TEST (2018) 27:95–121
https://doi.org/10.1007/s11749-016-0514-2

ORIGINAL PAPER

On the estimation of the characteristic function in finite populations with applications

M. D. Jiménez-Gamero[1] · J. L. Moreno-Rebollo[1] · J. A. Mayor-Gallego[1]

Received: 23 October 2015 / Accepted: 21 November 2016 / Published online: 30 November 2016

Abstract This paper studies the estimation of the characteristic function of a finite population. Specifically, the weak convergence of the finite population empirical characteristic process is studied. Under suitable assumptions, it has the same limit as the empirical characteristic process for independent, identically distributed data from a random variable, up to a multiplicative constant depending on the sampling design. Applications of the obtained results for the two-sample problem, testing for independence and testing for symmetry are given.

Keywords Finite population · Design-based inference · High entropy designs · Characteristic function · Test for the two-sample problem · Test for independence · Test for symmetry

Mathematics Subject Classification 62D05 · 62G10 · 62G09

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11749-016-0514-2) contains supplementary material, which is available to authorized users.

✉ M. D. Jiménez-Gamero
dolores@us.es

J. L. Moreno-Rebollo
jlmoreno@us.es

J. A. Mayor-Gallego
jmayor@us.es

[1] Departamento de Estadística e I.O., Universidad de Sevilla, Sevilla, Spain

1 Introduction

Let us consider a finite population $U = \{1, \ldots, N\}$ of size N. Let $y_U = (y_1, \ldots, y_N)^{\mathrm{T}}$, where $y_j \in \mathbb{R}$ is the value of the study variable Y for the jth population element. For simplicity in notation, along this paper it will be assumed that Y is scalar, nevertheless all methods and results presented here are valid for $Y \in \mathbb{R}^d$, for any fixed $d \in \mathbb{N}$.

The basic question in survey sampling is to estimate a linear parameter $\theta_\gamma = \gamma_U^{\mathrm{T}} y_U$, with $\gamma_U^{\mathrm{T}} = (\gamma_1, \ldots, \gamma_N)$ a known parameter vector. Usually, survey sampling focuses on the estimation of the finite population total, the mean on the whole population or on a domain and the population distribution function (df) at a fixed point or at all points,

$$F_N(t) = \frac{1}{N} \sum_{j \in U} \Delta(y_j \leq t), \quad t \in \mathbb{R},$$

where $\Delta(\mathcal{A}) = 1$ if $\mathcal{A}$ is true, and $\Delta(\mathcal{A}) = 0$ otherwise. The estimation of the df is of interest per se and because it provides a useful tool for making inferences on the population. For example, Wang (2012) proposed goodness-of-fit procedures for a study variable based on divergence measures between the design weighted estimator of the df and the hypothesized distribution; Conti (2014) studied the estimation of the df with applications to the construction of confidence bands, comparison of two populations and testing for the nullity of certain dependence measures of two variables. These articles study the limit in law of $\{\sqrt{n}(\hat{F}_{N,\pi}(t) - F_N(t)),\ t \in \mathbb{R}\}$, where $\hat{F}_{N,\pi}(t)$ denotes the Horvitz–Thompson (HT) estimator of $F_N(t)$, (also $\{\sqrt{n}(\hat{F}_{N,H}(t) - F_N(t)),\ t \in \mathbb{R}\}$, where $\hat{F}_{N,H}(t)$ denotes the Hájek estimator of $F_N(t)$). Wang (2012) did it under a superpopulation setting with F_N replaced by its superpopulation analogue, obtaining that the covariance of the limit process has two components, one resulting from finite population sampling and the other from the variation of the superpopulation df itself; by contrast, these processes were studied in Conti (2014) under a design-based framework, obtaining a different limit process.

In the context of sampling from a random variable (sfarv), it is well-known that the empirical distribution function (edf) provides a helpful device for making inferences on the random variable generating the data. The inferential procedures proposed in the articles in the above paragraph are sample survey versions/adaptations of some existing procedures based on the edf in the context of sfarv. Another quite valuable means for making inferences in such a context is the characteristic function (cf). The cf has been used in sampling from finite populations as a tool to derive results (see, for instance, the paper by Erdös and Rényi (1959), which uses the cf of the sum of the study variable in all samples to derive a central limit theorem for simple random sampling; or the book by Tillé (2006), that employs the cf of a sampling design to get properties of the sampling design). Nevertheless, to the best of our knowledge, the cf of a study variable has not been considered as a population parameter deserving to be estimated.

Proceeding as in the sfarv context, we define the cf associated to the study variable Y in the finite population U as

$$C_N(t) = \int \exp(\mathrm{i}yt)\mathrm{d}F_N(y) = \frac{1}{N}\sum_{j\in U}\exp(\mathrm{i}y_jt) = R_N(t) + \mathrm{i}I_N(t), \quad t \in \mathbb{R},$$

with $R_N(t) = \frac{1}{N}\sum_{j\in U}\cos(ty_j)$, $I_N(t) = \frac{1}{N}\sum_{j\in U}\sin(ty_j)$. Note that for each fixed t, $R_N(t)$ and $I_N(t)$ are linear parameters, therefore, given a sample s from U, they can be unbiasedly estimated by their HT estimators,

$$\hat{R}_{N,\pi}(t) = \frac{1}{N}\sum_{j\in s}\frac{\cos(ty_j)}{\pi_j}, \quad \hat{I}_{N,\pi}(t) = \frac{1}{N}\sum_{j\in s}\frac{\sin(ty_j)}{\pi_j},$$

where $\{\pi_j = P(j \in s),\ j \in U\}$ are the first order inclusion probabilities. These estimators can be alternatively obtained as follows: $C_N(t)$ can be estimated by replacing $F_N(y)$ by its HT estimator

$$\hat{F}_{N,\pi}(y) = \frac{1}{N}\sum_{j\in s}\frac{\Delta(y_j \le y)}{\pi_j}, \quad y \in \mathbb{R},$$

obtaining

$$\hat{C}_{N,\pi}(t) = \int \exp(\mathrm{i}yt)\mathrm{d}\hat{F}_{N,\pi}(y) = \frac{1}{N}\sum_{j\in s}\frac{\exp(\mathrm{i}y_jt)}{\pi_j} = \hat{R}_{N,\pi}(t) + \mathrm{i}\hat{I}_{N,\pi}(t),$$

$t \in \mathbb{R}$. The problem with the estimator $\hat{C}_{N,\pi}(t)$ is that since, in general, $\hat{F}_{N,\pi}(y)$ is not a df, it may not be a proper cf (it happens when $\hat{C}_{N,\pi}(0) = \frac{1}{N}\sum_{j\in s}{\pi_j}^{-1} \neq 1$). This can be avoided by considering the Hájek estimator of $F_N(y)$,

$$\hat{F}_{N,H}(y) = \frac{1}{\hat{N}_\pi}\sum_{j\in s}\frac{\Delta(y_j \le y)}{\pi_j}, \quad y \in \mathbb{R},$$

with $\hat{N}_\pi = \sum_{j\in s}\pi_j^{-1}$, which is a true df. We can thus define the Hájek estimator of $C_N(t)$ as

$$\hat{C}_{N,H}(t) = \int \exp(\mathrm{i}yt)\mathrm{d}\hat{F}_{N,H}(y) = \frac{1}{\hat{N}_\pi}\sum_{j\in s}\frac{\exp(\mathrm{i}y_jt)}{\pi_j} = \hat{R}_{N,H}(t) + \mathrm{i}\hat{I}_{N,H}(t),$$

$t \in \mathbb{R}$, with $\hat{R}_{N,H}(t) = N\hat{R}_{N,\pi}(t)/\hat{N}_\pi$ and $\hat{I}_{N,H}(t) = N\hat{I}_{N,\pi}(t)/\hat{N}_\pi$, which is a proper cf.

The estimation of the population cf at a single point has little (or no) concern since $C_N(t)$ is seldom an interesting parameter. Nevertheless, it will be seen that $C_N(t)$ can be a useful tool for making inferences in finite populations. With this aim, this paper studies the performance of $\hat{C}_{N,\pi}(t)$ and $\hat{C}_{N,H}(t)$ as estimators of the whole cf for general sample designs. Under certain conditions on the sample design, it will be seen that the asymptotic behavior is quite similar to that obtained in the sfarv. Specifically, for high entropy designs and for the Hájek estimator, it is shown that the empirical characteristic function (ecf) process converges in law to a limit which is, up to a multiplicative constant depending on the design, equal to the one obtained in the sfarv setting. This let us propose procedures for testing several hypotheses of interest, in parallel to some well-established cf-based procedures in the sfarv context such as tests for the two-sample problem (see Meintanis 2005; Henze et al. 2005; Alba-Fernández et al. 2008; Hušková and Meintanis 2008, for sfarv), testing for independence of two variables (see Csörgő 1985; Székely et al. 2007; Meintanis and Iliopoulos 2008; Hlávka et al. 2011, for sfarv) and testing for symmetry (see Feuerverger and Mureika 1977; Neuhaus and Zhu 1998; Henze et al. 2003, for sfarv).

The paper is organized as follows. Section 2 describes the setting and lists the assumptions used to prove the results in next sections. Section 3 studies the limit law of the ecf process when the cf is estimated by means of the HT estimator and of the Hájek estimator. Sections 4–6 give applications of the obtained results to the problems of testing for the equality of two (or more) populations, testing for the independence of two (or more) study variables and testing for the symmetry of the study variable about an known or unknown point, respectively. The proposed procedures are illustrated with numerical simulations. Section 7 contains some concluding remarks. All proofs are sketched in the supplementary material.

Before ending this section, we introduce some notation: along this paper M denotes a generic positive constant taking many different values; $\xrightarrow{P}$ denotes convergence in probability; for any complex number $x = a + \mathrm{i}b$, $a, b \in \mathbb{R}$, $|x| = (a^2+b^2)^{1/2}$ denotes its modulus; if w is a nonnegative function satisfying

$$0 < \int w(t)\mathrm{d}t < \infty, \tag{1}$$

where an unspecified integral denotes integration over $\mathbb{R}$, then $L_2(w) = \{g : \mathbb{R} \to \mathbb{C} : \|g\|_w^2 = \int |g(t)|^2 w(t)\mathrm{d}t < \infty\}$; if $g_1, g_2 \in L_2(w)$ then, $\langle g_1, g_2\rangle_w = \int g_1(t)\bar{g}_2(t)w(t)\mathrm{d}t$; $C(K)$ denotes the space of continuous complex-valued functions defined on K, endowed with the usual supremum norm.

