A MODEL-BASED JUSTIFICATION FOR SURVEY WEIGHTS

Charles H. Alexander. U.S. Bureau of the Census

1. INTRODUCTION

Estimates based on data from sample surveys usually incorporate weights which adjust for various differences between the sample and the population. The explanation of the weighted estimators has traditionally relied on the theory of sampling from a finite population. Accordingly the stated goal has been to make inferences about the finite population, and estimators have been evaluated in terms of their variance and bias under the distribution of all possible samples from the population under the sample design. (See Hansen, Hurwitz and Madow (1953) or Cochran (1977).)

Various aspects of the theory behind the use of weights have been criticized, especially in the past decade. The relevance of the bias and variance over all possible samples has been challenged in a series of papers by Royall. (For example, see Cumberland and Royall (1981).) The basic inferential role of information about the sampling mechanism has been brought into question by the work of Rubin (1978), among others. An overview of the debate, with relevant references, is given by the discussions following Hansen, et al (1983). A more recent discussion is given by Hoem (1986). This theoretical dispute has caused doubt about whether the weights supplied with most survey data need to be used at all, especially by analysts making model-based inferences. The impression may be left that weights are useful only for making inferences about the finite population and have no role in inferences about the process which produced the population.

The present paper offers a model-based justification for using weights in certain circumstances to make inferences about the "superpopulation distribution" from which the population was generated. The discussion considers three alternative models for how the sample is selected, corresponding to different situations which lead to differential sampling rates for household surveys. Under these simple models, factors very much like the traditional survey weights are needed to calculate maximum likelihood estimators for the parameters of the model.

The discussion is restricted to categorical variables. The models do not reflect the systematic cluster sampling which is characteristic of many surveys. The scope of the paper is also limited in that the different situations which may call for weighting adjustment are discussed separately, while in practice they must be addressed simutaneously.

Although somewhat simplistic, the models in the paper may provide a framework for discussing when weighting is appropriate, without reference to the usual finite population sampling theory.

2. A MODEL FOR THE POPULATION.

The N units in the population will be assumed to be a "simple random sample with replacement" from a superpopulation. In other words, the variables of interest for each population unit will be assumed to be random variables following some joint probability distribution. The goal of the analysis is to make inferences about this underlying distribution.

The model will include four discrete random variables:

Y, a dichotomous (0 or 1) response variable
S, the "analytic" subdomain;
R, the "unknown relevant information" subdomain;
T, the sampling stratum.

Sample units are selected separately from the different sampling strata in the population, possibly with different sampling rates. Models for the sample selection will be described in the next section. It will be assumed that Y is conditionally independent of T given S and R, i.e.

(2.1) $P(Y=1|\ S=s,\ R=r,\ T=t) = P(Y=1|\ S=s,\ R=r,\ T=t')$, for all possible s, r, t, and t'

It is assumed that the goals of the analysis are estimate $P(Y=1|\ S=s)$ and $P(Y=1)$.

EXAMPLE:

Let Y = 1 if a household in the population has income above the poverty level, and Y = 0 otherwise. Let the analytic subdomain S be equal to one of the six possible values of the random vector $(S_1,\ S_2)$ where

S_1 = number of persons in the household, either 1,2, or 3+
S_2 = urban / rural location of the person's household.

The analyst's interest is in estimating $P(Y=1|\ S=s)$ for s=1,....6, and $P(Y=1)$.

Regardless of how many variables are included in the analysis, there is always the possibility that relevant variables are omitted. This may be an "oversight" or it may be because the surveyers do not know how to measure the variables. In our example, let R take one of the four possible values of the random vector (R_1, R_2) where

R_1 = 1 if the household has a telephone.
0 Otherwise
R_2 = 1 if the household members have a a strong work ethic
0 otherwise.

The sample is assumed to be selected at a constant rate within certain sampling strata. The sampling stratum may involve information already included in S and R, plus additional variables. In our example, let there be eight possible values of T corresponding to the possible values of $(T_1,\ T_2,\ T_3)$ where

$T_1 = S_2$
$T_2 = R_1$
T_3 = 1 if household is in an area where it is hard to recruit interviewers
0 otherwise

These sampling strata might correspond to a planned oversampling of urban households, a dual frame telephone/address sample , and a need to reduce the sampling rate in certain areas where interviewers are hard to recruit.

