Journal of Official Statistics, Vol. 13, No. 2, 1997, pp. 123–142

Testing of Distribution Functions from Complex Sample Surveys

Abba M. Krieger[1] *and Danny Pfeffermann*[2]

Testing the parametric family of distributions is a classical problem in theoretical and applied statistics. However, when the sample is selected with unequal selection probabilities which are related to the values of the response variable, standard methods no longer apply. In this article we consider two alternative approaches for taking account of the sample selection effects. Under the first approach, the range of the response variable is divided into a fixed number of intervals and large-sample Wald statistics and other related statistics are constructed from design-based estimators of the interval probabilities. Under the second approach, the parametric distribution of the sample data is extracted as a function of the hypothesized population distribution and the sample inclusion probabilities. The extracted distribution is then tested using standard test statistics. The two approaches are compared in a simulation study which indicates that the second approach performs better overall in terms of the achieved significance levels and powers against alternative distributions considered.

Key words: Chi-squared statistics; inclusion probabilities; Kolmogorov-Smirnov; probability integral transformation; randomization distribution; Wald statistics.

1. Introduction

Testing parametric families of distribution functions is a classical problem in theoretical and applied statistics. Several goodness of fit statistics based on different distance functions between the hypothesized distribution and the empirical sample distribution have been explored in the literature for their theoretical properties, and are in common use in applied work. Well known examples are the Kolmogorov-Smirnov and the Chi-square test statistics. For review and discussion of these and other test statistics, see, for example, the books by Kendall and Stuart (1973, Vol. 2, Chapter 30) and Pratt and Gibbons (1981, Chapter 7).

A common assumption underlying the use of these statistics is that the sample measurements are independent realizations of the population distribution. This assumption is violated in a typical sample survey where the sample units are often selected with unequal selection probabilities, at least at some stages of the selection process. When the selection probabilities are related to the values of the response variable, the empirical sample distribution is in general not consistent with the distribution of the population measurements, implying that standard testing procedures no longer apply.

For example, consider the testing of an income distribution; a familiar problem in economic studies. A major source for income data are household surveys like for example

[1] University of Pennsylvania, The Wharton School, 3620 Locust Walk, Philadelphia, PA 19104-6302, U.S.A.
[2] Hebrew University, Mount Scopus, 91905 Jerusalem, Israel.

the Panel Study of Income Dynamics (PSID) in the U.S. and the Survey of Consumer Finances (SCF) in Canada. The PSID oversamples low-income families, whereas the SCF uses incomes for the formation of strata. Under both designs the income distribution in the sample is not representative of the income distribution in the population. Testing income distributions based on the Canadian SCF prompted the present study.

In this article we consider two classes of test statistics that can be used in such situations. The first class consists of large-sample Wald (1943) statistics, constructed by dividing the range of the response variable into a fixed number of intervals and computing the Mahalanobis (1936) distance between the estimated probabilities of these intervals and the true probabilities under the hypothesized distribution. The notable feature of this class is that the estimators are chosen so that they are design-consistent under repeated sampling for the corresponding probabilities that would have been computed if all the population values had been observed. Alternatives to the Wald statistics are considered as well.

The second class contains standard test statistics and a moments-based statistic. These statistics, however, are applied to the distribution of the sample measurements under the null hypothesis, derived as a function of the hypothesized population distribution and known characteristics of the sampling design, such as the first order sample inclusion probabilities.

In Section 2 we discuss the rationale for the use of design-based test procedures and define several statistics that fall into this class. Section 3 defines the sample distribution of survey data, illustrating its dependency on the population distribution and the sample selection probabilities. The use of this distribution for testing hypotheses on the population distribution is considered in Section 4. Section 5 presents the results of a simulation study aimed to compare the performance of the two classes of test statistics. The general conclusion from this study is that the test statistics of the second class outperform the test statistics of the first class both in terms of type I error probabilities and in terms of power against the alternatives considered. Section 6 contains some general remarks with suggestions for further research.

2. Design-Based Test Statistics

2.1. Rationale for the use of design-based test statistics

Survey data may be viewed as the outcome of two random processes: the process generating the values in the (finite) population $U = \{1, \ldots, N\}$ from which the sample is taken, (the 'superpopulation' model), and the process generating the sample data from the finite population values $\mathbf{Y}'_U = \{Y_1 \ldots Y_N\}$.

In what follows we assume that the population values are independent measurements from a continuous probability density function (pdf) $f_p(y;\boldsymbol{\theta})$, indexed by the (possibly vector) parameter $\boldsymbol{\theta}$. The subscript "p" is used to distinguish the population distribution from the sample distribution defined in Section 3. The problem investigated in this study is the testing of hypotheses of the general form

$$H_0 : f_p(y;\boldsymbol{\theta}) = f^*(y;\boldsymbol{\theta}^*) \qquad (1)$$

with $f^*(y;\boldsymbol{\theta}^*)$ fully specified.

Let $\tilde{F}(t) = \frac{1}{N}\Sigma_{i=1}^{N} D(t - y_i)$ denote the empirical distribution in the finite population where $D(a) = 1$ for $a \geq 0$ and $D(a) = 0$ otherwise. If $\tilde{F}(t)$ was known, (as in the case of a census), the null hypothesis defined by (1) could be tested by testing the significance of the distance between $\tilde{F}(t)$ and $F^*(t; \boldsymbol{\theta}^*)$, the cumulative distribution under H_0, using standard test procedures. The distribution $\tilde{F}(t)$ is, however, not computable in the case of a sample and so the idea behind the use of design-based test statistics is to estimate percentiles of $\tilde{F}(t)$ by design-consistent estimators, and use those and their design-based variances to construct appropriate tests.

Let $\mathbf{T}_N(\mathbf{Y})$ define a known vector function of the finite population values $Y_1, \ldots, Y_N$. In our application $\mathbf{T}_N(\mathbf{Y})$ will stand for the proportions of values $\{Y_j\}$ falling in given intervals. Suppose that a sample $S = \{1 \ldots n\}$ with measurements $\mathbf{y}' = \{y_1 \ldots y_n, n < N\}$ is selected from the finite population U by a well defined probability sampling design $P(S)$.

Definition 1: The sample statistic $\mathbf{t}_n(\mathbf{y})$ is *design-consistent* for $\mathbf{T}_N(\mathbf{Y})$ if $Plim_{n \to \infty, N \to \infty}[\mathbf{t}_n(\mathbf{y}) - \mathbf{T}_N(\mathbf{Y})] = 0$ where '*Plim*' stands for limit in probability under the randomization distribution as induced by repeated sampling from the finite population using the sampling design $P(S)$.

Comment 1: The definition of design consistency requires a formulation of the manner by which the sample and the population mutually increase. Such a formulation is given, for example, by Isaki and Fuller (1982).

Proposition 1: *Assuming that $\mathbf{T}_N(\mathbf{Y})$ converges in probability to some functional $\tilde{\tau}$ under the model distribution, the design-consistent sample statistic $\mathbf{t}_n(\mathbf{y})$ is consistent for $\tilde{\tau}$ under the mixed $R * M$ (randomization and model) distribution generating the sample data.*

The variance-covariance (V-C) matrix of $\mathbf{t}_n(\mathbf{y})$ as an estimator of $\tilde{\tau}$ can be decomposed as

$$V_{R*M}[\mathbf{t}_n(\mathbf{y})] = E_M\{V_R[\mathbf{t}_n(\mathbf{y})|\mathbf{Y}]\} + V_M\{E_R[\mathbf{t}_n(\mathbf{y})|\mathbf{Y}]\} = E_M\{V_R[\mathbf{t}_n(\mathbf{y})|\mathbf{Y}]\} + O(N^{-1}) \tag{2}$$

Notice that $E_R[t_n(\mathbf{y})|\mathbf{Y}]$ is a population quantity which suggests that its variance under the model is $O(N^{-1})$. In a cluster sample, however, it is often the case that the number of clusters, rather than the number of ultimate sampling units has to be increased for the asymptotics to hold. This will be the case, for example, when estimating the between cluster (group) variance in a variance components model.

