C. J. Skinner and N.Shlomo

Estimating frequencies of frequencies in finite populations

Article (Accepted version)
(Refereed)

Original citation:
Skinner, Chris J. and Shlomo, N. (2012) *Estimating frequencies of frequencies in finite populations. Statistics & probability letters*, 82 (12). pp. 2206-2212. ISSN 0167-7152

DOI: 10.1016/j.spl.2012.07.023

Estimating Frequencies of Frequencies in Finite Populations

C. J. Skinner and N.Shlomo

28 July 2012

Abstract

Given a sample from a finite population partitioned into classes, we consider estimating the distribution of the class frequencies. We propose first to estimate certain moments of this distribution, assuming Poisson sampling with unequal inclusion probabilities, and then to adapt these estimates using modelling assumptions. A simulation study illustrates the bias-robustness of the approach to departures from these assumptions.

1 Introduction

If a finite population is partitioned into classes, the frequency distribution of the class frequencies is sometimes called the frequency of frequencies distribution (Good, 1953; Bishop et al., 1975, sect. 9.8). If class membership is only observed for a sample from the population, various inferential problems arise. We focus on the case when the number of classes is large and where the frequencies of many of these classes may be small. A widely studied inferential problem in this setting is how to estimate the number of classes when this is unknown (Goodman, 1949; Bunge and Fitzpatrick, 1993). In this paper we consider the different problem of estimating the frequencies of frequencies from a sample when the number of classes is known, as for the case when the classes are formed by cross-classifying several discrete variables, each with a known number of categories.

Motivation for this estimation problem comes from statistical disclosure risk assessment in the release of survey microdata, where there is concern about the possible identification of individuals through rare combinations of discrete variables which could be used to link microdata records to external information (Bethlehem et al., 1990; Skinner and Shlomo, 2008). In this setting, classes are defined by the combinations of values of the variables. The single individual in a class with frequency one is unique in the population and the survey record for this individual could be identified with certainty if matched exactly to a known individual using these variables, assuming no misclassification of the variables. An individual in a class with frequency r could also be identified with probability $1/r$ from such a match. The more classes there are with such small counts the greater is likely to be the concern. The numbers of classes in the population with small frequencies of 1, 2 or 3, say, are sometimes therefore used to measure the 'risk' of identification. Since these numbers are generally unobserved, there is interest in estimating them from sample-based data from the survey.

Standard design-based estimators of population totals from survey sampling may be used to estimate population frequencies when the number of classes is small. However, the application of such approaches to estimating frequencies of frequencies breaks down as the number of classes increases relative to the sample size and when increasingly many of the sample class frequencies become small. Under simple random sampling without replacement, Goodman (1949, Theorem 4) showed that it is possible to obtain a unique design-unbiased estimator of the population frequencies of frequencies even in this case, provided that the sample size is at least as large as the maximum population class frequency. However, he and subsequent authors (Bunge and Fitzpatrick, 1993) found that the sum of these estimators, which estimates the total number of classes, tends to have a very high variance. Such a purely design-based approach to our problem does not seem promising and we shall not pursue it further.

A purely model-based approach is more straightforward, at least if we may assume the population class frequencies obey a compound Poisson distribution of known parametric form and Bernoulli sampling is employed so that the sample class frequencies also obey a compound Poisson distribution with the mixing distribution rescaled by the sampling probability. Bethlehem et al. (1990) proposed to estimate the number of population uniques under a Poisson-gamma model in the context of statistical disclosure control. They expressed the expected number of population uniques under the model in terms of the model parameters and then estimated these by the method of moments. Their approach may be extended to the estimation of the frequencies of other population frequencies. It is of concern, however, that such an approach will be sensitive to the assumption about the mixing distribution.

In this paper we propose a hybrid model/design-based approach. We show that features of the frequency of frequencies distribution can be estimated robustly in a design-based way. We then propose to use the model for estimating the residual aspects of the distribution not captured by these features. Our approach still depends on the specification of a parametric mixing distribution, judged to be realistic, but aims to be more robust to departures from this assumption than a purely parametric model-based approach. Another new feature of our approach is that it handles unequal probability sampling. We present simulation evidence regarding the relative robustness of the proposed approach.

