Journal of Official Statistics
Vol. 8. No. 2, 1992, pp. 183–200

Weighting for Unequal P_i

Leslie Kish[1]

Abstract: Four distinct sources for unequal selection probabilities P_i of elements are distinguished concerning their origins, their effects, and their need for weights $k_i \propto 1/P_i$. Three other types of weighting for estimation are also identified. Survey sampling theory is for unbiased estimation with weights k_i but model based theory is against. The main disadvantage of weighting is the increase in variances from S^2/n to $S^2(1 + C_k^2)/n$ for weighted estimates $\bar{y}_w$, where C_k^2 is the relvariance of the k_i. This is balanced against the increase of the mean square error of the unweighted estimate $\bar{y}_u$ from S^2/n to $(S^2/n + R_{ky}^2 C_k^2 S^2)$, where $R_{ky} C_k S$ is the bias $= \bar{y}_u - \bar{y}_w$ of $\bar{y}_u$. This comparison of the mean square errors is explored for reasonable choices between $\bar{y}_w$ and $\bar{y}_u$. Very recently (1990–91) some compromises are being suggested, especially "trimming" extreme weights, and "shrinkage" estimators. The problem becomes difficult for multipurpose surveys, which are much more common than a single purpose $\bar{y}_w$.

Key words: Selection probabilities; unequal selections; selection biases; self-weighting.

1. Introduction

Fundamental questions about weighting seem to be *the* most common during the analysis of survey data and I encounter them almost every week, requiring prompt and practical actions. The requests come from social researchers of all kinds. But I cannot find textbooks or references for them, because we "lack a single, reasonably comprehensive, introductory explanation of the process of weighting" (Sharot 1986), readily available to and usable by survey practitioners, who are looking for simple guidance, and this paper aims chiefly to meet some of that need. Some partial treatments have appeared in the survey literature (e.g., Bailar, Bailey, and Corby 1978; Kish 1965, 11.7, 1987, 7.4, 1989), but the topic seldom appears even in the indexes. However, we can expect growing interest, as witnessed by many publications since 1987 listed in the references. With their concentrations and with their style they aim at technical statisticians, whereas I address social researchers and statisticians who want advice for applied problems.

This paper aims to help researchers to find reasonable solutions to practical problems of weighting their data. Here follow some typical questions posed to sampling consultants by client researchers concerned with their data of sample cases. (a) Now that you (or we) have found that the cases had different selection probabilities (P_i), should these data be weighted by $w_i \propto P_i^{-r}$? (b) In

[1] Professor, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106-1248, U.S.A.

Acknowledgement: This version was considerably improved by critical remarks from Stuart Scott, Paul Flyer, Frank Potter, Vijay Verma, and an anonymous referee.

which situations is it "proper," or "necessary," or "important" to weight data? (c) Should we weight for nonresponse rates $(1 - r_h)$ which differ between classes h in the data? (d) When does weighting make a considerable difference in estimates? (e) How can we calculate proper and accurate weights? (f) How can we apply weights to cases, tapes, and estimates? (g) How do we apply weights in formulas and software?

The long neglect of weighting as a distinct topic in statistics textbooks reveals an interesting schism in our literature, I propose. Most of statistics deals with *identically and independently* distributed (IID) random variables, where differential weighting need not occur as a topic. On the other hand, sampling methods deal mostly with *selection* procedures; and this concentration is realistic, because in practice the statistical analysis of survey data is often removed in time and personnel from the selection process. But definitions of sample design often include both selection and estimation, and the two aspects cannot be completely separated. Weighting clearly pertains to estimation but it is also related to selection probabilities. Nevertheless, and despite their emphasis on "unbiased estimators," most sampling books refer to weighting only separately in connection with two or three distinct problems.

Some kind of weighting is frequently involved in the analysis of many survey reports, and *ad hoc* explanations appear sometimes, usually hidden in appendices behind project reports. On the other hand, we can also find articles with theoretical discussions that are concentrated only on some single specific aspect of weighting, such as stratification, or post-stratification, or nonresponses, or variance reductions.

We can also encounter misleading statements, even among some theoretical discussions based on diverse models, which are opposed to weighting. Researchers with sample data based on unequal selection probability must face the question whether to use weighted estimates like $\bar{y}_w = \Sigma w_i y_i / \Sigma w_i$ or simple unweighted (equal weighted) estimates like $\bar{y}_u = \Sigma y_i / n$. They are often confused by misleading statements resembling those below, though these are extreme forms of common misconceptions. (a) Weighting data by $w_i \propto P^{-1}$ is a "simple" process that should be "always" applied to samples with unequal P_is (according to "design-based" theory). (b) We cannot find any justification for weighting in "model-based" theory. (c) Weighting is needed for means (like $\bar{y}_w$) but not for testing hypotheses or for regressions, because these are model-based. (d) It is unethical to weight sample cases, because the process can be misused to produce biased results. We cannot fully explore all the deep implications of such misleading statements. Rather, we address the practitioners who want to know WHY, WHEN, and HOW to weight their data.

I must avoid those arguments in this simple and brief treatment, which aims to be general and useful (but see Brewer and Mellor 1973; Hansen, Madow, and Tepping 1983). In order to satisfy those two criteria, to be both simple and general, I had to forsake any attempt at profundity and precision. Anybody who tries to satisfy all three criteria of simplicity, generality, and profundity is bound to fail, probably on all three, I believe. Greater length would be especially needed to also treat the technical subjects of "optimal" weights for estimation. But here, as in the references cited, we are concerned with questions of whether and how to compensate with inverse weights for unequal selection probabilities, as clarified in Section 2. The basic problem is most simply stated by Spencer

and Cohen (1991) in introducing "shrinking" for a compromise (but their unweighted Z_m is my $\bar{y}_u$ and their unbiased Z_u is our $\bar{y}_w$):

"A longstanding question in making inferences from unequal probability samples is whether to use an unweighted or other model-based estimator, say Z_m, or whether to use an approximately unbiased estimator Z_u that uses sampling weights reflecting the unequal selection probabilities. (Moments are defined with respect to the sampling design unless otherwise noted.) An unweighted estimator of a population mean often will have smaller variance than a weighted estimator but it will have a bias proportional to the correlation between the characteristic of interest and the sample weights (Rao 1966). For many sampling strategies, the variances of Z_m and Z_u alike decrease to zero as the sample size increases, but although the bias of Z_u is zero or approaches zero the bias of Z_m does not. In such cases, for sufficiently large samples Z_u will have smaller mean square error. On the other hand, for small samples Z_m may have a smaller mean square error (Cochran 1977, p. 296–297). If one could know the mean square errors of Z_m and Z_u one could easily choose the optimal one. Fortunately, it is possible to use the sample itself to estimate the mean square errors, as DuMouchel and Duncan (1983) proposed in a different context."