It follows from equation (2) that as $n \to \infty$ and $(n/N) \to 0$, the V-C matrix under the mixed $R * M$ distribution can be estimated consistently by estimating consistently the randomization V-C matrix (n and N may represent the number of sample and population clusters, see the previous comment). For a more rigorous discussion with examples of the use of design-consistent estimators for inference about model parameters, see Pfeffermann (1993).

2.2. *General structure of design-based test statistics of distribution functions*

Let $[a_0 < a_1 < \ldots < a_{K-1} < a_K]$ define a division of the range of Y into K exclusive and exhaustive intervals with $a_0 = -\infty$, $a_K = \infty$. Let $P_k^* = F^*(a_k) - F^*(a_{k-1}) = \Pr_{F^*}[a_{k-1} \leq Y \leq a_k]$, $k = 1 \ldots K$ denote the computed interval probabilities under H_0.

Let $\{\hat{P}_k\}$ be design-consistent for $\{P_k^*\}$ with randomization variance $C_{kk} = \text{Var}_R(\hat{P}_k)$ and covariances $C_{k\ell} = \text{Cov}_R(\hat{P}_k, \hat{P}_\ell)$. Denote $\mathbf{P}^{*\prime} = [P_1^* \ldots P_{K-1}^*]$,

$$\hat{\mathbf{P}}' = [\hat{P}_1 \ldots \hat{P}_{K-1}],\ C_R = [C_{k\ell}];\ 1 \leq k,\ \ell \leq K-1$$

The large-sample Wald statistic (1943) is defined as

$$W^2 = n(\hat{\mathbf{P}} - \mathbf{P}^*)'\hat{C}_R^{-1}(\hat{\mathbf{P}} - \mathbf{P}^*) \tag{3}$$

where $\frac{1}{n}\hat{C}_R$ is design-consistent for C_R. (Here and throughout this article we assume a single stage sample of size n). Assuming that $\hat{\mathbf{P}}$ is asymptotically normal (see, e.g., Fuller 1975; and Binder 1983, for central limit theorems applicable to complex survey data), the statistic W^2 is distributed asymptotically as $\chi^2_{(K-1)}$ under H_0, Stroud (1971).

The use of the Wald statistic and modifications thereof has been studied extensively in the literature on categorical data analysis. The modifications discussed below are borrowed from that literature.

Suppose that $\hat{C}_R$ is replaced by the multinomial sampling V-C matrix $\Delta^* = [\text{diag}(\mathbf{P}^*) - \mathbf{P}^*\mathbf{P}^{*\prime}]$. The statistic W^2 then has the form

$$X^2 = n\sum_{k=1}^{K}(\hat{P}_k - P_k^*)^2/P_k^* \tag{4}$$

which is in the form of the Pearson Chi-Squared test statistic. Rao and Scott (1981) establish the asymptotic distribution of X^2 under general sampling designs as the distribution of the weighted sum $\Sigma_{k=1}^{K-1}\lambda_k W_k$ of independent $\chi^2_{(1)}$ random variables W_k, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{K-1}$ are the eigenvalues of the matrix $D = n\Delta^{*-1}C_R$. Calculating percentage points of the distribution of X^2 requires therefore the knowledge of all the eigenvalues $\{\lambda_k\}$ or consistent estimators of them.

In practice, it is often the case that reliable estimates of the full V-C matrix C_R are not available because of the complexity of the sampling design and/or limitations in access to microdata files required for the computation of such estimates. Rao and Scott (1981) propose simple modifications to X^2 applicable in such situations. Two such modifications are

$$X_M^2 = X^2/M(\hat{\lambda}) \tag{5}$$

where $M(\hat{\lambda}) = \Sigma_{k=1}^{K-1}\hat{\lambda}_k/(K-1)$; $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_{K-1}$ are the eigenvalues of $\hat{D} = n\hat{\Delta}^{-1}\hat{C}_R$; $\hat{\Delta} = [\text{diag}(\hat{\mathbf{P}}) - \hat{P}\hat{P}']$, and

$$X_S^2 = X_M^2/[1 + V(\hat{\lambda})/M^2(\hat{\lambda})] \tag{6}$$

where $V(\hat{\lambda}) = \Sigma_{k=1}^{K-1}[\hat{\lambda}_k - M(\hat{\lambda})]^2/(K-1)$.

The distribution of X_M^2 under H_0 is approximated by $\chi^2_{(K-1)}$. The distribution X_S^2 is approximated under H_0 as $\chi^2_{(\nu)}$ where $\nu = (\text{K} - 1)/[1 + V(\hat{\lambda})/M^2(\hat{\lambda})]$. The idea behind the approximations is to have the first moment of X_M^2 and the first and second moments of X_S^2 approximately equal to the corresponding moments of the X^2 distributions used for the approximations. As discussed by Rao and Scott, the computation of X_M^2 only requires estimates for the variances $C_{kk} = \text{Var}_R(\hat{P}_k)$ whereas the use of X_S^2 requires essentially an estimate of the full V-C matrix C_R. Nonetheless, the use of X_S^2 is often recommended as an alternative to W^2 in situations where C_R is unstable, as reflected by large differences in the eigenvalues $\hat{\lambda}_k$. Thomas and Rao (1987) also consider F

transformations of the statistics X_M^2 and X_S^2 for the case of a cluster sample with equal selection probabilities. The validity of this transformation in the case of unequal selection probabilities is, however, not clear.

The design-based estimators $\hat{P}_k$ and the associated variance estimators $\hat{C}_{kk}$ can be used also to construct a test procedure based on Bonferroni probability bounds. The procedure consists of computing the statistic

$$BON = \max_{1 \leq k \leq K} \left| \frac{\hat{P}_k - P_k^*}{\hat{C}_{kk}^{1/2}} \right| \tag{7}$$

and compare it to the α/2K percentage point of the standard normal distribution. The use of BON is known to be conservative but it has the advantage that it does not require the estimation of the covariances $C_{k\ell}$.

2.3. *Design-based estimators considered*

The following design-based estimators $\{\hat{P}_k\}$ have been considered in the empirical study for construction of the Wald statistics and the modifications discussed in Section 2.2. The notation $I_k(z)$ is used to define an indicator variable taking the value of 1 when $a_{k-1} \leq z \leq a_k$.

1) The modified Horvitz-Thompson (1952) estimator

$$\hat{P}_{k,HT} = \sum_{i=1}^{n} I_k(y_i)w_i \bigg/ \sum_{i=1}^{n} w_i \tag{8}$$

where $w_i = 1/\pi_i$ is the sampling weight associated with unit i.

2) Rao, Kovar, and Mantel (1990) Difference and Ratio estimators

$$\hat{P}_{k,D} = N^{-1}\left\{\sum_{i=1}^{n} I_k(y_i)w_i + \left[\sum_{i=1}^{N} I_k(\hat{R}x_i) - \sum_{i=1}^{n} I_k(\hat{R}x_i)w_i\right]\right\} \tag{9}$$

$$\hat{R}_{k,R} = N^{-1}\left\{\left[\sum_{i=1}^{n} I_k(y_i)w_i\right]\left[\sum_{i=1}^{N} I_k(\hat{R}x_i)\right] \bigg/ \left[\sum_{i=1}^{n} I_k(\hat{R}x_i)w_i\right]\right\} \tag{10}$$

where X is an auxiliary variable related to Y with known values for every unit in the population, $R = \Sigma_{i=1}^{N} y_i / \Sigma_{i=1}^{N} x_i$ and $\hat{R} = \Sigma_{i=1}^{n} y_i w_i / \Sigma_{i=1}^{n} x_i w_i$. Note that when $y_i = Rx_i$ for all i, $\hat{P}_{k,D} = \hat{P}_{k,R} = [\tilde{F}(a_k) - \tilde{F}(a_{k-1})]$, the proportion of units in the population with y values falling into the kth interval.

