Journal of Official Statistics, Vol. 11. No. 1, 1995, pp. 55–77

Methods for Design Effects

Leslie Kish[1]

1. Introduction

I aim here to provide a simple, practical manual on where, why, and how values of *deft* should be computed for the sampling errors of statistics from complex survey samples. *Deft*, or design effects, are "only" tools rather than a theory or even a method. However, they are based on concepts that are theoretical, perhaps even philosophical, based on a different paradigm from the random variables (IID) of prevailing mathematical statistics, as we shall discuss briefly.

Deft are used to express the effects of sample design beyond the elemental variability (S^2/n), removing both the units of measurement and sample size as "nuisance parameters." With the removal of s, the units, and the sample size n, the design effects on the sampling errors are made generalizable ("transferable") to other statistics and to other variables, within the same survey, and even to other surveys.

I shall restrict this exposition to basic, essential approximations, which are sufficient in most cases. I must also give simple advice that may hold only in most common situations, say 0.95 or 0.99 of practical situations. I shall note with (Ex) possible exceptions relegated to the Appendix; for example (E4) denotes a remark proposed from Section 4 to the Appendix. Thus forward momentum on basic concepts can be retained.

Users are computing design effects more often and in greater volume now, since the arrival of several computing packages. This overview should help these users to compute and to use them correctly – if not always, more often than now.

The exposition is divided into the following brief sections.

1. Introduction
2. Definitions of Design Effects
3. When *Deft* are Unnecessary
4. Necessary *Deft*
5. Other Needed Sampling Errors
6. *Deft* for Subclasses and Differences
7. *Deft* for Complex Statistics
8. Weighting and Generalization
9. Computing *Deft*.

[1]Institute for Social Research, The University of Michigan, Ann Arbor, MI 48106, U.S.A.
Acknowledgement: I gratefully acknowledge careful and creative suggestions from James M. Lepkowski and Thomas Piazza, and this would be a better paper with more time to accept all of them.

2. Definitions of Design Effects

Definitions should be public servants. Servants rather than our masters: thus rather than arguing about their titles and what they "really" are, we can designate what they should do for us. Public rather than private: to avoid confusion we users must agree on how we use them.

Design effects have been first defined and commonly used for sample means ($\bar{y}$), and based on *properly* computed actual variances, $var(\bar{y})$, as

$$deff = \frac{var(\bar{y})}{(1-f)(s^2/n)}. \tag{2.1}$$

But, nowadays I and many others prefer a slightly different definition

$$deft = \sqrt{\frac{var(\bar{y})}{s^2/n}}. \tag{2.2}$$

Here s^2 denotes computed element variances, based on an achieved sample size n, selected with *EPSEM*, i.e., equal (fixed) sampling rate (probability) of f. Unequal probabilities and weighted samples are postponed to Section 8. Furthermore, applications to other, more complex (analytical) statistics are postponed to Sections 6 and 7.

We must note some strategic choices made here. First, the technique must be made simple (after the complex computations of variances) in order to allow easy computations of $deft(\bar{y})$ values for *all*, or most, or many survey means. This is necessary, because of the large variations of $deft(\bar{y})$ we found for different variables *within* the same surveys (Section 4). Second, we propose methods for inferring *deft* for more complex statistics from the $deft(\bar{y})$, based on regularities found empirically (Sections 6 and 7). Third, computations of $var(\bar{y})$ include the effects of clustering, stratification and weighting, and separation of the effects would be difficult (Section 8). The methods for computing *proper* estimates $var(\bar{y})$ of the true variances of ($\bar{y}$), and of other statistics are beyond the scope of this paper. This is a central topic of most textbooks and many articles on survey sampling.

These strategies indicate changes (by me and others), based on empirical results found during the past 40 years. Earlier we computed values of $deft(\bar{y}_i)$ for a few variables i, with the stated aim of generalizing from their averages to all other variables on the same survey. Then we had assumed that the average design effect would allow us to generalize to other variables. But decades of computations with ever increasing numbers of *deft* facilitated by faster computers convinced many of us that some of the variables can have much larger values of *deft*, and generalizing to them is not safe. Second, however, we also found, that values of $deft(\bar{y})$ were most useful for generalizing to $deft(b)$ for other statistics (b): Sections 6 and 7. Third, we also found that weighting posed more difficult and more common problems than we had anticipated: Section 8.

We also found that (2.2) has four advantages over (2.1), although the numerical differences are commonly trivial.

1. *Deft* is expressed in the same units as the factors in the intervals $\bar{y} \pm tste(\bar{y})$,

which it must chiefly serve. Thus it can appear directly as a multiplier either in t or $ste(\bar{y})$.

2. It is easier to type $deft^2$ than $\sqrt{deff}$, when one of these is needed; and *deft* is needed more often.
3. The factor $(1-f)$ when computed for the numerator (not often), may be considered as part of the design effects; the bases are variances of "unrestricted" sampling; i.e., simple random with replacement (IID).
4. The factor $(1-f)$ may be difficult to compute when the selection is not *EPSEM* (equal probability selection method) with f.

These minor differences should not be taken seriously here, because *deft* should be viewed as rough measures for large effects. Similarly, the factor $(n-1)/n$ for computing s^2 is neglected here; and pq/n may be more convenient than $pq/(n-1)$ for s^2/n for proportions when $\bar{y} = p$.

However, we should distinguish the population parameters *Deft* and *Deff* from the statistics (2.2 and 2.1) based on sample results, hence subject to sampling variability; often very large variability (E.2). Similarly we distinguish $Ste(\bar{y}) = \sqrt{Var(\bar{y})}$ from $ste(\bar{y}) = \sqrt{var(\bar{y})}$. Also *Deff* should refer to the concept and theory of "design effects," not to specific cases. Most of the time here I use *deft* both for the singular and the plural, as in many *deft* values or many *deft*.

We shall also discuss $deft(b)$ for various other statistics (b)

$$deft(b) = \sqrt{[var(b)/SRS\, var(b)]}. \qquad 2.3$$

I wish to alert you to two other decisions in defining *DEFF* which have philosophical natures. First, we chose *SRS* variance S^2/n for the standardizing denominator, because of its fundamental, classic character. Fisher's "efficiency" uses the "optimal" design for the denominator, but this concept is situationally restricted, like the choice of zero on the Fahrenheit scale, whereas our choice of *SRS* resembles the zero of the Celsius scale from the fundamental freezing point of water. Second,

$$Deff - 1 = \frac{Var(b) - SRS\, Var(b)}{SRS\, Var(b)}$$

would be often a more convenient base concept (as we shall see). However, these could result in negative values; e.g., for proportionate stratified random sampling. $Deft^2$ has a minimum of zero for $Var(b) = 0$, and this resembles the absolute zero of the Kelvin scale. (E.2)

3. When Deft Are Unnecessary

It is not necessary to compute *deft* when *any one* of the following conditions holds.
A. When the *population* closely resembles a well mixed urn, its distribution is random; any portion can be regarded as I.I.D., identically and independently distributed. Examples are a well mixed urn, as in a good lottery, or a well mixed deck of cards. But real populations, whether astronomical, physical, biological, or social, are never thoroughly mixed but are "grainy"; clustered-in my experience, model, and philosophy.

This clustered status is measured by *deft*. This paradigm is not shared or is ignored by many, including "model-dependent" statisticians, mathematical statisticians, and econometricians; also opportunistic nonstatisticians in academic or market research. It is true that some populations come close enough to random for practical purposes, for example, the last four digits of the U.S. Social Security numbers; or the output of some of the better randomization programs on computers. Nevertheless, advice from a sampling expert should be sought to judge those populations.

B. The *sample design* may be *SRS* or close enough for practical purposes. Here again, consulting a sampling expert (not any statistician) may help. For example, a systematic sample from a good list of elements (individuals) may yield a sample with *deft* only slightly lower than 1, due to mild stratification. Stronger stratifications are possible, but then they should be foreseen or detected with *deft*. However, even a telephone list may result in diverse "frame problems" (Kish 1965, 2.7). Furthermore, selecting single adults from household telephones may result in nonnegligible *deft*, due to weighting (Section 8).