The initial goal of the analysis is to estimate the probability that a household has income above the poverty level, given the household's size and urban/rural location.

For this example, (2.1) requires only the assumption that Y is conditionally independent of T_3, given S_1, S_2, R_1, and R_2. The conditional independence of Y and T then follows since T_1 and T_2 have already been included in the conditioning variables.

DISCUSSION OF THE MODEL

In some situations T is a function of (S,R), so that (2.1) follows automatically. In other situations, T may depend on features of the sampling process which have little direct relevance to the characteristics of the population. Some examples:

A. If there are multiple list frames the units which appear on several lists will usually be sampled at a higher rate than the set of units which appear on only one list. Thus, the properties of the sampling strata may depend primarily on the process by which the lists were generated. This also occurs when list, area, or telephone-number frames are combined.

B. Sampling rates often vary based on the cost or difficulty of interviewing in various geographic areas. The cost and difficulty of interviewing may depend primarily on the organizational structure of the sampling organization.

C. Even when households are subsampled or oversampled based on variables of analytic interest, the sampling is often based on a "quick and dirty" screening question, which may give erroneous results. The variable of analytic interest is the "true" characteristic determined by a subsequent detailed interview.

In situations like these, it may be inconvenient to include T among the analytic variables. Consequently, S and T are distinguished in our model.

Because there may be relevant variables not included in S, the two variables Y and T are not necessarily conditionally independent given S, in spite of (2.1). I.e., the sampling mechanism is not necessarily "ignorable" for estimating $P(Y|S)$.

The distinction of R, S and T is not necessary to obtain our mathematical results. The same results can be obtained if R is dropped from the model, provided assumption (2.1) is eliminated. The distinction is made for heuristic purposes to emphasize that the sampling strata may be relevant either because they are directly relevant to Y, or because within the different strata there occur different distributions of other, possibly unknown, relevant variables.

3. THREE MODELS FOR THE SAMPLE SELECTION

Three basic models for sampling from the finite population need to be considered. A sample of n units will be selected from the N population units. The three models are:

Model 1: (Known population size) In each of the k_T sampling strata, n_t units are selected randomly without replacement from the N_t population units in the stratum. Both n_t and N_t are known. The sampling interval is $w_t = N_t/n_t$.

Model 2: (Unknown population size) In the t^{th} stratum, n_t^* units are selected randomly without replacement from a list of N_t^* units. The list includes the N_t actual population units, plus an additional U_t "invalid" units. Whether a unit is valid or invalid is only discovered after the unit has been selected. U_t and N_t are unknown. However, $w_t = N_t^*/n_t^*$ is known, as are n_t and u_t, the number of invalid sample units. Model 2 describes the most common situation in sample selection for household surveys from an address frame: the address lists inevitably contain vacant or demolished housing units which are not part of the eligible population. Although the actual sampling process may be more complicated than a one-stage simple random sample, it is usually the case that only the sampling interval w_t and sample size n_t are known. Unknown population sizes are also common in area sampling, or telephone sampling.

Model 3: A simple random sample of n^* units is selected from the population. From the n_t^* sample units in the t^{th} sampling stratum, a final subsample of n_t units is selected. The subsampling interval $w_t = n^*_t/n_t$ is known.

4. ADDITIONAL NOTATION AND LIKELIHOOD FUNCTIONS

Let

$q_{srt} = P(S=s,\ R=r,\ T=t)$,

$q_{s(t)} = P(S=s \mid T=t)$,

$p_{srt} = P(Y=1 \mid S=s,\ R=r,\ T=t)$,

$p_{st} = P(Y=1 \mid S=s,\ T=t)$,

where

$s = 1, \ldots, k_S;\ r=1, \ldots, k_R,\ t=1, \ldots, k_T$

The following are random variables:

n_{st} = number of sample units with S=s and T=t

m_{st} = number of sample units with S=s, T=t and Y=1

N_{srt}= number of population units with S = s, R = r, and T=t.

Note that $n_t = \sum_s n_{st}$ and $N_t = \sum_s \sum_r N_{srt}$.

Let $q_t = \sum_s \sum_r q_{srt}$.