2. Reasons for Weighting

I distinguish here seven separate main sources of weighting, because they usually have very different effects, and also because they need different strategies and treatments. Of these seven the first three arise from different selection probabilities P_i for the sample cases. Compensatory weights ($w_i \propto 1/P_i$) are the main concern of this and many papers, and for much applied research.

The general and most useful form of weighting is to assign the weights w_i to the sample cases i, with $w_i = 1/P_i$. The selection probabilities P_i for all sample cases must be known for all probability samples by definition. (Obtaining the actual numbers is often a nontrivial but necessary task.) Then the weighted mean is computed as $\bar{y}_w = \Sigma w_i y_i / \Sigma w_i$, and similar "consistent" estimates are discussed in Section 3, as well as the use of convenient weights $w_i \propto 1/P_i$ permitted in averages by the "normalizing" sum of weights $(\Sigma w_i)^{-1}$. The probabilities P_i may have to be computed from complex multistage (or even multiphase) processes. The weights may also be used to compensate for nonresponses, so that $w_i = 1/(P_i r_i)$ where r_i is a response rate often calculated for classes of response.

1. *Disproportional sampling fractions* f_h can be introduced deliberately to decrease either variances or costs. Often these are made with "optimal allocations" to distinct strata h, according to the well known allocation formula $f_h \propto S_h/\sqrt{C_h}$. But they may also result from two (or multi) phase selections. Samplers often achieve spectacular gains in variances (and costs) with these methods, especially applied to surveys of establishments. Gains in household surveys are less spectacular and frequent, but possible (Kish 1961). These deliberate differences in the sampling fractions f_h should be large to be effective, by factors from 2 to 10 and even greater; smaller differences seldom produce large enough effects to be worthwhile (Kish 1987, 4.5; 1965 11.7). The differences among f_h should also be highly related to survey variables. The f_h may be simple integral multiples of a basic sampling rate f, like $2f$ or

$10f$. These should (always) be compensated with inverse weights (e.g., 1/2 or 1/10) in order to avoid bad biases in combined statistics.

2. *Allocation to domains* of different sampling fractions f_h happens commonly and for distinct reasons. It is common to increase the sampling fraction from f to kf ($k > 1$) in order to reduce sampling errors in one or more domains, especially for small provinces. Or the rates in one or two provinces may be reduced to f/k to save overall costs. Sometimes equal sample sizes n^* are designed for unequal domain sizes N_h, so that the sampling rates are $f_h = n^*/N_h$. These inequalities in f_h are commonly compensated with unequal weights $w_h \propto 1/f_h = N_h/n_h$. However, the need for weights may not be quite as compelling as in A, because the weights may be less extreme, and relations of weights for domains to survey variables less strong. Estimates for means may be weighted but not for regressions perhaps (Section 6).
3. *Frame problems* may induce inequalities in selection probabilities P_i that may need compensating with $w_i \propto 1/P_i$. There are four basic classes of frame problems (Kish 1967, 2.7). (1) *Small clusters* of unequal sizes N_i are common; for example, dwellings are commonly selected with equal f and then a single adult from the N_i adults in the dwelling, then $P_i = f/N_i$. Since the number of adults are mostly few ($N_i = 1, 2, 3, 4$ mostly), the biases may be moderate for unweighted means, and the weighted means have variance increases of 1.05 to 1.20 mostly. In one case buildings were selected with equal f, then a single dwelling with $1/N_i$, so that for the dwellings $P_i = f/N_i$ and the N_i ranged from 1 to 62; the biases for unweighted estimates were large, and the variance increase for weighted estimates was 2.6 (Kish 1977). (2) *Duplicate (replicate) listings* may result in $P_i = d_i f$ when elements with d_i listings are selected with f applied to listings. If replications are common and uncorrected, considerable bias may result, and compensation with $w_i \propto 1/d_i$ is needed. (3) *Blanks and foreign elements* among listings selected with f cause no inequalities if they are simply disregarded. But a common mistake of substituting the "next valid" listing often causes $P_i = L_i f$ where $(L_i - 1)$ represent the invalid blanks before the valid listing. (4) *Missing units* (incomplete frame, noncoverage) refer to elements (units) missing from the sampling frame, hence $P_i = 0$. Since this would mean that w_i is not defined $(1/P_i = 1/0)$, obviously other measures are needed, and they receive much attention, though never satisfaction.
4. *Nonresponses* present problems that differ from those of 1, 2, and 3, which can be compensated with "inverse P_i" weighting. Weights for nonresponses must involve models or assumptions of some kind, explicit or implicit. It is common practice to assume implicitly that nonresponses arise randomly within response subclasses, though they differ between these subclasses. Thus differential response rates r_h are computed within those subclasses. Thus the sample cases receive weights $w_i = 1/(P_i r_h)$. The subclasses h of the sample are formed with auxiliary variables, such as age, gender, geography, etc. These variables must be (a) available for response cases, (b) somehow also for the nonresponse cases, and (c) also related to the survey variables. It is difficult and rare to obtain data either from the sample or from check statistics that closely satisfy the last two requirements safely and to a high degree. When nonresponses are not high, the differences between subclasses tend to be small, and then small

differences in weights will not have large effects on combined results.

For *item nonresponses* compensations seem more often justified, and they are usually made with imputation (replication) of responses (Kalton 1983a; Little and Rubin 1987; Rubin 1987). Corrections and weighting for *noncoverage* are much more difficult than for nonresponses, because coverage rates cannot be obtained from the sample itself, but only from outside sources. These may be done with "post-stratification," discussed below, where they more properly belong.

The sources above concern unequal probabilities of selection P_i, known from determinate selection operations. They are the chief subject of controversies about weighted versus unweighted estimates. However, the next three sources and types are motivated by estimation, rather than selection, and use models and auxiliary data sources. Some may question whether "weighting" is an appropriate term for these methods. Nevertheless the procedure can be summarized with factors c_i so that $w_i = c_i/(P_i r_i)$ can be used for the ith case. However, there are so many possible methods that we must limit our discussion to a few examples.