Rao, Kovar, and Mantel (1990), in fact consider the estimation of the empirical population distribution $\tilde{F}(t)$, rather than the finite population interval proportions but the modification of their estimators to the latter case is trivial. As discussed in Section 2.1, these estimators, as well as $\hat{P}_{k,HT}$ can be viewed as estimating also the 'superpopulation' interval probabilities $P_k = [F(a_k) - F(a_{k-1})]$. The authors provide general formulae for the asymptotic randomization variances of the three estimators and estimators for the variances based on linearization. Estimators for the randomization covariances $C_{k\ell} = \mathrm{Cov}_R(\hat{P}_k, \hat{P}_\ell)$ can be obtained from the estimators of the variances of the corresponding differences $(\hat{P}_k - \hat{P}_\ell)$.

It should be noted that several other estimators of the empirical population distribution

$\tilde{F}(t)$ have been proposed in the literature, see e.g., Chambers and Dunstan (1986), Kuo (1988), Rao, Kovar, and Mantel (1990), Kuk (1993) and Silva and Skinner (1995). The common feature to all of these estimators is that they utilize auxiliary population information as in the Difference and Ratio estimators defined by (9) and (10). In fact, Silva and Skinner (1995) found the other estimators to perform better than the three estimators considered in our study with respect to an average MSE criterion, but their simulation study was restricted to simple random sampling with small sample sizes ($n \leq 50$). We did not consider these other estimators in our study for two reasons:

1) Some of the estimators are purely model-based in the sense that they assume certain relationships between the response variable and the auxiliary variable, or that they require the specification of a kernel density and a corresponding bandwidth. The estimator proposed by Silva and Skinner (1995) requires a division of the population into poststrata based on the ascending values of the auxiliary variable. (Their estimator is defined as $\Sigma_{g=1}^{G}(N_g/N)\hat{P}_{k,HT}^{g}$ where $\hat{P}_{k,HT}^{g}$ is the *H-T* estimator (8) derived from poststratum g of size N_g.) The authors found that for simple random sampling a division into strata of equal size performs well, but the specification of an appropriate division when the selection probabilities depend on the auxiliary values, (allowing also for stable variance estimators), has yet to be investigated.
2) Estimation of the randomization variances of these estimators for sampling designs with unequal selection probabilities is either computationally intensive, even for small samples, and/or it requires the computation of third order sample inclusion probabilities which is not feasible under most sampling designs in common use.

3. Distributions of Complex Survey Data

The sample pdf of Y_i, the measurement associated with unit i, is defined as the conditional pdf of Y_i, given that $i \in S$, and is obtained as

$$f_s(y_i;\lambda) = f(y_i|i \in S) = \Pr(i \in S|Y_i = y_i;\gamma)f_p(y_i;\theta)/\Pr(i \in S) \qquad (11)$$

where $\lambda = (\theta, \gamma)$ and $\Pr(i \in S) = \int_y \Pr(i \in S|Y_i = y_i;\gamma)f_p(y;\theta)dy$. The subscript "$s$" is used to distinguish the sample pdf from the pdf $f_p(y_i;\theta)$ holding in the population, prior to sampling. Notice that the probabilities $\Pr(i \in s|Y_i = y_i)$ may depend, in general, on some vector parameter γ, see equations (12) and (14) below. In what follows we denote $\Pr(i \in S|Y_i = y_i;\gamma)$ as $\Pr(i \in S|y_i;\gamma)$.

Proposition 2: *The sample pdf is different from the superpopulation pdf generating the finite population values, unless $Pr(i \in S|y_i;\gamma) = Pr(i \in S)$ for all y_i, in which case the sampling design is noninformative.*

Let I_i be the sample indicator variable such that $I_i = 1$ if $i \in S$ and $I_i = 0$ otherwise and let $\pi_i = \Pr(I_i = 1)$ be the sample inclusion probability of unit i under the sampling design $P(S)$. In this study we consider the case where π_i is a measure of size. By viewing the probabilities $\{\pi_i; i = 1, \ldots, N\}$ as random outcomes with conditional pdf $g_p(\pi_i|y_i;\gamma)$, (Smith 1988), the probability $\Pr(i \in S|y_i;\gamma)$ can be expressed alternatively as

$$\Pr(i \in S|y_i;\gamma) = \int_{\pi_i} \Pr(I_i = 1|y_i, \pi_i)g_p(\pi_i|y_i;\gamma)d\pi_i = E_M(\pi_i|y_i;\gamma) \qquad (12)$$

since $\Pr(I_i = 1|y_i, \pi_i) = \pi_i$. Substituting (12) into (11) yields

$$f_s(y_i;\lambda) = E_M(\pi_i|y_i;\gamma)f_p(y_i;\theta) \Big/ \int_{y_i} E_M(\pi_i|y_i;\gamma)f_p(y_i;\theta)dy_i. \tag{13}$$

Notice that the denominator of (13) is $E_M(\pi_i;\lambda)$.

Proposition 3: *For a given pdf $f_p(y_i;\theta)$, the sample pdf $f_s(y_i;\lambda)$ is fully specified by the conditional expectation $E_M(\pi_i|y_i;\gamma)$.*

Comment 2: The relationship in (13) can be extended straightforwardly to the case of conditional pdfs $f_p(y_i|\boldsymbol{x}_i;\theta)$.

Comment 3: It is important to emphasize that even when $E_M(\pi_i|y_i;\gamma)$ depends on y_i, the sample pdf may still be of the same parametric family as the population pdf, with only some of the parameter values being changed. Pfeffermann, Krieger, and Rinott (1995) consider several such examples and define more general invariance conditions under which the parametric distribution of the population measurements is "closed under sampling," a term borrowed from the Bayesian literature, where it refers to conjugate prior distributions.

Under the assumption that $E_M(\pi|y;\gamma) = m(y;\gamma)$ is a continuous function, a general class of sample pdfs is obtained by approximating $m(y;\gamma)$ by the Taylor approximation

$$m(y) \hat{=} \sum_{j=0}^{J} A_j y^j \tag{14}$$

where the $\{A_j; j = 0, \ldots, J\}$ are functions of γ with J appropriately specified. (See Section 4.) Notice that $0 < m(y;\gamma) < 1$ and for a fixed sample size n, $\Sigma_{i=1}^{N} m(y_i;\gamma) = n$. It is not required, however, to scale the expression on the right-hand side of (14), since it appears in both the numerator and the denominator. Substituting (14) into (13) yields

$$f_s(y;\lambda) \hat{=} \sum_{j=0}^{J} A_j E_M(y^j) h_p^{(j)}(y;\theta) \Big/ \sum_{j=0}^{J} A_j E_M(y^j) \tag{15}$$

where $h_p^{(j)}(y;\theta) = y^j f_M(y;\theta)/E_M(y^j)$. The pdf (15) is a mixture of the pdfs $h_p^{(j)}$, $j = 0, 1, \ldots, J$ with mixture coefficients $c_j = A_j E_M(y^j)/\Sigma_{j=0}^{J} A_j E_M(y^j)$.