2 The setting and assumptions

Throughout this paper a sample s is a subset of n distinct units from U, where $n \in \mathbb{N}$ is a constant, called the sample size. Let $\mathcal{S}$ be the set of all samples s from U. Any function P on $\mathcal{S}$ satisfying $P(s) \geq 0$, $\forall s \in \mathcal{S}$, and $\sum_{s\in\mathcal{S}} P(s) = 1$ is called a fixed size sampling design (without replacement). For each $j \in U$, let $\delta_j(s) = \Delta(j \in s)$, that is, δ_j is a Bernoulli random variable taking the value 1 when the unit j is included in the sample. Note that $\sum_{j\in U} \delta_j = n$.

Let $p_1, \ldots, p_N$ be N positive numbers so that $\sum_{j\in U} p_j = n$. A Poisson sampling design with parameters $p_1, \ldots, p_N$ is a sampling design such that the random variables $\delta_1, \ldots, \delta_N$ are independent with $P(\delta_j = 1) = p_j$, $1 \leq j \leq N$. The rejective sampling is a Poisson sampling conditionally on $\sum_{j\in U} \delta_j = n$. The entropy of a sampling design is defined as

$$H(P) = -\sum_{s\in\mathcal{S}} P(s)\log\{P(s)\},$$

with $0\log(0) = 0$. The rejective sampling has maximal entropy among the sampling designs of constant sample size n and fixed first order inclusion probabilities (Hájek 1981, Theorem 3.4). The type of designs considered in this work are (asymptotically) close to the rejective sampling. The closeness will be measured as follows: let P be a sampling design and let P_R be the rejective sampling design having the same first order inclusion probabilities as P, then the Kullback–Leibler divergence between P and P_R,

$$D(P||P_R) := \sum_{s\in\mathcal{S}} P(s)\log\left\{\frac{P(s)}{P_R(s)}\right\},$$

will be used to quantify how close is P to P_R.

Next we list some assumptions required to derive the results in the subsequent sections.

Assumption A.1 The population and the sampling design belong to a sequence of populations and fixed size sampling designs, respectively, indexed by ν. The sample size n_ν, the population size N_ν and the sampling designs $\{P_\nu(s),\, s \subseteq U_\nu\}$ (and hence the associated inclusion probabilities) also vary with ν. N_ν increases with ν. All limits are taken when $\nu \to \infty$ but, to simplify notation, ν will be suppressed. All convergence results are to be interpreted as being with respect to the sequence of sampling designs.

Assumption A.2 $\frac{d_N}{N} \to d$, for some $0 < d < \infty$, where $d_N = \sum_{j\in U}\pi_j(1-\pi_j)$.

Assumption A.3 $\frac{n}{N} \to f$, for some $0 < f < 1$.

Assumption A.4 $\min_{j\in U}\pi_j \geq M$, $\forall N$.

Assumption A.5 $A_N = \frac{1}{N}\sum_{j\in U}\frac{1}{\pi_j} \to A$, for some $A > 0$.

Assumption A.6 For each population $\{U_N,\ N \geq 1\}$, let P be the actual sampling design and let P_R be the rejective sampling design having the same first order inclusion probabilities as P. Then, $D(P||P_R) \to 0$.

Assumption A.7 $\frac{1}{N}\sum_{j\in U}\pi_j\cos(ty_j) - \frac{n}{N}R_N(t) \to 0$, $\frac{1}{N}\sum_{j\in U}\pi_j\sin(ty_j) - \frac{n}{N}I_N(t) \to 0$, $\forall t$.

Assumption A.8 $\frac{1}{N}\sum_{j\in U}\frac{1}{\pi_j}\cos(ty_j) - A_N R_N(t) \to 0$, $\frac{1}{N}\sum_{j\in U}\frac{1}{\pi_j}\sin(ty_j)$ $-A_N I_N(t) \to 0$, $\forall t$.

Assumption A.9 $C_N(t) \to C(t)$, $\forall t$, where $C(t) = R(t) + \mathrm{i}I(t)$ is a cf.

Some comments are in order: Assumption A.1 is commonly assumed in design-based inference (see, for example, Isaki and Fuller 1982). As stated at the end of this assumption, all convergence results will be interpreted with respect to the sequence of sampling designs. For instance, we say that $K = o_P(n^{-\sigma})$ if $P_\nu(n_\nu^\sigma|K| > \varepsilon) \to 0$, $\forall\varepsilon > 0$, when $\nu \to \infty$, and we say that $K = O_P(n^{-\sigma})$ if $\forall\varepsilon > 0\ \exists M = M(\varepsilon) > 0$

and $\nu(\varepsilon)$ such that $P_\nu(n_\nu^\sigma|K| \leq M) \geq 1-\varepsilon, \forall \nu \geq \nu(\varepsilon)$. Recall that the subindex ν will be omitted.

Assumption A.2 implies that $d_N \to \infty$. Since $d \leq n$, Assumption A.2 implies that $n \to \infty$. It also follows that $f \geq d$.

Let $f_N(t) = \frac{1}{N}\sum_{j\in U} \pi_j \cos(ty_j) - \frac{n}{N} R_N(t)$. Then

$$|f_N(t)| = \left|\frac{1}{N}\sum_{j\in U}\left(\pi_j - \frac{n}{N}\right)\cos(ty_j)\right| \leq \frac{1}{N}\sum_{j\in U}\left(\pi_j + \frac{n}{N}\right) = 2\frac{n}{N}.$$

The same bound is valid for $f_N(t) = \frac{1}{N}\sum_{j\in U} \pi_j \sin(ty_j) - \frac{n}{N} I_N(t)$. Therefore, the dominated convergence theorem and Assumption A.7 imply that

$$\int \left\{\frac{1}{N}\sum_{j\in U}\pi_j\cos(ty_j) - \frac{n}{N}R_N(t)\right\}^2 w(t)\mathrm{d}t \to 0,$$
$$\int \left\{\frac{1}{N}\sum_{j\in U}\pi_j\sin(ty_j) - \frac{n}{N}I_N(t)\right\}^2 w(t)\mathrm{d}t \to 0,$$

for any nonnegative function w satisfying (1). Analogously, Assumptions A.8 and A.9 imply

$$\int \left\{\frac{1}{N}\sum_{j\in U}\frac{\cos(ty_j)}{\pi_j} - A_N R_N(t)\right\}^2 w(t)\mathrm{d}t \to 0,$$
$$\int \left\{\frac{1}{N}\sum_{j\in U}\frac{\sin(ty_j)}{\pi_j} - A_N I_N(t)\right\}^2 w(t)dt \to 0$$

and

$$\int |C_N(t) - C(t)|^2 w(t)\mathrm{d}t \to 0,$$

respectively, for any nonnegative function w satisfying (1).

Assumption A.9 holds if, for each N, $y_1, \ldots, y_N$ are realizations of independent, identically distributed (iid) random variables with common cf $C(t)$.

Assumptions A.7 and A.8 say that $\{\cos(ty_j),\ j \in U\}$ (also $\{\sin(ty_j),\ j \in U\}$) and $\{\pi_j,\ j \in U\}$ (also $\{1/\pi_j,\ j \in U\}$) are asymptotically uncorrelated $\forall t$. This is true for the simple random sampling and for sampling designs with $\pi_j \propto x_j$, where $x_1, \ldots, x_N$ denote the values in the population units of a variable X, and there is no relationship among the y_js and the x_js. This requirement is also called for in Conti (2014).

Hájek (1964) showed the asymptotic normality of the HT estimator of the total of a population for the rejective sampling. For equal first order inclusion probabilities, the rejective sampling coincides with the simple random sampling, which can be easily carried out. For unequal first order inclusion probabilities, the rejective sampling is hard to implement in the sense that it is very time consuming. Because of this reason Berger (1998) investigated other designs for which such estimator keeps on being asymptotically normal. Assumption A.2 is required in both papers. Berger (1998, Theorem 5) showed that Assumption A.6 is necessary for the asymptotic normality, that is, the designs must be close to the rejective design in the sense stated by this assumption. Examples of designs satisfying Assumption A.6 are the Rao-Sampford sampling and the successive sampling (see Berger 1998).

3 Limit of the ecf process

Let us consider the ecf process $\{Z_n(t) = \sqrt{n}(\hat{C}_{N,\pi}(t) - C_N(t)),\ t \in \Upsilon\}$, for some $\Upsilon \subseteq \mathbb{R}$ (the choice of Υ will depend on the distance to be considered), that is a finite population version of the ecf process for iid data which, as observed before, is the basis of a number of inferential procedures in sfarv. To propose a finite population version of such procedures, this section is devoted to study this process. Observe that the realizations of Z_n are elements in $C[t_1, t_2]$, for any finite $t_1, t_2 \in \mathbb{R}$, with $t_1 < t_2$ $(\Upsilon = [t_1, t_2])$; they can be also seen as elements in $L_2(w)$, for any nonnegative w satisfying (1) $(\Upsilon = \mathbb{R})$. We will study the convergence of $\{Z_n(t),\ t \in \Upsilon\}$ in both spaces, with special emphasis on the second one, since most proposed procedures in the iid framework are based on $L_2(w)$ norms of appropriate functions of the ecf process.

For the process $\{Z_n(t),\ t \in [t_1, t_2]\}$ to converge in law in $C[t_1, t_2]$ to a process, say $\{Z(t),\ t \in [t_1, t_2]\}$, we must check the convergence in law of the finite-dimensional distributions (fidis) and the tightness of $\{Z_n(t),\ t \in [t_1, t_2]\}$. In the statistical literature on iid data, several authors have given conditions for the sequence of ecfs to be tight. For example, in Feuerverger and Mureika (1977), it is shown that $E|Y|^{1+\alpha} < \infty$, for some $\alpha > 0$, is a sufficient condition for the tightness; Marcus (1981) gave a necessary and sufficient condition in terms of the integrability of certain function involving the covariance function of $\{Z(t),\ t \in [t_1, t_2]\}$ (observe that such condition is also necessary in our setting because it is equivalent to the a.s. sample-continuity of $Z(t)$); Csörgő (1981) gave a sufficient condition in terms of the behaviour of the tails of the population df. Next we give a sufficient condition which is analogous to the one in Feuerverger and Mureika (1977). It assumes that the study variable has finite $(1+\alpha)$-order moment, for some $\alpha > 0$.