C. Accepting either populations or samples as "approximately random" should be easier for *small samples*, where both the demands and resources are more restricted than for large samples. We are balancing possible biases of mild non-randomness against sampling errors, which are larger for smaller samples. Sampling errors also increase in decreasing sized subclasses of larger samples; and these subclasses are often important objectives of large samples.

D. When *only descriptive statistics* are needed, and inferential (second-order) statistics will not be computed or used for the survey; that is, when standard errors and confidence intervals are ignored and only "point estimates" are made.

E. *Sampling errors* and inferential statistics are needed only for one or *a few statistics*. For these few intervals $\bar{y} \pm t \cdot ste(\bar{y})$ may be sufficient without transformations through *deft* values. This means that there is no need for averaging with other *deft* values from other statistics, from either the same survey or other surveys.

These five cases represent situations when sampling errors are not needed (D); or can be computed with classic *I.I.D.* formulas (A, B, C); or with complex $ste(\bar{y}_i)$ survey formulas, but without conversions into *deft*, (E).

4. Necessary Deft

Most populations are far from random, and many survey samples are based on selections that are far from *SRS*, and considerably clustered. Thus their sampling errors suffer considerable design effects from clustering. Therefore computing sampling variances that reflect appropriately the restrictions of the sample design is *necessary* for appropriate statistical (confidence) intervals and inference. These are the basis of "measurability" for survey samples (Kish 1987, 7.1E). Proper variances without *deft* would suffice for computing and constructing probability intervals for one or a few statistics. However, variances and standard errors are not *sufficient* when averaging and generalizing are needed for many statistics; and *deft* are widely used for these purposes.

The principal reasons for computing *deft* from standard errors can be listed briefly:

a. Averaging sampling error for different survey variables from the same survey. Averaging the standard errors would be meaningless for variables with diverse units of measurement. The averages of *deft* values over survey variables are considerably better and may be most popular now. However, I urge caution about these averages, because recent experience has shown great variations of *deft* between variables (Table 4). Variations from 1.0 to 3.0 of *deft* are common, but note that standard errors can easily vary by factors of 100 or more. (E.4)
b. Averaging over periodic surveys for the same variables. *Deft* values are better than standard errors when the effects of changes in sample sizes need to be removed.
c. Relating the errors for different statistics on the same survey. Thus errors for complex statistics (b) such as regression coefficients, may be inferred by generalization from simpler statistics ($\bar{y}$); as in Sections 6 and 7.
d. Generalization from past surveys for designing other surveys from the same frame. Clearly these generalizations involve increasing risks with distancing of either survey variables or of sample designs for the new surveys. These warnings should be strengthened when other survey organizations try to "borrow" *deft* values.
e. Checking for gross mistakes in variance computations is greatly facilitated by *deft*. Usually a quick look at the printouts of *deft* values reveals to experienced eyes anything unusual that needs further investigation.

We come now to our central question. For what variables and for what statistics must values of *deft* be computed?

Compute values of deft($\bar{y}_t$) for overall means of all survey variables (i). This advice differs from past practice when values of *deft* were computed for only a few variables, chosen to be the most "important" or the most "representative." This change from our past practice and past advice is based on three recent realistic considerations: (1) Computations of sampling errors have become *much* easier with available machines and programs. Where this is not true, some may retreat partially toward the older restrictions. (2) Much empirical work has shown us that there are large variations in values of *Deft* for diverse survey variables (Table 4). These variations are of course much greater for values of $deft^2$, hence of "effective sample size," ($n/deft$) also of probability (p_α) values, also of standard errors. For example, a $deft = 3$ denotes $deft^2 = 9$, and instead of probability statements of $t = 2$ and $\alpha = .05$, you get $t = 2/3$ and $\alpha = 0.50$! (3) Empirical results have also revealed reasonably dependent relations from *deft* values for means to *deft* values of more complex statistics (Sections 6 and 7).

The $deft(\bar{y}_t)$ refers to overall means for the entire sample. Less often the simple expansion totals $\hat{Y}$ may be of greater interest, as in census counts, but totals are often computed from means $\bar{y}$.

Proportions p (and percentages $100p$) are the most common forms of survey means. Whether for proportions or otherwise, the means from survey samples are commonly ratio means $\bar{y} = y/x$, the ratio of random variables, hence they are often denoted as $r = y/x$. The denominator x is the sample size. When proportions come from dichotomies, like gender, with $p_1 = 1 - p_0$, then $deft(p_1) = deft(p_0)$ and we compute only one *deft*. Often, however, proportions come from polytomies with $k > 2$ categories, such as religion, occupation, etc., which are unordered. But ordered categories like number

Table 4a. Eight fertility surveys, six classes of variables, high and low values of deft and mean roh values for each class

		Korea 1971	Korea 1973	Taiwan 1973	Peru 1969	USA 1970	Malaysia 1970 Rural	Urban	Metr.	All
Socio-economic	H	6.05	2.67	1.67	2.67	2.02	5.07	1.57	1.03	
	L	1.89	1.76	1.50	1.72	2.02	2.13	1.23	0.93	
	roh	.128	.081	.016	.126	–	–	–	–	.045
Demographic	H	1.78	1.73	1.86	1.35	–	3.21	1.48	1.23	
	L	1.13	1.06	1.86	1.23	–	1.22	0.78	0.71	
	roh	.014	.025	.025	.024	.105	–	–	–	.010
Fertility experience	H	2.01	1.38	1.83	1.67	1.66	1.60	0.88	0.95	
	L	1.08	1.02	1.21	1.00	1.03	1.35	0.65	0.70	
	roh	.016	.009	.014	.034	.019	–	–	–	.025
Contraceptive practice	H	4.49	1.90	2.88	2.02	1.72	2.17	0.90	0.94	
	L	1.31	0.80	1.19	1.08	1.25	1.67	0.71	0.66	
	roh	.047	.021	.030	.054	.029	–	–	–	.022
Birth preference	H	1.96	1.90	5.41	–	1.56	2.06	0.92	1.37	
	L	1.46	1.03	1.56	–	1.09	1.03	0.69	0.78	
	roh	.023	.024	.072	–	.019	–	–	–	.028
Attitudes	H	2.48	1.73	5.28	2.12	2.04	1.83	1.08	1.48	
	L	1.22	1.16	1.19	2.12	1.34	1.66	1.02	1.11	
	roh	.028	.026	.145	.094	.051	–	–	–	.017
Mean deft($\bar{y}$)		2.16	1.47	2.35	1.65	1.47	1.92	0.99	1.01	
Mean roh		.050	.033	.059	.062	.038	–	–	–	
roh_s/roh_t		1.19	1.36	1.33	1.15	1.37	–	–	–	1.15
Sample *n*		6284	1919	5588	3327	–	–	–	–	–
PSUs *a*		62	42	56	88	126				

[Kish, Groves, Krotki 1976]

Table 4b. Twelve fertility surveys, five classes of variables, high and low values of deft and median roh for each class