Example: In the empirical study, (Section 5), we consider the case where $f_p(y;\theta) = \text{Gamma}(\alpha, \beta)$. Under (14), the sample pdf is a mixture of Gamma densities,

$$f_s(y;\lambda) = \sum_{j=0}^{J} c_j^* \text{ Gamma } (\alpha + j, \beta) \Big/ \sum_{j=0}^{J} c_j^* \tag{16}$$

where $c_0^* = A_0$ and $c_j^* = A_j \alpha(\alpha + 1)\ldots(\alpha + j - 1)/\beta^j$.

Comment 4: For given values of α and β, a full specification of the sample pdf in (16) requires a specification (estimation) of the coefficients c_j^*. See Section 4.

4. Application to Testing Distribution Functions

4.1. Use of classical test procedures

By a classical test procedure we mean testing that the empirical sample distribution is consistent with the distribution postulated under the null hypothesis, viewing the sample measurements as independent identically distributed (iid). The application of such procedures therefore requires the specification of the sample distribution under the null hypothesis, with an added assumption of independence. The sample pdf under unequal sample selection probability schemes is defined in (13) and for the case where $E_M(\pi | y; \gamma)$ can be approximated by a polynomial in y, it takes the form defined by (15). Our proposed testing procedure therefore consists of the following stages:

I Specify the highest power J in the polynomial approximation defined by (14).

II Estimate the coefficients $\{A_j\}$ in (14) or the "mixture coefficients" $c_j = A_j E_M(y^j) / \Sigma_{i=0}^J A_i E_p(y^i)$ in (15).

III Apply classical test statistics to the cumulative distribution obtained from (15) under the null hypothesis $f_p(y; \boldsymbol{\theta}) = f^*(y; \boldsymbol{\theta}^*)$ with $A_j(c_j)$ replaced by their sample estimates.

The specification of J can be based on knowledge of the sampling design, or it can be carried out in conjunction with the estimation of the coefficients $A_j(c_j)$, using an appropriate stepwise selection algorithm. In theory, this could be achieved by regressing the probabilities π_i observed for the sample data against powers of y_i. The resulting estimators and hence the specification of J could however be severely biased this way since the sample is selected with these probabilities (the regression dependent variable). Two alternative approaches which account for the sample selection effects are:

a) Regress the probabilities π_i against powers of y_i but estimate the coefficients A_j by use of design-based estimators which incorporate the sampling weights. See Pfeffermann (1993) for a review of such methods.

b) Estimate the coefficients c_j by maximum likelihood techniques applied to the likelihood obtained from (15) – this can be carried out most conveniently by means of the EM algorithm. It amounts to iterating the set of equations $c_j^{(t)} = \frac{1}{n}\Sigma_{i=1}^n \{c_j^{(t-1)} B(i,j) / \Sigma_{\ell=0}^J c_\ell^{(t-1)} B(i,\ell)\}$ where $B(i,j) = y_i^j / E_M(y^j)$ such that $h_p^{(j)}(y_i; \boldsymbol{\theta}) = B(i,j) f_p(y_i; \boldsymbol{\theta})$, $j = 0, \ldots, J$; $i = 1, \ldots, n$.

In the empirical study the two approaches yield very similar estimates. We used the first approach for estimating the (randomization) variances of the estimated coefficients.

The use of maximum likelihood for estimating the coefficients c_j and the consequent application of classical test procedures to the estimated sample distribution assumes that the sample measurements are independent. In practice, the selection of units to the sample is not carried out independently, with the joint selection probabilities possibly related to the values of the target response variables. As a result, the independence of the sample measurements could be distorted, depending on the sampling design and the sampling rates.

Studying the effect of the sampling scheme on the interdependence of sample measurements is complicated since the sample inclusion probabilities often depend on several design variables, some of which may not be known to the analyst. Moreover, for most of the sampling designs in common use, only the second order inclusion probabilities, $\pi_{ij} = \Pr[(i,j) \in S]$, can be computed systematically. When the relationship between these probabilities and the response variable values is known, the effect of the sampling scheme on the mutual dependence of pairs of measurements can be assessed. See the Appendix for such an analysis in the case of systematic probability proportional to size (PPS) sampling.

For the case of PPS sampling *with replacement*, asymptotic independence of the sample measurements can be established theoretically under mild conditions on the population distribution. The asymptotic analysis assumes fixed sample size with the population size increasing to infinity. These results extend to three PPS selection methods without replacement in common use. See Pfeffermann, Krieger, and Rinott (1997) for details. The independence of the sample measurements under different methods of selection *without replacement* and given population distributions and sampling designs can be assessed by simulations.

This can be carried out as follows: If the variable of interest is discrete with possible values $y_1, ..., y_D$, one can calculate from the simulated samples the fraction of times that values $y_i, i = 1, ...D$, pairs $(y_i, y_j), 1 \leq i \leq j \leq D$ etc., are observed. These fractions can then be compared to the probabilities implied by the independence assumption. This approach is, however, not computationally feasible when D is large and it also does not apply if the variable of interest is continuous. It is more natural (and important) to consider, in such cases, sample statistics of interest (e.g., the likelihood function) and compare the sampling distribution of these statistics with the distribution implied under independent sampling from the extracted sample distribution. The latter can again be assessed by simulations if it is not feasible theoretically.

A simulation study along these lines, considering a variety of population distributions and sampling designs has been carried out by Pfeffermann, Krieger, and Rinott (1997), indicating that for the commonly used sampling designs, the independence of sample measurements is preserved. See also the empirical results in Section 5.

4.2. *Test statistics under proposed approach*

In this section we define four test statistics that can be used in conjunction with the proposed approach. These statistics are considered in the empirical study in the next section.

4.2.1. The Pearson chi-squared test statistic

$$CSQ = \sum_{k=1}^{K} [(n_k - n\tilde{P}_k)^2 / n\tilde{P}_k] \tag{17}$$

where the probabilities $\tilde{P}_k, k = 1...K$ are defined analogously to the definition of the probabilities P_k^* in Section 2.2, but based on the extracted cumulative sample distribution $F_s^*(y; \lambda^*)$ and the n_k are the corresponding observed counts.

4.2.2. The Kolmogorov-Smirnov test statistic

$$K - S = \sup_{y_1 \cdots y_n} [\tilde{F}_s(y_i) - F_s^*(y_i)] \tag{18}$$

where $\tilde{F}_s(t) = \Sigma_{i=1}^{n} D(t - y_i)/n$ is the empirical sample distribution.

4.2.3. The Bonferroni version of the CSQ statistic (see also Section 2.2)

$$BONCS = \max_k \left| \frac{n_k - n\tilde{P}_k}{\sqrt{n\tilde{P}_k(1 - \tilde{P}_k)}} \right| \tag{19}$$

4.2.4. A fourth test procedure considered in the empirical study utilizes the following two properties of distribution functions

(A) For a continuous random variable X with cumulative distribution function F, $F(X) \sim U(0,1)$

(B) Under very general conditions, the set of all the moments of a distribution, when they exist, determine the distribution (Kendall and Stuart Vol. 1, 1973, Chapter 4).

The proposed test procedure consists therefore of the following steps:

(a) Apply the probability integral transformation $T_i = F_s^*(y_i)$, $i = 1 \ldots n$.

(b) Compute the empirical moments $u_m = \Sigma_{i=1}^n T_i^m / n$, $m = 1 \ldots M$.

(c) Compute the large-sample Wald statistic based on the empirical moments in (b).

Note that for the $U(0,1)$ distribution, $\mu_m = E(u_m) = 1/(m+1)$; and $\sigma_{m\ell} = \text{Cov}(u_m, u_\ell) = m\ell/[(m+1)(\ell+1)(m+\ell+1)n]$ for all m and ℓ. Thus, assuming that $\boldsymbol{u}' = (u_1, \ldots, u_M)$ has an approximately normal distribution, under H_0

$$UNIF = (\boldsymbol{u} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{u} - \boldsymbol{\mu}) \rightsquigarrow \chi^2_{(M)} \qquad (20)$$

where $\boldsymbol{\mu}' = (\mu_1, \ldots, \mu_M)$ and $\Sigma = [\sigma_{k\ell}]$, $1 \leq k$, $\ell \leq M$. An important question underlying the use of *UNIF* is the choice of M. Since $\text{corr}^2(u_m, u_{m-1}) = [1 - (1/4m^2)]$, it is evident that high order moments add only marginally to the power of the test. The choice $M = 5$ was found in the empirical study to perform well with respect to both types of errors.