5. *Statistical adjustments* for improved estimates have diverse names: post-stratification, ratio estimators, and regression estimators are all described in the sampling literature, for reducing variances with controls that were not used in the selection process. In practice, however, post-stratification may denote a ratio estimator for reducing the biases of nonresponse and especially of noncoverage. Thus the ratio estimator $\Sigma X_h y_h/x_h$ with the auxiliary variable X is also $\Sigma N_h y_h/n_h$ in post-stratification for the population size N; these aggregates become means when divided by ΣX_h or ΣN_h. For example, check data from the census may be used for correcting for age-sex-race biases of surveys. As an extreme it adjusts for a noncoverage of young black males of 13% (USBC 1978, Ch V; Kish 1987, 4.7). For those methods the data and the software must both be appraised for integrity; large biases may be introduced with inadequate models or data. Other technical methods also appear in the literature, such as weighting cases proportional to their precision, $w_i = 1/\sigma_i^2$ (Kalton 1968).
6. *Adjustments to match controls* can have a variety of motivations. Whereas the reasons under 5 for post-stratification and ratio estimation concerned mostly sampling variations, here we refer mostly to adjustments of samples from one frame population to some other target (standard) population(s), often known as *standardization*. For example, a sample from one province (state) may be reweighted to the national population (Kish 1987, 4.5). Or we may reweight samples from one country or period to another country or period. Generally, the subclasses of the sample are reweighted to the domains of the target population, and these controls must be available both for the sample and for the target population. If there are too many cells, the control data may be unavailable and the sample cases too few for stability, and then marginal adjustments may be used, with iterated fitting.

   Reweighting may also be used to examine differences, like $(\bar{y}_a - \bar{y}_b)$ of two (sub)populations (a and b) free from the effects of the different "compositions" (N_{ha} and N_{hb}) of the two populations in strata h (Kish 1987, 4.51D).

   Adjustments of nonprobability samples to fit check data in subclass cells, (e.g., age, sex, and province) are also

made. These adjustments can hardly overcome the biases of nonprobability selections within the subclasses. They may be viewed as similar to "quota" sampling, but with weighting substituted for selection.

7. *Combining samples* is becoming more popular, more important, and more feasible because of increasing numbers of samples that are available for combinations. All combinations concern weighting in some form, and one should always be explicit about the weights; and also careful about possible differences in measurements. We note also that any national sample combines diverse domains, some like provinces, some like diverse social or demographic classes; and all those domains differ in the distributions of the survey variables. Nowadays, one may also combine standardized national samples from several countries, e.g., African birth rates from separate national samples of the World Fertility Surveys. Similarly to spatial integration, we may also combine periodic samples into *rolling samples* integrated over a longer time span; e.g., annual averages of influenza, or cancer rates, or unemployment, or incomes from weekly or monthly surveys (Kish 1990). *Meta-analysis* is a growing field for combining statistics, and already foreshadowed in 1924 by Yates and Cochran (1938). A special and simple method of combining can be the *cumulations* of individual cases (Kish 1987, 6.6).

3. Methods for Implementing Weighting

Four alternative procedures for weighting need individual attention because they require different techniques and also because they can have different effects on the variances.

1. *Individual case weights* (ICW) yield the most common, simple, practical, and flexible procedures, especially with modern computers and programs that can handle them. (Not all programs are equally adept.) The other procedures may be compared with and based on ICW, and they may increase variances more than ICW. The weights w_j for sample elements ($j = 1, \ldots n$) may reflect a product $p_j r_j$ of the element probabilities p_j from complex statistical multistage selections with the response rates r_j, which may also include coverage rates. The weights $w_j = 1/(p_j r_j)$ are inversely proportional to these products. Both p_j and r_j should be available for all elements of probability samples, and newcomers to surveys must be made aware that those values must be obtained with careful bookkeeping. It is also possible to incorporate weights W_h for post-stratification (ratio) estimators so that $w_j'' = W_h/(p_j r_j)$.

The basic statistic is the weighted sum of sample values $\Sigma w_j y_j = \Sigma y_j/p_j$. This is a desired unbiased estimator of the population sum $\Sigma Y_i = Y = N\bar{Y}$. For equal probability selections, or *epsem*, of n from N elements (whether simple random or more complex), $f = n/N$, we have the expected value $\text{Exp}(\Sigma y_i/f) = \Sigma\text{Exp}(y_i N/n) = \Sigma \bar{Y}N/n = N\bar{Y} = Y$. For weighted estimates we also have $\text{Exp}(\Sigma y_i/p_j) = N\bar{Y} = Y$. This is shown in all sampling books, sometimes as a "Horvitz-Thompson" estimator (Cochran 1977, 9A.7; Kish 1965, 2.8C). This simple expansion estimator is basic to probability sampling and should perhaps be called an "expectation estimator." In practice it is seldom used in this simple form and it needs adjustment for nonresponses, so that $w_j = 1/(p_j r_j)$.

The most common statistic is the weighted mean $\bar{y}_w = \Sigma w_j y_j/\Sigma w_j$. This is

Table 1. Disproportionate allocation to illustrate weights

W_h	N_h	f_h	r_h	n_h	$1/w_j$	w_j	$w_j n_h$
.900	90,000	.01	.90	810	.009	111.111	90,000
.09	9,000	.10	.85	765	.085	11.765	9,000
.01	1,000	1.0	.8	800	.80	1.25	1,000
1.00	100,000 = N					$\Sigma w_j n_h$ =	100,000

A population of $N = 100{,}000$ elements was divided into three highly unequal strata of N_h elements. The disproportionate selection rates f_h applied and the different response rates r_h obtained result in $\Sigma n_h = 2{,}375$ observations. From $p_j r_j = 1/w_j$ in the three strata the weights $w_j = 1/(p_j r_j)$ are obtained. Note that $\Sigma w_j = w_j n_h = N_h$ exactly for each stratum, though in practice minor irregularities would cause small variations. For random variables Y_i we do not get the Y_h exactly, but we get the expectation $E(\Sigma w_{hj} y_{hj}) = Y_h$.

not "unbiased" technically (because it is a ratio estimator) but it is "consistent"as are the other similar statistics. Thus with $\Sigma w_j y_j^2 / \Sigma w_j - \bar{y}_w^2$ for the element variance; also with $\Sigma w_j y_j x_j / \Sigma w_j$ replacing $\Sigma y_j x_j / n$ from *epsem* selections. Because they are normalized (standardized) with Σw_j, the weights can be any positive numbers proportional to the expansion weights $1/(p_j r_j)$. It may help to note that for *epsem* selections we have the expansion weights $w_j = 1/f = N/n$ and $\Sigma w_j = N$; whereas in common formulas $w_j' = f/f = 1$ and $\Sigma w_j' = n$.

When we can find appropriate, unbiased, and dependable values of N_h for the population, the sampling fractions n_h/N_h in domains h can be used sometimes for $w_j = N_h/n_h$, and this is justifiable when the elements j are selected with actual equal probabilities within the domains h. On the other hand, in many situations the selection probabilities f_h must be applied, because reliable N_h are not available. However, it is misleading to confuse a mere fraction of elements in the sample with a true sampling probability; e.g., that a sample of n_h is selected from a population of N_h one may perhaps refer to a sampling "fraction" of n_h/N_h. But probabilities of selection must be justified with probability operations: otherwise we are faced with judgment samples, "quota" samples, and other model dependent sampling.