Let $\{Z(t) = \text{Re}Z(t) + \text{iIm}Z(t),\ t \in \Upsilon\}$ be a zero-mean complex valued Gaussian process with covariance structure

$$Cov\{\text{Re}Z(t),\ \text{Re}Z(s)\} = \frac{1}{2}f(A-1)\{R(t+s)+R(t-s)\} - f\frac{(1-f)^2}{d}R(t)R(s),$$
$$Cov\{\text{Re}Z(t),\ \text{Im}Z(s)\} = \frac{1}{2}f(A-1)\{I(t+s)+I(t-s)\} - f\frac{(1-f)^2}{d}R(t)I(s),$$

$$Cov\{\mathrm{Im}Z(t),\ \mathrm{Im}Z(s)\} = \frac{1}{2}f(A-1)\{-R(t+s)+R(t-s)\}-f\frac{(1-f)^2}{d}I(t)I(s).$$

Proposition 1 *Suppose that the sampling design P satisfies Assumptions A.1–A.9. Let $t_1, t_2 \in \mathbb{R}$ with $t_1 < t_2$. If*

$$\frac{1}{N}\sum_{j\in U}|y_j|^{1+\alpha} \le M, \quad \forall N, \tag{2}$$

for some $0 < \alpha \le 1$, then $\{Z_n(t), t \in [t_1, t_2]\}$ converges in law to $\{Z(t), t \in [t_1, t_2]\}$ in $C[t_1, t_2]$.

Observe that for each N, exists $M = M(N)$ so that $\frac{1}{N}\sum_{j\in U}|y_j|^{1+\alpha} \le M(N)$. Condition (2) requires that the upper bound M does not depend on N. This condition is clearly satisfied if, for each N, $y_1, \ldots, y_N$ are realizations of iid random variables with finite moment of order $1+\alpha$.

The next result shows that $\{Z_n(t), t \in \mathbb{R}\}$ also converges in law to $\{Z(t), t \in \mathbb{R}\}$ in $L_2(w)$, for any nonnegative function w satisfying (1). In contrast to the result in Proposition 1, no additional assumption such as (2) is required.

Proposition 2 *Suppose that the sampling design P satisfies Assumptions A.1–A.9. Let w be a nonnegative function satisfying (1), then $\{Z_n(t), t \in \mathbb{R}\}$ converges in law to $\{Z(t), t \in \mathbb{R}\}$ in $L_2(w)$.*

As argued in the Introduction, it may be preferable to work with the process $\{W_n(t) = \sqrt{n}(\hat{C}_{N,H}(t) - C_N(t)),\ t \in \Upsilon\}$. Note that

$$W_n(t) = \sqrt{n}\{\hat{C}_{N,H}(t) - C_N(t)\} = \frac{N}{\hat{N}_\pi}Y_n(t),$$

with $Y_n(t) = \sqrt{n}\left\{\hat{C}_{N,\pi}(t) - \frac{\hat{N}_\pi}{N}C_N(t)\right\}$. From the proof of Proposition 1 it follows that if the sampling design P satisfies Assumptions A.1–A.6, then $N/\hat{N}_\pi = 1+o_P(1)$, and thus

$$W_n(t) = (1+o_P(1))Y_n(t).$$

Therefore, if $\{Y_n(t),\ t \in \Upsilon\}$ converges in law to some random variable, then $\{W_n(t),\ t \in \Upsilon\}$ converges to the same limit. We next study the convergence in law of $\{Y_n(t),\ t \in \Upsilon\}$.

Let $\{Y(t) = \mathrm{Re}Y(t) + \mathrm{i}\mathrm{Im}Y(t),\ t \in \Upsilon\}$ be a zero-mean complex valued Gaussian process with covariance structure

$$\begin{aligned}
Cov\{\mathrm{Re}Y(t),\ \mathrm{Re}Y(s)\} &= f(A-1)\,\{0.5[R(t+s)+R(t-s)] - R(t)R(s)\}\,,\\
Cov\{\mathrm{Re}Y(t),\ \mathrm{Im}Y(s)\} &= f(A-1)\,\{0.5[I(t+s)+I(t-s)] - R(t)I(s)\}\,,\\
Cov\{\mathrm{Im}Y(t),\ \mathrm{Im}Y(s)\} &= f(A-1)\,\{0.5[-R(t+s)+R(t-s)] - I(t)I(s)\}\,.
\end{aligned}$$

Note that the covariance kernel of $\{Y(t),\ t \in \Upsilon\}$ is $f(A-1)$ times the covariance kernel of the limit of the ecf process in the iid case (see, for example, Feuerverger and Mureika 1977). For simple random sampling the factor $f(A-1)$ becomes $1-f$.

The results below state analogous results to those in Propositions 1 and 2 for $\{Y_n(t),\ t \in \Upsilon\}$.

Proposition 3 *Suppose that the assumptions in Proposition 1 hold, then* $\{Y_n(t),\ t \in [t_1, t_2]\}$ *converges in law to* $\{Y(t), t \in [t_1, t_2]\}$ *in* $C[t_1, t_2]$.

Proposition 4 *Suppose that the assumptions in Proposition 2 hold, then* $\{Y_n(t),\ t \in \mathbb{R}\}$ *converges in law to* $\{Y(t), t \in \mathbb{R}\}$ *in* $L_2(w)$.

Corollary 1 *Suppose that the assumptions in Proposition 1 hold, then* $\{W_n(t),\ t \in [t_1, t_2]\}$ *converges in law to* $\{Y(t), t \in [t_1, t_2]\}$ *in* $C[t_1, t_2]$.

Corollary 2 *Suppose that the assumptions in Proposition 2 hold, then* $\{W_n(t),\ t \in \mathbb{R}\}$ *converges in law to* $\{Y(t),\ t \in \mathbb{R}\}$ *in* $L_2(w)$.

Therefore, asymptotically, $\{W_n(t),\ t \in \Upsilon\}$ behaves just like $\{\sqrt{f(A-1)}Q_n(t),\ t \in \Upsilon\}$, where $\{Q_n(t),\ t \in \Upsilon\}$ is the ecf process associated to a random sample $Y_1, \ldots, Y_n$ from a population with cf $C(t)$, that is, $Q_n(t) = \sqrt{n}\{\hat{C}_n(t) - C(t)\}$, with $\hat{C}_n(t) = n^{-1}\sum_{j=1}^{n} \exp(\mathrm{i}tY_j)$. As a consequence, all inferential procedures designed for iid data in the sfarv framework based on the ecf process could be easily adapted for the current setting. Specifically, next sections give applications to some testing problems.

At this point one may wonder if the results in the above propositions keep on being true when $C_N(t)$ is replaced by $C(t)$, the answer is not. In fact, it only makes sense under a superpopulation model where $y_1, \ldots, y_N$ are realizations of iid random variables with common cf $C(t)$ and recall that the inferences in this paper are design-based.

Remark 1 As observed in the Introduction, although for simplicity in notation the study variable Y has been assumed to be real, all above results remain valid for d-variate Y, with the appropriate changes (in Propositions 1 and 3 and Corollary 1, $[t_1, t_2]$ is replaced by an arbitrary compact set in $\mathbb{R}^d$; in Propositions 2 and 4 and Corollary 2, w is proportional to a probability density function on $\mathbb{R}^d$).

4 Application 1: the two-sample problem

Consider two finite populations (or subpopulations or strata from a finite population), say U_1 and U_2, with sizes N_1 and N_2. Let $y_{1,1}, \ldots, y_{1,N_1}$ and $y_{2,1}, \ldots, y_{2,N_2}$ be the values of the study variable in populations U_1 and U_2, respectively. Let $C_{N_k}(t)$ and $F_{N_k}(t)$ denote the cf and df, respectively, of population k, $k = 1, 2$. Let s_k be a sample of size n_k from population U_k, selected according to the sampling design P_k, with first order inclusion probabilities π_j^k, $j \in U_k$, $k = 1, 2$. Let A_k and f_k denote the limits of A_{N_k} and n_k/N_k, respectively, $k = 1, 2$.

The problem of testing whether two samples come from the same population is a classical one in Statistics, which has generated a considerable amount of papers with many different approaches, in the context of sfarv. As observed in Conti (2014), in case of survey data, the literature mainly focuses on categorical variables, where Wald-type or Chi-square type statistics can be used (see, for instance, Särndal et al. 1992), requiring the data to be grouped in classes. If such classes are not natural, it is unclear how to construct them and (even more seriously) what is the effect of data-based classes on the distribution of the resulting test statistic. To overcome these difficulties for non-grouped data, Conti (2014) proposed a finite population version of the Kolmogorov–Smirnov test in the context of design-based inference, and Wang (2012) proposed edf-based tests for the equality of two superpopulation distributions.

In the context of sfarv, the test in Meintanis (2005) (see also Anderson et al. 1994; Henze et al. 2005; Alba-Fernández et al. 2008), which is based on comparing the ecf of the samples, competes very satisfactorily with the Kolmogorov–Smirnov two-sample test. Specifically, he proposed to reject the null hypothesis for large values of $\|\hat{C}_{n_1} - \hat{C}_{n_2}\|_w^2$, where $\hat{C}_{n_k}$ is the ecf of $Y_{k,1}, \ldots, Y_{k,n_k}$, which are iid from a random variable with cf $C_k(t)$, $k = 1, 2$, and n_1, n_2 are such that

$$\frac{n_1}{n_1 + n_2} \to \tau \in (0, 1). \tag{3}$$

Under the null hypothesis, that is, when $C_1(t) = C_2(t) = C(t)$, he showed that $\frac{n_1 n_2}{n_1+n_2}\|\hat{C}_{n_1} - \hat{C}_{n_2}\|_w^2$ converges in law to $\|D_{0T}\|_w^2$, where $\{D_{0T}(t),\ t \in \mathbb{R}\}$ is a zero-mean complex valued Gaussian process described in that paper.

