in Cifarelli et al. (1996) and then use result (4) in Gill et al. (1989).

Before illustrating our simulation study is important to stress a couple of remarks. First of all, the conditioning design variable is supposed discrete for the sake of simplicity. In fact, estimating conditional distribution functions when the conditioning variable is discrete, does not involve different estimation techniques, but only focusing on a subgroup of the population than the whole population. Allowing the conditioning variable being continuous implies more complex estimator of the distribution function (like kernel estimators) that fall outside the spirit of the present dissertation. In addition, assuming the conditioning variable discrete is

quite usual in survey sampling. In fact, focusing on a subgroup of the population is what happens in stratified design or when considering domains in a multipurpose survey. The second remark is about the resampling procedure. In order to perform a test with resampling techniques, it is necessary to resample under the null hypothesis, thus in our case we need to resample under the hypothesis of conditional independence of the the two interest characters X, Y. To this purpose, the pseudo-population generation phase of our resampling technique has been modified as follows. According to the previous notation X, Y are variables of interest, and T takes values $T^1, \ldots, T^k$. In addition, let s be a sample of units selected from a $N - sized$ finite population $\mathcal{U}_N$ with a πps sampling design P, where the inclusion probabilities $\pi_i \propto T_j$. Define $s_j = \{i \in s | t_i = T^j\},\ j = 1, \ldots, k$, the set of sampled units with T-value equal to T^j. Let $n_1, \ldots, n_k$ be the size of $s_1, \ldots, s_k$. Firstly, a pseudo-population of N values $T_1^*, \ldots, T_N^*$ is generated where each unit is selected independently with probability $\pi_i^{-1} / \sum_{j \in s} \pi_j^{-1}$. Then for $l = 1, \ldots, N$ if $T_l^* = T^j,\ j = 1, \ldots, k$ we sample independently from s_j, with probability $1/n_j$, a X-value X_l^* and a Y-value Y_l^*. At the end of this procedure a pseudo population $\mathcal{U}_N^* = (X_l^*, Y_l^*, T_l^*,\ l = 1, \ldots, N)$ is obtained, where X^* and Y^* are independent conditionally on T^*. At this point the second phase of the resampling method shown in Chapter 3 can be used. The considered resampling scheme is able to recover the distribution of $\sqrt{n}(\hat{\gamma}_{g,H|T} - \gamma_{g|T})$ under the null hypothesis of independence, and hence to perform the test. In fact, it is sufficient to notice that a stratum can be considered itself as a population. Hence for each stratum all the results obtained in Chapters 2-3 are still valid. The only thing we have to show, is that the resampling procedure mimics effectively the strata generation process. Suppose that the variables T that defines the groups in the superpopulation, takes values $T^1, \ldots, T^k$ with probability $(p_1, \ldots, p_k)$. Clearly, defined $N_j = \{i \in U | t_i = T^j\}$ the size of the j-th stratum in the finite population (U_j in the sequel). We have that under assumption $H2$, $N_j \sim Bin(p_j, N)$. Hence we have that

$$\mathbb{E}_{\mathbb{P}} \left[\frac{N_j}{N} \right] \to p_j, \text{ a.s.} - \mathbb{P}. \tag{4.33}$$

that is, the weight in the finite population of the j-th stratum is a consistent estimator of the weight of the j-th stratum in the superpopulation.

Define also

$$\hat{N}_j = \sum_{i \in s_j} \frac{1}{\pi_i} = \sum_{i \in U_j} D_i \frac{1}{\pi_i} \tag{4.34}$$

$$\hat{N} = \sum_{i \in s} \frac{1}{\pi_i} = \sum_{i \in U} D_i \frac{1}{\pi_i} \tag{4.35}$$

it is clear by Definitions (4.34)-(4.35), result (4.33) and Lemma 1.1.1

$$\frac{\hat{N}_j}{\hat{N}} \xrightarrow{P-probability} \mathbb{E}_P \left[\frac{\hat{N}_j}{\hat{N}} \right] \approx \frac{N_j}{N} \xrightarrow{a.s.-\mathbb{P}} p_j \tag{4.36}$$

that is, the estimated weight of the j-th stratum is a (weak) consistent estimator of the real weight. Consider now $N_j^* = \{i \in \mathcal{U}_N^* | T_i^* = t_j\}$ the size of the j-th stratum in the pseudo-population generated following the procedure described few lines above. As we have already noticed in the previous chapter, N_j^* follows a Binomial distribution where the number of trials is N and the success probability in this case is $\hat{N}_j/\hat{N}$. Hence we have that

$$\frac{N_j^*}{N} \xrightarrow{\mathbb{P}^*-probability} \mathbb{E}_{\mathbb{P}^*} \left[\frac{N_j^*}{N} \right] = \frac{\hat{N}_j}{\hat{N}} \tag{4.37}$$

and by what we observed in (4.36), we can conclude that the proposed modification to the resampling scheme asymptotically mimics the mechanism that generates the strata in the superpopulation.

In the sequel, we will focus on a simulation study where the function $g(s) = s^2$ is used. With this choice of g, the coefficient γ_g become exactly the non-normalized version of the Spearman's rank coefficient ρ_s (cfr. Cifarelli et al. (1996)).

For the simulations study we assumed that in the superpopulation there are four strata, indexed by the discrete variable $T \in \{1, 2, 3, 4\}$. For each stratum we have

the interest variables (X, Y) distributed as a bivariate normal $N(\mu_T, \Sigma)$ where

$$\Sigma = \begin{pmatrix} 150^2 & 150 \cdot 60 \cdot 2 \cdot \sin(\frac{\pi}{6} \cdot \rho_s) \\ 150 \cdot 60 \cdot 2 \cdot \sin(\frac{\pi}{6} \cdot \rho_s) & 60^2 \end{pmatrix} \tag{4.38}$$

$$\mu_T = \begin{cases} (800, 300)' & \text{if } T = 1 \\ (900, 400)' & \text{if } T = 2 \\ (1000, 500)' & \text{if } T = 3 \\ (1100, 600)' & \text{if } T = 4 \end{cases} \tag{4.39}$$

In addition each stratum has a weight in the superpopulation equal to

$$\omega_T = \begin{cases} 0.4 & \text{if } T = 1 \\ 0.3 & \text{if } T = 2 \\ 0.2 & \text{if } T = 3 \\ 0.1 & \text{if } T = 4 \end{cases}.$$

Setting the covariance matrix as in (4.38), involves having exactly a Spearman's correlation coefficient between X, Y of ρ_s in each group (for a proof see for instance Kruskal (1958)). Thus we have an overall Spearman's correlation coefficient conditionally on T equal to ρ_s. The choice of the means (that are all on the line $y = x - 500$) is made in order to have a strong dependence between X and Y when not conditioning on T. In this setting our test becomes

$$\begin{cases} H_0 : & \rho_{s|T} = 0 \\ H_1 : & \rho_{s|T} \neq 0 \end{cases}. \tag{4.40}$$

The estimated region of rejection for test (4.40) based on resampling has the form:

$$\left\{ |\hat{\rho}_{s,H|T}| > z_{1-\frac{\alpha}{2}} \frac{S^*}{\sqrt{n}} \right\}, \tag{4.41}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1 - \dfrac{\alpha}{2}$-quantile of a standard normal distribution and S^* is the square root of the bootstrapped (under the null hypothesis) variance of $\hat{\rho}_{s,H|T}$.

For the simulations sample sizes $n = 50, 150$ and sampling fractions $f = 1/3, 1/10$ are considered. For each sample size and sampling fractions, $J = 1000$ finite populations have been generated and for each sample selected from these populations, $M = 1000$ bootstrap samples have been drawn. In addition, two sampling scenarios

have been concerned. The first one (CP-PA) where samples are selected according to a Conditional Poisson (CP) sampling design[2] and in resampling procedure a Pareto (PA) design has been used. The second one (PA-PA) where in both sampling and resampling a Pareto (PA) design is implemented.
A test of nominal level $\alpha = 5\%$ has been performed and to evaluate the performance several Monte Carlo estimates have been computed

1. Estimated Type 1 Error

$$\hat{\alpha} = \frac{1}{J}\sum_{j=1}^{J} I\left(|\hat{\rho}^{j}_{s,H|T}| > z_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}\right) \tag{4.42}$$

$$\tag{4.43}$$

   with $\rho_s = 0$ in (4.38).

2. Estimated Power

$$EP = \frac{1}{J}\sum_{j=1}^{J} I\left(|\hat{\rho}^{j}_{s,H|T}| > z_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}\right) \tag{4.44}$$

$$\tag{4.45}$$

   with $\rho_s = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ in (4.38).

3. Median of Estimated P-values $Me(\hat{P}^j)$, $j = 1, \ldots, J$ where the Estimated P-values are

$$\hat{P}^j = \frac{1}{M}\sum_{m=1}^{M} I(|Z^*_{n,m}| > |\rho^{j}_{s,H|T}|) \tag{4.46}$$

   with $Z^*_{n,m}$ is defined in (3.25).

4. The estimated Relative Bias RB for the Spearman's rho standard deviation under the null hypothesis of independence. It is computed as in (4.19), with the obvious modifications.

Results of our simulation study are summarized below.

[2] According to Hájek (1964) and Hájek and Dupac (1981) the Conditional Poisson sampling design is equivalent to the rejective sampling design

Sample size and Sampling fraction	$\hat{\alpha}$ (CP-PA)	$\hat{\alpha}$ (PA-PA)
$n = 50, \quad f = 0.1$	0.053	0.051
$n = 150, \ f = 0.1$	0.045	0.048
$n = 50, \quad f = 0.3$	0.06	0.051
$n = 150, \ f = 0.3$	0.05	0.048

Table 4.10. Estimated Type 1 Error for different sample sizes, sampling fractions and couples of sampling designs for the sampling and resampling stage. Nominal $\alpha = 5\%$.

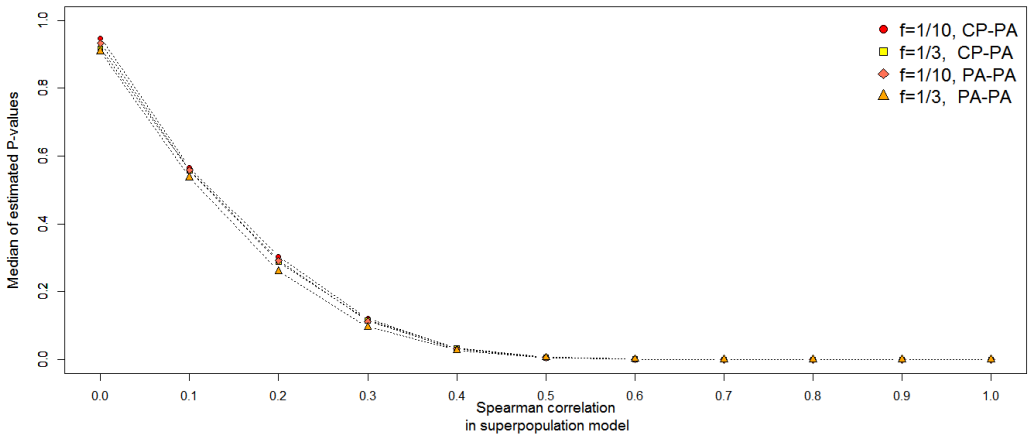


Fig.4.1. Median of estimated P-values for each level of correlation with $n = 50$, $f = 1/3$, $f = 1/10$. and considering CP-PA, PA-PA scenarios

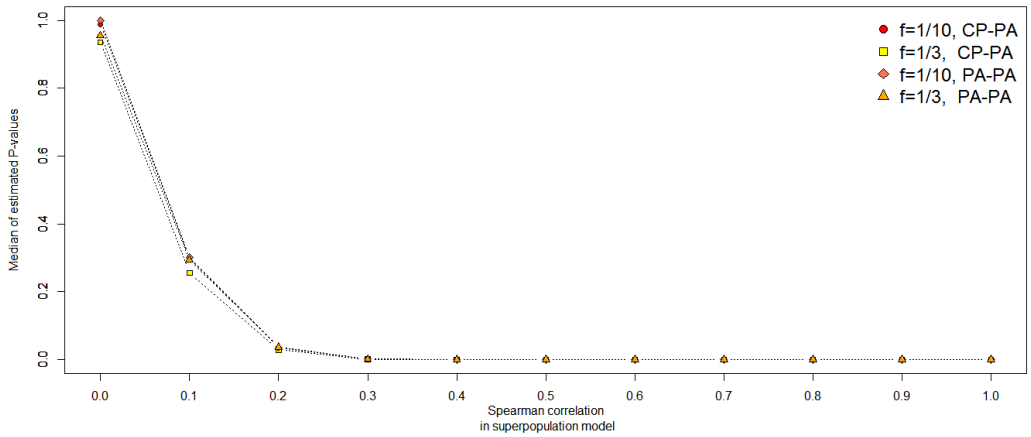


Fig.4.2. Median of estimated P-values for each level of correlation with $n = 50$, $f = 1/3$, $f = 1/10$. and considering CP-PA, PA-PA scenarios

Sample size and Sampling fraction	RB (CP-PA)	RB (PA-PA)
$n = 50, \quad f = 0.1$	0.0025	0.0209
$n = 150, \ f = 0.1$	-0.019	-0.0313
$n = 50, \quad f = 0.3$	0.0007	0.0179
$n = 150, \ f = 0.3$	0.0043	-0.01628

Table 4.11. Estimated Relative Bias in the different considered scenarios.