4.3. *Implementation of the procedure in practice*

The implementation of the procedure discussed in Sections 4.1 and 4.2 involves *three steps*:

A Evaluation (approximation) of the conditional expectation $E_M(\pi_i | y_i, \gamma)$. This step is needed for extracting the sample distribution under the null hypothesis.

B Determination of the number of intervals and the interval boundaries used for the various chi-square statistics.

C Specification of the number of moments used for the UNIF statistic.

For single stage sampling with the π_is as measures of size, (PPS sampling), implementation of the first step can be carried out by regressing the π_i against y_i, similarly to the analysis in Section 4.1. See also the empirical study in Section 5. Note that the focus in this article is on continuous distributions, but the ideas follow through to the case of discrete distributions, such as testing cell probabilities of a multinomial distribution.

A different situation arises in a (nonproportional) stratified sample where the sample selection probabilities are fixed within strata. In such cases the conditional expectations $E(\pi_i | y_i; \gamma)$ depend also on the formation (definition) of the strata. Krieger and Pfeffermann (1992) consider the case where the strata are defined based on the ascending values of a design variable Z, assumed to be a function of the target response variables and possibly *other* survey variables. The corresponding sample pdf of the survey variables is then different in different strata, see equation (3.8) of that article. The marginal sample distribution of the response variables can be obtained by integration. (Pfeffermann and Krieger consider the case where the joint population distribution of the survey variables is normal, but the extraction of the sample pdf is not restricted to the normal case.) A special case of this sampling design occurs when the stratification is based directly on the values of the response variable. Two actual surveys of this kind are the ''Gary Income Maintenance Experiment'' – Hausman and Wise (1981) and the ''National Maternal and Infant Health Survey'' – Korn and Graubard (1995).

The case of a multistage sampling design requires a different treatment because the sample selection effects may prevail at any of the selection stages. For example, in the case of a two stage cluster sample, it is often the case that where as the 'clusters', (primary sampling units), are selected with unequal selection probabilities, the ultimate sampling units are selected with equal probabilities. The sample distribution under a two stage cluster sampling design and its use for inference in relation to the mixed variance components models is considered in Pfeffermann, Krieger, and Rinott (1997).

The problems mentioned under step B above are not unique to the proposed procedure. In fact, both the specification of the number of intervals and the interval boundaries are still largely open questions and the reader is referred to Kendall and Stuart (1973, Vol. 2, Chapter 36) for a thorough discussion of these problems. In the empirical study we follow the recommendation made by the two authors and use intervals of equal probabilities under the null hypothesis. We consider two specifications for the number of intervals, $k = 5$ and $k = 10$, which are common choices in other studies.

As for the specification of the number of moments for use of the UNIF statistic, (step C), our experience based on the empirical study is that the choice of $M = 5$ performs well with respect to both types of error. See the discussion below equation (20).

5. Monte Carlo Simulation Results

5.1. Design of the Monte Carlo study

In order to illustrate and compare the performance of the various test statistics defined in Sections 2 and 4, we designed a simulation study by which samples were selected with unequal inclusion probabilities from populations generated randomly from given distributions. Specifically, populations were generated from the distributions

$$f_p(y;\boldsymbol{\theta}) = m \times \text{Gamma}(\alpha_1,\beta) + (1-m) \times U(0, 2\alpha_1/\beta) \tag{21}$$

with $m = 1$, 0.7. Fixing $m = 1$ allows one to assess the performance of the various test statistics under the null hypothesis $H_0 : f_p(y,\boldsymbol{\theta}) = \text{Gamma}(\alpha_1,\beta)$. The specification of $m = 0.7$ is used for power comparisons. Note that $E_M(Y) = (\alpha_1/\beta)$ for all m.

The samples were selected using systematic probability proportional to size (PPS) sampling (Cochran 1977, Section 9A.10) with the size variable, Z, defined as either

$$Z_i = A_0 + A_1 y_i + A_2[G_i - (\alpha_2/\beta)] \tag{22}$$

or

$$Z_i = \exp\{A_0^* + A_1^* y_i + A_2^*[G_i - (\alpha_2/\beta)]\} \tag{23}$$

where $G_i \sim \text{Gamma}(\alpha_2,\beta)$. Note that under (22), $E_M(\pi_i|y_i)$ is linear in y_i so that the approximation in (14) is exact with the corresponding sample distribution obtained from (16). The relationship (23) is considered in order to assess the robustness of the polynomial approximation (14) to the expectations $E_M(\pi_i|y_i;\gamma)$. The specification of the highest power J in the approximation has been carried out using a forward regression stepwise algorithm applied to the sample measurements, $\{\pi_i, y_i; i = 1, \ldots, n\}$. The regression coefficients and their randomization variances were estimated by probability weighted estimators, thus accounting for the sample selection effects (see Section 4.1).

The estimation of the randomization variances of the regression coefficients and the randomization variances and covariances of the estimators $\hat{P}_k$ of the interval probabilities, used for the construction of the design-based test statistics, (Section 2.2), requires the use of the second order inclusion probabilities $\pi_{ij} = \Pr(i,j \in S)$. These probabilities are unknown for the systematic PPS sampling scheme and we therefore used the approximation developed by Hartley and Rao (1962) which is of order $O(N^{-3})$, where N is the population size. Using this approximation, it is shown in the Appendix that under the systematic PPS sampling scheme

$$f_s(y_i, y_j; \lambda) = f(y_i, y_j | i, j \in S) = f_s(y_i; \lambda) f_s(y_j; \lambda) + O(N^{-1}), \qquad (24)$$

establishing an approximate independence of pairs of sample measurements. See also Pfeffermann, Krieger, and Rinott (1997) for simulation results illustrating an overall independence of the sample measurements.

5.2. *Results*

The results reported in this section are each based on four populations of size $N = 5{,}000$ and 150 samples selected independently from each population. We consider two sample sizes, $n = 300$ and $n = 500$. The size values $\{z_i, i = 1, \ldots, 5{,}000\}$ have been randomly ordered before each sample selection. The corresponding parameter values are:

$\alpha_1 = 2$, $\beta = 1 \rightarrow$ for the distribution of Y in the population, (equation 21), $A_0 = A_1 = 2$, $A_2 = 1$, $\alpha_2 = 2 \rightarrow$ for the case where Z_i is linearly related to Y_i, (equation 22), $A_0^* = -2$, $A_1^* = 0.25$, $A_2^* = 0.25$, $\alpha_2 = 2 \rightarrow$ for the case of a logistic relationship (equation 23). The number of intervals used for the construction of the design-based test statistics (Section 2) and the chi-squared statistic (equation 17) is either $K = 5$ (Tables 1–4) or $K = 10$ (Tables 5–8). In both cases the interval boundaries were determined such that