Model 1:

The observed data are the vectors (n_{st}), (m_{st}), (N_t). The parameters used to model the distribution of these random variables will be (p_{st}), $(q_{s(t)})$, (q_t). These parameters are subject to the constraints:

$$\sum_t q_t = 1$$

$$(4.1) \quad \sum_s q_{s(t)} = 1 \text{ for } t = 1, \ldots, k_T$$

The likelihood function is:

$$L((p_{st}),(q_{s(t)}),(q_t) \mid (n_{st}),(m_{st}),(N_t)) =$$

$$\left[\prod_{st} \begin{bmatrix} n_{st} \\ m_{st} \end{bmatrix} p_{st}{}^{m_{st}} (1-p_{st})^{n_{st} - m_{st}} \right].$$

$$(4.2) \quad \prod_t \left[(n_t ! \,/\, \prod_s n_{st} !) \prod_s q_{s(t)}{}^{n_{st}} \right].$$

$$\left[(N! \,/\, \prod_t N_t!) \prod_t q_t{}^{N_t} \right].$$

P(Y=1 |S=s) and P(Y=1) may be written as functions of the parameters of the model:

$$P(Y=1 |S=s) = (\sum_t p_{st}\, q_{s(t)}\, q_t)/(\sum_t q_{s(t)} q_t)$$

$$(4.3) \quad P(Y=1) = \sum_s \sum_t p_{st}\, q_s(t)\, q_t$$

The maximum likelihood estimates (MLEs) are easily obtained by differentiating the logarithm of (4.2), incorporating the constraints (4.1) through use of Lagrange multipliers. The MLEs are:

$$\hat{p}_{st} = m_{st} / n_{st};\ \hat{q}_{s(t)} = n_{st} / n_t$$

$$\hat{q}_t = N_t / N$$

Consequently, the MLE of P(Y=1 |S=s) is:

$$\hat{P}(Y=1 |S=s) =$$

$$(4.4) \quad = (\sum_t m_{st}\, w_t) / (\sum_t n_{st}\, w_t)$$

where $w_t = N_t/n_t$. Expression (4.4) is the weighted proportion of those sample units with S=s which have Y=1, where each unit's weight is w_t, if T=t is the unit's stratum.

$$\hat{P}(Y=1) = \sum_s \sum_t (m_{st}/n_{st})(n_{st}/n_t)(N_t/N)$$

$$(4.5) \quad = (\sum_s \sum_t m_{st}\, w_t) / \sum_s \sum_t n_{st}\, w_t$$

This is the weighted proportion of all sample units which have Y=1.

Model 2:

To specify a likelihood function for Model 2, the "invalid units" (U_t) will be included in the probability model. For each of the N_t^* units in stratum t, let the random variable V determine whether the unit is valid or invalid, with V=1 for valid units and V=0 for invalid units.

Let b_{tv} = P(T=t, V=v),

for v= 0,1; t=1,...,k_T

Let $b_t = b_{t0} + b_{t1}$.

In terms of the previously defined notation, $b_{t1} = q_t$

Let $c_{1(t)} = b_{t1} / b_t$ be the probability that a unit from stratum t is valid.

For model 2, the observed random variables are (n_{st}), (m_{st}), (N_t^*), (n_t), and (n_t^*). The parameters of the model for the distribution of these data are (b_t), $c_{1(t)}$, $(q_{s(t)})$, (p_{st}). These are respectively the probability that a unit is in stratum t, the probability that a unit in stratum t is valid, the probability that a valid unit in stratum t is in stratum s, and the probability that a unit in stratum t and stratum s has Y=1. The likelihood function is:

$$L((b_t),(c_{1(t)}),(q_{s(t)}),(p_{st}) |$$
$$(n_{st}),(m_{st}),(N_t^*),(n_t),(n_t^*))$$

$$= \left[(N^*! \,/\, \prod_t N_t{}^*!)(\prod_t b_t{}^{N_t^*}) \right]$$

$$\prod_t \left[\begin{bmatrix} n_t{}^* \\ n_t \end{bmatrix} c_{1(t)}{}^{n_t} \ (1-c_{1(t)})^{n_t{}^* - n_t} \right]$$

$$\prod_t \left[(n_t! \,/\, n_{st}!) \prod_s q_{s(t)}{}^{n_{st}} \right]$$

$$\prod_t \prod_s \left[\begin{bmatrix} n_{st} \\ m_{st} \end{bmatrix} p_{st}{}^{m_{st}} \ (1 - p_{st})^{n_{st} - m_{st}} \right]$$