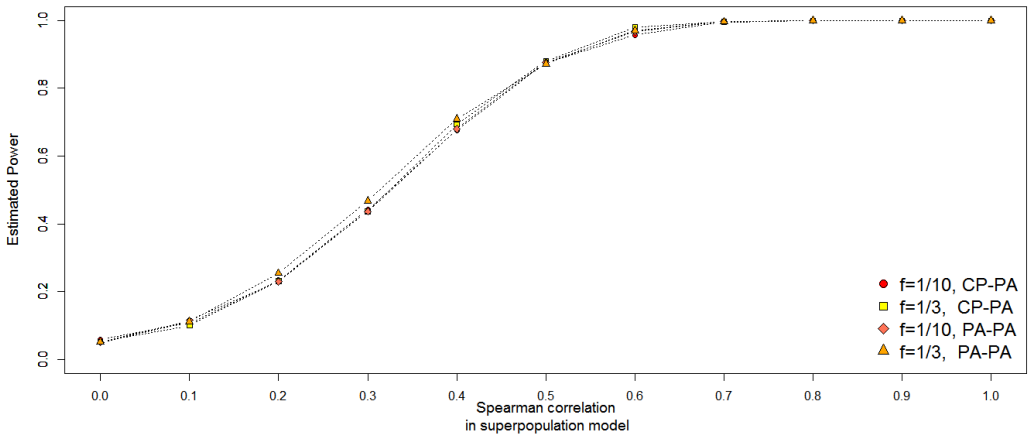


Fig.4.3. Estimated power function where $n = 50$, $f = 1/3$, $f = 1/10$. and considering CP-PA, PA-PA scenarios

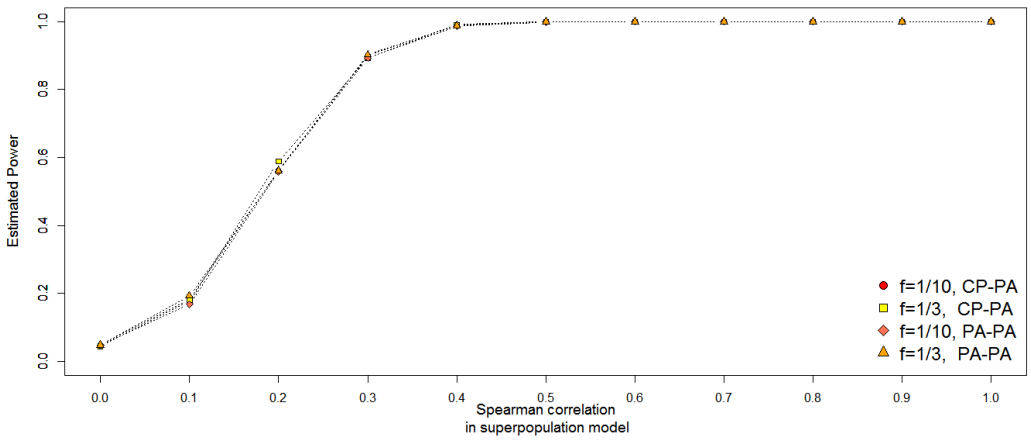


Fig.4.4. Estimated power function where $n = 150$, $f = 1/3$, $f = 1/10$. and considering CP-PA, PA-PA scenarios

From Table 4.10, it is immediately seen that our procedure works well in both situations of a small and big sampling fractions. In addition, we can notice that

using two different sampling designs for the sampling and resampling stages give results similar to those obtained by using the same sampling design. In both situations the estimated level $\hat{\alpha}$ is very close to the nominal level of 5%. Another important remark is that PA-PA scenario seems to be more stable with respect to the CP-PA one, in the sense that the estimated type 1 error fluctuates less in PA-PA scenario than in CP-PA. As far as the estimated P-values and the estimated power functions, are concerned, it is seen from Figures 4.1-4.2 that the differences for the same sample size in the different scenarios are really small, but we have generally lower P-values, thus a better performance, when considering $f = 1/3$. For a sample size of $n = 50$ the median of estimated P-values becomes zero when the (superpopulation) Spearman correlation is 0.6. Of course, increasing sample size implies a decrease of the Spearman correlation level beyond which the median of the estimated P-values is zero. The analysis of the estimated power functions leads to similar conclusions. In fact, for a sample size $n = 50$ the estimated power functions are very similar, but the power is higher in the case of a larger sampling fraction. This result is reasonable. In fact, the larger the sampling fraction the larger the information that the sample carries. In this particular case, a larger sampling fraction allows the sample to reconstruct more easily the correlation structure present in the finite population (and in the superpopulation).

Results about the variance estimation with our resampling methods are reported in Table 4.11. In all of the situations examined in our simulation study, we have that the "multinomial scheme" based variance estimator is essentially unbiased. The maximum absolute relative bias that we can observe is about 3% in the case of the PA-PA scenario with a sample size of $n = 150$ and sampling fraction of 0.10. The scenario where the samples are drawn according to a Rejective sampling (CP-PA) seems to perform well for variance estimation if compared to the case where a Pareto design is used in both sampling and resampling stages. In the PA-PA scenario the estimated relative bias, fluctuates more than in the CP-PA scenario.

4.3 Testing for marginal independence

The goal of the present section is to construct a test for the marginal independence of two (continuous) characters of interest Y, Z, without conditioning on the design variables T_js. For the sake of simplicity, in the sequel we will consider a single design variable T, say. The general framework is the same of the previous Section. We resort to the unconditioned version of the measure of monotone dependence defined

in (4.22) with $g(s) = s^2$, that is

$$\rho_s = \int_{\mathbb{R}^2} (F(x) + G(z) - 1)^2 - (F(x) - G(z))^2 \, dH(x, z), \tag{4.47}$$

and its sample version

$$\hat{\rho}_{s,H} = \frac{\sum_{i=1}^{N} D_i \pi_i^{-1} \left((\widehat{F}_H(x) + \widehat{G}_H(z) - 1)^2 - (\widehat{F}_H(x) - \widehat{G}_H(z))^2 \right)}{\sum_{i=1}^{N} D_i \pi_i^{-1}}, \tag{4.48}$$

to test the hypothesis

$$\begin{cases} H_0 : & \rho_s = 0 \\ H_1 : & \rho_s \neq 0 \end{cases}.$$

Clearly all the results derived in the previous paragraph are still valid, in particular we have that $\sqrt{n}(\hat{\rho}_{s,H} - \rho_s)$ is asymptotically normal with zero mean and a complex variance that depends on the Hadamard derivative of the functional θ_γ that brings a bivariate distribution function, in the associated measure of dependence γ_g.

Although from the theoretical point of view it seems to be an easy problem, from the practical point of view it presents more difficulties than the case analyzed before. In fact, performing a test with resampling requires the ability of sampling under the null hypothesis. In this framework, for each sampling unit we have a triplet (y_i, z_i, t_i); thus, a unique sample value of T is associated to each pair (y_i, z_i). In order to apply our resampling procedure to the testing problem, we have to generate a pseudo-population Y_i^*, Z_i^*, T_i^* from the sample values in such a way that Y^* and Z^* are marginally independent (null hypothesis). Independence can be obtained by sampling independently from the (Hajék) estimators of the marginal distribution functions of Y, Z. However, in this way it is not possible to uniquely associate a value T_i^* to each pair (Y_i^*, Z_i^*). To avoid this problem, we look at the testing problem as the inverse of an interval confidence problem: an asymptotic confidence interval of size $1 - \alpha$ provides an asymptotic test of size α. Of course this way of looking at the problem is simpler but has some limits. One of them is that we can not provide estimated p-values, because for their computation it would be necessary to resample under the null hypothesis.

With the previous notation, the following interval

$$\left[\hat{\rho}_{s,H} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{S}^{2*}}{n}}, \hat{\rho}_{s,H} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{S}^{2*}}{n}}\right] \tag{4.49}$$

where z_α is the α-quantile of a standard Normal distribution and $\widehat{S}^{2*}$ is the bootstrap estimate of the variance of Spearman coefficient, is a confidence interval for ρ_s of asymptotic size $1-\alpha$. The null hypothesis of independence is accepted if the bootstrapped confidence interval covers the 0 , and rejected otherwise.

Focusing on (4.47) it easy to see that the Spearman coefficient ρ_s depends only on the *copula* associated to the joint distribution $H(x,z)$ (this can proved by making the change of variables $(U,V) = (F(Y), G(Z))$ in (4.47)). This fact is quite intuitive. A "good" measure of dependence has to summarize only the association between two variables, being invariant to the marginal and the joint distributions. In virtue of what we have just observed, for our simulation study we assumed a copula as superpopulation model for our interest variables, without specifying any particular marginal distribution for the interest characters. In particular, (Y, Z) is assumed to be a bivariate Marshall-Olkin copula (for more see Marshall and Olkin (1967b), Marshall and Olkin (1967a), Mai and Scherer (2012)).

One of the advantages of the bivariate Marshall-Olkin copula is that it allows a Spearman's correlation coefficient that has an analytic form, that only depends on the parameter of the copula (as for the Gaussian copula used in the previous paragraph), and that takes value in the interval $[0,1]$. For the simulation study three different sample sizes, $n = 50,\ 150,\ 250$ have been considered, in both situations of a large ($f = 1/3$) and small ($f = 1/10$) sampling fractions. For each sample size and sampling fraction, $J = 1000$ finite populations have been generated, and for each sample selected from these populations, $M = 1000$ samples have been drawn. Samples were selected according to a Conditional Poisson design. As far as the resampling stage is concerned, a Pareto design was used. The inclusion probabilities π_i have been taken proportional to $T = f(U)W$, where $U = Y + Z$, $f(u) = u^3/3 - 0.5u^2 + 0.10u + 0.5$ and $W \sim \log N(0, \sigma^2)$ with $\sigma^2 = 0.4$ if $f = 1/10$ and $\sigma^2 = 0.08$ if $f = 1/3$. The design variable T possesses correlation with Y and Z in the finite population, ranging in between 0.4 and 0.5, and a broad range of variation of the inclusion probabilities (about $[0.02, 0.95]$). Tests of different sizes $\alpha = 0.1,\ 0.05,\ 0.01$ have been performed.