		Nepal	Mexico	Thailand	Indonesia	Colombia	Peru	Bangladesh	Fiji	Sri Lanka	Guyana	Jamaica	Costa Rica
Nuptiality	H	2.49	1.68	1.59	1.79	1.69	1.25	1.48	1.65	1.27	1.31	1.34	1.48
	L	1.14	1.12	0.98	1.14	1.07	1.02	1.04	0.97	1.06	1.09	1.06	0.99
All .02	roh	.02	.03	.01	.02	.01	.01	.01	.02	.02	.02	.04	.01
Fertility	H	3.02	2.08	1.98	1.56	1.63	1.50	1.20	1.30	1.29	1.07	1.21	1.12
	L	1.26	1.06	0.98	1.28	0.86	0.87	1.00	0.89	1.07	0.94	0.99	1.00
All .025	roh	.02	.06	.01	.02	.00	.04	.01	.00	.03	.05	.03	.02
Fertility	H	3.76	1.96	1.63	1.86	1.46	1.42	1.28	1.63	1.30	1.31	1.18	1.12
preference	L	2.13	1.06	1.15	1.31	1.06	0.97	1.10	0.94	1.04	0.93	1.00	0.91
All .03	roh	.05	.03	.02	.05	.01	.05	.02	.02	.03	.03	.07	.00
Contraceptive	H	4.19	3.20	2.77	2.60	2.84	2.10	1.80	1.66	1.47	1.61	1.35	1.17
knowledge	L	2.44	2.50	1.94	2.24	2.30	1.74	1.42	1.66	1.31	1.37	1.18	0.85
All .08	roh	.05	.17	.08	.14	.05	.14	.06	.04	.07	.08	.09	.01
Contraceptive	H	2.23	2.19	2.35	1.95	2.35	1.84	1.56	1.44	1.42	1.14	1.14	1.10
use	L	2.23	1.50	1.78	1.30	1.13	1.06	0.99	1.02	1.19	1.02	1.00	1.03
All .05	roh	.03	.11	.06	.06	.04	.09	.03	.02	.07	.03	.04	.02
Average .04	roh	.03	.07	.04	.05	.02	.05	.02	.02	.05	.03	.05	.01
Sample *n*		5940	6255	3820	9136	3302	5640	6513	4928	6810	3616	2765	3935
PSUs *a*		40	1.82	70	376	405	410	240	100	606	196	410	288

[Verma, Scott, O'Muircheartaigh 1980]

of children born, years of education, or income classes are often presented also as proportions, although these may also be presented as means.

Compute values of deft(p_t) for all k categories of categorical variables. This advice goes beyond the common practice of computing and presenting sampling errors and *deft*(p) values only for one category, chosen as most "important" or "representative." But empirical data have convinced me recently that there occur sometimes (not always) large variations of *deft*(p) values for different categories of the same variable. Where computers and programs are readily available, they permit printing vast amounts of data, including *deft* for all variables, and all categories (E.6B).

The printout of the values of *deft* and sampling errors needs some interested and knowledgeable expert to examine the entire output. The values of *deft* should range mostly from 1 to about 3 or 5 in most situations, hence yield clues about mistakes, alarms, and outliers more readily than variances or standard errors.

However, the display and presentation of data printed in research reports, articles, and books should be much more restricted, and pointed to a less specialized audience. There may not be space for all categories, perhaps not even all variables, in many large survey reports, because including each *ste*($\bar{y}$) along with each ($\bar{y}$) would unduly complicate them for most readers. The *deft* values may be presented in technical appendixes. One example presents all *deft* in decreasing order, with a code for variable classes, which the readers' eyes can examine (see Table 4a) (Kish, Groves, and Krotki 1976). In another publication, the classes have been formed by experts who present average *deft* to the reader (Table 4b) (Verma, O'Muircheartaigh, and Scott 1980). Even further averaging has been done for survey means represented mostly by proportions: tables of sampling errors are shown for a few proportions (0.5, 0.2, 0.1) with the sampling errors including average design effects, 2(ave *deft*)$\sqrt{(pq/n)}$. In addition to total n, some major limits of n for subclasses may be added. Problems arise due to differences of *deft* values between variables (Kish 1965, Section 14.1). A new paper makes abundantly clear, with 3 variables from 56 countries, the great differences of *deft* values between variables, and the similarities across countries (Verma and Le 1995; Le and Verma 1995).

In addition to *deft* values for the entire sample, separate *deft* values may also be computed for major regions. But this topic is better discussed with other subclasses, for which *deft* may also be computed but not necessarily published (Section 6). *Deft* for differences of means and of subclass means will also be discussed there. *Deft* for more complex, analytical statistics are discussed in Section 7.

5. Other Needed Sampling Errors

In addition to *deft*, some other functions of variances, also some other statistics related to them may be computed and presented at the same time. Most of these are needed for and computed as factors for *deft*; and others are easily available. Furthermore, these auxiliary statistics can also serve for interpreting *deft*, and sampling variability in general.

It is common to print out the mean ($\bar{y}$, or $r = y/x$, or p); and sometimes the sample

total y or the estimated population total $\hat{Y}$. Also worthwhile is printing the values of $ste(\bar{y})$ and $SRS\,ste(\bar{y})$, whose ratio becomes $deft(\bar{y})$, which is also printed.

But it is less clear that printing $\text{var}(\bar{y}) = ste^2(\bar{y})$ is needed; nor $2ste(\bar{y})$ and the two values $\bar{y} \pm 2ste(\bar{y})$ which have also been printed sometimes. This may be useful in some reports (perhaps medical or pharmaceutical) that publish only one or a few dozen statistics. Perhaps making graphs of the intervals $\bar{y} \pm 2ste(\bar{y})$ may be useful, for graphical interpretation.

It is usual to print n, usually the simple count of sample elements for *EPSEM* selections. This would not be necessary if the values of n for all variables were similar, because they come from the same *EPSEM* sample, with only small differences due to differential nonresponses. Then n and a, the number of PSUs, can be stated in the introduction, which may come from $a/2$ strata for paired selections. However the values of n can vary greatly if some variables have smaller bases, because other cases are not relevant; e.g., only homeowners or only registered (or intending) voters; only males, or females, etc., when subclasses are tabulated. Therefore, it is good to have the program print out n. I wish to point to some problems here, for which I cannot currently offer satisfactory solutions. Would it be better to print $n' = n/deft^2$, the "effective sample size?" This becomes more difficult for weighted samples. Sometimes $roh = (deft^2 - 1)/(n/a - 1)$, a synthetic but useful "ratio of homogeneity" is also printed. This practice is justified in Section 6 on subclasses. But its interpretation is more difficult for weighted samples (Section 8).

The coefficient of variation of the statistic $(\bar{y})$ is printed because it is available as $cv(\bar{y}) = ste(\bar{y})/\bar{y}$, or perhaps a percentage error $100cv(\bar{y})$. We must be careful with denominators near zero, because that would make the cv unstable. For example, for differences the $cv(\bar{y}_a - \bar{y}_b)$ varies around zero.

Our own OSIRIS programs have also included computing and printing $cv(x) = ste(x)/x$, the coefficient of variation of the denominator x of the ratio estimate y/x. This precaution is necessary to guard against estimates that may be unstable, when $cv(x) > 0.2$. (Hansen, Hurwitz, and Madow Vol. II, Section 4.12). But printing the values of $cv(x)$ proved insufficient safety once, when nobody looked at them. Thereafter we had the program compare $cv(x)$ against the floor level 0.1, and whenever the program found $cv(x) > 0.1$ it printed out in red letters STOP, LOOK, AND DO SOMETHING. We did not stop the printout, but the warning brought human help to the problem (E.5).

6. Subclasses and Differences

Here I join the treatment for differences (comparisons) of means to those for subclasses for practical, empirical reasons, not because of theoretical considerations. The most common reasons for the frequent use of subclass statistics come from their comparisons. These take most commonly the form of differences $(\bar{y}_c - \bar{y}_b)$ between subclasses c and b, though ratios $\bar{y}_c/\bar{y}_b$ are also used, and sometimes other forms, such as $[p_b/(1-p_b)]/[p_c/(1-p_c)]$.

Before dealing with subclasses, we may look at differences $(\bar{y}_1 - \bar{y}_2)$ between two entire samples, and we may distinguish two common types from among many possible types. First, we may compare two independent samples, such as two

regions, or two countries, then: $Deft^2(\bar{y}_1 - \bar{y}_2) = [Var(\bar{y}_1) + Var(\bar{y}_2)]/[SRS\ Var(\bar{y}_1) + SRS\ Var(\bar{y}_2)]$. If *Deft* and n are similar for both means, the $Deft^2$ of the difference will also be similar to the two. If the two differ, then the $Deft^2$ of the comparison will be the average of the two $Deft^2$, each weighted by $1/n_i$. For comparing regions from the same survey, it may be common and safest to use for the regions $(n_t/n_c)Deft^2(\bar{y}_t)$, where n_t and n_c are sample sizes for the total and the region.