The interest here is in P(Y=1| S=s,V=1) and P(Y=1| V=1), since invalid units are of no interest in modelling the population. These probabilities can be written as functions of the parameters as follows:

$$P(Y=1| S=s, V=1) = (\sum_t p_{st}\, q_{s(t)}\, c_{1(t)} b_t) /$$
$$(\sum_t q_{s(t)}\, c_{1(t)} b_t)$$

$$P(Y=1| V=1) = (\sum_s \sum_t p_{st}\, q_{s(t)}\, c_{1(t)}\, b_t) /$$
$$(\sum_t c_{1(t)}\, b_t)$$

The MLEs of the model parameters are

$$\hat{b}_t = N_t{}^*/N^*;\ \hat{c}_{1(t)} = n_t/n_t{}^*$$

$$\hat{q}_{s(t)} = n_{st}/n_t;\ \hat{p}_{st} = m_{st}/n_{st}$$

Thus, the MLE of P(Y=1| S=s,V=1) is:

$$\hat{P}(Y=1| S=s,V=1) =$$

$$(4.6) \quad = (\sum_t m_{st}\, w_t) / (\sum_t n_{st}\, w_t)$$

where $w_t = N_t{}^*/n_t{}^*$. Like (4.4), this is the weighted proportion of valid units with S=s, which have Y=1. Here the weight is the inverse of the probability of selecting a given valid unit from the finite population in stratum t.

$$\hat{P}(Y=1| V=1) =$$

$$(4.7) \quad = (\sum_s \sum_t m_{st}\, w_t)/(\sum_s \sum_t n_{st}\, w_t),$$

which is the weighted proportion of all valid sample units which have Y=1, using w_t as the weight.

Model 3

Recall that the N units of the population in Model 1 are N independent identically distributed observations from the superpopulation distribution. In Model 3, the n^* elements, selected as a simple random sample from the N population units, are n^* independent and identically distributed observations from the superpopulation distribution. Thus, Model 3 is mathematically identical to Model 1, except that n^* and n_t^* take the role of N and N_t. Consequently, the MLEs for P(Y=1| S=s) and P(Y=1) for Model 3 are given by (4.4) and (4.5), where $w_t = n_t^*/n_t$.

5. CIRCUMSTANCES WHEN WEIGHTS ARE NOT NEEDED

While weights appear in the MLEs for P(Y=1| S=s) for our model in general, there are two special cases where they are unnecessary. The first is the case when T is a function of S, i.e., when each analytic stratum s is contained within a single sampling stratum t. This corresponds to including the sampling strata in the analysis. Then in expressions (4.4) and (4.6) only the weight w_t for one sampling stratum appears, so the weights in these formulas cancel. In this case, the weighted estimator may be used, but the weighted estimator is equal to the unweighted estimator. Weights still are needed to estimate P(Y=1) by (4.5) and P(Y=1| V=1) by (4.7).

In second special case, the weighted and unweighted estimators differ. This is the case when there are no R strata, i.e., all relevant variables are assumed to be included in the analysis. Then (2.1) implies that p_{st} is the same for all t, so that this parameter may be written p_s. The MLE of p_s is $\hat{p}_s = m_s/n_s$, where $m_s = \sum_t m_{st}$ and $n_s = \sum_t n_{st}$. Then, instead of (4.4), we have

$$\hat{P}(Y=1|S=s) = m_s/n_s.$$

Thus in this case, the unweighted proportion gives the MLE for P(Y=1| S=s). Here the weighted estimator is less efficient. Weights still are used to estimate P(Y=1); in this case (4.5) becomes

$$\hat{P}(Y=1) = \sum_s \hat{p}_s \left(\sum_t n_{st} w_t/N\right).$$

Here the unweighted estimators for the analytic cells $\hat{p}_s$ are "weighted up" using weighted estimators of P(S=s).

A third "special case" goes beyond the assumptions of our model. This is when Y is conditionally independent of T given S, even though there may be unmeasured relevant variables R, i.e.,

$$(5.1) \quad P(Y=1|\ S=s, T=t) = P(Y=1|\ S=s,\ T=t'),$$

for all s,t and t', even though there exists a random variable R such that

$$(5.2) \quad P(Y=1|S=s, R=r) \neq P(Y=1|S=s, R=r'),$$

for some s, r, and r'.

