Testing for independence in analytic inference

Test di indipendenza nell'inferenza analitica

Pier Luigi Conti and Alberto Di Iorio

Abstract In analytic inference, data usually come from complex sampling designs, possibly with different inclusion probabilities, stratification, clustering of units. The effect of a complex sampling design is that sampling data are not *i.i.d.*, even if they are at a superpopulation level. This dramatically changes the probability distribution of usual test-statistics, such as Spearman's Rho. An approach based on a special form of resampling is proposed, and its properties are studied.
Abstract *Nell'inferenza analitica i dati generalmente provengono da disegni campionari complessi, che includono differenti probabilità di inclusione, stratificazione, grappoli di unità. L'effetto di un disegno di questo tipo è che i dati a livello campionario non sono* i.i.d.*, anche se lo sono a livello di superpopolazione. Di conseguenza, viene completamente modificata la distribuzione di probabilità delle statistiche-test comunemente utilizzate (come il Rho di Spearman). Nel presente lavoro viene studiato un approccio basato su una nuova forma di ricampionamento, di cui si studiano le proprietà.*

Key words: Independence tests, sampling design, asymptotics, empirical process, resampling.

1 Introduction

The use of superpopulation models in survey sampling has a long history, going back (at least) to [2], where the limits of assuming the population characteristics as *fixed*, especially in economic and social studies, are stressed. As clearly appears (cfr.

Pier Luigi Conti
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy, e-mail: pierluigi.conti@uniroma1.it

Alberto Di Iorio
Banca d'Italia; Via Nazionale, 91; 00184 Roma; Italy, e-mail: alberto.diiorio@uniroma1.it

[7], [5]), there are basically two forms of inference in a 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.

In the present paper attention is focused on a special problem of analytic inference, namely testing for independence between two characters. The main consequence of using a sampling design with possibly different inclusion probabilities is that commonly used test-statistics based on ranks, such as Spearman Rho rank correlation or Gini G cograduation statistics are not distribution free under independence, and in general do not have the same distribution (neither finite sample, nor asymptotic) as in the case of *i.i.d.* sample data. This calls for the need of developing new test-statistics, suitable for data collected through a complex design. Since their distribution, both for a finite sample size and asymptotically, does have a complicate form depending on the superpopulation, the need of approximating their distribution arises. Unfortunately, the widespread Efron's bootstrap does not work in the present case, again because of the use of a complex sample design. In the sequel, a new resampling scheme will be proposed, and its properties studied. In particular, it will be shown that it is asymptotically correct.

Let $\mathscr{U}_N$ be a finite population of size N, and let X, Y be two characters of interest, defined on the population $\mathscr{U}_N$. Let further x_i, y_i be the values of characters X, Y for unit i $(= 1, \ldots, N)$.

A sample s of size n is a subset of $\mathscr{U}_N$. The selection of s is performed according to a probabilistic sampling design. Formally speaking, for each unit i in $\mathscr{U}_N$, define a Bernoulli random variable (r.v.) D_i, such that the unit i is included in the sample if and only if (iff) $D_i = 1$, and let $\mathbf{D}_N = (D_1, \ldots, D_N)$. A (unordered, without replacement) sampling design P is the probability distribution of $\mathbf{D}_N$. In particular, $\pi_i = E_P[D_i]$ $(\pi_{ij} = E_P[D_i D_j])$ is the first (second) order inclusion probability of unit i (pair of units i, j). The suffix P denotes the sampling design used to select population units.

The first order inclusion probabilities are frequently taken proportional to an appropriate function of the values of the *design variables*. The design variables may include strata indicator variables, as well as qualitative variables measuring cluster and unit characteristics (cfr. [5]); in what follows they are denoted by $\mathscr{T}_1, \ldots, \mathscr{T}_L$, whilst $t_{i1}, \ldots, t_{iL}$ are their values for unit i. As already said, we will assume that $\pi_i \propto z_i$, where $z_i = h(t_{i1}, \ldots, t_{iL})$ is the value of $Z = h(\mathscr{T}_1, \ldots, \mathscr{T}_L)$ for unit i.