2. *Weighted statistics*, e.g., $\bar{y}_w = \Sigma W_h \bar{y}_h$, combine separate subpopulation statistics $\bar{y}_h$ with appropriate relative weights W_h, with $\Sigma W_h = 1$. This method may be preferred over ICW for: (a) combining published statistics when individual cases are not available; (b) combining a few strata based on disparate selection procedures; and for (c) relatively simple statistics, like means or totals. But they are not as useful for complex analyses of single surveys. Dependable weights W_h are needed from justifiable sources. These can also be used as $w_j = W_h$ with the ICW, as above.
3. *Duplication of cases* may be used instead of ICW in order to prepare self-weighting tapes for convenience in some situations. It is especially convenient for item nonresponses, and particularly for complex analytical statistics, because both reasons hinder individual weights (Kalton 1983a). Some compromise between random selection and "closest" matching to reduce bias is generally used for duplication within subclass cells. If the response rate is r_h in cell h, $(1 - r_h)$ cases

can be duplicated to fabricate $(1 - r_h)$ pseudo cases; either randomly selecting with probability $(1 - r_h)/r_h$ or by finding the "closest" matching of that fraction of cases. Duplication increases variances over individual weighting ICW, but those increases are not great for duplicating only a small proportion of the samples. Furthermore, these increases of variances can be almost eliminated with procedures of "multiple replications" (Little and Rubin 1987; Rubin 1987).

We must caution against the crude mistake of accepting from the computing programs the tape counts (or card counts) m, which contain $(m\text{-}n)$ replicates as well as n genuine cases. The n genuine cases can be "tagged" for counting. But the "effective number" may be further diminished by duplication to $n' = n/(1 + L)$, as noted in 4.2.

4. *Elimination of cases* can be justified in some situations, although throwing away information may appear statistically criminal and is seldom practiced. Nevertheless, consider three justifiable situations. (a) Large samples have been selected with different sampling rates for a nation's several provinces; then a self-weighting sample is designated for complex national analysis, with rates suited to the lowest provincial rates; "microtape" samples can be made self-weighting. (b) A small domain has been greatly oversampled for separate analysis, but a proportionate sample has been "tagged" from it for joint complex analysis, which could be difficult and not much more precise with the extra cases from the small domain. (c) Eliminating a small proportion of cases (say $< .05$) increases the variance only little more than duplication of a similar proportion. This counterintuitive result can be used for compromise adjustment for differential nonresponses between strata (Kish 1965, 11.7B).

4. Reasons Against Weighting

1. *Complications* often arise from weighting, even when good computing programs are available, and this factor is often neglected and difficult to quantify. That is why I put it first, though it should not be the most important. Mistakes arise in the man-machine system and they tend to increase for more complex analyses. Other complications are more basic: for complex, analytical statistics, and for inferential statistics, such as tests of significance, adequate methods may not be available for weighted estimators or for their sampling errors. Theoretical contributions are now fast developing, but they are not useful, general, and simple enough to be "available" for many survey practitioners (Kott 1991a, b; Rao and Scott 1981).
2. *Increased variances* can result from weighting for random, or haphazard, or irregular differences in selection probabilities, when these are not "optimal." For example, the inequalities due to frame problems or to nonresponses are generally of this kind. Furthermore, these increases of variances (unlike those due to clustering) tend to persist undiminished for most such subclasses and for all statistics, as if they were to increase the element variances from σ^2 to $(1 + L)\sigma^2$, or to decrease the "effective" number of elements from n to $n/(1 + L)$. Here L denotes relative loss, so that $L = 0.8$ means a factor of $1 + L = 1.8$ increase in the variances, explained below.

"Haphazard" sources of weights are most common in survey work, but they are counterintuitive to minds attuned to "optimal allocation," source 1 in Section 2. Those weights are highly related to

stratum values $\bar{y}_h$ and σ_h. But $(1 + L)$ refers to other sources (2, 3 and 4), small domains, frame problems, and nonresponses, which are hardly related (negatively or positively) to most survey variables. Therefore the best summary measure of their effect is a relative increase of $(1 + L)$ in variances, a statement based on much experience in multipurpose surveys. For example, even a sample with optimal allocation for *mean* incomes and assets turns out to be less efficient than proportional allocation for buying behavior and even for *median* income and assets in the same sample (Kish 1961; Verma, Scott, and O'Muircheartaigh 1980).

Three simple formulas yield adequate estimates of the increases $(1 + L)$ in element variances. The choice between these three alternatives depends on the situation and the data available for the weights to be used. In these formulas N_h represent population sizes and n_h sample sizes for strata h, and $W_h = N_h/\Sigma N_h$ and $w_h = n_h/\Sigma n_h$ denote relative sizes, so that $\Sigma W_h = \Sigma w_h = 1$. The weights are represented by k_h, which can be $1/f_h$, the inverse of the selection rates, but they may be only *relative* values proportional to them, $k_h \propto 1/f_h$. I use k_h or k_j for w_j here for easy comparison with references, where 4.1 to 4.5 are derived (Kish 1967 11.C, 1976, 1988).

a. In the design stage one may consider using sampling fractions and weights in the proportions k_h in strata with relative sizes W_h. If the element variances are roughly equal ($\sigma_h^2 \simeq \sigma^2$ approximately) then the variance of the mean (and many other statistics) will increase approximately by the factor

$$(1 + L) = (\Sigma W_h k_h)(\Sigma W_h/k_h). \quad (4.1)$$

b. In the analysis stage, if n_h cases have weights k_h, the increase in the variance (with conditions as above) is approximately

$$(1 + L) = \Sigma n_h \Sigma n_h k_h^2/(\Sigma n_h k_h)^2. \quad (4.2)$$

For individual element weights k_j when $n_j = 1$ for all n cases (4.2) becomes simply

$$(1 + L) = n\Sigma k_j^2/(\Sigma k_j)^2. \quad (4.3)$$

Note also that the relative increase or loss L may be viewed as the *relvariance* cv^2 = variance/ mean^2 of the relative weights k_j, because

$$cv^2 = n\Sigma k_j^2/(\Sigma k_j)^2 - (\Sigma k_j)^2/(\Sigma k_j)^2 = (1 + L) - 1 = L. \quad (4.4)$$

Thus the factor $1 + cv^2 = 1 + L$ depends on the relative variances of case weights k_j. It serves as good precaution to compute the cv^2 or $1 + L$, or the frequency distribution of the weights to estimate what the increase may be.

c. When the population sizes N_h and sample sizes n_h are both directly available they can be used directly without the relative weights k_h to compute

$$1 + L = (\Sigma N_h^2/n_h)n/N^2. \quad (4.5)$$

3. *Lower mean square errors* (MSE) may be achieved by unweighted, biased estimators, such as means $\bar{y}_u$. Comparisons with weighted means $\bar{y}_w$ can be based on MSE $(\bar{y}_u) = S^2(1 + B^2/S^2)$ versus Var$(\bar{y}_w) = S^2(1 + L)$. The bias ratio B/S for $\bar{y}_u$ can be estimated from the ratios $(\bar{y}_u - \bar{y}_w)/\text{ste}(\bar{y}_u)$ computed from survey data for several (many) survey variables.