To assume (5.1) when (5.2) holds requires restrictive assumptions about the joint distribution of Y, S, R, and T.

6. COMPARISON WITH TRADITIONAL SURVEY WEIGHTS

The sampling models described in Section 3 correspond to some of the stages of weighting for national household surveys such as the Current Population Survey (CPS), the Consumer Expenditure Survey (CE), and the National Crime Survey (NCS). Model 2 corresponds to the **basic weight** assigned to each sample unit. Sample households from different parts of the list or area frame are known to have been sampled at different rates. However, as described in Section 3, the total number of occupied households in the frame is not known. The usual basic weight is the inverse of the probability of selection, which agrees with the MLE (4.6).

Model 1 may be used to describe post-stratification. A sample is selected and the number of sample units n_t in each of several strata is observed. The number of units in the population in each stratum, N_t, is assumed to be known. **The post-stratification factor** N_t/n_t is applied to each unit in the stratum. This corresponds to (4.4). In practice, the sample sizes (n_t) in the strata are not fixed in advance. However, conditional on (n_t), the post-stratification factor may be regarded as an application of Model 1. Post-stratification is often applied to correct for systematic under coverage, i.e., when the sampling frame omits units in the population. If there is undercoverage, use of Model 1 requires the strong assumption that the omitted units in each stratum are a random sample from the population units in the stratum.

Model 3 can be used to describe unit nonresponse, if it is assumed that the n_t responding units in stratum t are a random sample from the n_t^* sample units in the stratum. (Here the stratum is commonly called the "noninterview cell".) The usual **noninterview adjustment factor** is n_t^*/n_t, which is the factor in the MLE for model 3. The inverse of this factor is the proportion of units in the cell which respond to the survey. This proportion may be viewed as an estimate of the probability of response for units in the cell. Note that under our model, the actual proportion would be used in preference to the underlying probability, even if the probability were known.

Model 3 can also be used to describe "field subsampling" or other cases where a "**special weight**" or "weighting control factor" is applied. For example, suppose the initial sample selection has assigned an interviewer $n_t^* = 25$ cases in a given block, which would lead to too great a workload. The rules may allow the sampling clerks to select 1 case in 2 using a systematic sample from a randomly ordered list of cases. This produces a sample of $n_t = 12$ or $n_t = 13$ cases, so Model 3 calls for a factor of 25/12 or 25/13 respectively.

In this field sampling situation, the usual practice differs from the MLE under Model 3. Usually, the inverse probability of selection, namely 1/(1/2) = 2 in our example, is used as the weight.

The major reason for this departure is that the list of cases is often sorted according to some possibly relevant variables, such as apartment number, before sampling. Consequently, Model 3 may not apply. Also the probability of selection is more readily available, with less record-keeping, than the actual proportion selected.

Except for this departure, the weights applied in "traditional" survey practice are in accordance with those prescribed by the MLEs in Section 4 under the relevant sampling model. The probability of selection comes into the weight under Model 2, when the actual sampling fractions in the various strata are unknown.

Other weights used in survey practice -- the "first-stage ratio adjustment factor" for CPS and NCS, and the "principal person household weight" used for CPS, NCS and CE -- do not fit into the framework of our model. These might be modelled more appropriately as regression estimators.

7. EXTENSIONS OF THE MODEL

More General Discrete Response Variables: The restriction that Y takes only two values, 0 or 1, can be weakened. Suppose that Y can take K possible values $y_1,\ldots,y_K$. Let the indicator variable $Y_k = 1$ if $Y=y_k$ and zero otherwise. Then $P(Y=y_k|\ S=s) = P(Y_k=1|\ S=s)$. Accordingly, the results of Section IV apply to estimation of $P(Y=y_k|\ S=s)$.

Continuous Response Variables: The above model is not the most natural way to approach the use of weights when Y has a continuous distribution. However, it is common for such continuous variables to be analyzed as a grouped frequency table. For a bounded random variable Y, k intervals are defined, of the form:

$$[0,\ t_1],\ (t_1,\ t_2],\ldots,\ (t_{k-1},\ t_k].$$

Then the random variable which is often analyzed is the random variable Y' defined by $Y' = (t_i + t_{i-1})/2$, if Y is in the i^{th} interval.