Next, we study a finite population version of such test. The objective is testing

$$H_{0T} : F_{N_1}(t) = F_{N_2}(t), \quad \forall t \in \mathbb{R} \Longleftrightarrow C_{N_1}(t) = C_{N_2}(t), \quad \forall t \in \mathbb{R}.$$

With this aim, in view of the results in Sect. 3, we consider the following test statistic

$$D = \|\hat{C}_{N_1,H} - \hat{C}_{N_2,H}\|_w^2,$$

for some nonnegative function w satisfying (1), where $\hat{C}_{N_k,H}(t)$ stands for the Hájek estimator of $C_{N_k}(t)$, $k = 1, 2$. The test statistic D satisfies the following.

Proposition 5 *Suppose that P_1 and P_2 satisfy the assumptions in Proposition 2, then $D = \|C_{N_1} - C_{N_2}\|_w^2 + r_1 + r_2$, with $r_k = o_{P_k}(1)$, $k = 1, 2$.*

Therefore, when H_{0T} is true $D = r_1 + r_2$, and thus D converges in probability to 0, but when H_{0T} is false D converges to a positive quantity, provided that w is such that $\|C_{N_1} - C_{N_2}\|_w^2 > 0$ whenever $C_{N_1} \neq C_{N_2}$. Thus, a reasonable test should reject H_{0T} for large values of D. To decide what are large values of D we need to know the null distribution of D or an approximation of it. A way to approximate the null distribution of D is by means of its asymptotic null distribution, which is given in the next result. To derive it, we will assume that the two samples are independent. Observe also that from Assumption A.9, $C_{N_k}(t) \to C_k(t)$, $\forall t$, $C_k(t)$ being a cf, $k = 1, 2$. Therefore, under H_{0T}, $C_1(t) = C_2(t) = C(t)$ (say), $\forall t$.

Proposition 6 *Suppose that P_1 and P_2 satisfy the assumptions in Proposition 2, that the samples are independent, that H_{0T} is true and that n_1, n_2 satisfy (3), then*

$$\frac{n_1 n_2}{n_1+n_2} D \xrightarrow{\mathcal{L}} \kappa \| D_{0T} \|_w^2,$$

where

$$\kappa = (1-\tau) f_1(A_1 - 1) + \tau f_2(A_2 - 1). \tag{4}$$

The distribution of $\|D_{0T}\|_w^2$ is unknown because it depends on the unknown common cf $C(t)$ (see, for example, Meintanis 2005). In the context of sfarv, Alba-Fernández et al. (2008) showed that it can be consistently approximated by means of permutation or bootstrap procedures. For the bootstrap approximation, two independent random samples, $Y_{1,1}^*, \ldots, Y_{1,n_1}^*$ and $Y_{2,1}^*, \ldots, Y_{2,n_2}^*$ are generated from the pooled sample and the null distribution of the test statistic is approximated by means of the conditional distribution, given the data, of $\|C_{n_1}^* - C_{n_2}^*\|_w^2$, where $C_{n_k}^*$ denotes the ecf of $Y_{1,1}^*, \ldots, Y_{1,n_k}^*$, $k = 1, 2$. The permutation approximation is analogously obtained, with the samples obtained by randomly permuting the pooled sample. Therefore, to approximate the null distribution of D, we treat the elements in s_1 and s_2 as if they were two random samples from independent random variables. Specifically, to estimate the p-value, p, of the observed value of the test statistic D, D_{obs}, we proceed as follows: generate B bootstrap (or permutation) replications of $D_{RV} = \|\hat{C}_{n_1} - \hat{C}_{n_2}\|_w^2$ as explained before, say $D_{RV}^{*1}, \ldots, D_{RV}^{*B}$, and then approximate p by

$$\hat{p} = \frac{1}{B} \text{card}\{b : \ D_{obs} \le \hat{\kappa} D_{RV}^{*b}\},$$

with

$$\hat{\kappa} = \frac{n_2}{n_1+n_2} \frac{n_1}{N_1} (A_{N_1} - 1) + \frac{n_1}{n_1+n_2} \frac{n_2}{N_2} (A_{N_2} - 1), \tag{5}$$

for some large B (it is usually taken equal to 1000).

Remark 2 For the construction of the test statistic D, we have considered the Hájek estimators of the population cfs. If we instead consider the HT estimators then the result in Proposition 5 continue to be true. The reason for considering the Hájek estimators is that the asymptotic null distribution coincides with that obtained in the iid case that, although also unknown, can be consistently estimated.

Remark 3 An expression, useful from a computational point of view, of the test statistic D is

$$D = \frac{1}{\hat{N}_{1,\pi}^2} \sum_{j,k \in s_1} \frac{u(y_{1,j} - y_{1,k})}{\pi_{1,j}\pi_{1,k}} + \frac{1}{\hat{N}_{2,\pi}^2} \sum_{m,v \in s_2} \frac{u(y_{2,m} - y_{2,v})}{\pi_{2,m}\pi_{2,v}}$$
$$- \frac{2}{\hat{N}_{1,\pi}\hat{N}_{2,\pi}} \sum_{j \in s_1,\, m \in s_2} \frac{u(y_{1,j} - y_{2,m})}{\pi_{1,j}\pi_{2,m}},$$

where $u(t) = \int \cos(xt) w(x) \mathrm{d}x$.

Remark 4 Since two distinct characteristic functions can be equal in a finite interval (Feller 1971, p. 479), a general way to ensure $\|C_{N_1} - C_{N_2}\|_w^2 > 0$ whenever $C_{N_1} \neq C_{N_2}$, is by taking $w(t) > 0, \forall t \in \mathbb{R}$.

Remark 5 The test can be extended to test for the equality of $k \geq 2$ populations following the procedure proposed in Hušková and Meintanis (2008) in the setting of sfarv.

Remark 6 As pointed out by an anonymous reviewer, the test statistic D is not scale invariant. Nevertheless, D can be done scale invariant by choosing the weight function depending on the common scale parameter, say $w(t) = w(t; \sigma)$, in a similar fashion to Lemma 2 in Jiménez-Gamero et al. (2009) for the goodness-of-fit problem, and considering $\hat{D} = \int |\hat{C}_{N_1,H}(t) - \hat{C}_{N_2,H}(t)|^2 w(t; \hat{\sigma})\mathrm{d}t$, for some adequate estimator $\hat{\sigma}$ of σ. Moreover, it is easy to check that if $\hat{\sigma}$ converges in probability to σ and $w(t; \sigma)$ is such that $|w(t; \sigma) - w(t; \gamma)| \leq w_0(t; \sigma)|\sigma - \gamma|, \forall \gamma$ in an open neighborhood of σ, with $\int w_0(t; \sigma)\mathrm{d}t < \infty$, then the asymptotics are not altered, they are the same as if the weight function $w(t; \sigma)$ is used.

To numerically study the performance of the proposed approximation to the null distribution, as well as the power of the test, we conducted a simulation study. With this aim, we generated a finite population $\{(y_j, x_j), 1 \leq j \leq N\}$ of size $N = 10{,}000$ as follows: $y_1, \ldots, y_N$ are iid from a law $N(0, 1)$ and $x_1, \ldots, x_N$ are iid from a law $U(0, 1)$; $y_1, \ldots, y_N$ are the values of the study variable, and $x_1, \ldots, x_N$ will be used to define the first order inclusion probabilities. To study the level of the proposed test, we generated two samples of the population with sizes $n_1 = n_2 = 200, 250$. Two designs were considered to draw the samples: simple random sampling without replacement (srs in the tables) and the Rao-Sampford sampling (sam in the tables) with inclusion probabilities proportional to $z = (1 - a)x + a$ for several values of a (note that $z_1, \ldots, z_N$ are iid from a law $U(a, 1)$). Observe that in the simple random sampling all first order inclusion probabilities are equal and that by moving the value of a in the Rao-Sampford sampling we can control how different are the first order inclusion probabilities. The column headed coc in Table 1 displays the quotient $\max_j \pi_j / \min_j \pi_j (\approx 1/a)$ for the Rao-Sampford sampling which, in a sense, measures how far is this sampling from the simple random sampling. Two weight functions were considered: w_1, the probability density function (pdf) of a Laplace distribution with variance 2, giving rise to $u(t) = 1/(1 + t^2)$, and w_2, the pdf of a standard normal distribution, giving rise to $u(t) = \exp(-0.5t^2)$. The associated statistics are denoted as D_1 and D_2, respectively. 2000 pairs of samples were generated from the population in each case. For each pair of samples, $B = 1000$ bootstrap samples were generated from the pooled sample to estimate the p value. We also included in our simulation study the Kolmogorov–Smirnov test proposed in Conti (2014) (denoted as KS in the tables). Its p-value was also approximated with bootstrap, multiplying the bootstrap replications of the test statistic by $\sqrt{\hat{\kappa}}$. As suggested by an anonymous referee, we also calculated the p-values by wrongly assuming that the samples come from sfarv, that is, by ignoring the factor $\hat{\kappa}$. Obviously, such effect will depend on the value of $\hat{\kappa}$, which is showed in Table 1 joint with the obtained results. In all cases two nominal levels were considered: 5 and 10 %.

Looking at Table 1, it can be concluded the following: as expected, the effect of ignoring the factor $\hat{\kappa}$ depends on the value of such factor, the further is $\hat{\kappa}$ from 1, the further from the nominal levels are the empirical percentages of rejection; the approximation is reasonably good for all test statistics, specially when $n_1 = n_2 = 250$, when the quotient $\max_j \pi_j / \min_j \pi_j$ is not too big, for rather large values of $\max_j \pi_j / \min_j \pi_j$ greater samples seems to be required.

To study the power, the above experiment was repeated with one of the samples from Y and the other from some deformation of Y, say Z. Table 2 displays the results for $Z_1 = 0.5Y$ if $Y < -1.7$, $Z_1 = 1.5Y$ if $Y > 1.7$, otherwise $Z_1 = Y$ and $Z_2 = Y + 0.5$. In the light of the results in Table 1, for the Rao-Sampford sampling, we only considered $a = 0.1, 0.5$. Looking at Table 2, we see that all tests are able to detect the studied alternatives, D_1 and D_2 behave quite closely, no test outperforms the others in all cases: in case (b) KS is a bit more powerful than D_1 and D_2; in case (a) D_1 and D_2 beat KS. The power increases with the sample size. It is also observed that as $\max_j \pi_j / \min_j \pi_j$ becomes larger, the power decreases.