Similarly, the factor $(1 + L)$ can be estimated with $(1 + C_k^2)$ from the sample with (4.3) or anticipated in the design

with (4.1), and this increase in the variance is rather constant for most statistics. The factor C_k^2 is also important for B^2/S^2 (4.8), but these bias ratios differ greatly for diverse survey variables of the same survey. Furthermore for subclasses and their comparisons the variances are much higher, hence the ratios much lower (Section 7).

4. *Model dependent arguments* have been advanced that weighting corrections for selection biases are not needed for regressions from surveys (Brewer and Mellor 1973; Hansen, Madow, and Tepping 1978).
5. *Public relations or ethics* may also hinder overt and differential weighting, because it is possible to misuse it to produce subjectively desired, prejudiced results (Sharot 1986). For example, the combined mean $\bar{y}_w = \Sigma W_h \bar{y}_h$ could be made to approach any of the components $\bar{y}_h$ with extreme weights W_h. Journalists, alas, do this commonly, by using the cost of either automobiles and TV sets, or housing and health care to contrast the cost of living in economies with contrasting price systems. The naively prejudiced weights tend to escape the public's scrutiny, but any explicit weighting system suffers from exposure, unfortunately.

5. Balancing Variance Increases Against Biases

We saw that the ratio of increase of variances due to (haphazard) weighting can be computed as

$$1 + L = 1 + C_k^2 = 1 + \sigma_k^2/\bar{k}^2 \tag{5.1}$$

from the data (4.3) or anticipated in the design (4.1). This has been shown to be true generally for departures from optimal allocating in linear sample designs (Kish 1976). The relative bias $-B = (\bar{Y}_w - \bar{Y}_u)/\bar{Y}_u$ of unweighted samples can also be estimated from the sample data. But it is remarkable and useful that these biases for means can also be shown to depend on the same C_k^2, and on the correlation R_{ky} between the weights and the survey variables (Kish 1987, 7.4.12). Thus:

$$\begin{aligned} -\text{Bias} &= \bar{Y}_w - \bar{Y}_u = \Sigma y_i k_i/\Sigma k_i - \bar{Y}_u \\ &= \bar{k}^{-1}[N^{-1}\Sigma y_i k_i - \bar{k}\bar{Y}_u] \\ &= \bar{k}^{-1}\text{Cov}(k_i y_i) \\ &= \bar{k}^{-1} R_{ky}\sigma_k\sigma_y, \end{aligned} \tag{5.2}$$

where $\bar{k} = \Sigma k_i/N$. These summations are to the population N, but they can be made to the sample n in sample estimates, which concern us most. The sample difference $(\bar{y}_w - \bar{y}_u)$ estimates the population difference $(\bar{Y}_w - \bar{Y}_u)$, with the expectations: $\text{Exp}(\bar{y}_w) = \Sigma(P_i Y_i/w_i)/N = \Sigma Y_i/N = \bar{Y}$, the population mean; but $\text{Exp}(\bar{y}_u) = \Sigma P_i Y_i/N = \bar{Y}_u = \bar{Y} + \text{Bias}$, with the entire population exposed to unequal selection.

Thus $\bar{Y}_u - \bar{Y}_w = \text{Bias}$, because $\bar{Y}_w = \bar{Y}$ is unbiased, but $\bar{Y}_u$ is biased; but I used (5.2) for convenience, because in the sample we have $\bar{y}_u$ most conveniently. Similarly relative values based on $\bar{Y}_w$ are preferable on theoretical grounds, for B, S, and C_y, as in (5.5). But I used $\bar{Y}_u$ for everyday convenience and because it has lower variances.

$$-B = (\bar{Y}_w - \bar{Y}_u)/\bar{Y}_u = R_{ky} C_k C_y$$

and

$$B^2 = R_{ky}^2 C_y^2 C_k^2. \tag{5.3}$$

In order to contrast the increase in variance $(1 + C_k^2)$, with the effects of biases, let us consider a mean with the effective sample size $n_d = n/\text{Deff}$, where Deff is the "design effect," often appreciably greater than 1; these effects, $\text{Deff} = \text{Var}(\bar{y})/(S_y^2/n) > 1$, have been computed and used in many studies (Kish 1976). Then the relative mean square errors (RMSE) for $\bar{y}_w$ and $\bar{y}_u$,

respectively, are:

$$S^2(1 + C_k^2) = C_y^2(1/n_d + C_k^2/n_d) \quad (5.4)$$

and

$$B^2 + S^2 = C_y^2(1/n_d + R_{ky}^2 C_k^2). \quad (5.5)$$

These relations were expressed in relative terms, with the biases and variances divided by $\bar{Y}_u^2$. I preferred these in order to make comparisons easier for many variables within the same survey and also between surveys. However, some may prefer to avoid that, especially for situations where division by $\bar{Y}_u^2$ is inappropriate, as when $\bar{Y}_u$ is near zero or is a proportion when P or $(1 - P)$ may be confused. But the same relation may be found in

$$\text{Var}(\bar{y}_w) = S_y^2(1/n_d + C_k^2/n_d) \quad (5.6)$$

$$\text{Bias}^2 + \text{Var}(\bar{y}_u) = S_y^2(1/n_d + R_{ky}^2 C_k^2). \quad (5.7)$$

Thus the relative increase C_k^2/n_d for the variance of $\bar{y}_w$ decreases along with the variance itself. But the effects of the biases in $\bar{y}_u$ do not decrease, hence come to dominate for large samples; and these functions of R_{ky}^2 tend to vary greatly between variables.

From the relations above we construct some useful guidelines for practical work.