To evaluate the performance of our procedure, we have computed the same estimators as for the previous case of conditional independence. Exception is made for the median of estimated P-values, because as specified above, in the present situation we are unable to recover the null distribution via resampling. Clearly in (4.42) and (4.44) some modification in order to take into account how many times the resampled confidence interval does not cover the "zero" . To give the idea on the goodness of the variance estimation via resampling, we report here the relaive bias only for $\rho_s = 0, 0.3, 0.7$ that express a situation of independence, average correlation and strong correlation.

Results are summarized below.

	$\boldsymbol{\alpha = 0.1}$		$\boldsymbol{\alpha = 0.05}$		$\boldsymbol{\alpha = 0.01}$	
	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
$n = 50$	0.124	0.116	0.074	0.065	0.02	0.017
$n = 150$	0.126	0.11	0.064	0.062	0.021	0.012
$n = 250$	0.109	0.1	0.055	0.061	0.012	0.016

Table 4.12. Estimated type I error probability $\hat{\alpha}$, for different sample sizes and sampling fractions.

	$\boldsymbol{\rho_s = 0}$		$\boldsymbol{\rho_s = 0.3}$		$\boldsymbol{\rho_s = 0.7}$	
	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
$n = 50$	0.0196	0.0135	0.069	0.0493	0.0614	0.0727
$n = 150$	0.0260	−0.0069	0.0306	0.0092	0.0357	0.0341
$n = 250$	−0.0047	0.0174	0.0263	−0.0081	0.0534	0.0270

Table 4.13. Estimated relative bias for the standard deviation of $\hat{\rho}_s$ in different scenarios.

For the sake of brevity, only graphs of estimated power functions for a nominal level $\alpha = 0.05$ are shown.

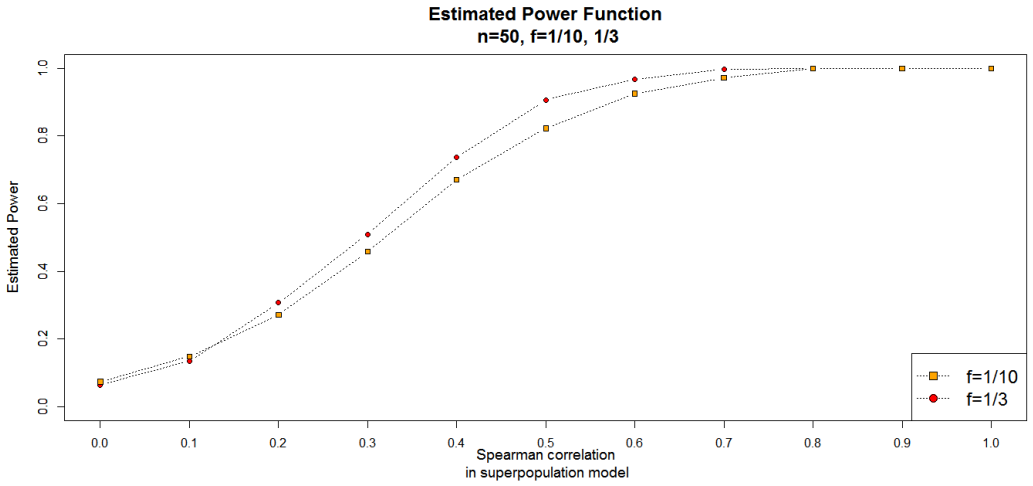


Fig.4.5. Estimated power function where $n = 50,\ f = 1/3,\ f = 1/10$.

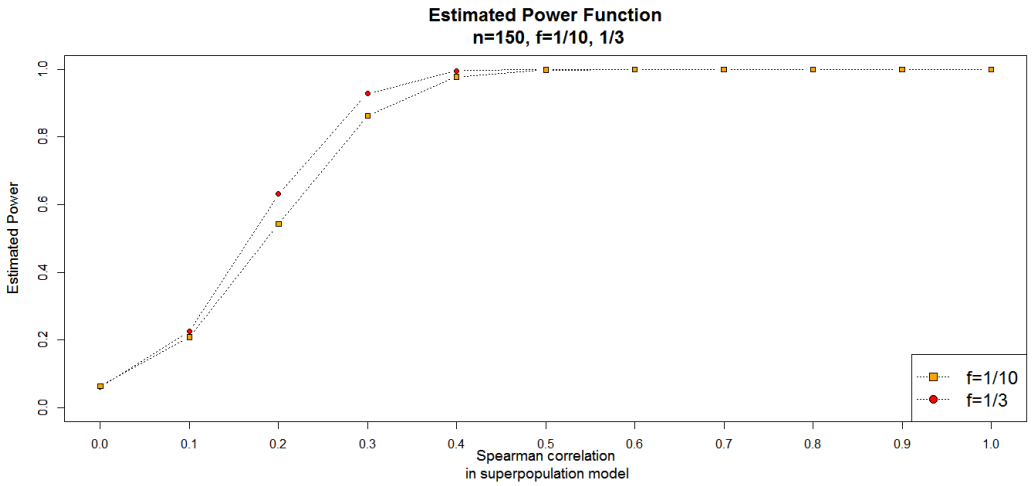


Fig.4.6. Estimated power function where $n = 150,\ f = 1/3,\ f = 1/10$.

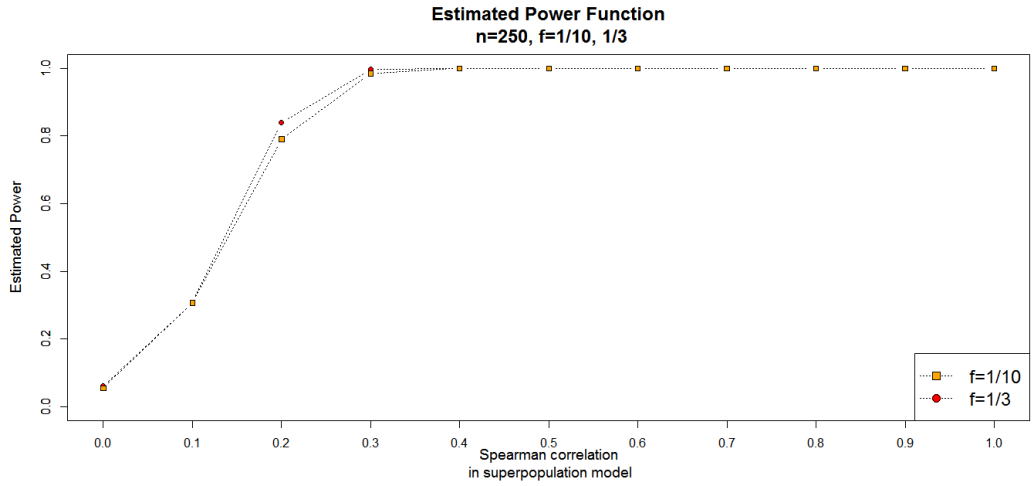


Fig.4.7. Estimated power function where $n = 250,\ f = 1/3,\ f = 1/10$.

From table 4.12, it is seen that estimated type I error is very close to the nominal α. As expected, the largest error corresponds to the smallest sample size ($n = 50$) with a maximum absolute difference between α and $\hat{\alpha}$ of 2.4%. Of course, these errors decrease when the sample size increases. As far as the sampling fractions are concerned, results in the cases $f = 1/3$ and $f = 1/10$ are similar; hence, the sampling fraction seems to play no special role. The estimated power functions (figures 4.5-4.7) exhibit a behavior similar to that of the estimated power functions studied in the previous section. In fact, the estimated power function when $f = 1/3$ dominates the estimated power function when $f = 1/10$ for all sample sizes. Furthermore, differences between power functions decrease as the sample size increases. This suggests that the tests asymptotically have the same power, whatever the sampling fraction may be. For what concerns estimated relative bias reported in Table (4.13), we have a good performance in independence situation ρ_s with a relative bias equal to 2.6%. There is no evident difference between the two considered sampling fractions. The worst case in results summarized in Table (4.12) is when a sample size of $n = 50$, a correlation of 0.7 and a sampling fraction of 1/3 are considered. In such case the estimated relative bias is about 7%. When the sample size is increased the situation improves in all of the considered scenarios, except for the case $n = 250,\ f = 1/10,\ \rho_s = 0.7$ that shows a larger relative bias if compared to the other analyzed cases.

4.4 Confidence Bands For Lorenz Curves

When dealing with economics data, especially when analyzing income distribution, Lorenz curve and related quantities (such as Gini's index) are useful tools of investigation. For instance, Lorenz curve is one of the most known measure of inequality used to understand (or forecast) the consequences of economic policies of a nation. Due to the importance of this measure, in this section and in the next one, we aim at providing some easily-implementable inferential tools, with special attention for *statistics practitioners* that may hardly handle analytic computation.

The aim of this section is to provide asymptotic confidence bands of fixed size, for a given (superpopulation) Lorenz curve. This means, finding two random curves that upper and lower bound the superpopulation Lorenz curve with a fixed probability. It is clearly the extension of the confidence intervals to the functional case.

Let now fix some notation. Given the superpopulation distribution function F of a non-negative interest character Y (this is the classical setting when dealing with Lorenz curves, where negative incomes are not allowed) the superpopulation generalized Lorenz curve is defined as

$$G(p) = \int_0^p Q(u)du,\ 0 \leq p \leq 1 \tag{4.50}$$

where $Q(\cdot)$ is the quantile function defined in (4.1). The usual Lorenz curve is the normalized version of (4.50), that is purged by the effect of the mean of the considered distribution. Formally

$$L(p) = \frac{\int_0^p Q(u)du,}{\int_0^1 Q(u)du} = \frac{G(p)}{G(1)} = \frac{\int_0^p Q(u)du,}{\mathbb{E}_{\mathbb{P}}[Y]},\ 0 \leq p \leq 1. \tag{4.51}$$

Clearly, the finite population and the sample counterparts of (4.50) and (4.51) are defined as

$$G_N(p) = \int_0^p Q_N(u)du \quad \text{and } L_N(p) = \frac{\int_0^p Q_N(u)du}{\int_0^1 Q_N(u)du} = \frac{G_N(p)}{G_N(1)}, \quad 0 \leq p \leq 1 \tag{4.52}$$

$$\widehat{G}_H(p) = \int_0^p \widehat{Q}_N(u)du \quad \text{and } \widehat{L}_H(p) = \frac{\int_0^p \widehat{Q}_H(u)du}{\int_0^1 \widehat{Q}_H(u)du} = \frac{\widehat{G}_H(p)}{\widehat{G}_H(1)}, \quad 0 \leq p \leq 1 \tag{4.53}$$

Let us now focus on the functional $\theta_L(\cdot) : D[0, +\infty] \to C[0, 1]$ that brings a dis-

tribution function F in its Lorenz curve. In order to apply results obtained in the present dissertation, we need fot θ_L to be an Hadamard-differentiable functional. The Hadamard differentiability at F of the map θ_L is proved, under different assumptions, in both Donald et al. (2004) and Bhattacharya (2007). In Donald et al. (2004) it is required that: *i*) F has to be twice continuously differentiable and *ii*) $0 < \inf \psi(y) < \sup \psi(y)$ where $\psi(y) = F'(y)$ is the density associated to F. In Bhattacharya (2007) the author relaxes the assumptions on F in order to have the Hadamard differentiability of θ_L at F. The assumptions made in Bhattacharya (2007) are listed below.

i) F is differentiable with strictly positive derivative on the compact subset of $(0, 1)$ and the moments up to order 2 must be finite.

ii)
$$\lim_{y \to +\infty} \frac{(1 - F(y))^{1+\delta}}{\psi(y)} = 0 \tag{4.54}$$

iii)
$$\lim_{y \to 0} \frac{F(y)^{\delta}}{\psi(y)} \tag{4.55}$$

for some $\delta \in (0, 1)$

Conditions (4.54)-(4.55) control the tail behavior of the density function when the density approaches 0. As noticed in Bhattacharya (2007) these conditions are satisfied by the Pareto and the lognormal family of distributions that are two of the most important parametric families of distributions widely used to model income. By virtue of what we have just observed, conditions (4.54)-(4.55) hold also for distributions that have thinner tails than the Pareto, like exponential ecc.