Remark 7 To keep the notation as simple as possible, we have only considered the case of a univariate study variable. Nevertheless, the proposed test can be applied to testing the equality of distributions of d-variate variables, for arbitrary $d \geq 1$. Although the Kolmogorov–Smirnov test can be also applied to data with arbitrary dimension, its practical implementation is computationally difficult (see, for example, Xiao 2017 and the references therein).

5 Application 2: independence

Now, we deal with the problem of constructing a test for the independence of two study variables, say X and Y, defined on the same finite population. This hypothesis is of particular interest because auxiliary information is often used to improve the precision of estimators. With this aim, it is assumed that the study variable and the auxiliary information are related, which makes sense if the null hypothesis of independence is rejected.

Consider a finite population, U, with size N. Let $(x_1, y_1), \ldots, (x_N, y_N)$ be the values of the study variable (X, Y) in the population. Let $C_N(t_1, t_2)$ denote the joint cf associated to (X, Y),

$$C_N(t_1, t_2) = \frac{1}{N} \sum_{j \in U} \exp\{\mathrm{i}(t_1 x_j + t_2 y_j)\}.$$

The marginal cfs of X and Y are $C_N(t_1, 0)$ and $C_N(0, t_2)$, respectively. The hypothesis of absence of relationships between X and Y can be written as

$$H_{0I} : C_N(t_1, t_2) = C_N(t_1, 0) C_N(0, t_2), \quad \forall t_1, t_2 \in \mathbb{R}.$$

The problem of testing for independence in the context of sfarv using the familiar equation linking the joint cf and the product of component cfs has been exploited in

Table 1 Empirical percentages of rejection for testing homogeneity when $s_1,\ s_2$ come from Y

		a	coc	$n_1 = n_2 = 200$							$n_1 = n_2 = 250$						
				$\hat{\kappa}$	D_1		D_2		KS		$\hat{\kappa}$	D_1		D_2		KS	
					5%	10%	5%	10%	5%	10%		5%	10%	5%	10%	5%	10%
srs	srs			0.980	5.8	11.6	4.8	12.0	5.8	11.9	0.975	5.3	10.6	4.9	10.6	5.8	9.9
					5.5	10.7	4.7	11.4	5.5	11.3		5.1	10.0	4.7	10.1	5.6	9.5
srs	sam	0.5	1.999	0.999	4.4	11.3	4.4	11.3	6.1	11.4	0.995	4.0	8.6	3.8	9.2	5.5	10.4
					4.4	11.3	4.4	11.3	6.1	11.4		4.0	8.5	3.7	9.1	5.4	10.4
sam	sam			1.020	5.4	10.5	5.3	10.0	5.2	11.9	1.015	5.1	10.0	4.8	10.1	5.6	9.5
					5.5	11.0	5.7	10.1	5.5	12.4		5.1	10.1	5.2	10.7	5.8	9.9
srs	sam	0.2	4.998	1.083	5.6	10.9	5.6	10.2	5.8	10.7	1.078	5.2	10.2	5.0	10.4	5.7	11.0
					6.8	12.7	7.0	12.5	7.0	13.3		6.1	12.5	6.6	12.6	7.3	12.9
sam	sam			1.187	5.1	9.6	5.6	9.9	5.3	11.6	1.182	5.6	10.9	5.6	10.2	5.8	10.7
					8.7	15.8	7.9	14.8	9.7	16.4		6.8	12.7	7.0	12.5	7.0	13.3
srs	sam	0.1	9.995	1.183	5.4	9.8	5.4	10.6	5.3	11.0	1.178	5.5	9.5	5.2	9.7	5.8	11.9
					8.4	16.0	8.0	14.7	9.7	16.7		7.9	14.5	7.8	13.9	9.7	16.4
sam	sam			1.386	5.3	9.5	5.3	9.4	4.9	10.1	1.380	4.5	10.9	4.5	9.8	4.6	9.9
					11.4	20.6	10.7	19.2	13.1	22.0		12.5	19.3	11.2	19.1	12.7	21.3
srs	sam	0.01	99.595	1.653	6.0	10.5	5.4	10.4	6.1	11.1	1.648	5.9	9.1	6.1	9.1	5.5	8.7
					20.3	28.2	18.0	26.5	23.0	29.6		15.3	25.5	14.4	23.0	17.9	26.0
sam	sam			2.327	4.8	9.3	4.6	9.1	4.0	8.7	2.322	5.6	10.7	4.9	10.6	5.4	9.3
					29.7	40.7	27.0	36.7	33.6	45.4		31.8	43.0	26.8	38.0	35.1	47.3
srs	sam	0.001	962.343	2.185	3.8	7.0	3.6	6.1	3.8	5.4	2.180	4.1	6.4	4.0	6.7	3.5	6.4
					19.4	26.7	17.9	23.8	20.7	28.9		18.2	27.9	16.7	24.5	21.2	31.8

Table 1 continued

		a	coc	$n_1 = n_2 = 200$							$n_1 = n_2 = 250$						
				$\hat{\kappa}$	D_1		D_2		KS		$\hat{\kappa}$	D_1		D_2		KS	
					5%	10%	5%	10%	5%	10%		5%	10%	5%	10%	5%	10%
sam	sam			3.390	3.2	4.9	3.2	4.7	2.5	3.5	3.385	2.4	4.6	2.7	4.8	2.2	4.1
					32.5	42.6	28.1	37.3	36.9	47.6		32.2	43.8	27.4	37.8	37.1	47.8
srs	sam	0.0	25667.76	3.172	1.9	2.8	1.9	2.9	1.1	2.1	2.815	1.9	3.3	2.1	3.7	1.5	2.9
					18.5	27.1	17.1	24.5	20.6	29.2		20.6	28.0	18.6	26.7	21.8	30.2
sam	sam			5.363	1.3	2.0	0.8	1.9	0.8	1.1	5.358	0.7	1.7	0.6	1.9	0.6	1.2
					30.8	42.0	27.1	37.2	34.2	47.4		32.2	43.7	27.8	37.8	35.9	48.2

For each design, the second line is calculated by ignoring the factor $\hat{\kappa}$

Table 2 Empirical percentages of rejections for testing homogeneity in cases: (a) s_1 comes from Y, s_2 comes from Z_1; (b) s_1 comes from Y, s_2 comes from Z_2

		a	$n_1 = n_2 = 200$						$n_1 = n_2 = 250$					
			D_1		D_2		KS		D_1		D_2		KS	
			5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
(a)														
srs	srs		13.4	24.0	12.7	22.6	4.7	10.9	16.7	29.6	16.2	26.5	6.8	12.2
srs	sam	0.5	13.3	24.2	11.7	22.8	5.4	13.0	15.5	29.3	14.3	25.9	5.2	10.7
sam	sam		13.0	24.2	12.5	23.0	5.2	11.4	14.7	29.9	14.6	27.3	5.8	11.5
srs	sam	0.1	11.6	22.0	10.5	19.4	4.0	10.7	13.7	25.3	12.5	24.0	4.9	11.3
sam	sam		11.0	19.9	11.2	18.8	4.7	9.1	13.4	24.8	13.0	23.1	5.4	12.6
(b)														
srs	srs		13.4	20.8	13.8	21.3	16.6	24.6	16.8	25.4	18.1	26.2	18.9	28.9
srs	sam	0.5	12.4	20.1	13.4	21.2	14.2	23.7	14.1	22.2	14.8	23.0	16.2	25.1
sam	sam		11.7	21.0	13.7	22.0	13.8	23.3	13.2	21.1	14.4	22.7	15.6	24.1
srs	sam	0.1	11.4	18.9	12.3	19.5	13.1	19.9	12.9	22.4	13.8	23.3	15.2	25.6
sam	sam		11.5	17.8	11.1	18.2	11.8	18.0	11.9	20.2	12.3	21.4	13.0	22.1

several papers (see, for example, Csörgő 1985; Székely et al. 2007; Meintanis and Iliopoulos 2008 and the references therein). Here, we follow the approach in Székely et al. (2007) and Meintanis and Iliopoulos (2008), that proposed to reject the null hypothesis for large values of $\|\hat{C}_n(t_1,t_2)-\hat{C}_n(t_1,0)\hat{C}_n(0,t_2)\|_w^2$, where, $\hat{C}_n$ is the ecf of $(X_1,Y_1),\ldots,(X_n,Y_n)$, which are iid from a random variable with cf $C(t_1,t_2)$. The main difference between the methods in these papers is that while the weight function in the second paper satisfies (1), the one in the first paper does not. The payment one must do for considering a non-integrable weight function is that the application of the resulting test requires the population to have finite first order moment.

Let us first assume that w is any nonnegative function so that $\int_{\mathbb{R}^2} w(t)\mathrm{d}t < \infty$. Under the null hypothesis, from the results in Csörgő (1985) it follows that $n\|\hat{C}_n(t_1,t_2)-\hat{C}_n(t_1,0)\hat{C}_n(0,t_2)\|_w^2$ converges in law to $\|D_{0I}\|_w^2$, where $\{D_{0I}(t), t\in\mathbb{R}^2\}$ is a zero-mean complex valued Gaussian process described in that paper. Next, we study a finite population version of such test statistic, specifically,

$$T=\|\hat{C}_{N,H}(t_1,t_2)-\hat{C}_{N,H}(t_1,0)\hat{C}_{N,H}(0,t_2)\|_w^2.$$

Proposition 7 *Suppose that P satisfies the assumptions in Proposition 2, then* $T=\|C_N(t_1,t_2)-C_N(t_1,0)C_N(0,t_2)\|_w^2+o_P(1)$.

Therefore, reasoning as in the previous subsection, a reasonable test should reject H_{0I} for large values of T. To try to approximate the null distribution of T, the next result derives its asymptotic null distribution. Observe also that from Assumption A.9, $C_N(t_1,t_2)\to C(t_1,t_2)$, $\forall(t_1,t_2)\in\mathbb{R}^2$, C being a cf on $\mathbb{R}^2$. Therefore, under H_{0I} we have that $C(t_1,t_2)=C(t_1,0)C(0,t_2)$, $\forall(t_1,t_2)\in\mathbb{R}^2$.

Proposition 8 *Suppose that P satisfies the assumptions in Proposition 2 and that H_{0I} is true, then* $nT\xrightarrow{\mathcal{L}} f(A-1)\|D_{0I}\|_w^2$.