Let p_i = P(Y is in the i^{th} interval) for $i=1,\ldots,k$. Then Y' is a discrete random variable whose probability distribution depends only on $p_1,\ldots,p_k$. Our model could therefore be used to justify weights in estimates pertaining to the distribution of Y'. This suggests that it may be possible to view weighting of continuous data as a nonparametric approach to estimating an approximation to the underlying probability distribution.

Longitudinal Estimates: Our model would apply to some simple longitudinal estimates. For example, let S=0 or 1 indicate whether a household is above poverty at time t and let Y=0 or 1 indicate whether the household is above poverty at time t + 1. According to Section 4, weights are needed to estimate $P(Y=1|\ S=0)$, etc, in certain circumstances.

Other Parametric Models: In general, the analyst may view Y as a random variable whose distribution is a member of some specified parametric family $f(y|\theta)$ and the goal may be to estimate θ under this model. (Here θ and Y may both be vectors. Our saturated model relating Y and S is a special case where Y now takes the role that (Y,S) took in Section 2.) Ordinarily, based on data from N independent and identically distributed observations from the distribution $f(y|\theta)$, the MLE of θ can be written as some function of the empirical distribution of Y, i.e.,

$$(7.1)\quad \hat{\theta} = g_n(e(y_1),\ldots,e(y_K)),\ \text{where}$$

$$e(y_k) = (\text{number of units with } Y=y_k)/N.$$

However, if data only for a sample of the N units are observed, estimation of θ may become more complicated, depending how the sample is selected. If the sample is selected at different rates according to some random variable T which was observed for the original observations, then the distribution of Y for any given units in the final sample is no longer $f(y|\theta)$, and (7.1) can no longer be used; it is necessary to include T.

One way to do this is to model the joint distribution of Y and T as some parametric family $h(y,t|\gamma)$. The MLE $\hat{\gamma}$ may be obtained from the available observations of Y and T. If the original parameter θ can be expressed as a function of the new parameter γ, this may satisfy the goals of the analysis.

This approach requires the analyst to specify at least certain aspects of the relationship of Y and T, which may be difficult to do in some situations, as discussed at the end of Section 2. (With respect to our saturated multinomial models for Y and T, one version of this approach is simply to assume (5.1).)

A second approach is to use (7.1), but to replace $e(y_k)$ by a weighted estimate of $P(Y=y_k)$, using the weighted estimator in (4.4) or (4.6). Under suitable conditions on f, i.e., on g, the resulting $\hat{\theta}$ will be a consistent estimator of θ, as N and n grow large. This second approach is quite similar, at least in spirit, to the method of "pseudo-maximum-likelihood" estimation in Gong and Samaniego (1981).

The second approach avoids the dangers of misspecifying the relationship of Y and T. This relationship is often poorly understood. By contrast, the sampler has control over the assumptions necessary to estimate $P(Y=y_k)$ using Model 2 for the "basic weight", or Model 3 for the "special weight". In this sense, weighting may be said to protect against model misspecification. (The assumptions needed for the post-stratification factor and the noninterview factor are not under the sampler's control.)

Naturally, use of weights to adjust for sampling given no protection from misspecification if the model $f(y|\theta)$, which describes the population, is not correct.

For many surveys, use of either approach is complicated by the fact that there are no simple analytic expressions for the variance of the estimator, and replication methods or other special methods must be used to estimate variances. This problem is most often due to cluster sampling of households, or multiple stages of sample selection, rather than to the weights.

8. DISCUSSION OF THE ROLE OF THE SUPERPOPULATION

When there is unequal probability sample selection, either intentional or unintentional, the distribution of the observed data is not necessarily the distribution of the real-world process of interest. The sample distribution may have been altered by operational considerations of no intrinsic interest to the analysis. This makes the role of unmeasured relevant variables more crucial then it is in the usual modelling situation.

Consider the simple case of a Bernoulli random variable. In our example, Y is a Bernoulli random variable conditional on S=s, with probability of "success" $P(Y=1|S=s)$. For s=1, this is the probability that a single-person urban household is above the poverty level. This Bernoulli probability has two different interpretations. Ordinarily it does not matter which interpretation is adopted, but with differential sampling rates it may matter.