Table 1. Proportion of significant results, m = 1, linear relationship, K = 5

Nominal levels	$W^2(H\text{–}T)$	$W_M^2(H\text{–}T)$	$W_S^2(H\text{–}T)$	$W^2(D)$	$W_M^2(D)$	$W_S^2(D)$
0.15	0.242	0.213	0.207	0.208	0.175	0.135
0.10	0.180	0.167	0.155	0.155	0.128	0.083
0.05	0.102	0.092	0.078	0.077	0.073	0.027
0.025	0.065	0.048	0.040	0.040	0.045	0.010
0.01	0.035	0.033	0.017	0.020	0.018	0.003
	$W^2(R)$	$W_M^2(R)$	$W_S^2(R)$	CSQ	*K-S*	UNIF
0.15	0.225	0.348	0.290	0.175	0.152	0.180
0.10	0.172	0.295	0.218	0.112	0.115	0.125
0.05	0.107	0.227	0.140	0.067	0.068	0.060
0.025	0.063	0.170	0.102	0.037	0.032	0.033
0.01	0.023	0.127	0.082	0.012	0.012	0.020

Table 2. Proportion of significant results, $m = 1$, logistic relationship, $K = 5$

Nominal levels	$W^2(H–T)$	$W^2_M(H–T)$	$W^2_S(H–T)$	$W^2(D)$	$W^2_M(D)$	$W^2_S(D)$
0.15	0.200	0.177	0.173	0.192	0.177	0.147
0.10	0.140	0.122	0.108	0.137	0.123	0.083
0.05	0.075	0.055	0.053	0.083	0.067	0.031
0.025	0.040	0.037	0.032	0.042	0.035	0.015
0.01	0.023	0.017	0.013	0.018	0.017	0.002
	$W^2(R)$	$W^2_M(R)$	$W^2_S(R)$	CSQ	*K-S*	UNIF
0.15	0.210	0.170	0.138	0.178	0.173	0.168
0.10	0.148	0.127	0.078	0.125	0.102	0.100
0.05	0.088	0.073	0.048	0.062	0.065	0.055
0.025	0.052	0.055	0.035	0.032	0.028	0.032
0.01	0.030	0.033	0.027	0.008	0.008	0.012

the theoretical interval probabilities under H_0 are equal, ($P_k \equiv 0.2$ for 5 intervals and $P_k = 0.1$ for 10 intervals). The number of moments used for the test statistic UNIF (equation 20) is 5.

In Tables 1–8 we show the proportion of significant results as obtained for the various test statistics under the linear and logistic relationships. Tables 1–2 and 5–6 correspond to the case where $m = 1$, i.e., when the population data were generated from Gamma(α_1, β) as hypothesized under H_0. Tables 3–4 and 7–8 correspond to the case where $m = 0.7$ in which case the null hypothesis is incorrect. The proportion of significant results were calculated for five critical values $C(\alpha)$; $\alpha = 0.15, 0.10, 0.05, 0.025, 0.01$ of the corresponding distributions of the test statistics under the null hypothesis. Thus, for the case of $m = 1$ they estimate the significance levels of the test statistics whereas for $m = 0.7$ they estimate the respective powers.

The test statistics are denoted in the eight tables as follows:

$W^2(H - T)$, $W^2(D)$, $W^2(R)$ – The design-based Wald statistics (equation (3) with the interval probabilities estimated by the Horvitz-Thompson estimator (equation 8), the

Table 3. Proportion of significant results, $m = 0.7$, linear relationship, $K = 5$

Nominal levels	$W^2(H–T)$	$W^2_M(H–T)$	$W^2_S(H–T)$	$W^2(D)$	$W^2_M(D)$	$W^2_S(D)$
0.15	0.275	0.220	0.203	0.387	0.182	0.147
0.10	0.200	0.157	0.137	0.288	0.133	0.087
0.05	0.127	0.080	0.070	0.200	0.075	0.027
0.025	0.088	0.058	0.050	0.138	0.033	0.013
0.01	0.050	0.022	0.015	0.100	0.017	0.003
	$W^2(R)$	$W^2_M(R)$	$W^2_S(R)$	CSQ	*K-S*	UNIF
0.15	0.352	0.362	0.320	0.837	0.565	0.710
0.10	0.255	0.298	0.245	0.783	0.430	0.625
0.05	0.175	0.218	0.148	0.675	0.258	0.502
0.025	0.113	0.162	0.132	0.575	0.165	0.397
0.01	0.080	0.138	0.108	0.468	0.085	0.270

Table 4. Proportion of significant results, m = 0.7, logistic relationship, K = 5

Nominal levels	$W^2(H{-}T)$	$W^2_M(H{-}T)$	$W^2_S(H{-}T)$	$W^2(D)$	$W^2_M(D)$	$W^2_S(D)$
0.15	0.283	0.248	0.243	0.265	0.175	0.130
0.10	0.217	0.185	0.180	0.195	0.113	0.068
0.05	0.138	0.110	0.100	0.115	0.062	0.027
0.025	0.083	0.063	0.057	0.072	0.030	0.008
0.01	0.047	0.022	0.020	0.032	0.010	0.000
	$W^2(R)$	$W^2_M(R)$	$W^2_S(R)$	CSQ	*K-S*	UNIF
0.15	0.265	0.222	0.188	0.550	0.410	0.648
0.10	0.203	0.158	0.127	0.445	0.275	0.573
0.05	0.120	0.103	0.083	0.325	0.152	0.437
0.025	0.072	0.078	0.065	0.245	0.080	0.327
0.01	0.043	0.058	0.052	0.167	0.042	0.222

Difference estimator (equation 9) and the Ratio estimator (equation 10), respectively. The values of the auxiliary variable X, used for the construction of the Difference and Ratio estimators were taken to be the corresponding values of the design variable Z, defined in equations (22) and (23).

$W^2_M(H\text{-}T)$, $W^2_M(D)$, $W^2_M(R)$ – The modified chi-squared statistics of Rao and Scott (1981), defined by equations (4) and (5), with the interval probabilities estimated by the Horvitz-Thompson estimator, the Difference estimator and the Ratio estimator, respectively.

$W^2_S(H\text{-}T)$, $W^2_S(D)$, $W^2_S(R)$ – The modified chi-squared statistics of Rao and Scott (1981), defined by equations (4) and (6), with the interval probabilities estimated by the Horvitz-Thompson estimator, the Difference estimator and the Ratio estimator, respectively.

CSQ – The Pearson chi-squared statistic (equation 17).

K-S – The Kolmogorov-Smirnov statistic (equation 18).

UNIF – The uniform moments based test statistic (equation 20).

To save space we omit from the tables the results obtained for the test statistics based on the Bonferroni probability bounds (equations 7 and 19). As expected, the use of these test

Table 5. Proportion of significant results, m = 1, linear relationship, K = 10

Nominal levels	$W^2(H{-}T)$	$W^2_M(H{-}T)$	$W^2_S(H{-}T)$	$W^2(D)$	$W^2_M(D)$	$W^2_S(D)$
0.15	0.290	0.233	0.203	0.233	0.100	0.057
0.10	0.230	0.172	0.150	0.168	0.057	0.020
0.05	0.150	0.102	0.075	0.098	0.030	0.007
0.025	0.102	0.065	0.048	0.063	0.017	0.003
0.01	0.058	0.043	0.025	0.027	0.003	0.000
	$W^2(R)$	$W^2_M(R)$	$W^2_S(R)$	CSQ	*K-S*	UNIF
0.15	0.305	0.228	0.168	0.173	0.152	0.180
0.10	0.248	0.203	0.127	0.127	0.115	0.125
0.05	0.150	0.153	0.087	0.070	0.068	0.060
0.025	0.095	0.128	0.068	0.032	0.032	0.033
0.01	0.058	0.117	0.047	0.012	0.012	0.020

Table 6. Proportion of significant results, m = 1, logistic relationship, K = 10

Nominal levels	$W^2(H\text{–}T)$	$W^2_M(H\text{–}T)$	$W^2_S(H\text{–}T)$	$W^2(D)$	$W^2_M(D)$	$W^2_S(D)$
0.15	0.237	0.192	0.187	0.247	0.100	0.050
0.10	0.182	0.127	0.115	0.180	0.045	0.025
0.05	0.105	0.067	0.065	0.097	0.020	0.005
0.025	0.065	0.035	0.033	0.055	0.005	0.002
0.01	0.040	0.020	0.017	0.027	0.003	0.000
	$W^2(R)$	$W^2_M(R)$	$W^2_S(R)$	CSQ	*K-S*	UNIF
0.15	0.273	0.073	0.053	0.150	0.173	0.168
0.10	0.205	0.055	0.042	0.102	0.102	0.100
0.05	0.143	0.045	0.025	0.050	0.065	0.055
0.025	0.085	0.033	0.023	0.027	0.028	0.032
0.01	0.053	0.027	0.022	0.015	0.008	0.011

statistics is very conservative with much too low rejection probabilities in all cases.