Take now N real numbers $0 < p_i < 1$, $i = 1, \ldots, N$, with $p_1 + \cdots + p_N = n$. The sampling design is a *Poisson design* with parameters $p_1, \ldots, p_N$ if the r.v.s D_is are independent with $\pi_i = p_i$ for each unit i. The *rejective sampling*, or *normalized conditional Poisson sampling* ([4], [8]) corresponds to the probability distribution of the random vector $\mathbf{D}_N$, under Poisson design, conditionally on $n_s = n$.

The *Hellinger distance* between a sampling design P and the rejective design is

$$d_H(P, P_R) = \sum_{D_1,\ldots,D_N} \left(\sqrt{Pr_P(D_N)} - \sqrt{Pr_R(D_N)} \right)^2 . \tag{1}$$

Our basic assumptions are listed below.

A1. $(\mathscr{U}_N;\ N \geq 1)$ is a sequence of finite populations of increasing size N.

A2. For each N, $(y_i, x_i, t_{i1}, \ldots, t_{iL})$, $i = 1, \ldots, N$ are realizations of a superpopulation $\{(Y_i, X_i, T_{i1}, \ldots, T_{iL}),\ i = 1, \ldots, N\}$ composed by *i.i.d.* $(L+2)$-dimensional r.v.s. In the sequel, the symbol $\mathbb{P}$ will denote the (superpopulation) probability distribution of r.v.s $(Y_i, X_i, T_{i1}, \ldots, T_{iL})$s, and $\mathbb{E}$, $\mathbb{V}$ are the corresponding operators of mean and variance, respectively. Furthermore, if $Z_i = h(T_{i1}, \ldots, T_{iL})$, the joint superpopulation d.f. of (Y_i, X_i, Z_i) will be denoted by

$$K(y, x, z) = \mathbb{P}(Y_i \leq y, X_i \leq x, Z_i \leq z), \tag{2}$$

and

$$H((y, x|z) = \mathbb{P}(Y_i \leq y, X_i \leq x|Z_i = z), \tag{3}$$

$$F(y|z) = \mathbb{P}(Y_i \leq y|Z_i = z),\ \ G(x) = \mathbb{P}(X_i \leq x|Z_i = z) \tag{4}$$

are the joint and marginal superpopulation d.f.s of Y_i and X_i (given Z).

A3. For each population $\mathscr{U}_N$, sample units are selected according to a fixed size sample design with positive first order inclusion probabilities $\pi_i \propto z_i$, with sample size $n = \pi_1 + \cdots + \pi_N$, and $z_i = h(t_{i1}, \ldots, t_{iL})$, $i = 1, \ldots, N$. It is assumed that

$$\lim_{N,n\to\infty} \mathbb{E}[\pi_i(1-\pi_i)] = d > 0. \tag{5}$$

Furthermore, the notation $x_N = (x_1, \ldots, x_N)$ is used.

A4. The sample size n increases as the population size N does, with

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

A5. For each population $(\mathscr{U}_N;\ N \geq 1)$, let P_R be the rejective sampling design with inclusion probabilities π_1, …, π_N, and let P be the actual sampling design (with the same inclusion probabilities). Then

$$d_H(P, P_R) \to 0 \text{ as } N \to \infty,\ \ a.s. - \mathbb{P}.$$

A6. $\mathbb{E}[X_1^2] < \infty$, so that the quantity in (5) is equal to:

$$d = f\left(1 - \frac{\mathbb{E}[X_1^2]}{\mathbb{E}[X_1]^2}\right) + f(1-f)\frac{\mathbb{E}[X_1^2]}{\mathbb{E}[X_1]^2} > 0. \tag{6}$$

2 The problem

As already said, our goal is to construct an independence test for the two characters X, Y (conditionally on the design variables $T_j s$). For the sake of simplicity we will consider a single, discrete design variable T, taking values $T^1, \ldots, T^k$. Hence, the hypothesis problem takes the form

$$H_0 : H(x,y|T) = F(x|T)G(y|T)$$
$$H_1 : H(x,y|T) \neq F(x|T)G(y|T)$$

A simple approach could consist in using a rank based test-statistic, such as the Spearman's Rho statistic or the Gini's cograduation statistic. Unfortunately, due to the use of the sample design, such statistics are not distribution free under H_0, neither exactly nor asymptotically. Hence, their use is inappropriate under general sampling designs.