However, for differences of two time periods (1 and 2) of the same survey we use

$$Deft^2(\bar{y}_1 - \bar{y}_2) = [Var((\bar{y}_1) + Var(\bar{y}_2) - 2Cov(\bar{y}_1\bar{y}_2)]/[SRS\ Var(\bar{y}_1) + SRS\ Var(\bar{y}_2)]$$

because there are appreciable covariances that reduce significantly the value of $Deft^2$ for the difference (Kish 1965, Table 14.1. IV). The covariances and reductions of $Deft^2$ have been found often from using the same clusters (primary, secondary, and lower), even if the elements (and final segments) differ. The covariances can also be computed and presented as correlations R_{12}.

One type of subclass, called "proper" or "domain" subclass contains independent samples and resembles the first type above. Two examples are regions, also urban and rural subclasses and their comparisons, which are often presented. The computations of variances, hence *deft*, may differ between subclasses in these two examples. Methods of selection and clustering for urban samples may differ greatly from those for the rural sample, and the *deft* for the two may be quite distinct. Also there are usually enough primary selections, so that separate estimation of urban and rural variances and *deft* can be justified. However, for regions the situation may be quite different: when regions are small (and numerous) the number of primary selections (PSUs, "ultimate clusters") are few, the degrees of freedom fewer; and the variances and *deft* highly unstable. It is then preferable to use the overall $deft^2(\bar{y}_t)$ as an average and infer $(n_t/n_c)deft^2(\bar{y}_t)$ as the regional $deft^2$ with subsample size n_c.

The situation is quite different for the second type of subclasses, called "crossclasses," which are much more common: age and other demographic classes; education and other social classes; income, occupation and other economic classes; behavioral, attitudinal, and psychological classes, including those created by the survey questions, etc. These could not be and were not, part of the clustering and stratification effect. They "cut across" the sample design more or less evenly, or randomly, and tend to be found in all or many of the primary selections. Therefore, crossclasses are based on (almost) the same number of primary selections and have as much (or little) stability as the entire sample. Thus the sizes of sample clusters n_c/a decrease on the average with sample size n_c, hence the "design effect" (also) decreases proportionately (almost).

This follows from $deft^2 = 1 + roh(n_c/a - 1)$, to the extent that *roh* remains the same for the crossclasses, and that is an empirical question. It does not follow necessarily or mathematically. However, it has been shown empirically for many and very diverse situations and survey designs and for many survey variables, that generally *Deff* decreases toward 1 with the decrease in the cluster sizes n_c/a. The decrease is not entirely smooth nor complete, due partly to increasing relative variances of cluster sizes. These irregularities are greater for subclasses that are not true "crossclasses"; instead their cluster sizes have greater variations than random. Socio-

economic subclasses were indeed found to have greater variation and greater *rohs* than demographic subclasses (Kalton and Blunden 1973; Kish, Groves, and Krotki 1976). Hence we recommend using 1.2 *roh* or 1.3 *roh*; see below and Table 4a.

With the experience of vast amounts of empirical data one can use a reasonably good model for inferring subclass variances $var(\bar{y}_c) = ste^2(\bar{y}_c)$ from $var(\bar{y}_t)$ for the entire sample for the same variable.

a. The variance is increased by (n_t/n_c) inversely proportional to the sample sizes. But this *SRS* adjustment needs modification when $deft(\bar{y}_t)$ is not close to 1, because $deft(\bar{y}_c)$ should then be increased, as below.
b. The size of the sample cluster is changed from n_t/a to n_c/a and the value of roh_t is increased by $k_c > 1$, so that

$$deft^2(\bar{y}_c) = 1 + k_c roh_t(n_c/a - 1). \tag{6.1}$$

Strict proportionality would imply $k_c = 1$ but a better value of k_c seems to point to $k_c = 1.2$ for some subclasses, but to $k_c = 1.3$ for socio-economic subclasses, which are less evenly distributed because they are clustered. Another way to compute (6.1) would be

$$deft^2(\bar{y}_c) = 1 + k_c \frac{p_c n/a - 1}{n/a - 1}[deft^2(\bar{y}_t) - 1] \tag{6.2}$$

where $p_c = n_c/n$, the proportion size of subclass c. Somewhat simpler is still another approach (E6B)

$$deft^2(\bar{y}_c) = 1 + p_c[deft(\bar{y}_t) - 1]. \tag{6.3}$$

Differences between crossclass means $(\bar{y}_c - \bar{y}_b)$ are often principal objectives of sample surveys and the problems of design effects are somewhat different than for the crossclass means themselves. The following generalization has been found in many and diverse computations

$$SRS\, var(\bar{y}_c) + SRS\, var(\bar{y}_b) < var(\bar{y}_c - \bar{y}_b) < var(\bar{y}_c) + var(\bar{y}_b).$$

The left term is essentially $s^2/n_c + s^2 n_b$. The model behind this empirical generalization is similar to "additivity" in ANOVA: the primary clusters (within strata) that are high/low on variable y for crossclass c are also high/low for crossclass b. The wealth of empirical evidence is convincing, and the covariances tend to be positive and large so that most of the variances tend to fall near the lower *SRS* limits of $Deft^2$, near to 1. For most data even the upper limits are low because for small crossclasses they are mostly not much above 1. Thus the $Deft^2$ are squeezed to slightly above 1. Variances below the lower *SRS* limits also occur frequently, denoting $deft^2$ below 1, but both theory and experience teach us to attribute these to random variations and curtail $deft^2$ at 1. Instability of $var(\bar{y}_c - \bar{y}_b)$ is high, because it is the sum of three components, often each unstable.

In spite of my strong confidence in variances and *deft* for crossclasses from models based on the $deft^2(\bar{y}_t)$ computed for the entire sample, I strongly urge computation from actual data. Thus knowledge and confidence can be built. If meaningful contradictions are found, the results and their causes should discover better, even if more

complex, models. Good programs exist which facilitate computations of variances, *deft*, and other sampling errors. Computing them for all subclasses may not be feasible, but perhaps we can find those that are most "important," and for the survey variables with the highest *deft* values. These would provide the strongest tests for our models.

7. Deft for Complex Statistics

I must begin this section with two personal statements which I need as necessary cautions to the reader, and necessary defenses for myself. Other practicing samplers support these views, although it is difficult to find clear, unequivocal, written support. First, the existence of *Deft* and the need for probability selections for complex statistics are closely and causally linked. I cannot imagine a world where probability selection is irrelevant but which would also yield the empirical evidence of *deft*, as we have argued for four decades (Kish 1987, 1.4–1.8) (E.7).

Second, the needs are widespread, as witnessed by computing programs for regressions, etc., with automatic standard errors based on I.I.D., whereas probably most data going into them come from clustered samples, especially in the social sciences. The needs are greater than the reasonable conjectures we can offer; these outpace sufficient empirical results and are way ahead of solid mathematical theory. (Needs > conjecture > data base $\gg$ mathematics). Let us view current conjectures with those *caveats*.

What are complex, analytical statistics? Let us go beyond subclasses and differences, already seen in Section 6. Instead of attempting a definition, let us view major examples and how we treat them. We deal here with conjectures for upper and lower limits for values of $deft(b)$ for complex statistics based on $deft(\bar{y}_t)$ available for the same variables from the same survey. There are also computing programs for some complex statistics and the researchers should be encouraged to use them (Section 9). But the researchers may not be using these because they: (a) either have no access to the program; (b) or need features (weighting?) that the program lacks; (c) or lack time, ability, assistance to use them; (d) or have and need statistics for which no program of variances and *deft* exists; (e) or have a sample design that baffles the programs.