The main conclusions from the eight tables are as follows:

1) The three design-based Wald statistics $W^2(H\text{-}T)$, $W^2(D)$ and $W^2(R)$ perform poorly under H_0, yielding proportions of significant results that are way too high in all cases. The powers of these statistics are high in the case of ten intervals but very low in the case of five intervals.
2) The modified chi-squared statistics of Rao and Scott yield proportions of significant results that are close to the nominal levels in the case of the Horvitz-Thompson and the Difference estimators and five intervals, but not in the case of ten intervals where the proportions of significant results of $W^2_M(H\text{-}T)$ and $W^2_S(H\text{-}T)$ are in most cases still too high and those of $W^2_M(D)$ and $W^2_S(D)$ are always way too low. The statistics $W^2_M(R)$ and $W^2_S(R)$ show an even more erratic behavior, yielding extremely high proportions of significant results in the linear case with five intervals and

Table 7. Proportion of significant results, m = 0.7, linear relationship, K = 10

Nominal levels	$W^2(H\text{–}T)$	$W^2_M(H\text{–}T)$	$W^2_S(H\text{–}T)$	$W^2(D)$	$W^2_M(D)$	$W^2_S(D)$
0.15	0.748	0.487	0.452	0.753	0.252	0.163
0.10	0.678	0.403	0.358	0.678	0.172	0.072
0.05	0.577	0.278	0.227	0.567	0.067	0.022
0.025	0.480	0.193	0.142	0.477	0.032	0.005
0.01	0.355	0.122	0.080	0.343	0.015	0.000
	$W^2(R)$	$W^2_M(R)$	$W^2_S(R)$	CSQ	*K-S*	UNIF
0.15	0.708	0.377	0.287	0.788	0.565	0.710
0.10	0.632	0.323	0.217	0.718	0.430	0.625
0.05	0.522	0.272	0.165	0.597	0.258	0.502
0.025	0.413	0.232	0.138	0.505	0.165	0.397
0.01	0.308	0.203	0.102	0.383	0.085	0.270

Table 8. Proportion of significant results, $m = 0.7$, logistic relationship, $K = 10$.

Nominal levels	$W^2(H{-}T)$	$W_M^2(H{-}T)$	$W_S^2(H{-}T)$	$W^2(D)$	$W_M^2(D)$	$W_S^2(D)$
0.15	0.733	0.592	0.577	0.617	0.238	0.153
0.10	0.668	0.495	0.488	0.543	0.155	0.085
0.05	0.548	0.370	0.362	0.403	0.075	0.020
0.025	0.440	0.275	0.252	0.285	0.025	0.003
0.01	0.352	0.177	0.150	0.207	0.008	0.000
	$W^2(R)$	$W_M^2(R)$	$W_S^2(R)$	CSQ	*K-S*	UNIF
0.15	0.550	0.202	0.133	0.543	0.410	0.648
0.10	0.470	0.158	0.080	0.450	0.275	0.573
0.05	0.345	0.095	0.043	0.345	0.152	0.437
0.025	0.238	0.063	0.023	0.253	0.080	0.327
0.01	0.150	0.037	0.017	0.153	0.042	0.222

extremely low proportions in the logistic case with ten intervals. The statistic $W_S^2(R)$ performs well under H_0 in the other two cases.

The most striking outcome regarding the use of the modified chi-squared statistics is their very low powers in all the cases considered. These powers are lower, and in many cases much lower, than the powers of the corresponding design-based Wald statistics.

3) The overall poor performance of the design-based test statistics is mainly attributed to the erratic behavior of the estimated randomization V-C matrices $\hat{C}_R$ of the interval probabilities estimators. These matrices are occasionally ill conditioned, although when comparing their arithmetic mean over the 600 simulations with the corresponding empirical V-C matrices calculated from these simulations, the differences are in most cases very mild and do not indicate any systematic biases. We ran the same simulations with sample size $n = 500$. Increasing the sample size increases the power of all the design based test statistics, (the powers are still very low for the case of five intervals) but has little effect on their behavior under H_0. To save space we do not report the detailed results obtained for the larger sample sizes.

   Another possible explanation for the overall bad performance of the design-based test statistics is occasional biases of the estimators $\hat{P}_{k,H-T}$, $\hat{P}_{k,D}$ and $\hat{P}_{k,R}$ of the interval probabilities. For example, for the case of five intervals and the linear relationship, ($m = 1, n = 300$), ten out of the 15 differences $[Av(\hat{P}_k) - 0.2]$ are positive and most of the differences are larger than twice the corresponding standard errors. (The averages are over the 600 samples.) The biases are, however, very small and except for the first interval where for the Ratio estimator $Av(\hat{P}_{1,R}) = 0.212$ with standard error of 0.003, for all other estimators and intervals $[Av(\hat{P}_k) - 0.2| \leq 0.006$.

4) In contrast to the use of the design-based test statistics, the use of the standard test statistics for testing the extracted hypothesized sample distribution, as described in Section 4, seems to perform well in the simulation study. For all three test statistics considered, the proportions of significant results under H_0 are sufficiently close to the nominal levels and except in a few cases, the differences are not significant. (For a given nominal level α, the standard error of the differences is $[\alpha(1-\alpha)/600]^{1/2}$.)

The three statistics perform equally well in the case of a linear relationship and the case of a logistic relationship, indicating that the use of the polynomial approximation to the conditional expectation of the inclusion probabilities defined by (14) works well in this case. We mention also that the differences between the empirical and nominal significance levels of the three estimators become even smaller, under both relationships, when increasing the sample size to $n = 500$.

5) The powers of CSQ and UNIF are high in the case of a linear relationship but lower for the logistic relationship. The powers of all three test statistics increase, however, quite substantially when increasing the sample size to $n = 500$. For the logistic relationship with a nominal significance level under H_0 of 0.15, for example, the powers of CSQ, K-S and UNIF are 0.738, 0.678 and 0.857, respectively. The powers for the other nominal levels increase accordingly. (Unlike the case of the design based Wald statistics, the powers of CSQ do not increase when increasing the number of intervals from five to ten.)

The relatively high powers of UNIF are quite surprising. This statistic yields higher powers than K-S in all the cases and it outperforms even CSQ in the case of the logistic relationship. (CSQ yields higher powers in the case of a linear relationship.) The relatively low powers of K-S on the other hand are not surprising, as this property of the K-S is known from other empirical studies.

6. Concluding Remarks

The results of our study illustrate that it is possible in principle to extract the distribution of measurements, for units selected with unequal selection probabilities which are related to the response variable. The prominent advantage of extracting the sample distribution is that it permits the use of efficient inference tools like maximum-likelihood estimation and hypothesis testing as considered in the present article.

The practical implementation of the proposed approach requires a specification of the distribution of the population measurements and the modelling of the conditional expectations of the inclusion probabilities, given the values of the response variable. We mention with respect to the first issue that we have assumed in this article that the population distribution is fully specified, including all its parameters. In practice, it is often the case that only the family of distributions is specified under the null hypothesis without further specification of the parameter values. This is a well known problem in the classical theory of hypothesis testing and we plan to investigate the use of various modifications proposed in the literature to deal with this case like, for example, reducing the number of degrees of freedom by the number of estimated parameters in the case of the chi-squared statistic.