Assume that the conditions for the Hadamard differentiability of the map $\theta_L(\cdot)$ are satisfied. Considering the Lorenz process $\mathcal{L}^H$ defined as

$$\mathcal{L}^H(p) = \sqrt{n} \left(\theta_L(\widehat{F}_H)(p) - \theta_L(F)(p) \right), \ 0 \leq p \leq 1 \tag{4.56}$$

by Theorem 1.2.1 and Proposition 2.2.2 we have that

$$\mathcal{L}^H(p) \xrightarrow{weak} \mathcal{L}(p) = \theta'_L(W)(p) = -\int_0^p \frac{W(Q(u))}{\psi(Q(u))} du, \ 0 \leq p \leq 1 \tag{4.57}$$

where the expression for θ'_L is given in Donald et al. (2004). One of the possible ways to obtain an asymptotic confidence band of a fixed size, is to find the distribution

of the quantity

$$\sup_{p\in(0,1)} |\mathcal{L}(p)|. \tag{4.58}$$

In fact, defining q_α the α-quantile of the distribution of (4.58) you have that

$$Pr\left(\sup_p |\mathcal{L}(p)| < q_{1-\alpha}\right) = 1-\alpha \tag{4.59}$$

$$\Rightarrow Pr\left(-q_{1-\alpha} < \mathcal{L}(p) < q_{1-\alpha}\right) = 1-\alpha \approx \tag{4.60}$$

$$\approx Pr\left(-q_{1-\alpha} < \sqrt{n}(\widehat{L}_H(p) - L) < q_{1-\alpha}\right). \tag{4.61}$$

Hence, the region

$$\left[\widehat{L}_H(p) - \frac{q_{1-\alpha}}{\sqrt{n}}; \widehat{L}_H(p) + \frac{q_{1-\alpha}}{\sqrt{n}}\right],\ 0<p<1 \tag{4.62}$$

is a confidence band of fixed size for L of asymptotic confidence level $1-\alpha$.

Remark 4.4.1. It is clear that our aim is recovering via resampling, the distribution of the quantity (4.58). What we have as a consequence of the theory exposed in Chapter 2 is resumed by (4.57), but we are interested in the convergence of the supremum of the Lorenz process. The convergence holds, observing that the Lorenz curve is a continuous function, hence the supremum map is continuous and the weak convergence of (4.58) is a consequence of the continuous mapping theorem (cfr. Billingsley (1968)).

In the sequel, we resort to the Monte Carlo approach introduced in Chapter 3 to perform the resampling. Obviously, in order to approximate the distribution of (4.58), instead of (3.25) we will compute the quantity

$$\sup_p |\theta_L(\widehat{F}_H^*)(p) - \theta_L(\widehat{F}_H)(p)|. \tag{4.63}$$

Following the same approach of Proposition 3.3.1 it is easy to see that

$$\sup_z \left| R_{n,M}^* - Pr\left\{\sup_p |\mathcal{L}(p)| \le z\right\}\right| \xrightarrow{a.s.} 0 \tag{4.64}$$

$$R_{n,M}^{*-1}(\alpha) \xrightarrow{a.s.} q_\alpha, 0<u<1 \tag{4.65}$$

where $R_{n,m}^*$ it is the ECDF of the quantities (4.63) over the bootstrap samples, and

$R^{*-1}_{n,M}$ is its inverse. Thanks to (4.65) we have that

$$\left[\widehat{L}_H(p) - \frac{R^{*-1}_{n,M}(1-\alpha)}{\sqrt{n}}; \widehat{L}_H(p) + \frac{R^{*-1}_{n,M}(1-\alpha)}{\sqrt{n}}\right], \ 0 < p < 1 \tag{4.66}$$

is a confidence band for L of asymptotic size $1-\alpha$

We are now in a position to introduce our simulation study. For our simulations we have assumed as superpopulation model a variable Y distributed as a lognormal distribution $LogN(\mu, \sigma)$ with $\mu = 0.85$ and $\sigma = 0.6$. These choices for parameters are the same used for the simulation study in Barrett and Donald (2003) From the superpopulation model we have generated $J = 1000$ finite populations and for each sample drawn from these populations, $M = 1000$ bootstrap samples are drawn. In both sampling and resampling we have used a successive sampling design. We have computed confidence bands for three level of confidence $1-\alpha = 0.90, 0.95, 0.99$ and for two different sampling fractions $f = 1/10, f = 1/3$. In order to have inclusion probabilities correlated with the interest character Y we have generated a size variable X in two different ways, depending on the considered sampling fraction. If $f = 1/10$ $X = Y^{0.5}U$ where $U \sim LogN(0, 0.52)$. If $f = 1/3$ $X = Y^{0.2}U$ where $U \sim LogN(0, 0.16)$. With this choice of the size variable, the correlation between Y and X in the finite population is about 0.40, and the inclusion probabilities are spread over the unit interval. To test the goodness of our proposal we have computed the following Monte Carlo estimates:

1. Estimated Coverage Probability

$$CP = \frac{1}{J}\sum_{j=1}^{J} I(\sup|\widehat{L}_H - L| > \hat{d}_{1-\alpha}) \tag{4.67}$$

   where $\hat{d}_{1-\alpha}$ is the $1-\alpha$ quantile of the Monte Carlo approximation of the resampling distribution of $\sup \mathcal{L}^H$, as in 3.28, and $I(a)$ takes value 1 if a is true, 0 otherwise.

2. Estimated Relative Bias for the quantiles q_α

$$RB = \frac{1}{J}\sum_{j=1}^{J} \frac{\hat{q}^{MC}_\alpha - \hat{d}^j_\alpha}{\hat{q}^{MC}_\alpha} \tag{4.68}$$

   where q^{MC}_α is obtained as empirical quantile of the distribution of the supremum M^j computed for each of the $J = 1000$ samples selected from the

$J = 1000$ finite populations. Formally:

$$M^j = \sup_p \sqrt{n}|\hat{L}_H^j(p) - L(p)|, j = 1, \ldots, 1000$$
$$R(m) = \frac{1}{J}\sum_{j=1}^{J} I_{(M^j \leq m)}$$
$$\hat{q}_\alpha^{MC} = R^{-1}(\alpha)$$

Results of our simulations are summarized in Table 4.14-4.15

Estimated Coverage Probability

	$1-\alpha = 0.90$		$1-\alpha = 0.95$		$1-\alpha = 0.99$	
	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
$n = 50$	0.884	0.847	0.930	0.904	0.972	0.961
$n = 150$	0.0.897	0.881	0.943	0.931	0.983	0.981
$n = 250$	0.891	0.894	0.946	0.944	0.987	0.985

Table 4.14. Estimated coverage probabilities for different nominal levels $1 - \alpha = 0.90, 0.95, 0.99)$ and sampling fractions $f = 1/3, 1/10$.

Estimated Relative Bias

	$1-\alpha = 0.90$		$1-\alpha = 0.95$		$1-\alpha = 0.99$	
	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$	$f = 1/10$	$f = 1/3$
$n = 50$	0.0070	0.0628	0.0284	0.0264	0.0313	0.0717
$n = 150$	-0.0219	0.0020	0.0056	-0.0087	0.0286	0.0073
$n = 250$	-0.0227	-0.0105	-0.0179	-0.0235	-0.0368	-0.0107

Table 4.15. Estimated relative bias for quantiles of different order $1-\alpha = 0.90, 0.95, 0.99)$ and sampling fractions $f = 1/3, 1/10$.

As we can see from Table 4.14 the coverage probabilities are very close to the nominal values of the confidence level. The worst case is when the sampling fraction is $f = 1/3$ and the sample size is $n = 50$. When the sample size increases, we have a good performance in both the case of a smaller and larger sampling fraction. As expected when the estimated relative bias is concerned (Table 4.15), the situation is better when the sample size increases, although we have some fluctuations of the relative bias when the sample size goes from $n = 150$ to $n = 250$. Probably, these variations are due to the closeness of some inclusion probabilities to the upper limit of 1, when $n = 250$.

4.5 Testing For Stochastic Dominance

As already said in the previous section, Lorenz curve is a useful tool of investigation of economic inequality. In addition to the analysis of a single Lorenz curve, it is sometimes more interesting to compare two (or more) Lorenz curves. For example, this is the case of analyzing inequality in different subgroups of the population or analyzing if a taxes increment for the richest people of a nation facilitates an inequality reduction and so on. Clearly, comparisons between Lorenz curves can be made in a descriptive way. This is not our purpose, we want an inferential tool that is able to distinguish if the differences between the considered Lorenz curve are due to the sample or if they actually exist. This kind of inferential tool is a test of stochastic dominance. To this purpose we will extend the previous results about confidence bands for a single Lorenz curve, to the case where we have two Lorenz curves.

Being more formal, let $\mathcal{U}_{N1} = (\mathbf{y}_{N1}, \mathbf{x}_{N1})$ and $\mathcal{U}_{N2} = (\mathbf{y}_{N2}, \mathbf{x}'_{N2})$ (of size $N_1,\ N_2$, respecively) be two independent finite populations generated as in H2 by two independent superpopulation models. The interest variable is Y, say the income, and X and X' are the size variables in the two populations. Assume that $Y \sim F$ in the first superpopulation and $Y \sim G$ in the second one. With the notation of the previous chapter, the (superpopulation) Lorenz curve for the two populations are

$$L_F(p) = \frac{\int_0^p Q_F(u)du,}{\int_0^1 Q_F(u)du},\ 0 \leq p \leq 1 \tag{4.69}$$

$$L_G(p) = \frac{\int_0^p Q_G(u)du,}{\int_0^1 Q_G(u)du},\ 0 \leq p \leq 1 \tag{4.70}$$

where Q_F and Q_G are the quantile functions associated to F and G respectively.

We say that the Lorenz curve L_1 weakly dominates L_2 if

$$L_F(p) - L_G(p) \geq 0,\ \forall p \in [0,1] \tag{4.71}$$

that is, the income distribution in the first population exhibits a level of inequality at most as that in the second population. Our aim is to infer the curve $L_1 - L_2$ and test for the presence of weak dominance. Formally, the hypothesis in which we are interested are

$$\begin{cases} H_0: & L_F(p) - L_G(p) \geq 0,\ \forall p \in [0,1] \\ H_1: & L_F(p) - L_G(p) < 0,\ \text{for some } p \in [0,1] \end{cases} \tag{4.72}$$

The null hypothesis in (4.72), is the hypothesis of weak dominance of curve L_1 over L_2. This choice of null hypothesis, is supported from the econometric literature (for example, cfr. Barrett et al. (2014) and reference therein). The null hypothesis incorporates also the case were $L_1 = L_2$. This situation occurs only when $F(y) = G(\alpha y)$ with $\alpha \geq 0$ (see Lambert (1993)).