Our focus is on point estimation, viewing robustness primarily in terms of limiting model misspecification bias. Given the importance of such bias relative to standard errors, we do not attempt in this paper to develop statistical inference any further, such as to confidence interval estimation.

2 Preliminaries

Let U denote the set of units in a finite population, partitioned into mutually exclusive classes $C_1, \ldots, C_J$ with $C_1 \cup \ldots \cup C_J = U$. Let the population frequency in class C_j and the size of the population be denoted $F_j =\mid C_j \mid$ and $N =\mid U \mid$, respectively. The frequency of frequency r is defined as $N_r = \sum_{j=1}^{J} I(F_j = r)$, for $r = 0, 1, 2, \ldots$, where $I(\cdot)$ is the indicator function. Note that

$$\sum_{r=0}^{\infty} N_r = J, \quad \sum_{j=1}^{J} F_j = \sum_{r=1}^{\infty} r N_r = N. \tag{1}$$

Suppose that a sample $s \subset U$ of size n is drawn from U by a probability sampling design, where unit $i \in U$ is included in the sample with probability π_i. We shall mainly assume a Poisson design (Hájek, 1981). Let $C_{js} = C_j \cap s$ and $f_j =\mid C_{js} \mid$ denote the set of sample units and the sample frequency, respectively, in class C_j and let $n_r = \sum_{j=1}^{J} I(f_j = r)$ be the frequency of sample frequency r. Corresponding to (1), we have

$$\sum_{r=0}^{\infty} n_r = J, \quad \sum_{j=1}^{J} f_j = \sum_{r=1}^{\infty} r n_r = n. \tag{2}$$

Suppose that class membership is observed for units $i \in s$ and, thus, that the values n_r for $r = 1, 2, \ldots$ are known. As noted in the introduction, we assume that J is known and so n_0 is also known from the first equation in (2). Since class membership is not observed for unsampled units, the values N_r are generally unknown.

We take the primary problem to be the estimation of the frequencies of frequencies N_r for small positive values of r, such as 1, 2, 3 or 4. As a preliminary stage to constructing an estimator of N_r, we consider the estimation of moments of the class frequencies F_j.

3 Design-Based Estimation of Moments of Class Frequencies

We first consider estimation of the first two finite population moments of the class frequencies $J^{-1}\sum_{j=1}^{J} F_j$ and $J^{-1}\sum_{j=1}^{J} F_j^2$. Let the second order sample inclusion probabilities be denoted $\pi_{ik} = pr(i \in s, k \in s)$, where π_{ik} reduces to π_i if $i = k$. Design-unbiased estimators of $J^{-1}\sum_{j=1}^{J} F_j$ and $J^{-1}\sum_{j=1}^{J} F_j^2$ are given in the following lemma, subject to multiplication by the known value of J^{-1}.

Lemma 3.1 *The estimators* $\sum_{i\in s} \pi_i^{-1}$ *and* $\sum_j \left(\sum_{i\in C_{js}} \sum_{k\in C_{js}} \pi_{ik}^{-1}\right)$ *are design-unbiased for* $\sum_{j=1}^{J} F_j$ *and* $\sum_{j=1}^{J} F_j^2$ *respectively.*

The proof is straightforward, following, for example, Result 2.8.3 of Särndal et al. (1992). Note that under Poisson sampling $\pi_{ik} = \pi_i \pi_k$ if $i \neq k$ and $\pi_{ik} = \pi_i$ if $i = k$ so both estimators are well-defined for this design provided $\pi_i > 0$ for each $i \in U$. This condition will be assumed for the remainder of this section.