1. Values of $(1 + C_k^2)$ should be estimated in the design (4.1) or from the sample (4.3). When C_k^2 is moderate its effects on both biases and variances will be small, except for very large n_d and large R_{ky}^2. For example, when $C_k^2 < 1$, $C_k^2/100 < 0.01$ and $R_{ky}^2 < 0.01$ if $R_{ky} < 0.1$. Furthermore $(1 + L)$ can be often guessed well enough from Table 2 and values of the range $K = k_{\max}/k_{\min}$ of the relative k_i. For example, for $K = 1.3$, L is between 0.01 and 0.02; and this may be true for many nonresponse weights; for $K = 1.5$, L is between 0.015 and 0.04. For $K = 2$, L is between 0.04 and 0.125 and may be worth computing; and this is true for $K \geqslant 3$. For example, the value of $L = 0.33$ I computed for the Current Population Surveys would explain why only mild differences $(\bar{y}_w - \bar{y}_u)$ were found by Bloom and Idson (1991).

2. Note that C_k^2 depends on both the range of the k_j and the frequency distribution represented by W_h or $n_h = W_h/k_h$. Hence C_k^2 will be large only when $k_{\max}$ is large for large portions, W_h or n_h. Large values of L in Table 2 occur for large K, and for dichotomies, especially for W_h (or U) = 0.5.

3. C_k^2 is stable for the sample, but C_k^2/n_d is much increased for small n_d in subclasses. On the contrary, the values of R_{ky}^2 can vary by orders of magnitude between variables of the same survey. Therefore the effect of biases can be large for some variables, especially for the entire sample with large n_d; but it may be small for subclasses with small n_c, or small R_{ky}^2. Weighted estimates should be preferred when $R_{ky}^2 > 1/n_d$. Perceptible biases are seldom found, because the squared correlations have small effects.

4. We have disregarded here

a. the extra costs (troubles) of weighting;
b. the advantages of self-weighting samples;
c. the advantages of compromises between weighted and unweighted estimates, noted below in 7;
d. implications also for more complex statistics, such as regression coefficients.

6. Epsem Selections for Self-Weighting Samples

Self-weighting samples are often preferred, because they possess considerable advantages in reduced variances, in simplicity, and in robustness. Statistical theory also, from the lowest to the highest, overwhelmingly assumes self-weighting samples in one form

Table 2. Relative losses (L) for six models of population weights (U_i); for discrete (L_d) and continuous (L_c) weights: for relative departures (K_i) in the range from 1 to K[a,b]

Models	K	1.3	1.5	2	3	4	5	10	20	50	100
Dichotomous $U(1 - U)$											
(0.5)(0.5)		0.017	0.042	0.125	0.333	0.562	0.800	2.025	4.512	12.005	24.50
(0.2)(0.8)		0.011	0.027	0.080	0.213	0.360	0.512	1.296	2.888	7.683	15.68
(0.1)(0.9)		0.006	0.015	0.045	0.120	0.202	0.288	0.729	1.624	4.322	8.82
Rectangular	L_d	0.017*	0.042*	0.125*	0.222	0.302	0.370	0.611	0.889	1.295	1.620
$U_i \propto 1/K$	L_c	0.006	0.014	0.040	0.099	0.155	0.207	0.407	0.656	1.036	1.349
Linear decrease	L_d	0.017*	0.040*	0.111*	0.203	0.283	0.353	0.616	0.940	1.437	1.917
$U_i \propto K + 1 - K_i$	L_c	0.006	0.014	0.040	0.097	0.153	0.205	0.409	0.680	1.127	1.514
Hyperbolic decrease	L_d	0.017*	0.040*	0.111*	0.215	0.312	0.404	0.807	1.466	3.014	5.076
$U_i \propto 1/k_i$	L_c	0.006	0.014	0.041	0.103	0.171	0.235	0.528	1.011	2.138	3.621
Quadratic decrease	L_d	0.016*	0.036	0.080*	0.150	0.211	0.264	0.460	0.696	1.048	1.333
$U_i \propto 1/k_i^2$	L_c	0.006	0.014	0.040	0.099	0.155	0.207	0.407	0.656	1.036	1.349
Linear increase	L_d	0.017*	0.040*	0.111*	0.167	0.200	0.222	0.273	0.302	0.320	0.327
$U_i \propto k_i$	L_c	0.006	0.013	0.037	0.088	0.120	0.148	0.223	0.273	0.308	0.320

[a]From Kish, 1976 and 1987.

[b]Dichotomous, $1 + L = 1 + U(1 - U)(K - 1)^2/K$. Also all*. Discrete, $1 + L_d = (\Sigma U_i k_i)(\Sigma U_i/k_i)$, with $k_i = i = 1, 2, 3, \ldots, K$. Continuous, $1 + L_c = \int Uk\, dk \int (U/k)dk$, with $1 \leqslant k \leqslant K$. Only two values, 1 and K, were used for L_d for $K = 1.3$, 1.5, and 2.

or another. Furthermore, selections of elements with equal probabilities, *epsem* for short, often seems desirable and reasonable when the survey variables are more or less evenly distributed over the population. Voting by all adults springs readily to mind, but there are other behaviors, attitudes, and opinions which are also democratically, evenly distributed, or at least roughly so.

Self-weighting samples for analysis is a goal for many surveys and epsem selection is the principal means toward that end. Because we often confuse the two, let me clarify how they exist in common practice. Disproportionate "optimal allocations" (source 1 in Section 2) clearly are not epsem, and hardly anybody would use them for self-weighted (i.e., unweighted) analysis. Samples with oversampled (or undersampled) domains (source 2) are not epsem and they also are not self-weighting. When frame problems (source 3) result in unequal P_i of selection the question of weighting becomes quantitative (Sections 4, 5, and 7). For example, fertility studies may find that 1 to 5% of an epsem of households have two women of childbearing age; both can be selected to maintain the epsem f. If one is selected at random her probability is reduced to $f/2$, but in most cases the analysis may disregard the small factor $(1 + B^2/S^2)$ and proceed with self-weighting. On the contrary the example in Section 2 of single dwellings selected from buildings with N_i dwellings $(N_i = 1, 2 \ldots 62)$ presented gross violations with $P_i = f/N_i$; the variance factor was $(1 + L) = 2.6$, but the bias factors $(1 + B^2/S^2)$ were highly variable and much worse. It is best to avoid them, an important part of the sampler's art.

Nonresponses are impossible to avoid, but much can be done to reduce them and their effects (too much to detail here). Because of their omnipresence they should not cancel the label epsem, but if weights are used for compensation the analysis is not self-weighting. Thus epsem is not sufficient for self-weighting, and we saw above that epsem is not necessary in cases of trivial frame problems. The two are closely related but keeping the two concepts distinct clarifies both.

Similarly we must avoid the common confusion of epsem with simple random sampling (srs). Probably most survey samples are epsem, but very few are srs (outside academic writing). That popularity of epsem and self-weighting samples is due to their "robustness," I claim (Kish 1977). Three reasons were given in Section 4: avoiding complexity, the variance factor $(1 + L)$, and the bias factors $(1 + B^2/S^2)$. The fourth reason comes from the multipurpose nature of most surveys, and that self-weighting "satisfices" most purposes and comes close to optimizing ("satisficing") many or even most purposes (Kish 1976, 1988).