The distribution of $\|D_{0I}\|_w^2$ is unknown because it depends on the unknown common cf $C(t_1,t_2)$ (see Csörgő 1985). In the context of sfarv, to apply their test, Meintanis and Iliopoulos (2008) assume that the underlying distribution of the data is continuous in order to derive an asymptotically distribution free test statistic (see Kankainen and Ushakov 1998 for a theoretical justification). In our setting, the continuity assumption may not be adequate. Nevertheless, it can be easily shown that the following bootstrap estimator provides a consistent approximation to the null distribution of their test statistic (for any sort of data, continuous or not): given $(X_1,Y_1),\ldots,(X_n,Y_n)$, iid from a random variable with cf $C(t_1,t_2)$, let $X_1^*,\ldots,X_n^*$ and $Y_1^*,\ldots,Y_n^*$ be two independent random samples from the edf of $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_n$, respectively; let $\hat{C}_n^*(t_1,t_2)$ be the ecf of $(X_1^*,Y_1^*),\ldots,(X_n^*,Y_n^*)$; then estimate the null distribution of $n\|\hat{C}_n(t_1,t_2)-\hat{C}_n(t_1,0)\hat{C}_n(0,t_2)\|_w^2$ by means of the conditional distribution, given the data, of $n\|\hat{C}_n^*(t_1,t_2)-\hat{C}_n^*(t_1,0)\hat{C}_n^*(0,t_2)\|_w^2$. Now, proceeding as in the previous subsection, to approximate the null distribution of T we treat the elements in s as if they were a random sample from a bivariate random variable and estimate the p-value, p, of the observed value of the test statistic T, T_{obs}, as follows: calculate B bootstrap replications of $T_{RV}=\|\hat{C}_n(t_1,t_2)-\hat{C}_n(t_1,0)\hat{C}_n(0,t_2)\|_w^2$, say $T_{RV}^{*1},\ldots,T_{RV}^{*B}$, and then approximate p by

$$\hat{p}=\frac{1}{B}\text{card}\left\{b:\, T_{\text{obs}}\leq \frac{n}{N}(A_N-1)T_{RV}^{*b}\right\},$$

for some large B.

Remark 8 An expression, useful from a computational point of view, of the test statistic T is

$$T=\frac{1}{\hat{N}_\pi^2}\sum_{j,k\in s}\frac{u(x_j-x_k,y_j-y_k)}{\pi_j\pi_k}+\frac{1}{\hat{N}_\pi^4}\sum_{j,k,v,m\in s}\frac{u(x_j-x_k,y_v-y_m)}{\pi_j\pi_k\pi_v\pi_m}$$
$$-2\frac{1}{\hat{N}_\pi^3}\sum_{j,k,v\in s}\frac{u(x_j-x_v,y_k-y_v)}{\pi_j\pi_k\pi_v},$$

where $u(t_1,t_2)=\int_{\mathbb{R}^2}\cos(t_1x+t_2y)w(x,y)\mathrm{d}x\mathrm{d}y$.

Next, we deal with a finite population version of the test in Székely et al. (2007) that considered as weight function $w=m$ with

$$m(t_1,t_2)=m(t_1)m(t_2),\quad m(t)=\frac{1}{\pi t^2},\quad t\in\mathbb{R},$$

which clearly is non-integrable (note that we have used the same letter m for the univariate and the bivariate case with the aim of not using heavier notation—such as m_2 and m_1; the dimension will become evident from the arguments employed). However, if (X,Y) satisfies $E|X|<\infty$, $E|Y|<\infty$, then the associated test statistic satisfies similar properties to those given for integrable w. Specifically, $nV_n(X,Y)$, with $V_n(X,Y)=\|\hat{C}_n(t_1,t_2)-\hat{C}_n(t_1,0)\hat{C}_n(0,t_2)\|_m^2$, converges in law to $\|D_{0I}\|_m^2$, where $\{D_{0I}(t),\ t\in\mathbb{R}^2\}$ is as before. Let T_m denote the finite population version of such test statistic, defined as T with $w=m$. The properties of T_m cannot be directly derived from the theory in Sect. 3, since (1) was assumed in the results there. To derive similar properties for T_m as those given in Propositions 7 and 8 for T, we will have to assume stronger conditions.

Proposition 9 *Suppose that P satisfies the assumptions in Proposition 2 and that*

$$\frac{1}{N}\sum_{j\in U}|x_j|^{1+\alpha}\leq M,\quad \frac{1}{N}\sum_{j\in U}|y_j|^{1+\alpha}\leq M,\quad \forall N,\tag{6}$$

for some $\alpha>0$. Then,

(a) $T_m=\|C_N(t_1,t_2)-C_N(t_1,0)C_N(0,t_2)\|_m^2+o_P(1)$.

(b) *If in addition H_{0I} is true, then $nT_m\xrightarrow{\mathcal{L}}f(A-1)\|D_{0I}\|_m^2$.*

To be precise, the test statistic proposed in Székely et al. (2007) is not exactly $V_n(X, Y)$, but $R_n(X, Y) = V_n(X, Y)/\sqrt{V_n(X, X)V_n(Y, Y)}$ which, in a sense, imitates the definition of the usual linear correlation coefficient. $R_n(X, Y)$ is called the empirical distance correlation. A similar correlation version can be considered when a general weight function w is used. Analogous results to those stated in Propositions 7–9 can be given for the finite population version of such correlation type statistics. In addition, parallel comments to those given after Propositions 7 and 8 for T can be given for T_m as well as for their correlation versions.

Remark 9 An expression, useful from a computational point of view, of the test statistic T_m is that given in Remark 8 for T with $u(t_1, t_2) = |t_1||t_2|$.

Remark 10 To derive the results in Proposition 9, it was assumed that (6) holds for some $\alpha > 0$. If the finite population can be considered as a random sample from a random vector (X, Y), then it suffices to assume that $E|X| < \infty$, $E|Y| < \infty$.

To numerically study the performance of the proposed approximation to the null distribution, as well as the power of the tests, we conducted a simulation study. With this aim, we generated a finite population $\{(y_j, w_j, x_j), 1 \le j \le N\}$ of size $N = 10,000$ as follows: $y_1, \ldots, y_N, w_1, \ldots, w_N$ are iid from a law $N(0, 1)$ and $x_1, \ldots, x_N$ are iid from a law $U(0, 1)$; $(y_1, w_1 + ry_1), \ldots, (y_N, w_N + ry_N)$ are the values of the study variable, for $r = 0$ (null hypothesis), 0.1, 0.2, 0.3 (alternatives) and $x_1, \ldots, x_N$ will be used to define the first order inclusion probabilities. The Pearson correlation coefficient between of the study variables is $\rho = r/\sqrt{1 + r^2}$ (taking values 0, 0.0995, 0.1961 and 0.2873, respectively). To study the level of the proposed tests, we generated a sample of the population for $\rho = 0$ with size $n = 200$. As in the previous subsection, two designs were considered to draw the sample: simple random sampling without replacement (srs) and the Rao-Sampford sampling (sam) with inclusion probabilities proportional to $(1-a)x+a$, $a = 0.1, 0.5$. The weight functions considered are products of the same ones used in the previous section: $w_1(t, s) = w_1(t)w_1(s)$ and $w_2(t, s) = w_2(t)w_2(s)$. The associated statistics are denoted as T_1 and T_2, respectively. We also calculated T_m as well as their correlations versions, denoted as R_1, R_2 and R_m, respectively. 2000 samples were generated from the population in each case. For each sample, $B = 1000$ bootstrap samples were generated as described above to estimate the p-value. The whole experiment was repeated for $n = 250$ and $r = 0.1, 0.2, 0.3$. Table 3 reports the percentage of rejections in all experimental situations. Looking at this table, we see that, in terms of level ($r = 0$), the approximation is reasonably good for all test statistics. As for the power ($r = 0.1, 0.2, 0.3$), all tests are able to detect the studied alternatives. The tests based on the correlated versions of the test statistics have, in all cases, powers quite close to the ones based on the original statistics. The test based on T_2 is a bit more powerful than the one based on T_1; nevertheless, the one based on T_m has the highest power against all considered alternatives. The power increases with the sample size. It is again observed that as $\max_j \pi_j / \min_j \pi_j$ becomes larger, the power decreases.

Remark 11 To keep the notation as simple as possible, we have only considered the case of a bivariate study variable. Nevertheless, the proposed tests can be applied

Table 3 Empirical percentages of rejections for testing independence

r		a	T_1		T_2		T_m		R_1		R_2		R_m	
			5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
$n = 200$														
0	srs		5.52	11.09	5.31	10.95	5.44	10.82	5.46	11.24	5.29	10.81	5.40	10.89
	sam	0.5	5.54	10.45	5.39	10.33	5.33	10.81	5.47	10.34	5.25	10.31	5.43	11.16
	sam	0.1	5.70	10.95	5.55	11.10	6.00	11.20	5.75	11.20	5.80	11.15	5.55	10.45
0.1	srs		19.50	29.55	22.10	31.95	31.30	41.60	19.10	29.25	21.90	31.65	31.45	42.45
	sam	0.5	18.70	28.15	21.20	31.35	30.50	42.85	18.30	28.20	20.80	30.85	31.10	42.35
	sam	0.1	14.40	22.65	16.85	24.45	23.50	33.55	14.75	22.45	17.15	24.25	23.60	33.75
0.2	srs		52.70	65.55	58.90	70.95	76.30	84.55	51.60	64.45	58.00	69.60	78.05	86.40
	sam	0.5	52.60	64.35	58.20	70.35	76.25	83.35	51.40	63.10	57.40	70.05	77.45	84.10
	sam	0.1	39.25	50.95	45.25	56.55	60.70	71.35	38.85	49.35	43.95	55.20	62.25	72.45
0.3	srs		87.10	92.30	90.75	94.75	97.10	98.55	85.75	91.40	89.90	93.90	97.75	98.85
	sam	0.5	86.35	91.70	90.10	94.05	96.80	98.10	85.00	91.30	89.15	93.30	97.45	98.55
	sam	0.1	71.60	81.00	77.25	86.00	90.65	95.20	69.85	79.35	75.95	84.75	92.35	95.85
$n = 250$														
0	srs		5.62	10.95	5.20	10.90	5.47	10.66	5.55	10.75	5.13	10.75	5.59	10.87
	sam	0.5	5.69	11.02	5.71	11.22	5.58	11.11	5.77	11.09	5.69	11.05	5.70	11.00
	sam	0.1	5.60	10.55	5.40	10.95	5.60	10.30	5.70	11.10	5.40	10.65	5.05	10.35
0.1	srs		21.40	31.65	24.20	35.70	37.10	49.20	21.65	31.70	24.60	35.20	37.35	49.25
	sam	0.5	20.55	31.50	24.00	34.40	34.80	46.25	20.35	31.60	23.60	34.45	34.85	46.75
	sam	0.1	18.05	25.70	19.70	28.40	28.35	38.40	18.05	25.55	19.35	28.65	28.35	39.00