We next consider estimating the first two conditional moments of the population class frequencies F_j among classes with a given sample frequency f_j, that is we consider estimating $n_r^{-1} \sum_{j \in D_{s,r}} F_j$ and $n_r^{-1} \sum_{j \in D_{s,r}} F_j^2$, where $D_{s,r} = \{j : f_j = r\}$ is the set of indices of classes containing r sample units, for $r = 0, 1, 2, \ldots$ and $\mid D_{s,r} \mid = n_r$. For the first moment with $r = 1$, the problem reduces to estimating the number of population units in classes which are sample unique. This quantity is of some interest in disclosure risk assessment and Skinner and Elliot (2002) and Skinner and Carter (2003) showed that it may be estimated in a design-unbiased way for the cases of Bernoulli and Poisson sampling respectively. We shall now generalize their results. The notion of design-unbiasedness here is non-standard since the conditional moments are sample-dependent. We say that an estimator $\hat{\theta}(s)$ of a sample-dependent estimand $\theta(s)$ is design-unbiased if $E\{\hat{\theta}(s) - \theta(s)\} = 0$, where the expectation is across samples generated by the probability design. To derive our results, it will be mathematically convenient to transform the conditional moments to

$$\mu_{r1} = \sum_{j \in D_{s,r}} (F_j - r), \mu_{r2} = \sum_{j \in D_{s,r}} (F_j - r - 1)(F_j - r). \tag{3}$$

We propose the following estimators of these quantities

$$\hat{\mu}_{r1} = \sum_{j \in D_{s,r+1}} \sum_{i \in C_{js}} (\pi_i^{-1} - 1), \hat{\mu}_{r2} = \sum_{j \in D_{s,r+2}} \sum_{i \in C_{js}} \sum_{k \in C_{js}, k \neq i} (\pi_i^{-1} - 1)(\pi_k^{-1} - 1).$$

Their design-unbiasedness is established in the following two theorems, the proofs of which are given in Appendix A.

Theorem 3.1 *Under Poisson sampling,* $E(\hat{\mu}_{r1}) = E(\mu_{r1})$ *for* $r = 1, 2, \ldots$.

Remark 3.1 To aid interpretation, $\hat{\mu}_{r1}$ may be expressed as $\hat{\mu}_{r1} = (\bar{d}_{r+1} - 1)(r+1)n_{r+1}$, where $\bar{d}_{r+1} = (\sum_{j \in D_{s,r+1}} \sum_{i \in C_{js}} \pi_i^{-1})/\{(r+1)n_{r+1}\}$ is the mean design weight across sample units in classes with sample frequency $r+1$.

Remark 3.2 A curious feature of $\hat{\mu}_{r1}$ is that it uses data from classes with sample frequency $r+1$ to estimate a characteristic of a disjoint set of classes, those with sample frequency r.

Remark 3.3 The estimator $\hat{\mu}_{r1}$ respects the constraint that μ_{r1} is bounded below by zero since $\pi_i^{-1} \geq 1$ for all i. Furthermore, $\hat{\mu}_{r1}$ has the same aggregation relationship with the Horvitz-Thompson estimator $\hat{N} = \sum_{i \in s} \pi_i^{-1}$ of $\sum F_j = N$ in Lemma 1, as μ_{r1} has with N, that is, we may write:

$$\sum_{r=0}^{\infty} \hat{\mu}_{r1} = \hat{N} - n, \qquad \sum_{r=0}^{\infty} \mu_{r1} = N - n. \tag{4}$$

The first expression in (4) follows by straightforward derivation. The second expression follows from (2) and (3).

Remark 3.4 In the case of Bernoulli sampling when $\pi_i = \pi$, $\hat{\mu}_{r1}$ reduces to $(\pi^{-1} - 1)(r+1)n_{r+1}$, which generalizes Proposition 2 of Skinner and Elliot (2002). The implied estimator of $n_r^{-1} \sum_{j \in D_{s,r}} F_j$ is $r + (\pi^{-1} - 1)(r+1)n_{r+1}/n_r$. Multiplying by $\pi \approx n/N$, this is closely related to the formula $(r+1)n_{r+1}/n_r$ which (Good, 1953, equation (2)) presents as an approximate conditional expectation of $(n/N)F_j$ given $f_j = r$. A difference, however, is that Good (1953) defines the expectation with respect to a class j drawn with equal probability from all classes, whereas in our Bernoulli set-up it is the units which are drawn with equal probabilities.

Remark 3.5 In the case of Poisson sampling, Theorem 3.1 generalizes a result of Skinner and Carter (2003) for $r = 1$.

Theorem 3.1 *Under Poisson sampling,* $E(\hat{\mu}_{r2}) = E(\mu_{r2})$ *for* $r = 0, 1, 2, \ldots$.