Before introducing our idea to perform test (4.72) we have to extend a bit our asymptotic results. Consider now the two samples s_1 and s_2 selected from $\mathcal{U}_{N1}$ and $\mathcal{U}_{N2}$, according to high entropy sampling designs. We want to characterize the asymptotic behavior of the process

$$W^H_{F,G}(y) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\widehat{F}_H(y) - \widehat{G}_H(y) - F(y) + G(y) \right), \ y \in \mathbb{R} \tag{4.73}$$

where $\widehat{F}_H$ and $\widehat{G}_H$ are the Hàjek estimator of F, G and $n_1,\ n_2$ are the sizes of the samples $s_1,\ s_2$. Allowing different sizes of the samples involved in (4.73), makes it necessary a remark. In order to study the asymptotic behavior of $W^H_{F,G}$ we have to "jointly" limit the asymptotic behavior of the sample sizes, in addition to condition H4. Hence, we require that

$$\lim_{N_1, N_2} \frac{n_2}{n_1 + n_2} = \gamma, \ 0 < \gamma < 1. \tag{4.74}$$

It is evident, from definition of $W^H_{F,G}$ that

$$W^H_{F,G} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\widehat{F}_H - F \right) - \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\widehat{G}_H - G \right) = \sqrt{\frac{n_2}{n_1 + n_2}} W^H_F - \sqrt{\frac{n_1}{n_1 + n_2}} W^H_G \tag{4.75}$$

where $W^H_F,\ W^H_G$ have the same form of (2.21). Clearly, as a consequence of Proposition 2.2.2, we have that

$$W^H_F \xrightarrow{weak} W_F \tag{4.76}$$

$$W^H_G \xrightarrow{weak} W_G \tag{4.77}$$

where W_F and W_G are the limiting processes obtained by Proposition 2.2.2 Hence, by the independence of the considered populations, (4.74) and the symmetry of the Gaussian processes

$$W^H_{F,G} \xrightarrow{weak} \sqrt{\gamma} W_F + \sqrt{1 - \gamma} W_G. \tag{4.78}$$

Consider now the corresponding Lorenz process

$$\mathcal{L}_{F,G}^{H} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\widehat{L}_F^H - \widehat{L}_G^H - L_F + L_G \right) \tag{4.79}$$

where L_F^H and L_G^H are the Hàajek estimators of L_F and L_G, computed as in (4.53). By Theorem 1.2.1 we have that

$$\mathcal{L}_{F,G}^{H} \xrightarrow{weak} \sqrt{\gamma}\theta_L'^F(W_F) + \sqrt{1-\gamma}\theta_L'^G(W_G) = \mathcal{L}_{F,G} \tag{4.80}$$

where $\theta_L'^F$ and $\theta_L'^G$ are the Hadamard derivative of the functional θ_L introduced in the previous section, at F and G respectively. Clearly by Proposition 3.2.1 we have that also the resampled Lorenz process $\mathcal{L}_{F,G}^{H*}$ converges to the same limit of (4.80). We will use these asymptotic result to implement a decision rule in order to perform, *via* resampling, the test (4.72).

Following the same approach of the previous section, we have that for $0 \leq p \leq 1$,

$$\left[\widehat{L}_F^H(p) - \widehat{L}_G^H(p) - \frac{\sqrt{n_1+n_2}R_{n,M}^{*-1}(1-\alpha)}{\sqrt{n_1 n_2}}; \widehat{L}_F^H(p) - \widehat{L}_G^H(p) + \frac{\sqrt{n_1+n_2}R_{n,M}^{*-1}(1-\alpha)}{\sqrt{n_1 n_2}} \right] \tag{4.81}$$

is a confidence band of asymptotic size $1-\alpha$ for the difference of the Lorenz curves $L_F - L_G$.

We now come back to the original problem of testing the weak stochastic dominance of two Lorenz curves. Our decision rule for test (4.72) is very intuitive. If the difference of $L_F(p) - L_P(p)$ is "sufficiently" negative for some $p \in (0,1)$ (this means that in some points L_F is really smaller than L_G), we reject the null hypothesis, otherwise we do not. Of course, we have to specify what "sufficiently" means. To this purpose we come back to confidence band defined in (4.81). We will reject the null hypothesis if and only if for some $\tilde{p} \in (0,1)$ we have that

$$\widehat{L}_F^H(\tilde{p}) - \widehat{L}_G^H(\tilde{p}) + \frac{\sqrt{n_1+n_2}R_{n,M}^{*-1}(1-\alpha)}{\sqrt{n_1 n_2}} < 0. \tag{4.82}$$

Remark 4.5.1. It is clear that if (4.82) holds for a $\tilde{p} \neq \min_p(\widehat{L}_F^H(p) - \widehat{L}_G^H(p))$ it holds also for $\min(\widehat{L}_F^H - \widehat{L}_G^H)$. Vice-versa, if condition (4.82) holds for the minimum, by the continuity of the Lorenz curves $\widehat{L}_F^H$ and $\widehat{L}_G^H$, it holds for some point $\tilde{p} \neq \min_p(\widehat{L}_F^H(p) - \widehat{L}_G^H(p))$. This proves that to perform the test (4.72) with the above described decision rule, it is sufficient to look if the confidence band (4.81) is under

the zero line at the minimum of $\widehat{L}_F^H(p) - \widehat{L}_G^H(p)$.

Let now see how to obtain a test with type I error equal to a fixed $\alpha \in (0,1)$. Firstly, it is worth to be noticed, that from the decision rule that we have assumed, the bigger is the difference (that is positive or at least equal to zero, under the null hypothesis) between the two Lorenz curves, the lower is the probability of making a type I error. This implies that the supremum under the null hypothesis of that probability (that is the definition of the type I error), is reached when the equality holds in the null hypothesis. In order to compute the asymptotic type I error of our procedure it is sufficient to observe that, from (4.81) and assuming $L_F = L_G$, the relationships

$$\alpha \approx Pr\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup_{p\in[0,1]} |\widehat{L}_G^H - \widehat{L}_F^H| > R_{n,M}^{*-1}(1-\alpha)\right\} = \tag{4.83}$$

$$= Pr\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup_{p\in[0,1]} \left(\widehat{L}_G^H - \widehat{L}_F^H\right) > R_{n,M}^{*-1}(1-\alpha)\right\} \tag{4.84}$$

$$+ Pr\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup_{p\in[0,1]} \left(\widehat{L}_F^H - \widehat{L}_G^H\right) > R_{n,M}^{*-1}(1-\alpha)\right\}$$

$$= Pr\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} \inf_{p\in[0,1]} \left(\widehat{L}_F^H - \widehat{L}_G^H\right) < -R_{n,M}^{*-1}(1-\alpha)\right\} \tag{4.85}$$

$$+ Pr\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup_{p\in[0,1]} \left(\widehat{L}_F^H - \widehat{L}_G^H\right) > R_{n,M}^{*-1}(1-\alpha)\right\}$$

$$= 2Pr\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} \inf_{p\in[0,1]} \left(\widehat{L}_F^H - \widehat{L}_G^H\right) < -R_{n,M}^{*-1}(1-\alpha)\right\} \tag{4.86}$$

hold, where (4.86) is due to the symmetry of the process

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}} \left(\widehat{L}_F^H - \widehat{L}_G^H\right)$$

under the null hypothesis. Hence we have that

$$Pr\left\{\inf_{p\in[0,1]} \left(\widehat{L}_F^H - \widehat{L}_G^H\right) + \frac{\sqrt{n_1+n_2} R_{n,M}^{*-1}(1-\alpha)}{\sqrt{n_1 n_2}} < 0\right\} \approx \frac{\alpha}{2}. \tag{4.87}$$

Thus, if we want to perform a test of asymptotic type I error equal to α we have to compute a confidence band of level $1-2\alpha$. It is worth to be noticed that such a procedure gives us the ability of performing two tests. The one where the null hypothesis assumes the weak dominance of L_F over L_G and clearly the one where the null hypothesis assumes that L_G dominates L_F.

We are now able to introduce our simulation study. For our simulations we considered three different cases as in Donald et al. (2004)

Case 1.

$$F = LogN(0.85, 0.6) \tag{4.88}$$

$$G = LogN(0.85, 0.6) \tag{4.89}$$

Clearly this case corresponds to the null hypothesis

Case 2.

$$F = LogN(0.85, 0.6) \tag{4.90}$$

$$G = LogN(0.7, 0.5) \tag{4.91}$$

In this case we have that L_G (strictly) dominates L_F

Case 3.

$$F = LogN(0.85, 0.6) \tag{4.92}$$

$$G = I(U \geq 0.2)LogN(0.6, 0.2) + I(U < 0.2)LogN(1.8, 0.3) \tag{4.93}$$

where U is a uniform variable on $[0, 1]$ and $I(a)$ takes value 1 if a is true. In this latter case we have two crossing Lorenz curves.

For each one of the cases above, we have generated $J = 1000$ couples of finite populations $\mathcal{U}_{N1}$ $\mathcal{U}_{N2}$. From each population we have selected samples according to a successive sampling design. For what concerns sample and finite population sizes $(n_1,\ n_2,\ N_1,\ N_2)$ we have simulated two different scenarios. A balanced one, in which both $n_1 = n_2$ and $N_1 = N_2$ and an unbalanced one, where the sizes are different. The sampling fraction f is constantly equal to $1/10$. The inclusion probabilities are taken proportional to a size variable equal to $Y^(0.5) \times LogN(0, 0.52)$. With this choice the correlation between the inclusion probabilities and the interest characters is about 0.40 in the finite populations. For each of the considered samples we have selected $M = 1000$ bootstrap samples according to a successive sampling design. In computing confidence bands (4.81) confidence levels of $1 - \tilde{\alpha} = 0.90, 0.95, 0.99$ have been considered, corresponding to tests of nominal sizes $\alpha = 0.05, 0.025, 0.005$.

In order to test the performance of our proposal for each one of the cases above

we have performed both the tests:

$$T^F = \begin{cases} H_0^F : L_F(p) - L_G(p) \geq 0, \ \forall p \in [0,1] \\ H_1^F : L_F(p) - L_G(p) < 0, \ \text{for some } p \in [0,1] \end{cases} \tag{4.94}$$

$$T^G = \begin{cases} H_0^G : L_G(p) - L_F(p) \geq 0, \ \forall p \in [0,1] \\ H_1^G : L_G(p) - L_F(p) < 0, \ \text{for some } p \in [0,1] \end{cases} \tag{4.95}$$

and we have computed these indicators, naturally obtained as Monte Carlo estimates of the correspondent analytic quantity:

1. Estimated Type I Errors (when the Case 1. is considered)

$$\hat{\alpha}_F = \frac{1}{J}\sum_{j=1}^{J} I\left(\inf_{p\in[0,1]}\left(\widehat{L}_F^H - \widehat{L}_G^H\right) + \frac{\sqrt{n_1+n_2}R_{n,M}^{*-1}(1-\tilde{\alpha})}{\sqrt{n_1 n_2}} < 0\right) \tag{4.96}$$

$$\hat{\alpha}_G = \frac{1}{J}\sum_{j=1}^{J} I\left(\inf_{p\in[0,1]}\left(\widehat{L}_G^H - \widehat{L}_F^H\right) + \frac{\sqrt{n_1+n_2}R_{n,M}^{*-1}(1-\tilde{\alpha})}{\sqrt{n_1 n_2}} < 0\right) \tag{4.97}$$

2. Estimated Power, that is formally computed as (4.96) and (4.97) when Case 3 is assumed instead of Case 1.

Results are summarized in tables below.

Estimated Type 1 Error, Case 1

	$\boldsymbol{\alpha = 0.05}$		$\boldsymbol{\alpha = 0.025}$		$\boldsymbol{\alpha = 0.005}$	
	T^F	T^G	T^F	T^G	T^F	T^G
$n_1 = 50, \ n_2 = 50$	0.048	0.048	0.028	0.023	0.008	0.007
$n_1 = 150, \ n_2 = 150$	0.054	0.045	0.029	0.018	0.01	0.003
$n_1 = 250, \ n_2 = 250$	0.055	0.046	0.03	0.027	0.01	0.004

Table 4.16. Estimated type I error in the balanced scenario. In columns T^k are reported the quantity $\hat{\alpha}_k$ with $k = F, \ G$.