Table 3 continued

r		a	T_1		T_2		T_m		R_1		R_2		R_m	
			5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
0.2	srs		63.20	74.10	68.75	78.20	85.15	90.75	62.40	73.45	67.75	77.40	85.70	92.05
	sam	0.5	60.45	71.05	66.30	76.30	83.60	89.95	59.70	70.40	65.35	75.45	85.10	91.20
	sam	0.1	47.50	59.25	53.05	65.15	70.95	80.00	46.50	57.35	51.95	64.30	72.80	81.80
0.3	srs		94.40	97.05	96.10	98.10	99.40	99.75	93.60	96.80	95.65	97.75	99.55	99.70
	sam	0.5	92.65	96.50	95.70	98.10	99.20	99.35	91.35	95.60	94.45	97.80	99.20	99.50
	sam	0.1	80.15	87.75	85.65	91.30	94.35	97.00	78.05	86.65	84.25	90.15	95.50	97.50

to test the independence of any collection of subvectors from vectors with arbitrary dimensions.

6 Application 3: symmetry

Another hypothesis which may be of interest in sample survey is that of symmetry around a point. In his book, Kish (1965, pp 410–411) underlines that sampling from highly skewed population strains the assumptions about the normality of the distributions of estimates and affects the coverage of the confidence intervals. More recently, Conti and Marella (2015) assert that in last years there is an increasing demand from official and private institutions of statistical data regarding poverty and inequality. Poverty and inequality measures are functions of quantile estimates. Clearly, the estimation of quantiles can be simplified for symmetric populations; specifically, the poverty index depends on the median, which could be estimated by the mean for symmetric populations. This subsection is devoted to the problem of testing such hypothesis based on the results in Sect. 3.

Consider a finite population, U, with size N. Let $y_1, \ldots, y_N$ be the values of the study variable Y in the population. The objective is to construct tests for the hypothesis that $Y - \mu_N$ and $\mu_N - Y$ both have the same distribution function, where μ_N is a constant that maybe known or unknown. Let $X = Y - \mu_N$ and let $C_N(t) = R_N(t) + \mathrm{i}I_N(t)$ denote de cf of X. The hypothesis of symmetry can be written as follows

$$H_{0S}: \; I_N(t) = 0, \quad \forall t.$$

In the context of sfarv, several authors have suggested tests for the hypothesis of symmetry about a possibly unknown value μ, whose test statistics are functions of the ecf process. Specifically, here we consider the tests in Feuerverger and Mureika (1977) (for known μ), Neuhaus and Zhu (1998) (μ known or unknown) and Henze et al. (2003) (μ unknown). These authors proposed to reject the null hypothesis for large values of $\|\hat{I}_n\|_w^2$, where $\hat{I}_n(t)$ is the imaginary part of the ecf of $Y_1 - \mu, \ldots, Y_n - \mu$, if μ is known and with μ replaced by a consistent estimator, say $\hat{\mu}$, when unknown, where $Y_1, \ldots, Y_n$ are iid. The limit law of this test statistic depends on whether μ is known or not. Let $X = Y - \mu$ and let $C(t) = R(t) + \mathrm{i}I(t)$ denote the cf of X. If μ is known and the null hypothesis is true, then $n\|\hat{I}_n\|_w^2$ converges in law to $\|S_{01}\|_w^2$, where $\{S_{01}(t), t \in \mathbb{R}\}$ is a zero-mean Gaussian process with covariance kernel $K_1(t, s) = 0.5\{R(t-s) - R(t+s)\}$; if μ is unknown and it is estimated by means of the sample mean of the observed data, $\hat{\mu} = n^{-1}\sum_{j=1}^n Y_j$, then under the null hypothesis $n\|\hat{I}_n\|_w^2$ converges in law to $\|S_{02}\|_w^2$, where $\{S_{02}(t), t \in \mathbb{R}\}$ is a zero-mean Gaussian process with covariance kernel $K_2(t, s) = 0.5\{R(t-s) - R(t+s)\} + tR(t)R'(s) + sR(s)R'(t) + stR(t)R(s)$, and $R'(t) = \frac{\partial}{\partial t}R(t)$.

Next, we study finite population versions of these tests. Let us first assume that μ_N is known and consider the test statistic

$$S_1 = \|\hat{I}_{N,H}\|_w^2,$$

for some nonnegative function w satisfying (1). The test statistic S_1 satisfies the following.

Proposition 10 *Suppose that P satisfies the assumptions in Proposition 2, then $S_1 = \|I_N\|_w^2 + o_P(1)$.*

Therefore, reasoning as in the previous subsections, a reasonable test should reject H_{0S} for large values of S_1. To try to approximate the null distribution of S_1, the next result gives its asymptotic null distribution. Observe also that from Assumption A.9, $C_N(t) \to C(t)$, $\forall t$, $C(t) = R(t) + \mathrm{i}I(t)$ being a cf. Therefore, under H_{0S}, $I(t) = 0$, $\forall t$.

Proposition 11 *Suppose that P satisfies the assumptions in Proposition 2 and that H_{0S} is true, then $nS_1 \xrightarrow{\mathcal{L}} f(A-1)\|S_{01}\|_w^2$.*

The distribution of $\|S_{01}\|_w^2$ is unknown, because it depends on the unknown real part of $C(t)$. In the context of sfarv, Neuhaus and Zhu (1998) have proposed to approximate the null distribution of $n\|\hat{I}_n\|_w^2$ by means of the conditional distribution, given $Y_1, \ldots, Y_n$, of $n\|\hat{I}_n^*\|_w^2$, where $\hat{I}_n^*(t)$ is the imaginary part of the ecf of $e_1(Y_1-\mu), \ldots, e_n(Y_n-\mu)$, with $e_1, \ldots, e_n$ iid, independent of the data and such that $P(e_1=-1) = P(e_1=1) = 0.5$. They showed that this approximation provides a consistent estimator of the null distribution of the test statistic. Now, proceeding as in the previous subsections, to approximate the null distribution of S_1, we treat the elements in s as if they were a random sample from a random variable and estimate the p-value, p, of the observed value of the test statistic S_1, $S_{1,\mathrm{obs}}$, as follows: generate B replications of $n\|\hat{I}_n^*\|_w^2$, say $S_1^{*1}, \ldots, S_1^{*B}$, and then approximate p by

$$\hat{p} = \frac{1}{B}\mathrm{card}\left\{b : \, S_{1,\mathrm{obs}} \le \frac{n}{N}(A_N - 1)S_1^{*b}\right\},$$

for some large B.

Remark 12 An expression, useful from a computational point of view, of the test statistic S_1 is

$$S_1 = \frac{0.5}{\hat{N}_\pi^2} \sum_{j,k\in s} \frac{u(x_j - x_k) - u(x_j + x_k)}{\pi_j \pi_k},$$

where $u(t) = \int \cos(tx) w(x) \mathrm{d}x$.

Now, assume that μ is unknown and that it is estimated by means of the HT estimator of the population mean $\bar{y}_N = \frac{1}{N}\sum_{j\in U} y_j$, $\hat{\bar{y}}_{N,\pi} = \frac{1}{N}\sum_{j\in s}\frac{y_j}{\pi_j}$. From now on $x_j = y_j - \bar{y}_N$, $j \in U$. Let us consider the test statistic S_2 defined as S_1 with μ replaced by $\hat{\bar{y}}_{N,\pi}$. The next result states that S_2 satisfies a result similar to that in Proposition 10 for S_1.

Proposition 12 *Suppose that P satisfies the assumptions in Proposition 2, that $S_Y^2 = \frac{1}{N}\sum_{j\in U}(y_j - \bar{y}_N)^2 \to \sigma_Y^2 < \infty$ and that $\int t^2 w(t)dt < \infty$, then $S_2 = \|I_N\|_w^2 + o_P(1)$.*

Reasoning as before, if μ_N is unknown, a sensible test should reject H_{0S} for large values of S_2. To try to approximate the null distribution of S_2, the next proposition gives its asymptotic null distribution. With this purpose, we need a further assumption.

Assumption A.10 $\frac{1}{N}\sum_{j\in U}\pi_j x_j \to 0$, $\frac{1}{N}\sum_{j\in U}\frac{x_j^2}{\pi_j} - A_N\frac{1}{N}\sum_{j\in U}x_j^2 \to 0$, $\frac{1}{N}\sum_{j\in U}\frac{x_j\sin(tx_j)}{\pi_j} - A_N R_N'(t) \to 0$, $\forall t$, $R_N'(t) \to R'(t)$, $\forall t$, where $R_N'(t) = \frac{\partial}{\partial t}R_N(t)$, $R'(t) = \frac{\partial}{\partial t}R(t)$.

Proposition 13 *Suppose that P satisfies the assumptions in Proposition 12 and A.10. If H_{0S} is true and w is such that $\int t^4 w(t)dt < \infty$, then $nS_2 \xrightarrow{\mathcal{L}} f(A-1)\|S_{02}\|_w^2$.*

The distribution of $\|S_{02}\|_w^2$ is again unknown. In the context of sfarv, Henze et al. (2003) (see also Neuhaus and Zhu 1998) have proposed to approximate the null distribution of $n\|\hat{I}_n\|_w^2$ by means of the conditional distribution, given $Y_1, \ldots, Y_n$, of $V^* = \|\hat{V}_n^*\|_w^2$, where

$$\hat{V}_n^*(t) = \frac{1}{\sqrt{n}}\sum_{j=1}^n e_j\left\{\sin(tX_j) - \left(\frac{1}{n}\sum_{k=1}^n \cos(tX_k)\right)tX_j\right\},$$

$e_1, \ldots, e_n$ are as before, $X_j = Y_j - \bar{Y}$, $1 \le j \le n$, and $\bar{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$. They showed that this approximation provides a consistent estimator of the null distribution of the test statistic. Now, proceeding as in the previous subsections, to approximate the null distribution of S_2, we treat the elements in s as if they were a random sample from a random variable and estimate the p value, p, of the observed value of the test statistic S_2, $S_{2,\mathrm{obs}}$, as follows: generate B replications of V^*, say $V^{*1}, \ldots, V^{*B}$, and then approximate p by

$$\hat{p} = \frac{1}{B}\mathrm{card}\left\{b : S_{2,\mathrm{obs}} \le \frac{n}{N}(A_N - 1)V^{*b}\right\},$$

for some large B.