In order to construct a test-statistic for the above problem, the general measure of monotone dependence proposed in [1] is extended to the present case. Given two continuous variables X, Y, let F and G be their marginal distributions, respectively, and let H be the joint distribution of the bivariate variable (X,Y). A general measure of the monotone dependence γ_g between X and Y, is a real-valued functional γ_g of the bivariate distribution $H(x,y)$ defined as follows

$$\gamma_g = \int_{\mathbb{R}^2} g(|F(x|T) + G(y|T) - 1|) - g(|F(x|T) - G(y|T)|)\, dH(x,y|T), \quad (7)$$

where $g : [0,1] \to \mathbb{R}$ is a strictly increasing, continuous and convex function, with $g(0) = 0$ snd continuous first derivative. Under the null hypothesis of independence the latter quantity is equal to zero. If $g(s) = s^2$, then (7) reduces to the (un-normalized) Spearman's coefficient. If $g(s) = s$, then (7) reduces to the (un-normalized) Gini cograduation coefficient. In general, $\gamma_g = 0$ whenever X and Y are independent.

The basic idea is to estimate first H, F, G by their Hájek estimators

$$\widehat{H}(y, xvertt) = \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(x_i \leq x)} I_{(y_i \leq y)} I_{(T_i = t)} \Bigg/ \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(T_i = t)} \quad (8)$$

$$\widehat{F}(y) = \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(y_i \leq y)} I_{(T_i = t)} \Bigg/ \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(T_i = t)}, \quad (9)$$

$$\widehat{G}(x) = \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(x_i \leq x)} I_{(T_i = t)} \Bigg/ \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(T_i = t)} \quad (10)$$

and then in estimating the quantity γ_g with a plug-in approach, by replacing the distribution function is the distributions functions in (7) with their Hájek estimators. In this way, the test-statistic

$$\widehat{\gamma}_g = \frac{\sum_{i=1}^{N} \frac{1}{\pi_i} \left(g(|\widehat{F}(x_i|T_i) + \widehat{G}(y_i|T_i) - 1|) - g(|\widehat{F}(x_i|T_i) - \widehat{G}(y_i|T_i)|) \right) D_i}{\sum_{i=1}^{N} \frac{1}{\pi_i} D_i}. \quad (11)$$

is obtained.

Proposition 1. *Suppose that the conditions* $A1$*-*$A6$ *are met, and assume that the null hypothesis* H_0 *holds true. Then, the r.v.*

$$\sqrt{n}\widehat{\gamma}_g \quad (12)$$

tends in distribution to a normal r.v. with zero mean and variance σ_0^2*, as* N*,* n *increase.*

The asymptotic variance σ_0^2 of (12) does have a complicate form, and cannot be estimated on the basis of sample data. The basic idea is to approximate the distribution of (11) under H_0 by resorting to resampling.

Define S_j as the subset of sample units having value T^j of the design variable, and let n_j be the size of S_j, $j = 1, \dots, k$.

0. Repeat M times steps 1-4 below.
1. Generate a pseudo-population of size N by selecting unit i in the sample with probability $\pi_i^{-1} / \sum \pi_i^{-1} D_i$. To each unit i^* of the pseudo-population a value $T_{i^*}^*$ is attached, such that $T_{i^*}^* = T_i$ whenever $i^* = i$.
2. If $T_{i^*}^* = T^j$, then sample independently from S_j, and with probability $\pi_i^{-1} / \sum_{k \in S_j} \pi_k^{-1}$, a X-value $X_{i^*}^*$ and a Y-value $Y_{i^*}^*$. As a result, a pseudo-population $\mathscr{U}_N^*$ is obtained, such that for each unit $i^* \in \mathscr{U}_N^*$ a triplet $(Y_{i^*}^*, X_{i^*}^*, T_{i^*}^*)$ is defined. Furthermore, $Y_{i^*}^*$ and $X_{i^*}^*$ are independent conditionally on $T_{i^*}^*$.
3. Draw a pseudo-sample of size n from the population $\mathscr{U}_N^*$, using a high entropy sampling design P^* with first order inclusion probabilities π_{i^*} according to $A3$.
4 Compute the value $\widehat{\gamma}_g^*$ of the statistic (11) for the pseudo-sample drawn at step 3.