Now on to conjectures from $deft(\bar{y}_t)$ to $deft(\bar{y}_b)$ for several statistics.

a. *Ratios of ratio means and index numbers:* $r_1/r_0 = (y_1/x_1)/(y_0/x_0)$ may be used by researchers, also "odds ratios," sometimes in addition and sometimes instead of differences $(r_1 - r_0)$ of the ratio means of two surveys; these may be periodic surveys and r_0 the "base year." This may be somewhat difficult (though not impossible) to feed into programs of sampling errors. However, the variances for (r_1/r_0) and $(r_1 - r_0)$ are similar, except that the former have "relvariance" terms like $var(r)/r^2$ instead of variances. Therefore, we conjectured and found that the *deft* for (r_1/r_0) is similar to the *deft* for $(r_1 - r_0)$, which are easier to compute. Similarities may also be measured for linear combinations of these double ratios, such as $(r_2/r_0 - r_1/r_0)$ and *indexes* $\Sigma(r_{i1}/r_{i0})$ (Kish 1965, 12.11; Kish 1968).

b. *Medians and other quantiles* are often used by economists, sociologists and other researchers (in addition or in preference to means) for skewed distributions like income, wealth, time spent (in hospitals, prisons, on welfare, in queues), money spent, etc. Computing variances and *deft* may be difficult (even for the *SRS* variances), though possible (Woodruff 1952; Kish 1965, 12.9). However, it has been conjectured and found that *deft* for medians should be similar to *deft* for proportions near 0.5, and these are easy to compute (similarly for other quantiles). Also similarly, the difference of medians (e.g., between two subclasses, or two years) should have *deft* similar to those for differences between proportions from the two samples, also easy to compute.

c. *Differences* $(p_i - p_j)$ *of the proportions of pairs of categories* (i,j) from the same variable, with $k > 2$ categories have been investigated recently. For example, difference of preferences between two candidates, automobiles, religions, contraceptives, etc. It was found regularly on diverse surveys from several countries that to a good approximation in each situation $deft(p_i - p_j) = \frac{1}{2}[deft(p_i) + deft(p_j)]$. This holds even when the $deft(p_i)$ and $deft(p_j)$ are not close. Furthermore the $(p_i - p_j)$ may represent the net change (+ − minus − +) from two waves of a panel (Kish, Frankel, Verma, and Kaciroti, 1995).

d. *Coefficients for linear regressions* are the best known and most important of methods for multivariate statistical analysis. *Deft* have been computed and theory developed recently by a few investigators, since the development of computing programs and of resampling methods, as we shall see. For forty years, however, I had to argue with econometricians, mathematical statisticians and others, for the very *existence* and validity of *Deft* for regression. Their arguments were "model-based," or based on Bayesian and likelihood arguments, but happily some changed their minds since then. I believe that in a regression $\Sigma \pm b_i x_i$ $(i = 0, 1, 2 \ldots)$ the choice of the variables x_i, their exponents (1, or 0.5 or 2, or −1), their signs (+ or −) all come from the model of the researcher (economist, etc.) with little help from the statistician, although some statistical tests may help with the choices. However, the values of the b_i must come from empirical data, hence from specified samples from specified populations. The values of the b_i are conditional on the populations from which they are sampled, they are subject to sampling errors, and subject to *deft*, which must be measured (Kish 1987, 1.4–1.8).

Here follow several suggestions for estimating *deft* for coefficients in linear regressions:

1. The usual programs for computing linear regressions display estimates for the diverse coefficients (regression, partial, and simple bivariate) that are acceptable for complex samples. The standard errors, however, based on IID assumptions, serve as denominators for the *deft* values [E.2].
2. Computing programs exist for linear regressions from complex samples, with either the Taylor (delta, linear) approximations or one of three resampling methods BRR, JRR, or Bootstrap (but for bootstrap no programs seem yet to be useful for complex surveys). The relative advantages are discussed in Section 9.
3. Two cautions should be sounded here. First, weighting may be difficult (or

impossible?) with some programs. Some model-dependent econometricians denied the need for weighting, but I do not share those views (E.7). Second, long multivariate equations may overtax some computing programs; perhaps an abbreviated program of the most important predictors will yield enough *deft* for reasonable conjectures.

4. Conjectures from $deft(\bar{y})$ to $deft(b)$ of diverse coefficients have been made for three decades (Kish and Frankel 1974).

Table 7. Values of $\sqrt{Deff}$ for five types of estimators from three complex samples. Set A from Table 2, Set B from Table 3 of Kish and Frankel (1970). Set C from Table E-1 of Frankel (1971)

	Sample set		
	A	B	C
Ratio means	1.106	1.800	1.438
Simple correlations	1.096	1.262	1.355
Regression coefficients	1.015	1.295	1.106
Partial correlation coefficients	1.041	1.400	1.360
Multiple correlation coefficients	NA	1.465	1.894

i. $Deff(b) > 1$. In general, design effects for complex statistics are greater than 1. Hence standard errors based on simple random assumptions tend to underestimate the standard errors of complex statistics.
ii. $Deff(b) < Deff(\bar{y})$. The design effects for complex statistics tend to be less than those for means of the same variables. The latter, more easily computable than the former, tend to be "safe" overestimates. (We noted earlier the "pathology" of multiple *R*.)
iii. $Deff(b)$ is related to $Deff(\bar{y})$. For variates with high $Deff(\bar{y})$, values of $Deff(b)$ tend also to be high. See Kish and Frankel (1970, Section 7) for a set of striking results.
iv. $Deff(b)$ tends to resemble the *Deff* for differences of means. The latter is a simple measure of relations for which values of *deff* are easily computed, and for which (i)–(iii) also hold.
v. $Deff(b)$ tends to have observable regularities for different statistics. This is a hope based on theoretical considerations; confirming results would help us make useful conjectures.
From Kish and Frankel 1974.

A simple model of the above would be

$$Deff(b_g) = 1 + f_g\{Deff(\bar{y}) - 1\}$$

with $Deff(\bar{y}) > 1, 0 < f_g < 1$ and f_g specific to the variables and statistic denoted by g.

Notice that the lowest *deft* appear for regression coefficients in all three studies, though we do not know why. Nevertheless, even these *deft* of 1.106 and 1.295 are not negligible for statistical inference. We should like to see further research that would link $deft(b_i)$ to $deft(\bar{y}_i)$ for specific variables, to achieve tighter inference.

e. *For dummy variables and for categorical data regressions*, the above results and conjectures should also be helpful. With LISREL and similar sophisticated programs I have no experience and no guidance to offer.

f. *For Chi square tests* – such as *kxm* tests– from survey data I also have little experience or interest, but nevertheless useful conjectures, based on the essential similarity of 2×2 tests to differences of two proportions. For proportionate stratified element sampling *DEFF* goes to 1 (Kish and Frankel 1974). For clustered samples the *Deft* are much reduced and can be computed. For $k > 2$ and $m > 2$ we can argue by analogy from the *deft* values of the pairs of differences. (Nathan, Rao, and Scott, 1987).

8. Weighting and Generalization

All statistics involve generalizations; and the overall means $\bar{y}_t$, variances $var(\bar{y}_t)$, and also $deft^2(\bar{y}_t)$ for the entire sample ignore and average variations between its separate domains. For example, in a country's sample, the urban domain of metropolitan and large cities may have very different values of $deft^2$ than the rural, because both the cluster sizes b and the homogeneity roh can be different in $[1 + roh(\bar{b} - 1)]$. The design can be further complicated if a different (higher or lower) sampling rate was used in the urban domain than in the rural. The capital's area in a developing country may contain a minor fraction of the population for which a larger sample is desired; but in developed countries the small rural areas may need increased sampling. Perhaps separate values of the $deft^2$ should be computed and used.

The needs for generalizations from computed values of *deft* lead to conflicts that are especially difficult when complicated by weighting. *Within* the same survey *Deff* carries the effects of clustering and stratification, often in several stages. After standardizing for S^2/n the $Deft^2 = Var(\bar{y})/(S^2/n)$, we must recognize for each variable distinct *Deft* values, due to varying factors of homogeneity *roh*. However, we must distinguish four major *sources* of unequal selection probabilities (P_i) and hence of weighting ($W_i \propto 1/P_i$), because they need distinct treatments for computing *deft* (Kish 1992).

a. *Nonresponses* may be compensated with differential weighting in classes. These differences should be either small or rare, or both, if nonresponses are under reasonable control.
b. *Frame problems* result in unequal selection probabilities, because these could not be measured or controlled before selection. These also are (or should be) relatively small or rare. An example is weighting with number N_i of adults when one is selected at random from each sample household; we have seen values of *deft* from 1.05 to 1.20.
c. *Allocation to separate domains* of different sampling fractions occurs often, either to increase sample sizes in some entire domains, or to reduce costs in others; and sometimes to decrease variances in the total sample. These distinct domains may be regions, or urban/rural strata, large/small units, etc.
d. *Disproportionate sampling fractions* may be introduced into crossclasses, which are not domains, deliberately for "optimal allocation," or in order to increase the sample sizes of some subclass. For example, households found to have an ethnic or age group may be oversampled.
e. *Post-stratification and ratio estimates* for population control and adjustments faces us with difficulties. Often the range of weights is not great, mostly within $w(max)/w(min) < 2$; for these the unweighted *deft* values may suffice. I do not have a simple, general solution.