	$\boldsymbol{\alpha = 0.05}$		$\boldsymbol{\alpha = 0.025}$		$\boldsymbol{\alpha = 0.005}$	
	T^F	T^G	T^F	T^G	T^F	T^G
$n_1 = 50, \ n_2 = 150$	0.032	0.077	0.012	0.051	0.000	0.015
$n_1 = 150, \ n_2 = 50$	0.063	0.029	0.034	0.014	0.01	0.002
$n_1 = 100, \ n_2 = 250$	0.028	0.071	0.012	0.037	0.002	0.013
$n_1 = 250, \ n_2 = 100$	0.071	0.035	0.04	0.018	0.005	0.005

Table 4.17. Estimated type I error in the unbalanced scenario. In columns T^k are reported the quantity $\hat{\alpha}_k$ with $k = F, \ G$.

From Tables 4.16-4.17 it is evident that the balanced case performs better than the unbalanced one. In the balanced case we have estimated type I errors very close

Estimated Type 1 Error and Estimated Power, Case 2

	$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.005$	
	T^F	T^G	T^F	T^G	T^F	T^G
$n_1 = 50,\ n_2 = 50$	0.28	0.005	0.182	0.002	0.061	0.001
$n_1 = 150,\ n_2 = 150$	0.561	0.001	0.459	0.001	0.241	0.000
$n_1 = 250,\ n_2 = 250$	0.757	0.000	0.637	0.000	0.383	0.000

Table 4.18. Balanced scenario. In T^G columns estimated type I error is reported because in Case 2, H_0^G holds. In columns T^F is reported the estimated power, because in Case 2, H_1^F holds.

	$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.005$	
	T^F	T^G	T^F	T^G	T^F	T^G
$n_1 = 50,\ n_2 = 150$	0.301	0.005	0.192	0.003	0.057	0.001
$n_1 = 150,\ n_2 = 50$	0.431	0.000	0.331	0.000	0.163	0.000
$n_1 = 100,\ n_2 = 250$	0.503	0.000	0.373	0.000	0.160	0.000
$n_1 = 250,\ n_2 = 100$	0.497	0.000	0.368	0.000	0.153	0.000

Table 4.19. Balanced scenario. In T^G columns estimated type I error is reported because in Case 2, H_0^G holds. In columns T^F is reported the estimated power, because in Case 2, H_1^F holds.

Estimated Power, Case 3.

	$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.005$	
	T^F	T^G	T^F	T^G	T^F	T^G
$n_1 = 50,\ n_2 = 50$	0.046	0.31	0.017	0.198	0.002	0.063
$n_1 = 150,\ n_2 = 150$	0.162	0.678	0.073	0.536	0.007	0.297
$n_1 = 250,\ n_2 = 250$	0.336	0.892	0.155	0.803	0.022	0.564

Table 4.20. Estimated power in balanced scenario. In columns T^F the estimated power for test (4.94) is reported. Columns T^G refer to test (4.95)

	$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.005$	
	T^F	T^G	T^F	T^G	T^F	T^G
$n_1 = 50,\ n_2 = 150$	0.049	0.486	0.013	0.378	0.001	0.2
$n_1 = 150,\ n_2 = 50$	0.065	0.311	0.032	0.183	0.004	0.039
$n_1 = 100,\ n_2 = 250$	0.143	0.677	0.043	0.566	0.002	0.357
$n_1 = 250,\ n_2 = 100$	0.115	0.629	0.053	0.489	0.013	0.213

Table 4.21. Estimated power in balanced scenario. In columns T^F the estimated power for test (4.94) is reported. Columns T^G refer to test (4.95)

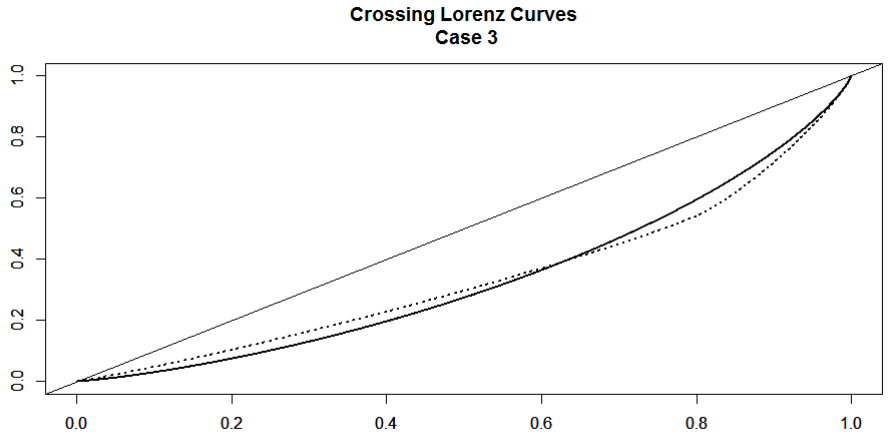


Fig.4.8. Crossing Lorenz Curves. The solid line is L_F and the dotted one is L_G.

to the nominal level and quite balanced. There are some fluctuation of estimated α when the sample size is increased, probably due to the bigger variability of the inclusion probanilities. The unbalanced scenario performs also well, but if compared to the balanced scenario the performance is lower. In fact, by the symmetry of the confidence band, the type I errors α_F and α_G have to be equal,but if we look at Table 4.17 we see some differences. Moving to Case 2. where the curve L_G dominates L_F we have results concordant to what intuition suggests. In fact as we can see from Tables 4.18-4.19, we have an estimated power that increases with the size in both the scenarios. In addition, when the sample size increases, the estimated type I error decreases to zero. The balanced case performs well in terms of power, being able to reject the null hypothesis a larger number of times if compared to the unbalanced case. To better understands case 3, we have to look Figure 4.8. As we can see, the biggest difference between the two curves, is around 0.8, when the curve L_F is over the curve L_G. This observation explain why, the estimated power in Tables 4.20-4.21 is bigger when the test (4.95) is considered. However, increasing the size of the samples makes our procedure able to reject both H_0^F and H_0^G, even if the strength of rejection for H_0^F is not as high as the one for H_0^G. At the end we want to highlight that from Tables 4.16-4.17 in Case 1., the size of the confidence band is well estimated in both the balanced and the unbalanced scenario. Thus the procedure of constructing a confidence band for the difference of two Lorenz curves performs well when two sample size are considered, while testing the stochastic dominance has a bit lower performance in the unbalanced scenario.

Conclusions And Additional Considerations

Nowadays the superpopulation approach is widespread in survey sampling. It allows more general inferential results and sometimes it is a necessity like in *small area estimation*, where the (direct) *design-based* inference is unfeasible. In addition, the use of a superpopulation in a *model-assisted* inference, can improve the results obtained by the classical *design-based* inference. For all of these reasons the present dissertation focused on deriving some inferential tools to use when the statistician interest is about the superpopulation.

In this work we followed a non-parametric approach in the sense that we have no parametric assumptions on the superpopulation model and in addition the inference procedure is totally *design-based*. This choice is made in order to obtain a *robustness* to violations of the assumed superpopulation model.

Assuming samples selected by a finite population with a complex sample design, it makes classical inferential results not valid, even if a superpopulation model is assumed. In this work the first main contribution is an extension of the Donsker's theorem to finite population framework with superpopulation approach. As discussed in Chapter 2 there are some recent parallel results, but they are derived under different assumptions. Making a bridge with classical empirical processes theory, we have shown the convergence of the Hàjek estimator of the superpopulation distribution function (opportunely scaled and centered), to a Gaussian process. Our functional central limit theorem fully characterizes the asymptotic distribution of the Hàjek estimator, providing an analytic tool for the inference about the superpopulation. In addition this characterization is also needed in order to prove the validity of our resampling scheme, following the same spirit of classical nonparametric statistics. In fact, our second main contribution is about the consistency of

the "multinomial" resampling scheme. The consistency is proved showing that the asymptotic distribution of the Hàjek estimator based on the resampled units is the same of the Hàjek estimator based on the original units. To our knowledge it is not available in the literature a resampling procedure that allows for inferring the superpopulation, and that is justified by asymptotic considerations. The proposed resampling procedure is tested in different situation *via* simulations. Several different applications are proposed, providing also an intuitive test procedure for the stochastic dominance of Lorenz curve. The resampling procedure showed a really well behavior in each one of the simulation studies. Clearly a next stage of this research will be testing the performance of our procedure on real data.

Before concluding, we want to add some considerations. Our functional limit theorem is derived assuming that the finite population is obtained as $i.i.d.$ replications of the superpopulation model. This result is proved showing the convergence of two processes, one that takes into account the sampling variability and one that considers the superpopulation variability. It is worth to notice that the convergence of the first process does not require in any way the $i.i.d.$ assumption. This condition is necessary to use the classic Donsker's Theorem. Thus, the whole convergence can be still proved if the $i.i.d.$ assumption is replaced by one that makes, a relaxed version of Donsker's Theorem, hold. Of course, as a consequence it changes the way the convergence happens. The last consideration is about one of the possible further developments of this work. In fact, in this work the auxiliary information is involved only at the sampling stage. It could be interesting to verify if our methodology is improved using it also in the estimation stage. One possibility to incorporate the auxiliary information in the estimation stage is represented by the distribution function estimator proposed in Rao et al. (1990).

Appendix

Appendix 1.

Proof of Lemma 1.1.1. Conditionally on $\mathcal{Y}_N$, $\mathcal{T}_N$, the expectation of (1.14) w.r.t the sampling design P is equal to 1. The variance of (1.14) w.r.t. the sampling design P, conditionally on $\mathcal{Y}_N$, $\mathcal{T}_N$, is equal to

$$\begin{aligned}\mathbb{V}_P\left(\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{\pi_i}\,\middle|\,\mathcal{Y}_N,\,\mathcal{T}_N\right) &= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\frac{1}{\pi_i^2}\mathbb{V}_P(D_i\,|\mathcal{Y}_N,\,\mathcal{T}_N)\right.\\ &\qquad\left.+\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{\pi_i\pi_j}\mathbb{C}_P(D_i,\,D_J\,|\mathcal{Y}_N,\,\mathcal{T}_N)\right\}\\ &\leq \frac{1}{N^2}\left\{\sum_{i=1}^{N}\frac{1}{\pi_i}+\sum_{i=1}^{N}\sum_{j\neq i}\left|\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right|\right\}.\end{aligned}$$

From $\pi_i = nx_i/\sum_{j=1}^{N}x_j$ (with $x_i = g(t_{i1},\,\ldots,\,t_{iL})$) and the strong law of large numbers, it is not difficult to see that the $N^{-1}\sum_i \pi_i^{-1}$ converges for a set of (sequences of) y_is, t_{ij}s of $\mathbb{P}$-probability 1. Furthermore, from the assumption of maximal asymptotic entropy of the sampling design implies (cfr. Hájek and Dupac (1981), Th. 7.4) that

$$\left|\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right| \leq \frac{C}{N}$$

C being an absolute constant. This shows that (1.14) tends to 1 as N increases, for a set of (sequences of) y_is, t_{ij}s of $\mathbb{P}$-probability 1 and for a set of $\mathbf{D}_N$s of P-probability tending to 1. □

Appendix 2.