In this way, the M replicates

$$\sqrt{n}\widehat{\gamma}_{g,m}^*, \;\; m = 1, \dots, M \quad (13)$$

are obtained.

Consider next the empirical distribution function (edf) constructed on the basis of the M replicates (13):

$$\widehat{R}(u) = \frac{1}{M} \sum_{m=1}^{M} I_{(\widehat{\gamma}_{g,m}^* \leq u)} \quad (14)$$

and let $\widehat{R}^{-1}$ be the corresponding quantile function

$$\widehat{R}^{-1}(p) = \inf\{u : \widehat{R}(u) \geq u\}. \quad (15)$$

Let further Φ_{0,σ_0} be a normal distribution function with zero expectation and variance σ_0^2. Proposition 1 establishes that, as both N, n increase, under the null hypothesis H_0 the probability distribution function $Pr_{H_0}(\sqrt{n}\widehat{\gamma}_g \leq u)$ tends to $\Phi_{0,\sigma_0}(u)$, uniformly w.r.t. u. Next proposition establishes that, as the number M of replicates increases, the edf (14) tends to the *same* limiting law. This proves the (asymptotic) validity of the proposed resampling technique.

Proposition 2. *Suppose that the conditions* $A1$*-*$A6$ *are met, and assume that the null hypothesis* H_0 *holds true. Then, as* N*,* n*,* M *increase, the following results hold (with probability 1).*

$$\sup_u \left| \widehat{R}(u) - Pr_{H_0}(\sqrt{n}\widehat{\gamma}_g \leq u) \right| \rightarrow 0;$$
$$\sup_u \left| \widehat{R}(u) - \Phi_{0,\sigma_0}(u)) \right| \rightarrow 0;$$
$$\widehat{R}^{-1}(p) \rightarrow \sigma_0 z_{1-p}$$

where z_{1-p} *satisfies the relationship* $\Phi_{0,1}(z_{1-p}) = p$*.*

As a consequence of Proposition 2, the region

$$\widehat{R}^{-1}(\alpha/2) \leq \sqrt{n}\widehat{\gamma}_g \leq \widehat{R}^{-1}(\alpha/2) \tag{16}$$

is an acceptance region of approximative size α for testing independence.

A comparison between the proposed test and the one based on the "usual" Spearman statistic has been performed *via* simulation. The proposed test performs better than the "usual" one in terms of power, as well as in terms of closeness of the actual size to the nominal significance level.

References

1. Cifarelli D M., Conti P L., Regazzini E.: On the asymptotic distribution of a general measure of monotone dependence. The Annals of Statistics, **24**, 1386–1399 (1996)
2. Cochran W G.: The use of the analysis of variance in enumeration by sampling. Journal of the American Statistical Association, **34**, 492–510 (1939)
3. Efron B.: Bootstrap methods: another look at the jackknife. The Annals of Statistics, **7**, 1–26 (1979)
4. Hájek J.: Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population. The Annals of Mathematical Statistics, **35**, 1491–1523 (1964)
5. Pfeffermann D.: The role of sampling weights when modeling survey data. International Statistical Review, **61**, 317–337 (1993)
6. Pfeffermann D., Sverchkov M.: Prediction of finite population totals based on the sample distribution. Survey Methodology, **30**, 79-92 (2004)
7. Särdnal C -E., Swensson B., Wretman, J H.: Model Assisted Survey Sampling. Springer-Verlag, New York (1992)
8. Tillé, Y.: Sampling Algorithms. Springer Verlag, New York (2006)