Classes a and b may often be treated simply, because the effects of weighting on *deft* may be small. The effects on the descriptive (first-order) statistics, like $(\bar{y})$, may be larger than negligible (or weights would not need to be used). Yet their effect on the variances may be small. And beyond that their effects on *deft* should be similar on both sides of the ratio $Var(\bar{y})/(S^2/n)$, hence even smaller.

Class c needs a different treatment and separate values of $deft^2$ can be calculated for two or a few domains. If the $deft^2$ are approximately similar they can be combined for simpler joint presentation.

Class d faces us with the most difficult problem and confronts us with the conflict for which I know of no satisfactory simple solution. The treatment we offer here applies also to classes a, b, and c, when the simple treatments seem unsatisfactory, and must be treated, like d.

d1. For *internal* use and inference, the standard definition, $var(\bar{y})/(s^2/n)$ yields $deft^2(\bar{y})$ that combines and confounds the effects of the specific weighting used with those of clustering and stratification. The numerator $var(\bar{y})$ contains the weighted variance from the input data. In the denominator the *weighted* s^2 merely estimates S^2 in the population; thus s^2/n estimates the variance of an *SRS sample of size n*. Since all statistics and variables have the same weights over the entire sample these $deft(\bar{y})$ values yield the proper *Deff* corrections for standard errors of the sample for *these* weights and these weights only.

d2. *Internal generalizations within the survey*, such as conjectures from $deft(\bar{y}_t)$ to other statistics are possible, but only with caution. For example, the effects of weighting on "crossclasses," like age classes, will be similar ("inherited"), that is, they will be about the same in the subclasses as in the total sample, unlike cluster effects which decrease in subclasses. However, for subclasses correlated with the selection/allocation probabilities, the effects may be either increased or decreased. The effects of allocations may also differ for diverse statistics. For example, for some of the means the unequal selection rates may produce reductions of the variance; these may overcome the clustering effects and thus result in $deft < 1$. However for some statistics, especially proportions, the same allocation may result in losses (increases of variances). Thus, an "optimal allocation" for mean incomes resulted in losses for median incomes. These results came from an investigation that used a simplified analysis, which treated the first phase ratings of dwellings from a multistage sample, as if it were a stratified element sample (Kish 1961).

d3. *Deft* for *external* use are difficult to fashion, though they would be highly desirable, since generalization is the chief reason for computing *deft*. First, one may wish to design an *EPSEM* (equal f rates) or a different allocation using the same sampling frame. Second, one may wish to plan a survey for a different sampling frame. Separation of the effects of weighting from the effects of clustering is needed.

One approach is to compute values of unweighted

$$deft^2 = \frac{var_u(\bar{y})}{s_u^2/n} \tag{8.1}$$

without weights. This estimates the values of $deft^2$ in a population in which the frequency distribution has been biased by the selection probabilities of the selection factors of the sample. The values of *Deff* in this distribution should differ some from those in the actual population.

This population can be approached with a subsample selection that reduces the oversampled portions in order to produce an *EPSEM* selection. The (unweighted)

computation of values of $deft^2$ estimate *Deff* in the actual population. These can be compared both to the $deft^2$ above and to the weighted estimates. This approach requires a separate research project that will seldom be undertaken.

d4. Haphazard or Random Weights can be dealt with more easily. Weights due to frame problems (a) and to nonresponses (b) may often be considered approximately random. In these situations I assume that the variances are increased by a factor $1 + L = n\Sigma k_j^2/(\Sigma k_j)^2$, where L represents the "loss" due to the element weights k_j. This loss L is easily shown to equal the "relative variance" of the weights k_j (Kish 1992). Then in $deft^2 = var(\bar{y})/[1 + L)s^2/n]$, both terms contain losses for random weights, and we have estimates of *deft* without them.

9. Computing Deft

A. The most basic concept of "measurability," for sampling errors for variances and for *deft*, is *random replication*. In complex clustered samples, *paired* selections of paired primary selections (i.e., ultimate clusters) from strata form most commonly the bases for computations. But sometimes the $a/2$ pairs are actually "collapsed" from single selections from a strata. Or a systematic selections are treated as $a/2$ pairs; or as $a - 1$ nonindependent pairs. Larger strata with $a = \Sigma a_s$ and $a_h > 2$ are used less commonly. And "interpenetrating samples" of k independent selections are rare in practice, I believe. When they occur, the variance computations should be easy, but they may be highly unstable (variable) when k is not large.

Paired selections are not necessary (as has been stated sometimes in the past) but they are convenient and also often efficient. They satisfy approximately three conflicting needs in cluster sampling: (1) To restrict the spread of the sample to a primary selections, because these are costly; (2) to use more strata for greater efficiency and because information is available; and (3) two random replicates are needed for computing variances.

The sizes of clustered samples should be judged not only in terms of the numbers n of elements, but also in numbers of clusters at all stages. The numbers of primary clusters is particularly important, not only for reducing the standard errors and *Deff* of statistics, but also for reducing the instability of standard errors and of *deft*.

There are many articles and books dealing with aspects of unbiased estimators ("exact," and "best") of variances. Under practical conditions we generally must use only approximately unbiased estimates, and we must avoid large biases, such as *SRS* estimates of variances that ignore *Deff*. But, alas, too little is written about the precision or stability of estimates of variances, standard errors, and *deft*. The single basic fact is that the coefficient of variation of *deft* is no less than $1/\sqrt{2d}$, where d is the number of degrees of freedom; a little more in practice because the clusters are unequal, and the *SRS* denominators s^2/n also contribute a little to the instability (E.5). The famous jackknife design of ten replicates has 9 degrees of freedom, hence $cv(deft) > 1/\sqrt{18} \simeq 0.25$. But this 25% variation, or 50% for two sigmas, is of little use for evaluating *deft*. Statisticians then fall back on averaging *deft*, but we saw that *deft* vary greatly between variables. With $a = 60$ systematic selections, 30 paired selections yield $d = 30$ and $cv(deft) > 1/\sqrt{60} \simeq 0.13$ and $2cv(deft)$ of 26%. If all (60-1)

differences are used, d is about $(4/3)(a/2) = 40$, and $cv(deft) > 1/\sqrt{(80)} \cong 0.11$ and $2cv(deft)$ of 22% (Dumouchel, Govindarajulu, and Rothman 1973).

For each primary selection (ultimate cluster) for valid computation of variances, hence of *deft*, we need the sample totals $y_{h\alpha i} = \Sigma_j y_{h\alpha ij}$ for every variable i; for each primary selection α (often only two, $\alpha = a$ or b), within each stratum h; j is the element case count, which is commonly on the tape. However, the stratum and primary selection identification numbers are carelessly omitted often from survey data and for those surveys the computation of sampling errors and *deft* are not feasible. (Only sampling code numbers are needed, and deliberate omission of names of units for confidentiality can be pursued.) We may consider the existence on data tapes of data for the covariance matrix of primary selections as necessary and sufficient for computing $var(\bar{y})$ for clustered *EPSEM* samples. For weighted data the element case weight w_j are needed also.

For the denominator s^2/n, and generally for $SRSvar(b)$, the element values y_j and w_j are needed; but the identifications of the primary selections are not. This is also true for the mean$(\bar{y})$ or p or $r = y/x$ and for other descriptive statistics (Ex 9). The computation of s^2/n should cause no problems. Furthermore, for complex statistics, like regression coefficients, I believe that computing programs automatically print useable estimates of SRS standard errors.