Remark 3.6 The analogy of the curious feature noted below Theorem 3.1, is that $\hat{\mu}_{r2}$ uses data from the classes with sample frequency $r+2$ to estimate a characteristic of a disjoint set of classes, those with sample frequency r. The estimator $\hat{\mu}_{r2}$ also respects the constraint that μ_{r2} is bounded below by zero since $\pi_i^{-1} \geq 1$.

Remark 3.7 Using the results of Theorems 3.1 and 3.2, it may be shown that a design-unbiased estimator of $\sum_{j \in D_{s,r}} F_j^2$ is given by $\hat{\mu}_{r2} + (2r+1)\hat{\mu}_{r1} - r(r+1)n_r$ and that this estimator sums over $r = 0, 1, 2, \ldots$ to give the estimator of $\sum_{j=1}^{J} F_j^2$ in Lemma 1 for the case of Poisson sampling.

Remark 3.8 In the case of Bernoulli sampling when $\pi_i = \pi$, the simpler formula $\hat{\mu}_{r2} = (r+2)(r+1)(1-\pi)^2 n_{r+2}/\pi^2$ is obtained.

4 Estimation of Frequencies of Frequencies

We now turn to estimation of the frequencies of frequencies N_r. For our proposed method, we first express N_r as

$$N_r = \sum_{t=0}^{r} \sum_{j \in D_{s,t}} I(F_j = r) = \sum_{t=0}^{r} n_t p_t(r),$$

where $p_t(r) = n_t^{-1} \sum_{j \in D_{s,t}} I(F_j = r)$ is the proportion of classes with population frequency r among those classes with sample frequency t. Our proposed estimator of N_r is:

$$\hat{N}_r = \sum_{t=0}^{r} n_t \hat{p}_t(r), \tag{5}$$

where $\hat{p}_t(r)$ is an estimator of $p_t(r)$, to be discussed.

To construct $\hat{p}_t(r)$, note first that the distribution of F_j in $D_{s,t}$ is truncated below by t so that $p_t(r) = 0$ for $r < t$ and we set, correspondingly, $\hat{p}_t(r) = 0$ for $r < t$. We now view $p_t(r)$ as a probability distribution on $r = t, t+1, \ldots$. Its first two moments may be expressed in terms of μ_{t1} and μ_{t2} defined in (3) and, conversely, we may write:

$$\mu_{t1} = n_t \sum_{r=t}^{\infty} (r-t)p_t(r), \mu_{t2} = n_t \sum_{r=t}^{\infty} (r-t-1)(r-t)p_t(r). \tag{6}$$

In section 3 we derived design-unbiased estimators $\hat{\mu}_{t1}$ and $\hat{\mu}_{t2}$ of μ_{t1} and μ_{t2}, respectively. We now propose to estimate $p_t(r)$ by $p_t(r;\hat{\theta}_t)$, where $p_t(r;\theta_t)$ is a parametric form assumed for $p_t(r)$, $\theta_t = (\theta_{t1}, \theta_{t2})$ is a 2-dimensional vector of parameters and $\hat{\theta}_t$ is obtained by solving:

$$\hat{\mu}_{t1} = n_t \sum_{r=t}^{\infty} (r-t)p_t(r;\theta_t), \quad \hat{\mu}_{t2} = n_t \sum_{r=t}^{\infty} (r-t-1)(r-t)p_t(r;\theta_t). \tag{7}$$

For illustration, consider the case where $p_t(r;\theta_t)$ is assumed to have the Poisson-gamma or negative binomial form:

$$p_t(r;\theta_t) = \frac{\Gamma(r-t+\theta_{t2}\theta_{t1})\theta_{t2}^{\theta_{t2}\theta_{t1}}}{(r-t)!\Gamma(\theta_{t2}\theta_{t1})(1+\theta_{t2})^{r-t+\theta_{t2}\theta_{t1}}}, \tag{8}$$