Proofs of preparatory Lemmas 2.1.1-2.1.5 are given in Conti et al. (2015)

Proof of Proposition 2.1.1. The formal proof is equivalent to the proof of Proposition 1 contained in the appendix of Conti (2014). Here we just give an informal idea on how the proof works In order to prove the weak convergence to a Gaussian process, two steps are necessary. The first one is to prove the convergence of the finite-dimensional distribution to a Gaussian variable, then we have to prove that these distributions are tight. The first step is reached observing that, because of Lemmas 2.1.1-2.1.5 and applying Theorem 0.3.1 (under assumption of high entropy sampling designs), the quantity (for a fixed $y \in \mathbb{R}$

$$\sqrt{N}\frac{\widehat{F}_H(y) - F_N(y)}{S_N(y)} \xrightarrow{weak} N(0,1) \tag{4.98}$$

Then to prove the convergence of the finite-dimensional distributions, result (4.98) is extended to the multivariate case through the Cramer-Wold device. The tightness part is obtained resorting to a practical criterion contained in Billingsley (1968) p.133 and using the high entropy assumption. □

Proof of Proposition 2.1.2. The proof of the weak convergence is obtained mimicking the proof of Proposition 1 in Conti (2014). We limit our selves to prove the form of the covariance kernel 2.12. Define the quantities:

$$Z_{i,N}(y) = I_{(y_i \leq y)} - \pi_i \frac{\sum_{i=1}^{N}(1-\pi_i)I_{(y_i \leq y)}}{\sum_{i=1}^{N}\pi_i(1-\pi_i)}$$

$$S_N^2(y) = \sum_{i=1}^{N}\left(\frac{1}{\pi_i} - 1\right) Z_{i,N}(y)^2$$

Following the same approach as in Hájek (1964) it is clear that, pushing N to the infinity, $S_N^2(y)/N$ gives the asymptotic variance of $\widehat{F}_{HT}(y)$. We have that

$$\frac{S_N^2(y)}{N} = B_{1,N} + B_{2,N} + B_{3,N}$$

where

$$B_{1,N} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\pi_i}-1\right) I_{(y_i\le y)}$$

$$B_{2,N} = \frac{1}{N}\sum_{i=1}^{N}\pi_i(1-\pi_i)\left(\frac{\sum_{i=1}^{N}(1-\pi_i)I_{(y_i\le y)}}{\sum_{i=1}^{N}\pi_i(1-\pi_i)}\right)^2$$

$$B_{3,N} = -2B_{2,N}$$

By the strong law of large number and by the assumption $\pi_i = nx_i/\sum_{i=1}^{N} x_i = fx_i/\bar{x}_N$ where $\bar{x}_N$ is the arithmetic mean of x_is, we have that

$$B_{1,N} \to \left(\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}K_{-1}(y)-1\right)F(y),\ a.s.-\mathbb{P} \tag{4.99}$$

$$B_{2,N} \to F(y) - f\frac{K_{+1}(y)F(y)}{\mathbb{E}_{\mathbb{P}}[X]} \tag{4.100}$$

Hence,

$$\frac{S_N^2(y)}{N} \to \left(\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}K_{-1}(y)-1\right)F(y) - \frac{1}{d}\left(F(y)-f\frac{K_{+1}(y)F(y)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)^2 \tag{4.101}$$

Because of the presence of the scaling factor $\sqrt{N}$ in (4.101), instead of $\sqrt{n}$ (that is the scaling factor considered for the empirical process W_n^{HT}) the covariance kernel must be rescaled by a factor f and hence, it has the form expressed in (2.12). □

Proof of Proposition 2.2.2. We only have to prove that the asymptotic independence of the two sequences of processes W_n^H and W_N. To this purpose, it is sufficient to show the asymptotic independence of their finite-dimensional distributions. Let m, l be positive integers, and take $m+l$ points $y_1^{(1)},\dots,y_m^{(1)},y_1^{(2)},\dots,y_l^{(2)}$.

It is not difficult to see that

$$\lim_{N\to\infty} Pr\left\{W_n^H(y_1^{(1)}) \le z_1^{(1)},\dots,W_n^H(y_m^{(1)}) \le z_m^{(1)}, W_N(y_1^{(2)}) \le z_1^{(2)},\dots,W_N(y_l^{(2)}) \le z_l^{(2)}\right\} =$$
$$\lim_{N\to\infty} \mathbb{E}\left[I_{(W_n^H(y_1^{(1)})\le z_1^{(1)},\dots,W_n^H(y_m^{(1)})\le z_m^{(1)},W_N(y_1^{(2)})\le z_1^{(2)},\dots,W_N(y_l^{(2)})\le z_l^{(2)})}\right] =$$
$$\lim_{N\to\infty} \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_P[I_{(W_n^H(y_1^{(1)})\le z_1^{(1)},\dots,W_n^H(y_m^{(1)})\le z_m^{(1)},W_N(y_1^{(2)})\le z_1^{(2)},\dots,W_N(y_l^{(2)})\le z_l^{(2)})}|\mathcal{Y}_N,\mathcal{T}_N]\right] =$$
$$\lim_{N\to\infty} \mathbb{E}_{\mathbb{P}}\left[P\{W_n^H(y_1^{(1)}) \le z_1^{(1)},\dots,W_n^H(y_m^{(1)}) \le z_m^{(1)}|\mathcal{Y}_N,\mathcal{T}_N\}\cdot I_{(W_N(y_1^{(2)})\le z_1^{(2)})}\cdots I_{(W_N(y_l^{(2)})\le z_l^{(2)})}\right] =$$
$$\mathbb{E}_{\mathbb{P}}\left[\lim_{N\to\infty} P\{W_n^H(y_1^{(1)}) \le z_1^{(1)},\dots,W_n^H(y_m^{(1)}) \le z_m^{(1)}|\mathcal{Y}_N,\mathcal{T}_N\}\cdot \lim_{N\to\infty} I_{(W_N(y_1^{(2)})\le z_1^{(2)})}\cdots I_{(W_N(y_l^{(2)})\le z_l^{(2)})}\right] =$$
$$Pr\{W_1(y_1^{(1)}) \le z_1^{(1)},\dots,W_1(y_m^{(1)}) \le z_m^{(1)}\}\cdot \lim_{N\to\infty} \mathbb{E}_{\mathbb{P}}\left[I_{(W_N(y_1^{(2)})\le z_1^{(2)})}\cdots I_{(W_N(y_l^{(2)})\le z_l^{(2)})}\right] =$$
$$Pr\left\{W_1(y_1^{(1)}) \le z_1^{(1)},\dots,W_1(y_m^{(1)}) \le z_m^{(1)}\right\}\cdot \lim_{N\to\infty} \mathbb{P}\left\{W_N(y_1^{(2)}) \le z_1^{(2)},\dots,W_N(y_l^{(2)}) \le z_l^{(2)}\right\} =$$
$$Pr\left\{W_1(y_1^{(1)}) \le z_1^{(1)},\dots,W_1(y_m^{(1)}) \le z_m^{(1)}\right\} Pr\left\{W_2(y_1^{(2)}) \le z_1^{(2)},\dots,W_2(y_l^{(2)}) \le z_l^{(2)}\right\}$$

which proves the asserted asymptotic independence. □

Proof of Proposition 2.2.4. As shown in proof of Proposition 6 in Conti (2014), from Proposition 2.1.1 and the Skorokhod representation theorem (cfr. Billingsley (1968)), it follows that

$$\sup_y \left|\widehat{F}_H(y) - F_N(y)\right| \to 0 \text{ as } N\to\infty \tag{4.102}$$

for a set of $\mathbf{D}_N$s with P-probability tending to 1, and for a set of (sequences of Y_is, T_{ij}s having $\mathbb{P}$-probability 1. In the second place, from the "classical" Glivenko-Cantelli theorem, we have:

$$\sup_y |F_H(y) - F(y)| \to 0 \text{ as } N\to\infty \tag{4.103}$$

for a set of (sequences of) Y_is, T_{ij}s having $\mathbb{P}$-probability 1. Conclusion (2.33) easily follows from (4.102) and (4.103). □

Appendix 3.

Proof of Proposition 3.2.1. We want to show that Lemmas 2.1.1-2.1.5 hold for the predicted population conditionally on the sample and on the original population:

i) We want to show that $\frac{d_N^*}{N} \to f - f^2 \frac{\mathbb{E}_{\mathbb{P}}[X_1^2]}{\mathbb{E}_{\mathbb{P}}[X_1]^2},\ a.s. - \mathbb{P}^*$.

$$\frac{d_N^*}{N} = \frac{1}{N} \sum_{j \in \mathcal{U}_N^*} \frac{n x_i^*}{\sum\limits_{j \in \mathcal{U}_N^*} x_j^*} \left(1 - \frac{n x_i^*}{\sum\limits_{j \in \mathcal{U}_N^*} x_j^*} \right) = f - f^2 \frac{1}{N \bar{x}_N^*} \sum_{i=1}^N x_i^{*2} \tag{4.104}$$

where,

$$x_i^* = \begin{cases} x_1 & \text{with prob } \frac{D_1 \pi_1^{-1}}{\sum\limits_{\mathcal{U}_N} D_k \pi_k^{-1}} \\ x_2 & \text{with prob } \frac{D_2 \pi_2^{-1}}{\sum\limits_{\mathcal{U}_N} D_k \pi_k^{-1}} \\ \vdots & \vdots \\ x_N & \text{with prob } \frac{D_N \pi_N^{-1}}{\sum\limits_{\mathcal{U}_N} D_k \pi_k^{-1}} \end{cases} \tag{4.105}$$

and $\bar{x}_N^*$ is the mean over the pseudo population $\mathcal{U}_N^*$ of x_i^*s. Clearly, conditionally on $s, \mathcal{U}_N$, are $i.i.d$ random variables. In addition, observing that

$$\mathbb{E}_{\mathbb{P}^*}[x_i^*] = \frac{\sum_{i=1}^N x_i D_i \pi_i^{-1}}{\sum_{j=1}^N D_j \pi_j^{-1}} \xrightarrow{in\ P-probability} \frac{\sum_{i=1}^N x_i}{N} \xrightarrow{a.s.-\mathbb{P}} \mathbb{E}_{\mathbb{P}}[X_1] < \infty \tag{4.106}$$

$$\mathbb{E}_{\mathbb{P}^*}[x_i^{*2}] = \frac{\sum_{i=1}^N x_i^2 D_i \pi_i^{-1}}{\sum_{j=1}^N D_j \pi_j^{-1}} \xrightarrow{in\ P-probability} \frac{\sum_{i=1}^N x_i^2}{N} \xrightarrow{a.s.-\mathbb{P}} \mathbb{E}_{\mathbb{P}}[X_1^2] < \infty \tag{4.107}$$

using the weak law of large numbers we have that

$$\frac{\sum_{i=1}^N x_i^*}{N} \xrightarrow{in\ \mathbb{P}^*-probability} \frac{\sum_{i=1}^N x_i D_i \pi_i^{-1}}{\sum_{j=1}^N D_j \pi_j^{-1}} \xrightarrow{in\ P-probability} \frac{\sum_{i=1}^N x_i}{N} \xrightarrow{a.s.-\mathbb{P}} \mathbb{E}_{\mathbb{P}}[X_1] \tag{4.108}$$

$$\frac{\sum_{i=1}^N x_i^{*2}}{N} \xrightarrow{in\ \mathbb{P}^*-probability} \frac{\sum_{i=1}^N x_i^2 D_i \pi_i^{-1}}{\sum_{j=1}^N D_j \pi_j^{-1}} \xrightarrow{in\ P-probability} \frac{\sum_{i=1}^N x_i^2}{N} \xrightarrow{a.s.-\mathbb{P}} \mathbb{E}_{\mathbb{P}}[X_1^2] \tag{4.109}$$

hence,

$$\frac{d_N^*}{N} \to f - f^2 \frac{\mathbb{E}_{\mathbb{P}[X_1^2}}{\mathbb{E}_{\mathbb{P}}[X_1]^2} \tag{4.110}$$

ii) For what concerns the equivalent of Lemmas (2.1.2) for the resampling, it is

sufficient to notice that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i^*}\left(I_{(y_i^*\leq y)}-F_N^*\right)=\frac{\sum_{i=1}^{N}x_i^*}{f_N N}\sum_{i=1}^{N}\frac{1}{x_i^*}I_{(y_i^*\leq y)}-F_N^*\frac{\sum_{i=1}^{N}x_i^*}{f_N N}\sum_{i=1}^{N}\frac{1}{x_i^*} \tag{4.111}$$

and observing that, by the laws of large numbers (and the consistency of the Horvitz-Thompson estimator)

$$F_N^*\xrightarrow{in\ \mathbb{P}^*-probability}\hat{F}_H\xrightarrow{in\ P-probability}F_N\xrightarrow{a.s.-\mathbb{P}}F \tag{4.112}$$

iii) The proof of the resampling version of 2.1.3 is obtained replacing these quantities

$$Z_{i,N}^*(y)=\left(I_{(y_i^*\leq y)}-F_N^*\right)-\pi_i^*\frac{\sum_{i=1}^{N}(1-\pi_i^*)\left(I_{(y_i^*\leq y)}-F_N^*\right)}{\sum_{i=1}^{N}\pi_i^*(1-\pi_i^*)}$$

$$S_N^{*2}(y)=\sum_{i=1}^{N}\left(\frac{1}{\pi_i^*}-1\right)Z_{i,N}^{*2}(y)$$

in the Proof of Proposition 2.1.2 and using (4.108), (4.109) and (4.112)

iv) We have to prove that

$$\forall\epsilon>0,\exists N_\epsilon \text{ such that } |Z_{i,N}^*|\leq\epsilon\pi_i^*S_N^*(y),\forall N\geq N_\epsilon \tag{4.113}$$