Several good computing programs are available for computing sampling errors that also compute *deft* values. I cannot attempt to be comprehensive; that would be futile and also would soon become obsolescent. There are also many more being used privately but not publicly available. In the Appendix the few best known in the U.S.A. are checklisted [Ex 9]. (Francis 1981; Cohen, Burt, and Jones, 1986; Cohen, Xanthopoulos, and Jones 1988.)

10. Appendix

(E.2A) For the *SRS* variance s^2 or $\hat{s}^2 = s^2(n-1)/n$ in the denominator of $deft^2$, I am not concerned with small factors like $(n-1)/n$, or whether pq/n or $pq(n-1)$ should be used, or about sampling with/without replacement. But we should be concerned about using $\hat{s}^2 = \Sigma y_j^2/n - (y/n)^2$, for samples that are clustered and stratified, and $\hat{s}^2 = \Sigma W_j y_j^2/\Sigma w_j - (\Sigma W_j y_j/\Sigma w_j)^2$ when they are weighted also. This is obvious to a few, but unforeseen or surprising for most but in either case it is most important and convenient, that for all data from probability samples, we have simply

$$Exp(\hat{s}^2) = \sigma^2 - Var(\bar{y})$$

that s^2 is an almost unbiased estimate of $\sigma^2 = \Sigma Y_i^2/N - \bar{Y}^2$, the element variance in the population. It is only a slight overestimate, since $Var(\bar{y})^2$ is smaller than σ^2 by n^{-1}. For *any EPSEM* we have (Kish 1965, 2.8)

$$Exp(\bar{y}) = Exp(\Sigma y_j/n) = \Sigma Y_i/N = \bar{Y}$$

$$Exp(\Sigma y_j^2/n) = \Sigma Y_i^2/N$$

By definition

$$Var(\bar{y}) = E(\bar{y} - \bar{Y})^2 = E(\bar{y}^2) - \bar{Y}^2$$

Then

$$Exp(\hat{s}^2) = Exp(\Sigma y_j^2/n) - Exp(\bar{y}^2) = \Sigma Y_i^2/N - Exp(\bar{y}^2)$$
$$= \Sigma Y_i^2/N - \bar{Y}^2 - Var(\bar{y}) = \sigma^2 - Var(\bar{y})$$

a. Dividing by random variables (not fixed) n results in approximate or conditional expectations.
b. The same derivation holds for P_j weighted samples that are not *EPSEM* (Kish 1965, 2.8).
c. The same derivations hold for $\Sigma y_i x_i$ terms in the covariance matrix; also for higher moments. Therefore, also for estimates of b and $SRS(b)$ of analytical statistics in Section 10. From the above we deduce that $\hat{s}^2 + var(\bar{y}) = \hat{s}^2(1 + deft^2/n)$ will give adequate estimates of σ^2. Also that $s^2 = \hat{s}^2(1 + 1/n)$ will suffice for the denominator when $deft^2$ is near 1.

(E.2B) A *brief history* of the background of *Deff* may be useful, before this name was introduced (Kish 1965). Since then high speed computers and the spread of probability sampling have made *Deff*s well-known, widely computed, and used. The earliest reference to a similar concept I found in earlier textbooks (Yule and Kendall 1965) under the name "Lexis ratio" (Kendall and Buckland 1982), traced to the publications of W. Lexis in German at the end of the 19th century. (This dictionary still lacks "design effects.") The concepts of "intraclass correlation" by Fisher (1950) are related through the within/between components in the analysis of variance. From the late 40s and through the 50s there were a few of us computing "ratios of variances" (Hansen, Hurwitz, and Madow 1953, 12D); and values of true *var/srsvar* are given for five large scale samples (Kish 1957). The U.S. Census Bureau had computing programs in the 1950s, but the first published program with *deft* appeared in 1972 (Kish, Frankel, and Van Eck 1972).

(E.4A) An older alternative to *deft* for generalizing sampling errors are the coefficients of variation $CV(\bar{y}) = Ste(\bar{y})/\bar{Y}$ and $CV^2(\bar{y}) = Var(\bar{y})/\bar{Y}^2$, called relvariances. They also remove the units of measurement, and have the advantage of being easily understood, as 100 $CV(\bar{y})$ = percent variation in $(\bar{y})$. Thus CV's for means and totals of quantities (people, money, acres) can be compared and the technique has been used for proportions. The name Generalized Variance Function (GVF) has been coined for fitting curves for generalizing to diverse statistics within a survey (Hansen, Hurwitz, and Madow 1953, Vol. I, 12.B.15; Wolter 1985).

However, CV's have several drawbacks compared to *deft* for generalizing and inference. (1) CV's are functions of $\sqrt{n}$, but these are removed from *deft*. (2) CV's are also subject to *Deff*, but without explicit expression. (3) They are unstable when the denominator is small; e.g., net change or difference, small or rare values. (4) $CV(p) \gg CV(1-p)$ for small p, although $Ste(p) = Ste(1-p)$ (Kish 1965, 2.5).

(E.4B) The two tables are abstracted from two frequently quoted publications on sampling errors of fertility related variables. Each contained the broadest set of

sampling errors at its time (but a new "champion" will be presented by Verma and Le to the ISI in August 1995). Each table represents tens of thousands of cases (n) from hundreds of PSUs (a). Note the large variation between the highest and lowest *deft* in each class of variables. The highest *deft* in each table 6.05 and 4.19 mean increases of *deff* of 36.6 and 17.6 in actual/*srs* variances, and similar decreases in the effective sizes of n. Even modest average values of 2^2 have drastic effect on variances and effective sizes. Note some reasonable regularities by classes of variables, and by countries. Nevertheless, substantial differences are revealed between variables, classes, and countries. Note also the rather regular ratios from 1.2 to 1.3 for roh_t of the total sample over *roh* for subclasses (Table 4a).

(E.5) $CV(x) = \sigma_x/a$ is directly proportional to the CV of the primary unit sizes X_α and inversely proportional to the number a of those units in the sample. Therefore, it can be reduced with better control of units size X_α, e.g., with stratification, PPS selection, redefinitions; or by taking more primary units a.

Reduction of $CV(x)$ is also needed to control the technical bias of ratio estimates, and these estimates are common in surveys – though happily not the biases (Hansen, Hurwitz, and Madow 1953, Vol II; Kish 1965, 6.6B).

(E6A) This (6.3) has been proposed as not only simpler, but also as giving a better fit (Skinner, Holt, and Smith 1989, Ch 3). Skinner et al. present interesting discussions about *Deff* in chapters 2 and 3, and show that their formula (3.12) (our 6.3) is better than my (6.2) with $k = 1.0$, and their more complex (3.13) even a little better. But with $k = 1.2$ my (6.2) would perform as well, and seems preferable theoretically, because it will take $Deft(\bar{y}_c)$ to 1, when $n_c/a = 1$, I believe.

Another technique for imputing $deft(\bar{y}_c)$ from $deft(y_t)$ presents a novel and sophisticated method for modeling *deft* for subclasses (Verma, Scott, and O'Muircheartaigh 1980). All three techniques have in common some model that roh_c for subclasses will be somewhat higher than roh_t for the entire sample, but also that the two are strongly related. I believe that they all lead to useable values of $deft(\bar{y}_c)$. Also that deciding between their relative values will require some empirical investigation based on several diverse surveys, because we lack a strong theoretical model for deciding between them.

(E.6B) The need for computing $deft(p_i)$ separately, for each of the k categories ($i = 1, 2, \ldots k$) of categorical variables with $k > 2$, is illustrated in a recent article (Kish, Frankel, Verma, and Kaciroti 1995). They found in eight surveys from five countries that the values of $deft(p_i)$ as well as the $ste(p_i) = deft(p_i)\sqrt{[p_i(1-p_i)]}$ varied a great deal. Therefore, that choosing only one of the categories to represent as "typical" the diverse categories of that variable was not sufficiently accurate. This result has been found by others before, I believe, but not emphasized. The variation between the variables was even greater than between categories of the same variable. These are empirical results for which simple models would be difficult to construct.