Remark 13 Clearly, the expression in Remark 12 for S_1 is also true for S_2 with μ_N replaced by $\hat{\mu}_N$. As for V^*, we have that

$$V^* = \frac{1}{n}\sum_{j,k=1}^n e_j e_k q_n(x_j, x_k),$$

Table 4 Empirical percentages of rejections for testing symmetry when $\mu_N = \bar{y}_N$ is: (a) known, (b) unknown

γ		a	$n = 200$				$n = 250$			
			$S_{1,1}$		$S_{1,2}$		$S_{1,1}$		$S_{1,2}$	
			5%	10%	5%	10%	5%	10%	5%	10%
(a)										
1	srs		5.0	8.5	4.9	9.1	5.1	9.6	5.2	9.3
	sam	0.5	5.1	10.4	5.0	10.5	4.6	9.9	4.7	9.6
	sam	0.1	5.6	10.8	5.5	10.9	5.5	10.3	5.7	10.4
1.1	srs		13.4	21.2	10.2	17.5	15.8	26.0	12.3	20.60
	sam	0.5	11.8	21.2	9.3	18.0	14.0	24.6	11.4	18.75
	sam	0.1	10.6	18.0	8.9	15.1	11.4	19.9	9.4	16.45
1.2	srs		39.6	55.1	29.4	43.8	49.3	64.7	37.1	52.4
	sam	0.5	39.4	53.0	29.6	42.3	47.3	62.4	35.1	50.0
	sam	0.1	27.6	42.6	21.1	32.6	35.2	50.3	26.0	38.6
(b)										
1	srs		4.2	9.4	4.6	9.7	4.8	10.3	6.0	10.8
	sam	0.5	5.8	10.1	6.0	10.3	5.0	9.4	5.6	9.6
	sam	0.1	5.4	10.5	5.2	10.9	5.0	10.8	5.3	10.9
1.1	srs		25.9	37.8	25.1	37.5	31.8	44.0	31.2	43.5
	sam	0.5	23.6	35.0	24.1	34.7	31.5	43.9	31.1	42.9
	sam	0.1	19.2	28.9	20.0	29.8	23.8	34.8	23.4	34.9
1.2	srs		71.2	81.2	72.7	82.7	83.3	89.7	84.3	90.0
	sam	0.5	71.0	81.2	71.6	82.0	79.4	88.2	81.1	89.1
	sam	0.1	56.2	68.3	57.4	69.2	67.9	79.1	69.0	80.0

with

$$\begin{aligned} 2q_n(a,b) = u(a-b) - u(a-b) + a\frac{1}{n}\sum_{j=1}^{n}\left\{u'(b+x_j) + u'(b-x_j)\right\} \\ + b\frac{1}{n}\sum_{j=1}^{n}\left\{u'(a+x_j) + u'(a-x_j)\right\} \\ - ab\frac{1}{n^2}\sum_{j=1}^{n}\left\{u''(x_j+x_k) + u''(x_j-x_k)\right\}, \end{aligned}$$

$u(x) = \int \cos(tx)w(t)\mathrm{d}t$, $u'(x) = \frac{\partial}{\partial x}u(x)$, $u''(x) = \frac{\partial^2}{\partial x^2}u(x)$.

To numerically study the performance of the proposed approximation to the null distribution, as well as the power of the test, we conducted a simulation study. With this aim, we considered the population in Sect. 4. To study the level of the proposed tests, we generated a sample with size $n = 200$ and observed the study variable. As in the previous sections, two designs were considered to draw the sample: simple random sampling without replacement (srs) and the Rao-Sampford sampling (sam) with inclusion probabilities proportional to $(1 - a)x + a$, $a = 0.1, 0.5$. The weight functions considered were the same as in Sect. 4: w_1 and w_2. The associated statistics are denoted as $S_{1,1}$ and $S_{1,2}$, when μ_N is assumed to be known, and $S_{2,1}$ and $S_{2,2}$, when μ_N is unknown, respectively. 2000 samples were generated from the population in each case. For each sample, $B = 1000$ bootstrap samples were generated as described above to estimate the p-value. The whole experiment was repeated for $n = 250$. To study the power, we modified the original population as follows: if $y_j < \bar{y}_N$ then $y_{\gamma,j} = \gamma y_j$, otherwise $y_{\gamma,j} = y_j/\gamma$, for several values of γ. Note that $\gamma = 1$ corresponds to the null hypothesis. Table 4 displays the percentages of rejections. Looking at this table, we see that, in terms of level ($\gamma = 1$), the approximation is reasonably good for all test statistics. As for the power ($\gamma = 1.1, 1.2$), $S_{1,1}$ and $S_{1,2}$ ($S_{2,1}$ and $S_{2,2}$) behave very closely, the tests with $\mu_N = \bar{y}_N$ unknown are more powerful than when $\mu_N = \bar{y}_N$ is unknown. The power increases with the sample size. It is again observed that as $\max_j \pi_j / \min_j \pi_j$ becomes larger, the power decreases.

Remark 14 To keep the notation as simple as possible, we have only considered the case of a univariate study variable. Nevertheless, the test can be applied to test the symmetry about a point for d-variate variables, for arbitrary $d \geq 1$.

7 Conclusions

The weak convergence of the finite population empirical characteristic process has been studied. Under suitable assumptions, it has the same limit as the empirical characteristic process for independent, identically distributed data from a random variable, up to a multiplicative constant depending on the sampling design. Applications of the obtained results for the two-sample problem, testing for independence and testing for symmetry have been given.

Assumptions A.7 and A.8 (also A.10) play a key role in deriving the results. They say that $\{\cos(ty_j),\ j \in U\}$ (also $\{\sin(ty_j),\ j \in U\}$) and $\{\pi_j,\ j \in U\}$ (also $\{1/\pi_j,\ j \in U\}$) are asymptotically uncorrelated $\forall t$. It is a matter of future research to derive the weak convergence of the finite population empirical characteristic process when these assumptions fail to be true.

Acknowledgements The authors thank the anonymous referees and the Associate Editor for their valuable time and careful comments, which improved the quality of this paper. M. D. Jiménez-Gamero acknowledges financial support from Grant MTM2014-55966-P of the Spanish Ministry of Economy and Competitiveness.

References

Alba-Fernández V, Jiménez-Gamero MD, Muñoz-García J (2008) A test for the two-sample problem based on empirical characteristic functions. Comput Stat Data Anal 52:3730–3748

Anderson NH, Hall P, Titterington DM (1994) Two-sample tests for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. J Multivar Anal 50:41–54

Berger YG (1998) Rate of convergence to normal distribution for the Horvith–Thompson estimator. J Stat Plan Inference 67:209–226

Conti PL (2014) On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. Sankhyā 76:234–259

Conti PL, Marella D (2015) Inference for quantiles of a finite population: asymptotic versus resampling results. Scand J Stat 42:545–561

Csörgő S (1981) Limit behaviour of the empirical characteristic function. Ann Probab 9:130–144

Csörgő S (1985) Testing for independence by the empirical characteristic function. J Multivar Anal 16:290–299

Erdös P, Rényi A (1959) On the central limit theorem for samples from a finite population. Publ Math Inst Hung Acad Sci 4:49–61

Feller W (1971) An introduction to probability theory and its applications, vol 2. Wiley, New York

Feuerverger A, Mureika RA (1977) The empirical characteristic function and its applications. Ann Stat 5:88–97

Hájek J (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann Math Stat 35:1491–1523

Hájek J (1981) Sampling from a finite population. Marcel Dekker, New York

Henze N, Klar B, Meintanis SG (2003) Invariant tests for symmetry about an unspecified point based on the empirical characteristic function. J Multivar Anal 87:275–297

Henze N, Klar B, Zhu LX (2005) Checking the adequacy of the multivariate semiparametric location shift model. J Multivar Anal 93:238–256

Hlávka Z, Hušková M, Meintanis SG (2011) Tests for independence in non-parametric heteroscedastic regression models. J Multivar Anal 102:816–827

Hušková M, Meintanis SG (2008) Tests for the multivariate k-sample problem based on the empirical characteristic function. J Nonparametr Stat 20:263–277

Isaki C, Fuller W (1982) Survey design under the regression superpopulation model. J Am Stat Assoc 77:89–96

Jiménez-Gamero MD, Alba-Fernández V, Muñoz-García J, Chalco-Cano Y (2009) Goodness-of-fit tests based on empirical characteristic functions. Comput Stat Data Anal 53:3957–3971

Kankainen A, Ushakov NG (1998) A consistent modification of a test for independence based on the empirical characteristic function. J Math Sci 89:1486–1493

Kish L (1965) Survey sampling. Wiley, New York

Marcus MB (1981) Weak convergence of the empirical characteristic function. Ann Prob 9:194–201

Meintanis SG (2005) Permutation tests for homogeneity based on the empirical characteristic function. J Nonparametr Stat 17:583–592

Meintanis SG, Iliopoulos G (2008) Fourier methods for testing multivariate independence. Comput Stat Data Anal 52:1884–1895

Neuhaus G, Zhu LX (1998) Permutation tests for reflected symmetry. J Multivar Anal 67:129–153

Särndal CE, Swenson B, Wretman J (1992) Model assisted survey sampling. Springer, New York

Székely GJ, Rizzo M, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35:2769–2794

Tillé Y (2006) Sampling algorithms. Springer, New York

Wang J (2012) Sample distribution function based goodness-of-fit test for complex surveys. Comput Stat Data Anal 56:664–679

Xiao Y (2017) A fast algorithm for two-dimensional Kolmogorov–Smirnov two sample tests. Comput Stat Data Anal 105:53–58