where the parameters θ_{t1} and θ_{t2} are such that the mean and variance of the distribution of $r-t$ are θ_{t1} and $\theta_{t1}(1+\theta_{t2})/\theta_{t2}$ respectively (McCullagh and Nelder, 1989). Hence $\mu_{t1} = n_t\theta_{t1}$ and $\mu_{t2} = n_t[(\theta_{t1}/\theta_{t2}) + \theta_{t1}^2]$. The solutions of (7) are thus given by

$$\hat{\theta}_{t1} = \hat{\mu}_{t1}/n_t, \quad \hat{\theta}_{2t} = \frac{\hat{\mu}_{t1}/n_t}{(\hat{\mu}_{t2}/n_t) - (\hat{\mu}_{t1}/n_t)^2},$$

provided that $n_t > 0$. We plug $\hat{\theta}_{t1}$ and $\hat{\theta}_{t2}$ into $p_t(r;\theta)$ to obtain $\hat{p}_t(r) = p_t(r;\hat{\theta}_t)$. Note that the estimator in (5) does not require $\hat{\theta}_t$ to be defined if $n_t = 0$.

As a reference estimation method for comparison to the proposed method, we consider a purely model-based approach of the kind proposed by Bethlehem et al. (1990), under the assumption of Bernoulli sampling. Suppose that $F_j \mid \lambda_j \sim Poisson(\lambda_j)$ and that λ_j has a gamma distribution with $E(\lambda_j) = \theta_1$ and $\mathrm{var}(\lambda_j) = \theta_1/\theta_2$ so that F_j has a negative binomial distribution with:

$$Pr(F_j = r) = \frac{\Gamma(r+\theta_2\theta_1)\theta_2^{\theta_2\theta_1}}{r!\Gamma(\theta_2\theta_1)(1+\theta_2^{r+\theta_2\theta_1})}, \qquad r = 0,1,2,\ldots \tag{9}$$

Now if Bernoulli sampling with inclusion probability π is employed then $f_j \mid \lambda_j \sim Poisson(\pi\lambda_j)$ and f_j has a negative binomial distribution, as in (9), with the parameters (θ_1, θ_2) replaced by $(\theta_{s1}, \theta_{s2}) = (\pi\theta_1, \theta_2/\pi)$. The first two moments of f_j are thus

$$\mu_1 = E(f_j) = \theta_{s1} = \pi\theta_1, \quad \mu_2 = E[f_j(f_j-1)] = (\theta_{s1}/\theta_{s2}) + \theta_{s1}^2 = \pi^2(\theta_1/\theta_2) + \pi^2\theta_1^2.$$

The method of moments estimators of θ_1 and θ_2 are

$$\hat{\theta} = \hat{\mu}_1/\pi, \quad \hat{\theta}_2 = \hat{\theta}_1/(\pi^{-2}\hat{\mu}_2 - \hat{\theta}_1^2) = (\pi\hat{\mu}_1)/(\hat{\mu}_2 - \hat{\mu}_1^2) \tag{10}$$

where

$$\hat{\mu}_1 = \sum f_j/J = n/J, \quad \hat{\mu}_2 = J^{-1}\sum f_j(f_j-1). \tag{11}$$

The implied estimator of N_r is:

$$\hat{N_r} = \frac{J\Gamma(r+\hat{\theta}_2\hat{\theta}_1)\hat{\theta}_2^{\hat{\theta}_2\hat{\theta}_1}}{r!\Gamma(\hat{\theta}_2\hat{\theta}_1)(1+\hat{\theta}_2)^{r+\hat{\theta}_2\hat{\theta}_1}}, \quad r = 0,1,2,\ldots.$$

5 Simulation Study

We now present a simulation study designed to compare the properties of the pure model-based point estimator of N_r, for small values of r, with our proposed hybrid approach under departures from the basic parametric model, which we take to be the Poisson-gamma model. We represent departures using mixtures of the parametric model and a real population for which this model clearly fails. The real population is obtained from data from the 2001 UK population census on one region with $N = 632,077$ individuals aged 16-65. The classes are taken to be the cells in the six-way cross-classification of (with numbers of categories in parentheses): area (2), sex (2), age group (10), marital status (6), ethnicity (17) by economic activity (10), giving $J = 40,800$ classes. See Skinner and Shlomo (2008) for discussion of the disclosure risk assessment context of this example. We define the basic parametric model which generates population frequencies for these $40,800$ classes as a negative binomial distribution with parameters $\theta_1^{(S)} = 0.000137$ and $\theta_2^{(S)} = 8,928.6$ in (9), obtained by equating $E(F_j)$ for this distribution to $N/J = 15.49$ and equating the expected number of population uniques, $E(N_1)$, under this distribution to the number in the real population, $E(N_1^{(R)})$. Comparison of $E(N_r)$ for this model with real population frequencies $N_r^{(R)}$ for $r \neq 1$ shows clear evidence of lack of fit of the negative binomial. In particular, $N_0^{(R)} = 29,137$ is seriously underfitted since $E(N_0) \approx 16,015$ under the model.