In these eight surveys it was also found, surprisingly for most of us, that *deft* for the differences $(p_i - p_j)$ of two correlated proportions from the same variable is similar to the average of the two $deft(p_i)$ and $deft(p_j)$. That is, $deft(p_i - p_j) \simeq (1/2)[deft(p_i) + deft(p_j)]$ approximately.

	Clusters	SUPERCARP	SUDAAN	OSIRIS
Relative Cost	Low	Low	High	Moderate
Relative Simplicity	Easy?	Complex	Complex	Easy?
Data Input	ASCII	ASCII or SAS	ASCII or SAS	ASCII or OSIRIS
S.E. Methods	Taylor	Taylor	Taylor	Taylor or BRR, JRR
Analytical Stats		Regression Chi Square	Regression Log Regression Chi Square Survival	Regression

(E.7A) "A great need exists for mathematical bases of analytical statistics to deal with data originating in complex sample designs. At present, these analytical statistics are not computed or they are computed incorrectly under *SRS* assumptions. The latter results in gross mistakes chiefly because of the effects of "clustering" on the sample. Because of these mistakes the researcher often may be actually using confidence coefficients which are distorted (unknown to him or her) from $P = .99$ or $P = .95$ to $P = .50$! These problems require the urgent attention of mathematical statisticians, particularly to provide formulas – valid under complex clustered designs – for some of the most important statistics. Examples of these are: 1. The coefficients of multivariate analysis and their variances: . . ." From my talk Sept. 7, 1956 to a joint meeting of ASA and IMS.

Although right in calling attention to this important neglected problem (and in some other aspects) I was wrong and naive in several others:

1. "First, mathematical statistics has not and *will not* give us complete distribution theories that will be useful directly, because there are too many parameters in the double complexity of analytical statistics from complex surveys. Second, model builders cannot make those complexities vanish." (Kish 1984).

2. The problems of *inferential* statistics should be more clearly separated from those of *descriptive* statistics. Inferential (second order) statistics depend on pairwise probabilities P_{ij} of selection, hence need ultimate cluster and stratum identifications of "measurability"; but descriptive (first order) statistics depend only on element probabilities P_i and the usual estimates suffice (E2) (Kish and Frankel 1974).

3. Several methods exist now for computing useful estimates of sampling errors and *deft*. These depend on Taylor and repeated replication methods without mathematical distribution theories, and on computing developments (Section 9).

4. I failed to foresee that 40 years later only a score or two of mathematical statisticians will pay any attention to the theoretical problems of survey sampling. Their names are in 200 plus References (Skinner, Holt, and Smith, eds., 1989). There are many more names of statisticians designing, operating, and analyzing the growing body of probability samples around the world.

(E.7B) In Table 4 we related the average $deft(b)$ for multivariate coefficients to the average $deft(\bar{x})$ of means. But after much empirical evidence about large differences

between the $deft(x_i)$ for different means it would be better to find more specific relations of the $deft(b_i)$ to the respective $deft(\bar{x}_i)$.

(E.9) These are but four outstanding examples of available programs from the United States. There will be others in the future, and some that we have missed. In other countries there are still others. Also many more that have been and will be prepared for internal institutional use but not easily available to the outside.

Each of the programs below can deal with case weights, more or less easily. Each of them can handle means, proportions, and ratio estimates.

11. References

Cohen, S.B., Burt, V.L., and Jones, G.K. (1986). Efficiencies in Variance Estimation for Complex Survey Data. The American Statistician, 40, 157–164.

Cohen, S.B., Xanthopoulos, J.A., and Jones, G.K. (1988). An Evaluation of Statistical Software Procedures Appropriate for the Regression Analysis of Complex Survey Data. Journal of Official Statistics, 4, 17–34.

Dumouchel, W.H., Govindarajulu, Z., and Rothman, E. (1973). A Note on Estimating the Variance of a Sample Mean in Stratified Sampling. The Canadian Journal of Statistics, 1, 267–274.

Fisher, R.A. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd.

Francis, I. (1981). Statistical Software: A Comparative Review. New York: North Holland.

Frankel, M.R. (1971). Inference From Survey Samples. Ann Arbor: Institute for Social Research.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory. Vols I and II, New York: John Wiley.

Johnson, E.G. and King, B.V. (1987). Generalized Variance Functions for Complex Sample Surveys. Journal of Official Statistics, 3, 235–250.

Kalton, G. (1979). Ultimate Cluster Sampling. Journal of the Royal Statistical Society (A), 142, 210–222.

Kalton, G. and Blunden, R.M. (1973). Sampling Errors in the British General Household Survey. Bulletin of the International Statistical Institute, 45(3), 83–97.

Kendall, M.G. and Buckland, W.R. (1982). A Dictionary of Statistical Terms. London: Longman Group.

Kish, L. (1957). Confidence Intervals for Clustered Samples. American Sociological Review, 22, 154–165.

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

Kish, L. (1968). Standard Errors for Indexes from Complex Samples. Journal of the American Statistical Association, 63, 512–529.

Kish, L. (1969). Design and Estimation for Subclasses, Comparisons, and Analytical Statistics. In Johnson, N.L. and Smith, H., eds., New Developments in Survey Sampling, New York: John Wiley.

Kish, L. and Frankel, M. (1970). Balanced Repeated Replications for Standard Errors. Journal of the American Statistical Association, 65, 1071–1094.

Kish, L., Frankel, M., and Van Eck, M. (1972). SEPP: Sampling Error Programs Package. Ann Arbor: Institute for Social Research, 184 pp. (out of print).

Kish, L. and Frankel, M.R. (1974). Inference from Complex Samples. Journal of the Royal Statistical Society (B), 36, 1–37.

Kish, L., Groves, R.M., and Krotki, K. (1976). Sampling Errors for Fertility Surveys. Occasional Paper No. 17, World Fertility Survey, 61 pp.

Kish, L. (1984). Analytical Statistics from Complex Samples. Survey Methodology (Statistics Canada).

Kish, L. (1987). Statistical Design for Research. New York: John Wiley & Sons (also Spanish and Hungarian translations).

Kish, L. (1992). Weighting for Unequal P_i. Journal of Official Statistics, 8, 183–200.

Kish, L., Frankel, M.R., Verma, V., and Kaciroti, N. (1995). Design Effects for Correlated $(P_i - P_j)$. Submitted to Journal of the American Statistical Association.

Le, T. and Verma, V. (1995). Sample Designs and Sampling Errors for the DHS. Paper prepared for 50th Session of the International Statistical Institute, Beijing.

Lepkowski, J.M. (1980). Design Effects for Multivariate Categorical Interactions. Ann Arbor: University of Michigan, Ph.D. Thesis.

Molinari, G. (1993). Design Effects and Ratio of Homogeneity in Complex Sampling Designs. Statistica, 53, 633–646.

Rao, J.N.K. and Scott, A.J. (1987). On Simple Adjustments to Chi-square Tests with Sample Survey Data. Annals of Statistics, 15, 385–397.

Rao, J.N.K. and Wu, C.F.J. (1985). Inference from Stratified Samples. Journal of the American Statistical Association, 80, 620–630.

Rao, J.N.K. and Wu, C.F.J. (1987). Resampling Inference from Complex Survey Data. Journal of the American Statistical Association, 83, 231–241.

Rust, K.F. (1984). Techniques for Estimating Variances for Sampling Surveys. Ann Arbor: University of Michigan, Ph.D. Thesis.

Skinner, C.J., Holt, D., and Smith, T.M.F. (eds.) (1989). Analysis of Complex Surveys. New York: John Wiley. Ch. 3, Skinner, C.J., Domain Means, Regressions and Multivariate Analysis. Ch 12, Pfeffermann, D. and La Vange, L., Regression Models for Stratified Multi-Stage Cluster Samples.

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 (A), 143, 431–473.

Verma, V. and Le, T. (1995). Sampling Errors for the DHS Surveys. Calverton, MD: MACRO International (80 pp.).

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer Verlag.

Woodruff, R.S. (1952). Confidence Intervals for Median and Other Positional Measures. Journal of the Americal Statistical Association, 58, 454–467.

Yule, G.U. and Kendall, M.G. (1965). Introduction to the Theory of Statistics, 14th ed. London: Griffin.