The first interpretation is that, with respect to poverty status, single-person urban households are like so many "identical coins", having identical propensities towards poverty. Under this interpretation, for any other variable R,

(8.1) $P(Y=1|S=s,R=r) = P(Y=1|S=s,R=r')$, for all possible r and r'.

The second interpretation allows different single-person urban households to have different individual probabilities of poverty, depending on other characteristics of the households. Under this interpretation $P(Y=1|S=s)$ is a conditional probability corresponding to the mix of other variables generated by the process which produces the population. Accordingly if there are additional variables R, then $P(Y=1|S=s,R=r)$ need not equal $P(Y=1|S=s,R=r')$ but

$$P(Y=1|S=s) = \sum_r P(Y=1|S=s,R=r)\ P(R=r|S=s).$$

Under the first interpretation, weights are not needed to estimate $P(Y=1|S=s)$. The second interpretation is the one adopted in this paper; weights are needed in the cases described in Section 4. Under assumption (8.1), $P(Y=1|S=s)$ can be estimated without a representative sample of the population. Under the second interpretation, $P(Y=1|S=s)$ makes sense only with regard to the probability distribution which produced the population, and a representative sample of the population is necessary to estimate the probability. If the unweighted sample is not representative, weights are needed.

For the kinds of socio-economic variables measured by many national household surveys, the strict homogeneity of (8.1) does not seem reasonable to assume, even when a large number of variables are included in S.

The second ("classical") interpretation of probability is not unique to survey samplers. For example, see Cramer (1945, Chapter 13.)

9. CONCLUSION

Under some simplistic models for sampling, weights have been shown to be necessary, in certain circumstances, to make estimates pertaining to the process which produced the population from which a sample was drawn. The weights are not needed unless 1) the model omits some relevant variables, 2) the variables determining the probability of selection (or the sampling rate) are not all included in the model.

For complex multiple-frame surveys of such phenomena as unemployment, expenditures, and crime, proponents of weighting (such as the author) would assert that no model will include all the relevant variables, and that few analysts will wish to include in their model all the geographic and operational variables which determine sampling rates. It is difficult to object in principle with the goal of correctly modelling all relevant variables, including the variables relating to sampling. However, the theoretical and empirical tasks of deriving, fitting, and validating such models seem formidable for many complex national demographic surveys. Thus, the question comes down to the desirability of making restrictive assumptions (such as (5.1)) about the sampling mechanism. Without such assumptions, our models lead to weights.

The results of Section 4 for our multinomial model may seem obvious. However, some consequences are worth noting.

1. Weights may be justified in some cases when the interest goes beyond the actual finite population.
2. Weights may be needed even though inferences are based on a model.
3. Weights may be needed even though the analysis concerns longitudinal transitions or the relationships of two or more variables.
4. The unconditional probability of selection has a role in the weighting under one of the sampling models (Model 2), in which the sampling rate is known, but the total population size is not known.

ACKNOWLEDGEMENTS

The author wishes to thank several Census Bureau colleagues for their helpful comments, and Brenda Kelly for her help in preparing the manuscript.

References

Cochran, W.G. (1977). Sampling Techniques, Third edition. New York. John Wiley and Sons.

Cramer, H. (1945). Mathematical Methods of Statistics. Princeton. Princeton University Press.

Cumberland, W.G., and Royall, R.M. (1981). Prediction Models and Unequal Probability Sampling. Journal of the Royal Statistical Society, Ser. B, 43, 353-367.

Gong, G. and Sameniego, F.J. (1981). Pseudo Maximum Likelihood Estimation: Theory and Applications. Annals of Statistics, 9, 861-869.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sampling Survey Methods and Theory, two volumes. New York. John Wiley and Sons.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability Sampling Inferences in Sample Surveys. Journal of the American Statistical Association, 78, 776-807. Discussions by Royall, R.M., Little, R.J.A., Dalenius, T., Smith, T.M.F., and Rubin, D.R.

Hoem, J.M. (1986). The Issue of Weights in Panel Surveys of Individual Behavior. Presented at the International Symposium on Panel Surveys, Washington, D.C.

Rubin, D.B. (1978). Bayesian Inference for Causal Effects: the Role of Randomization; Annals of Statistics, 7, 34-58.