We consider a series of finite populations, which are mixtures of a population generated from the negative binomial model $NB(\theta_1^{(S)}, \theta_2^{(S)})$ and the real population. Specifically, for each $j = 1, \ldots, J$, we set $F_j = F_j^{(S,p)} + F_j^{(R,1-p)}$, where $F_j^{(S,p)}$ is generated from $NB(p\theta_1^{(S)}, \theta_2^{(S)}/p)$, $F_j^{(R,1-p)}$ is generated from $Bin\{F_j^{(R)}, (1-p)\})$, p is the mixing proportion, $0 \leq p \leq 1$, and $F_j^{(R)}$ is the frequency in the real population. The resulting F_j for $j = 1, \ldots, J$ are then combined to give the finite population values N_r to be estimated. For simplicity, we assume Bernoulli sampling with fixed inclusion probability $\pi = 0.01$. We repeatedly draw 1000 samples from each of the finite populations. For each sample we obtain estimates of the N_r using both the proposed and reference methods of estimation in Section 4. We focus on small values of r, which are of primary interest in disclosure risk assessment.

The errors of the estimators $\hat{N}_r$ of the N_r are summarised by the relative root mean squared error, that is the root mean square of the $\hat{N}_r - N_r$ across the 1000 samples divided by N_r. Values of these relative root mean squared errors are plotted in Figure **??** against the mixing proportion as line plots, interpolating the points obtained for $p = 0.0, 0.1, 0.2, \ldots, 0.5$ with straight lines, separately for the two estimators and for $r = 1, 2, 3$ and 4.

A decomposition of the root mean squared errors displayed in Figure **??** into biases and standard errors reveals that the former dominate the latter when $p \geq 0.1$. Only when $p = 0.0$ does the bias become negligible for each estimator, relative to the standard error. The principal message we draw from Figure **??** is that the bias of the model-based estimator increasingly exceeds that of the proposed estimator as p increases from 0.1, for each value of $r = 1, 2, 3$ and 4. Only when $p = 0.0$ and the basic parametric model holds does the model-based estimator tend to have smaller mean squared error through its smaller standard error.