Unequal probabilities can be justified by "optimal" allocation if the purposes (all, or preponderant) of the survey justify it. In other situations, however, the sampler should search for ways to avoid the dilemma between the biases of unweighted and the increased variances of weighted samples. Achieving epsem samples during the selection operation is a fundamental skill in the art of survey sampling. This includes complex multistage selections with probabilities proportional (first directly, then inversely) to measures of sizes. Often it also requires clever handling of imperfections in the sampling frame. After selection, achieving acceptable response rates often needs skillful and devoted care.

7. Diverse Effects for Different Statistics

Users must be warned about these wide differences between effects, because many may be misled by the mere phrasing of "THE

Bias." The biases can be estimated from the sample by $(\bar{y}_u - \bar{y}_w)$, and should be. They vary greatly between the variables and depend strongly on the correlations between the survey variables y_j and the weights w_j; the bias may be negligible for most variables but large for others. The standard errors can differ also, increasing for small subclasses with small $\sqrt{n}$. However, they tend to differ less than the biases, and especially for proportions when $\sigma = \sqrt{P(1-P)}$ remains rather constant for all but extreme values of P.

1. *Expansion totals* $\hat{Y} = y/f$ are most sensitive to biases from weights; for example, even uniform and random nonresponses can result in bad underestimates, if not adjusted. Also expansions like $\hat{Y} = \Sigma N_h(y_h/n_h)$ can be very sensitive to biases in the borrowed values of the N_h. But differences or ratios of such totals from periodic studies would be less sensitive, therefore biases may be more tolerated in such comparisons.
2. *Means* are usually less affected than totals. Sample surveys survive the terrible nonresponse rates now prevailing in the USA (and elsewhere) only because nonrespondents do not differ drastically from respondents for most survey variables. Large biases result only from combinations of differences in *both* weights and survey variables within subclasses. If *either* of these is uniform over subclasses the net bias tends to be small (Kish 1965, 13.4B).
3. For *subclass means* the variances increase in proportion to the decrease of the sample bases, roughly in a ratio that may be denoted by $S_c = \sqrt{(\text{Deff}_c \sigma_c^2/n_c)}$. The n_c, σ_c, and Deff_c all refer to the subclasses c; the "design effects" Deff_c in cluster samples tend to decrease slightly from Deff > 1 toward 1 with decreasing subclass size, especially for "crossclasses," but much more slowly than n_c. Crossclasses refer to the majority of subclasses which cut across the clusters of sample designs; thus the Deff = [1 + roh$(\bar{b} - 1)$] of clustering tends to be reduced toward 1 as the average size of clusters reduces from $\bar{b}$ to $\bar{b}_c$ along with n to n_c (Kish 1987, 2.3; Verma, Scott and O'Muircheartaigh 1980). Since the biases B_c tend to be the same, either generally or on the average, the bias ratio B_c/S_c tends to decrease as S_c increases with decreasing sizes of subclasses.
4. For *subclass differences* $(\bar{y}_c - \bar{y}_b)$ the above process is greatly enhanced, because the standard errors $S = \sqrt{(S_c^2 + S_b^2)}$ are greater than for one subclass, and even more because biases often tend to be in the same direction, and thus tend to cancel in the differences. Then the bias ratio B/S with a drastically reduced numerator, and increased denominator, becomes greatly reduced (Kish 1987, 2.4–2.6). Hence the mean-square-errors $S^2(1 + B^2/S^2)$ become dominated by the variances S^2. "Model dependent" inference may go further and claim that weights can be disregarded in estimating subclass differences. However, I reject that null limit for B on philosophical grounds (Kish 1987, 2.4, 1.8).
5. *Analytical statistics* can be of many kinds, and a general statement about B/S seems difficult. I share the "population bound" view of inference that weighting matters and data have shown this for such analytical statistics as regression coefficients (Holt, Smith, and Winter 1980). When considerable differences are found between weighted and unweighted statistics (e.g., regression coefficients), I trust the former. It is also true that computational, methodological, and mathematical problems may pose formidable prob-

lems for weighted estimates. *Sampling errors*, inferential statistics, and tests of significance can also pose severe problems of computation, methodology, and interpretation for weighted estimates.

8. To Weight or Not to Weight? Compromises and Strategies

The literature on this topic, including my references, concern themselves with *the* bias of some statistic, often a regression equation, as if this were the single purpose of the survey. But most surveys, all in my experience, have many purposes; they are multipurpose and in several dimensions: several variables, several statistics, subclasses, etc., (Kish 1988). All these exhibit different relative effects $(1 + B^2/S^2)$ of biases. Is it possible or desirable to treat each statistic separately, perhaps using weighted estimates for means, but unweighted estimates for regressions, as implied by some?

Another paradox arises when we contrast academic literature with practice. In the former there are sharp contrasts made between the "population-bound" and the "model dependent" approaches (Brewer and Mellor 1973; Hansen, Madow, and Tepping 1978; Kish 1987, 1.4, 1.8). Perhaps the contrasts are sharpened by exaggerations of the opponent's misstatements. Most important, actual practice often leaned toward compromise, but this needs the guidance of theory.

Some theory has been coming in the past few years as my references show, and we may confidently await growing interest; the trend is toward compromises and the criterion of mean-square-errors predominates. Though the terms of $(1 + L)$ and $(1 + B^2/S^2)$ for relative increases in variances and biases may be mine, most of the concepts are not contradictory to them. I find it interesting and worthwhile to distinguish four "levels" of compromises.

1. At the very least comes the recognition that choices should be made for specific situations rather than a blanket population-bound stand of always weighting for unequal P_i versus a blanket model-based position of never weighting, especially for relationships and regressions (Bloom and Idson 1991; DuMouchel and Duncan 1981; Graubard and Korn 1991; Iannacchione, Milne, and Folsom 1991; Kish 1965, p. 400). I propose, however, that the MSE criterion of $(1 + B^2/S^2)$ is preferable to tests of significance for $|B/S| > 1.96$ at the 5% level.
2. At the next level we find some practice of *trimming* small percentages of extreme weights to accept small biases against large reductions in the $(1 + C_k^2) = (1 + L)$. It may be especially useful to trim the extreme "right tail" where a few cases of large k_j from small p_j (some of these outliers may be mistakes, others haphazard events) may greatly increase $(1 + C_k^2)$ (Flyer, Rust, and Morganstein 1989; Hidiroglou and Srinath 1981; Kott 1991a; Lee 1991; Potter 1988). I wonder if some smoother transformation than trimming of the k_j may do even better.
3. The *shrinkage* of weights seems to be a natural development that holds promise (Spencer and Cohen 1991). Instead of the *ad hoc* nature of trimming decisions, it points to a generalized, uniform treatment with theoretical bases. However shrinkage weights are specific for each variable and they can vary widely on multipurpose surveys. Trimming is uniform, and it affects only a small portion of the sample.