By the resampling equivalent of Lemma 2.1.3 you have that $\forall\epsilon>0$ there exist a N_ϵ such that $\forall N\geq N_\epsilon$ it holds:

$$\begin{aligned}\frac{S_N^{*2}}{N}>&\left(\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}K_{-1}(y)-1\right)F(y)(1-F(y))-\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}(K_{-1}(y)-\mathbb{E}_{\mathbb{P}}[X_1^{-1}])F(y)^2\\&-\frac{f^2}{d}\left(1-\frac{k_{+1}(y)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)^2F(y)^2-\epsilon,\end{aligned} \tag{4.114}$$

$$|Z_{i,N}^*|\leq 1+\epsilon+\frac{f}{\mathbb{E}_{\mathbb{P}}[X_1]}X_i\frac{f(1-K_{+1}(y)/\mathbb{E}_{\mathbb{P}}[X_1])}{d}F(y). \tag{4.115}$$

From (4.114) the inequalities

$$\epsilon\pi_i^* S_N^*(y) \geq \frac{\epsilon}{2}\frac{f}{\mathbb{E}_{\mathbb{P}}[X_1]}X_i \left\{ \left(\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f} K_{-1}(y) - 1 \right) F(y)(1-F(y)) \right. \tag{4.116}$$

$$\left. -\frac{\mathbb{E}_{\mathbb{P}}[X_1]}{f}(K_{-1}(y) - \mathbb{E}_{\mathbb{P}}[X_1^{-1}])F(y)^2 - \frac{f^2}{d}\left(1 - \frac{k_{+1}(y)}{\mathbb{E}_{\mathbb{P}}[X_1]}\right)^2 F(y)^2 - \epsilon \right\}^{\frac{1}{2}} \sqrt{N} \tag{4.117}$$

$$\geq \left(1 + \epsilon + \frac{f}{\mathbb{E}_{\mathbb{P}}[X_i]} X_i \frac{f(1 - K_{+1}(y)/\mathbb{E}_{\mathbb{P}}[X_1])}{d} F(y)\right) N^{\gamma} \tag{4.118}$$

hold with $0 < \gamma < 1/2,\ \forall N \geq N_\epsilon$. Inequalities (4.115) (4.118) prove (4.113)

v) Is a consequence of *iii*) and *iv*).

The proof is completed observing that the sequences of y_i^*s and x_i^*s have P and $\mathbb{P}^*$-probability tending to 1 and $\mathbb{P}$-probability equal to 1. □

Proof of Proposition 3.2.1. This Proposition is composed of 3 claims.

Claim 1 The proof is obtained mimicking the proof of Proposition 1 in Conti (2014).

Claim 2 The unconditional convergence is obtained resorting to Lemma 2.2.1, as well as we have done for the original process W_n^H

Claim 3 The asymptotic independence is obtained mimicking the proof of Proposition 2.2.2 given few lines above.

□

Proof of Proposition 3.3.1. Let

$$R_n^*(z) = P^*\{Z_{n,m}^* \leq z | s, \mathcal{U}_N^*\}$$

be the true (resampling) distribution function of $Z_{n,m}^*$ (defined in (3.25)). By the two sided Dvoretzky-Kiefer-Wolfowitz inequality (for more see Dvoretzky et al. (1956) and Massart (1990)), we have

$$Pr\left\{\sup_{z\in\mathbb{R}}|\hat{R}_{n,M}^*(z) - R_n^*(z)| > \epsilon \;\middle|\; s, \mathcal{U}_N\right\} \leq 2e^{\{-2M\epsilon^2\}}. \tag{4.119}$$

Taking into account that by Glivenko-Cantelli theorem (see Theorem 19.1 Van der Vaart (2000) p. 266) R_n^* converges uniformly to Φ_{0,σ_θ^2}, and you have that (3.29) holds in probability. To obtain the almost sure convergence it is sufficient to use the Borel-Cantelli first lemma. Convergence (3.30) follows by the Skorohod's representation Theorem, observing that $\Pr\left\{R_{n,M}^{*-1} \leq z\right\} = R_{n,M}^*(z)$. □

Bibliography

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106:534–543.

Barrett, G. F. and Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, 71(1):71–104.

Barrett, G. F., Donald, S. G., and Bhattacharya, D. (2014). Consistent nonparametric tests for lorenz dominance. *Journal of Business & Economic Statistics*, 32(1):1–13.

Berger, Y. G. (1998). Rate of convergence to normal distribution for the horvitz-thompson estimator. *Journal of Statistical Planning and Inference*, 67:209–226.

Berger, Y. G. (2011). Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics*, 27(4):407–426.

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics*, 137:674–707.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9:1196–1217.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley, New York.

Boistard, H., Lopuhaä, H. P., and Ruiz-Gazen, A. (2015). Functional central limit theorems in survey sampling. *ArXiv e-prints*.

Bondesson, L., Traat, I., and Lundqvist, A. (2006). Pareto sampling versus sampford and conditional poisson sampling. *Scandinavian Journal of Statistics*, 33(4):699–720.

Booth, J. G., Butler, R. W., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89(428):1282–1289.

Brewer, K. and Donadio, M. E. (2003). The high entropy variance of the {H} orvitz-{T} hompson estimator. *Survey Methodology*, 29(2):189–196.

Chao, M.-T. and Lo, S.-H. (1985). A bootstrap method for finite population. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 399–405.

Chatterjee, A. (2011). Asymptotic properties of sample quantiles from a finite population. *Annals of the Institute of Statistical Mathematics*, 63:157–159.

Chauvet, G. (2007). *Méthodes de bootstrap en population finie*. PhD thesis, ENSAI.

Cifarelli, D. M., Conti, P. L., Regazzini, E., et al. (1996). On the asymptotic distribution of a general measure of monotone dependence. *The Annals of Statistics*, 24:1386–1399.

Cochran, W. G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34:492–510.

Conti, P. L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B*, 76:234–259.

Conti, P. L. and Marella, D. (2015). Inference for quantiles of a finite population: Asymptotic versus resampling results. *Scandinavian Journal of Statistics*, 42:545–561.

Conti, P. L., Marella, D., and Mecatti, F. (2015). Recovering sampling distributions of statistics of finite populations via resampling: a predictive approach. *Submitted for publication*.

Csörgő, S. and Rosalsky, A. (2003). A survey of limit laws for bootstrapped sums. *International Journal of Mathematics and Mathematical Sciences*, 2003:2835–2861.

Donald, S. G., Barrett, G. F., et al. (2004). Consistent nonparametric tests for lorenz dominance. In *Econometric Society 2004 Australasian Meetings*, number 321. Econometric Society.

Dvoretzky, A., Kiefer, J. C., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27:642–669.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Efron, B. et al. (2003). Second thoughts on the bootstrap. *Statistical Science*, 18(2):135–140.

Gill, R. D., Wellner, J. A., and Præstgaard, J. (1989). Non-and semi-parametric maximum likelihood estimators and the von Mises method (Part 1)[with discussion and reply]. *Scandinavian Journal of Statistics*, 16:97–128.

Golan, A., Judge, G., and Miller, D. (1996). *Maximum entropy econometrics: robust estimation with limited data.* Series in financial economics and quantitative analysis. Wiley.

Grafström, A. (2010). Entropy of unequal probability sampling designs. *Statistical Methodology*, 7:84–97.

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, volume 1814184. American Statistical Association Ithaca, NY.

Hájek, J. (1959). Optimal strategy and other problems in probability sampling. *Časopis pro pěstování matematiky*, 84(4):387–423.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35:1491–1523.

Hajek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, part on by d. basu. *Foundations of Statistical Inference*, page 326.

Hájek, J. and Dupac, V. (1981). *Sampling from a finite population.* Marcel Dekker, New York.

Holmberg, A. (1998). A bootstrap approach to probability proportional-to-size sampling. *Proceedings of the ASA Section on Survey research Methods*, pages 378–383.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53:814–861.

Lambert, P. (1993). *The Distribution and Redistribution of Income: A Mathematical Analysis*. The Distribution and Redistribution of Income: A Mathematical Analysis. Manchester University Press.

Lundquist, A. (2009). Contributions to the theory of unequal probability sampling.

Mai, J.-F. and Scherer, M. (2012). *Simulating copulas: stochastic models, sampling algorithms and applications*. Imperial College Press, London.

Marshall, A. W. and Olkin, I. (1967a). A generalized bivariate exponential distribution. *Journal of Applied Probability*, 4:291–302.

Marshall, A. W. and Olkin, I. (1967b). A multivariate exponential distribution. *Journal of the American Statistical Association*, 62:30–44.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283.

McCarthy, P. J. and Snowden, C. B. (1985). The bootstrap and finite population sampling.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3:167–195.

Patrick, B. (1999). Convergence of probability measures.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61:317–337.

Pfeffermann, D. and Sverchkov, M. (2006). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30:79–92.

Presnell, B. and Booth, J. G. (1994). Resampling methods for sample surveys. *Presnell, B. and Booth, JG (1994) Resampling methods for sample surveys. Technical Report 470, Department of Statistics, University of Florida, Gainesville, FL.*

Ranalli, M. G. and Mecatti, F. (2012). Comparing recent approaches for bootstrapping sample survey data: A first step towards a unified approach. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 4088–4099.

Rao, J., Kovar, J., and Mantel, H. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77(2):365–375.

Rao, J. N. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the american statistical association*, 83(401):231–241.

Ren, J.-J. and Sen, P. K. (1995). Hadamard differentiability on $D[0,1]^p$. *Journal of Multivariate Analysis*, 55:14–28.

Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83:698–708.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, 17:141–159.

Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62(2):135–158.

Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62(2):159–191.

Särdnal, C. E., Swensson, B., and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 & 623–656.

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87(419):755–765.

Smirnov, N. (1939a). Ob uklonenijah empiricheskoi krivoi raspredelenija. *Recueil MathMatSbornik, NS*, 6:13–26.

Smirnov, N. V. (1939b). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2).

Theil, H. (1967). *Economics and information theory*. Studies in mathematical and managerial economics. North-Holland Pub. Co.

Tillé, Y. (2006). *Sampling Algorithms*. Springer Series in Statistics. Springer.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press, Cambridge.

Wang, J. C. (2012). Sample distribution function based goodness-og-fit test for complex surveys. *Computational Statistics and Data Analysis*, 56:664–679.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 253–261.