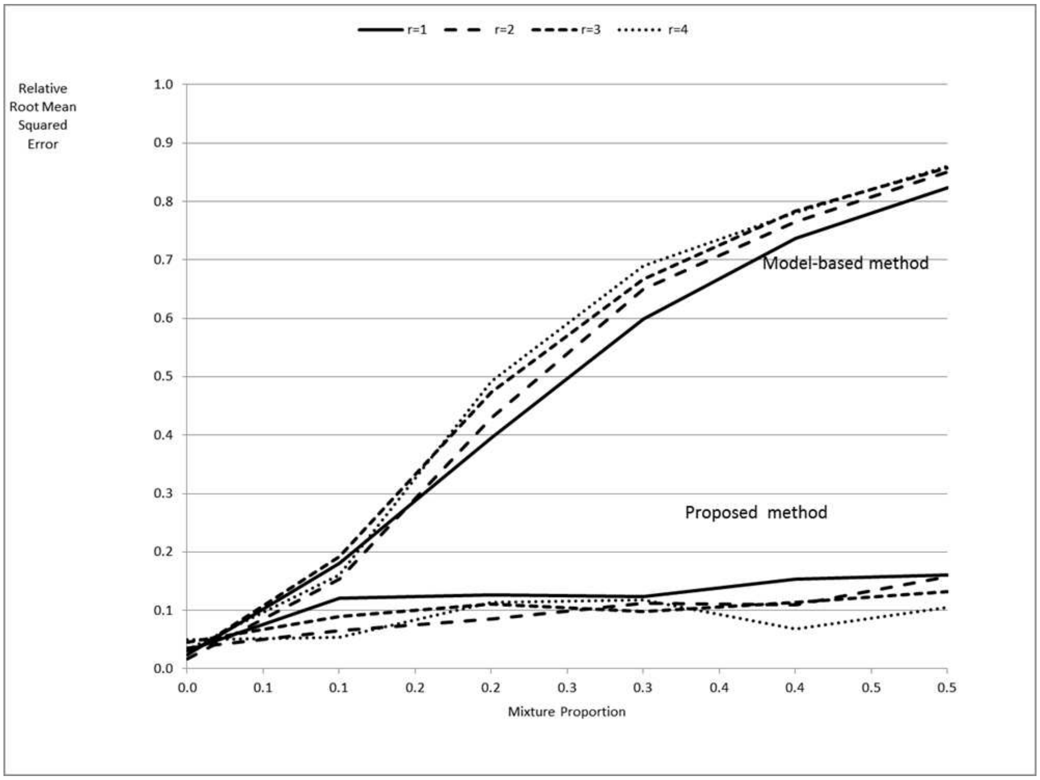

Figure 1: Simulation relative root mean squared errors by proportion of population derived from real population, separately for proposed and model-based estimation methods and for frequencies r=1,2,3 and 4.

6 Discussion

In this paper we have derived design-unbiased estimators for certain moments of the frequency of frequency distribution. We have made modelling assumptions to extend these estimators to estimators of the frequency of frequencies distribution. Our simulation study illustrates how our hybrid approach can be more robust than parametric model-based estimation in terms of bias and mean squared error under departures from the model.

The main results in this paper assume Poisson sampling, which reduces to Bernoulli sampling in the case of equal probabilities of selection. These designs, which lead to random sample sizes, may be viewed as approximations to the kinds of fixed sample size designs more commonly used in practice.

We did repeat our simulation study replacing Bernoulli sampling by simple random sampling without replacement, where the sampling fraction was set equal to Bernoulli sampling probability, and obtained results visually indistinguishable from those in Figure 1. This suggests that the bias properties of the proposed estimators under Poisson sampling are similar, in practice, for corresponding fixed sample size designs, at least for the kind of sampling fraction we considered in our simulation study.

This paper has only considered point estimation. Our justification is that model misspecification bias can be more important than standard errors in the kinds of large survey samples we are interested in, as illustrated in the simulation study. Nevertheless, it would be desirable in practice to have at least variance estimators to accompany our point estimators. The bootstrap may be the most natural approach.

Acknowledgement

We are grateful to insightful comments of a reviewer, including suggestions for simplifying the proofs of Theorems 3.1 and 3.2. Research was supported by UK Economic and Social Research Council.

Appendix A

Proof of Theorem 3.1 Let $Z_i = 1$ if $i \in s$ and $Z_i = 0$ otherwise and note that under Poisson sampling the Z_i are independent $B(1, \pi_i)$. Let $x_{ji} = 1$ if

$i \in C_j$ and $x_{ji} = 0$ otherwise. Using the fact that $f_j = \sum_{l \in U} x_{jl} Z_l$ we have

$$\begin{aligned}
E(\hat{\mu}_{r1}) &= \sum_{j=1}^{J} E\left[I\left(\sum_{l=1}^{N} x_{jl} Z_l = r+1 \right) \sum_{i=1}^{N} x_{ji} Z_i \left(\pi_i^{-1} - 1 \right) \right] \\
&= \sum_{j=1}^{J} \sum_{i=1}^{N} E\left[I\left(\sum_{l \neq i}^{N} x_{jl} Z_l = r \right) \right] E\left[x_{ji} Z_i \left(\pi_i^{-1} - 1 \right) \right] \\
&= \sum_{j=1}^{J} \sum_{i=1}^{N} E\left[I\left(\sum_{l \neq i}^{N} x_{jl} Z_l = r \right) \right] E\left[x_{ji} \left(1 - Z_i \right) \right] \\
&= \sum_{j=1}^{J} \sum_{i=1}^{N} E\left[I\left(\sum_{l \neq i}^{N} x_{jl} Z_l = r \right) x_{ji} \left(1 - x_{ji} Z_i \right) \right] \\
&= \sum_{j=1}^{J} \sum_{i=1}^{N} E\left[I\left(\sum_{l=1}^{N} x_{jl} Z_l = r \right) x_{ji} \left(1 - x_{ji} Z_i \right) \right] \\
&= \sum_{j=1}^{J} E\left[I\left(\sum_{l=1}^{N} x_{jl} Z_l = r \right) \left(F_j - r \right) \right] = E(\mu_{r1}), \text{as required.}
\end{aligned}$$