The modelling of the conditional expectations of the sample inclusion probabilities is more unique to our approach. The results obtained for the case of a logistic relationship between the size variable and the response variable suggest that the use of a polynomial relationship for modelling the conditional expectation is robust, but this property needs to be investigated further. Notice in this respect that in practice, it is often the case that the selection of the sample is carried out in several stages, with the final inclusion probabilities obtained as products of the selection probabilities at the various stages. As discussed in Pfeffermann, Krieger, and Rinott (1997), approximating the *log* of the conditional expectation by a polynomial relationship is more appropriate in such situations.

Appendix

In this Appendix we establish the approximate independence of pairs of sample measurements under the following conditions:

a. The population distribution, prior to sampling, satisfies

$$f_p(y_i, y_j; \boldsymbol{\theta}) = f_p(y_i; \boldsymbol{\theta}) f_p(y_j; \boldsymbol{\theta})$$

b. The size variable Z, used for the sample selection is of the form $Z_i = g(Y_i, \epsilon_i)$ where the ϵ_i are iid and $Z_i > 0$ for all i.
c. The sample is selected by systematic PPS sampling, (Cochran 1977, Section 9.A.10), such that $\pi_i = \Pr(i \in S) = nZ_i/N\bar{Z}$ where $\bar{Z} = \Sigma_{i=1}^N Z_i/N$ is the population mean
d. Z_i has all moments and $\lim_{N\to\infty} E(1/\bar{Z}^k) = \mu^{-k}$, for all k, where $\mu = E_p(Z_i)$. This condition requires interchange of limit and integral.

Hartley and Rao (1962) show that under condition c and some mild conditions on the first order sample selection probabilities, the joint inclusion probabilities $\pi_{ij} = \Pr(i, j \in S)$ can be approximated as

$$\pi_{ij} = n(n-1)[p_i p_j (1 - p^{(2)}) + p_i p_j (p_i + p_j)] + O(N^{-3}) \qquad (A1)$$

where $p_i = \pi_i/n = Z_i/(N\bar{Z})$ and $p^{(2)} = \Sigma_{i=1}^N p_i^2$

Following similar arguments to (3) and (4) in Section 2, the joint sample density can be written as

$$f_s(y_i, y_j; \lambda) = f(y_i, y_j | i, j \in S) = \frac{E_p(\pi_{ij} | y_i, y_j) f_p(y_i, y_j; \boldsymbol{\theta})}{\int_{y_i, y_j} E_p(\pi_{ij} | y_i, y_j) f_p(y_i, y_j; \boldsymbol{\theta}) dy_i dy_j} \qquad (A2)$$

Substituting (A1) in (A2) (ignoring the last term of equation A1) and multiplying the numerator and denominator by N^2 yields

$$f_s(y_i, y_j; \lambda) = \frac{f_p(y_i, y_j; \boldsymbol{\theta}) E_p\left(\frac{Z_i Z_j}{\bar{Z}^2} \middle| y_i, y_j\right)}{\int_{y_i, y_j} f_p(y_i, y_j; \boldsymbol{\theta}) E_p\left(\frac{Z_i Z_j}{\bar{Z}^2} \middle| y_i, y_j\right) dy_i dy_j} + O(N^{-1}) \qquad (A3)$$

It can be shown that under condition d,

$$E_p\left(\frac{Z_i Z_j}{\bar{Z}^2} \middle| y_i, y_j\right) = E_p(Z_i Z_j | y_i, y_j)/\mu^2 + O(N^{-1}) \qquad (A4)$$

Denoting $E_p(Z_i | y_i) = m(y_i)$, it follows from (A4) and conditions a and b that

$$\begin{aligned} f_s(y_i, y_j; \boldsymbol{\theta}) &= \frac{f_p(y_i; \boldsymbol{\theta}) m(y_i)}{\int_{y_i} f_p(y_i; \boldsymbol{\theta}) m(y_i) dy_i} \cdot \frac{f_p(y_j; \boldsymbol{\theta}) m(y_j)}{\int_{y_j} f_p(y_j; \boldsymbol{\theta}) m(y_j) dy_j} + O(N^{-1}) \\ &= f_s(y_i; \lambda) f_s(y_j; \lambda) + O(N^{-1}) \qquad (A5) \end{aligned}$$

7. References

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review, 51, 279–292.

Chambers, R.L. and Dunstan, R. (1986). Estimating Distribution Function from Survey Data. Biometrika, 73, 597–604.

Cochran, W.G. (1977). Sampling Techniques (third edition). New York, Wiley.

Fuller, W.A. (1975). Regression Analysis for Sample Surveys. Sankhyā, Series C, 37, 117–132.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with Unequal Probabilities and Without Replacement. Annals of Mathematical Statistics, 33, 350–374.

Hausman, J.A. and Wise, D.A. (1981). Stratification on Endogeneous Variables and Estimation; The Gary Income Maintenance Experiment. In Structure Analysis of Discrete Data with Econometric Applications, (eds. C.F. Mansky and D. McFadden). Cambridge, MA.: MIT Press, pp. 366–391.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Isaki, C.T. and Fuller, W.A. (1982). Survey Design Under a Regression Superpopulation Model. Journal of the American Statistical Association, 77, 89–96.

Kendall, M.G. and Stuart, A. (1973). The Advanced Theory of Statistics. New York, Hafner.

Korn, E.L. and Graubard, B.I. (1995). Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. The American Statistician, 49, 291–295.

Krieger, A.M. and Pfeffermann, D. (1992). Maximum Likelihood Estimation from Complex Sample Surveys. Survey Methodology, 18, 225–239.

Kuk, A.C. (1993). A Kernel Method for Estimating Finite Population Distribution Functions Using Auxiliary Information. Biometrika, 80, 385–392.

Kuo, L. (1988). Classical and Prediction Approaches to Estimating Distribution Functions from Survey Data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 280–285.

Mahalanobis, P.C. (1936). On the Generalized Distance in Statistics. Proceedings of the National Institute of Science of India, 12, 49–55.

Nascimento Silva, P.L.D. and Skinner, C.J. (1995). Estimating Distribution Functions with Auxiliary Information Using Poststratification. Journal of Official Statistics, 11, 277–294.

Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. International Statistical Review, 61, 317–337.

Pfeffermann, D., Krieger, A.M., and Rinott, Y. (1997). Parametric Distributions of Complex Survey Data Under Informative Probability Sampling. (Revised report to be published in Statistica Sinica.)

Pratt, J.W. and Gibbons, J.D. (1981). Concepts of Nonparametric Theory. New York, Springer-Verlag.

Rao, J.N.K., Kovar, J.G., and Mantel, H.J. (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information. Biometrika, 77, 365–375.

Rao, J.N.K. and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. Journal of the American Statistical Association, 76, 221–230.

Smith, T.M.F. (1988). To Weight or Not to Weight, That is the Question. In Bayesian Statistics 3, J.M. Bernardo, M.H. Degfoot, D.V. Lindley, and A.F.M. Smith (eds.). Oxford University Press, 437–451.

Stroud, T.W.F. (1971). On Obtaining Large Sample Tests from Asymptotically Normal Estimates. Annals of Mathematical Statistics, 42, 1412–1424.

Thomas, D.R. and Rao, J.N.K. (1987). Small-Sample Comparison of Level and Power for Simple Goodness of Fit Statistics Under Cluster Sampling. Journal of the American Statistical Association, 82, 630–636.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations Is Large. Transactions of the American Mathematical Society, 54, 426–482.

Received May 1995
Revised May 1996