   Briefly, a shrinkage mean $\theta\bar{y}_u + (1 - \theta)\bar{y}_w$, with $0 \leqslant \theta \leqslant 1$, in practice implies transforming the weights to $g_j = \theta\bar{k} + (1 - \theta)k_j, = k_j - \theta(k_j - \bar{k})$. On the other hand, with trimming we get $g_j = k_j$ for $k_j \leqslant K$, but $g_j = \bar{K}$ for

$k_j > K$, where K is some strategically chosen constant; for two sided trimming $|k_j| > K$ may be used. However, some other compromise transformation between these two transformations may be even better. For example, a square root transformation $g_j = \sqrt{k_j}$, or $g_j = k_j^c$ with $0 \leqslant c \leqslant 1$. Empirical investigations would be welcome in situations where C_k^2 is large enough to distinguish between the diverse gains.

4. Nevertheless the *multipurpose* nature of most surveys raises problems not addressed by my references. When are specific answers for each statistic of a survey feasible and desirable? Or is a general overall answer for all statistics on a survey more acceptable, and how does one arrive at such a compromise? Having had to answer these questions in practice convinces me that they need more theory. The good news is that there is need for much research, both theoretical and empirical, and especially combined. I also hope for compromise average weights, adapted from those for allocations (Kish 1976, 1988).

I end with a few pieces of advice.

a. Always compute estimates of the factor $1 + C_k^2$ (Section 5).
b. If this $(1 + L)$ is large, see how much you can reduce it with trimming.
c. Compute many (30 or 60?) estimates of $(\bar{y}_w - \bar{y}_u)$ for different variables and different statistics.
d. Make comparisons with the factors $(1 + B^2/S^2)$ and justify your decision.

9. References

Bailar, B.A., Bailey, L., and Corby, C. (1978). A Comparison of Some Adjustment and Weighting Procedures for Survey Data. In N.K. Namboodiri (ed.), Survey Sampling and Measurement, New York: Academic Press.

Bloom, D. and Idson, T. (1991). The Practical Importance of Sample Weights. Proceedings of the Section on Survey Research Methods, American Statistical Association (in press).

Brewer, K.R.W. and Mellor, R.W. (1973). The Effect of Sample Structure on Analytical Surveys. Australian Journal of Statistics, 15, 145–152.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed. New York: John Wiley.

DuMouchel, W.H. and Duncan, G.S. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association, 78, 535–543.

Flyer, P., Rust, K., and Morganstein, D. (1989). Complex Survey Estimators and Contingency Table Analysis Using Replication. Proceedings of the Section on Survey Research Methods, American Statistical Association, 253–259.

Graubard, G. and Korn, E. (1991). Testing Informativeness of Sample Weights for Multiple Regression Analysis. Proceedings of the Section on Survey Research Methods, American Statistical Association (in press).

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.

Hidiroglou, M.A. and Srinath, K.P. (1981). Some Estimators of a Population Total from Single Random Samples Containing Large Units. Journal of the American Statistical Association, 76, 690–695.

Holt, D., Smith, T.M.F., and Winter, P.D. (1980). Regression Analysis of Data From Complex Surveys. Journal of the Royal Statistical Society, ser. Ā, 143,

474–487.

Iannacchione, V., Milne, J., and Folsom, R. (1991). Response Probability Weight Adjustments Using Logistic Regression. Proceedings of the Section on Survey Research Methods, American Statistical Association (in press).

Kalton, G. (1968). Standardization: A Technique to Control for Extraneous Variables. Applied Statistics, 17, 118–136.

Kalton, G. (1983a). Compensating for Missing Survey Data. Ann Arbor, MI: Institute for Social Research.

Kalton, G. (1983b). Models in the Practice of Survey Sampling. International Statistical Review, 51, 175–188.

Kish, L. (1961). Efficient Allocation of a Multipurpose Sample. Econometrica, 29, 363–385.

Kish, L. (1965). Survey Sampling. New York: John Wiley, 11.7.

Kish, L. (1976). Optima and Proxima in Linear Sample Design. Journal of the Royal Statistical Society, ser. A, 139, 80–95.

Kish, L. (1977). Robustness in Survey Sampling. Bulletin of the International Statistical Institute, 47, 515–528.

Kish, L. (1987). Statistical Design for Research. New York: John Wiley, 7.4.

Kish, L. (1988). Multipurpose Sample Design. Survey Methodology, 14, 19–32.

Kish, L. (1989). Sampling Methods for Agricultural Surveys. Rome: FAO, 12.5–12.6.

Kish, L. (1990). Weighting: Why, When, and How. Proceedings of the Section on Survey Research Methods, American Statistical Association, 121–130.

Kott, P.S. (1991a). A Model-Based Look at Linear Regression in the Survey Data. The American Statistician, 45, 107–112.

Kott, P.S. (1991b). Estimating a System of Linear Equations with Survey Data. Proceedings of the Section on Survey Research Methods, American Statistical Association (in press).

Lee, H. (1991). Model Based Estimators That Are Robust to Outliers. Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census, 178–202.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley.

Potter, F.J. (1988). Survey of Procedures to Control Extreme Sampling Weights. Proceedings of the Section on Survey Research Methods, American Statistical Association, 453–458.

Potter, F.J. (1990). A Study of Procedures to Identify and Trim Extreme Sampling Weights. Proceedings of the Section on Survey Research Methods, American Statistical Association, 225–230.

Rao, J.N.K. (1966). Alternative Estimators in PPS Sampling for Multiple Characteristics. Sankhyā, A28, 47–60.

Rao, J.N.K. and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys. Journal of the American Statistical Association, 76, 221–230.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Sharot, T. (1986). Weighting Survey Results. Journal of Market Research Society, 28, 269–284.

Spencer, B. and Cohen, T. (1991). Shrinkage Weights for Unequal Probability Samples. Proceedings of the Section on Survey Research Methods, American Statistical Association (in press).

U.S. Bureau of the Census (1978). The Current Population Survey: Design and Methodology. Technical report, 40, ch. V.

Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample Designs and

Sampling Errors for the World Fertility Survey. Journal of the Royal Statistical Society, ser. A, 143, 431–473.

Yates, F. and Cochran, W.G. (1938). The Analysis of Groups of Experiments. Journal of Agricultural Science, 28, 556–580.

Received February 1991
Revised January 1992