Proof of Theorem 3.2

$$\begin{aligned}
E(\hat{\mu}_{r2}) &= \sum_{j=1}^{J} E\left[I\left(\sum_{l=1}^{N} x_{jl} Z_l = r+2 \right) \sum_{i \neq k} x_{ji} x_{jk} Z_i Z_k \left(\pi_i^{-1} - 1 \right) \left(\pi_k^{-1} - 1 \right) \right] \\
&= \sum_{j=1}^{J} \sum_{i \neq k} E\left[I\left(\sum_{l \neq i,k}^{N} x_{jl} Z_l = r \right) x_{ji} x_{jk} Z_i Z_k \left(\pi_i^{-1} - 1 \right) \left(\pi_k^{-1} - 1 \right) \right] \\
&= \sum_{j=1}^{J} \sum_{i \neq k} E\left[I\left(\sum_{l \neq i,k}^{N} x_{jl} Z_l = r \right) x_{ji} x_{jk} \left(1 - Z_i \right) \left(1 - Z_k \right) \right] \\
&= \sum_{j=1}^{J} \sum_{i \neq k} E\left[I\left(\sum_{l \neq i,k}^{N} x_{jl} Z_l = r \right) x_{ji} x_{jk} \left(1 - x_{ji} Z_i \right) \left(1 - x_{jk} Z_k \right) \right] \\
&= \sum_{j=1}^{J} E\left[I\left(\sum_{l=1}^{N} x_{jl} Z_l = r \right) \sum_{i \neq k} x_{ji} x_{jk} \left(1 - x_{ji} Z_i \right) \left(1 - x_{jk} Z_k \right) \right] \\
&= \sum_{j=1}^{J} E\left[I\left(\sum_{l=1}^{N} x_{jl} Z_l = r \right) \left\{ F_j(F_j - 1) - 2r(F_j - 1) + r(r-1) \right\} \right] \\
&= \sum_{j=1}^{J} E\left[I\left(\sum_{l=1}^{N} x_{jl} Z_l = r \right) \left(F_j - r - 1 \right) \left(F_j - r \right) \right] = E(\mu_{r2}), \text{as required.}
\end{aligned}$$

References

Bethlehem, J., Keller, W., Pannekoek, J., 1990. Disclosure control of microdata. J. Am. Statist. Assoc. 85, 38–45.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis; Theory and Practice. Cambridge Massachussetts: MIT Press.

Bunge, J., Fitzpatrick, M., 1993. Estimating the number of species: a review. J. Am. Statist. Assoc. 88, 364–73.

Good, I.J., 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40, 237–264.

Goodman, L.A., 1949. On the estimation of the number of classes in a population. Ann. Math. Statist. 20, 572–579.

Hájek, J., 1981. Sampling from a Finite Population. New York: Marcel Dekker.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models. Boca Raton: Chapman and Hall/CRC. second edition.

Särndal, C.E., Swensson, B., Wretman, J.H., 1992. Model Assisted Survey Sampling. New York: Springer.

Skinner, C.J., Carter, R.G., 2003. Estimation of a measure of disclosure risk for survey microdata under unequal probability sampling. Survey Methodology 29, 177–180.

Skinner, C.J., Elliot, M.J., 2002. A measure of disclosure risk for microdata. J. R. Statist. Soc. B 64, 855–867.

Skinner, C.J., Shlomo, N., 2008. Assessing identification risk in survey microdata. J. Am. Statist. Assoc. 103, 989–1001.