DEFFS: WHY, WHEN AND HOW? A REVIEW

Leslie Kish, The University of Michigan, Ann Arbor 48106

KEY WORDS: DEFFs, DEFTs, design effects, variances, complex samples, survey sampling, sampling errors

1. What Are DEFFs?

This is a quick and rough but broad and practical overview of "design effects," aimed at the many users of the results of sample surveys. Users are computing design effects more and more often, especially since the availability of several computing packages. This overview should help them to compute and to use them correctly - more often than now.

Design effects have been defined commonly for sample means ($\bar{y}$) in two similar ways:

$$deff(\bar{y}) = var(\bar{y})/\{(1-f)s^2/n\} \text{ and}$$
$$deft^2(\bar{y}) = var(\bar{y})/\{s^2/n\}. \qquad (1.1)$$

The second definition introduces three minor improvements on the earlier version. a) The root $deft(\bar{y})$ is used most often and it is easier to type than $\sqrt{deff(\bar{y})}$. b) The factor (1-f), when computed for the numerator, may be considered as part of the "design effects" and s^2/n estimates variances of "unrestricted," simple random sampling with replacement. c) The factor (1-f) may be difficult to compute when the selection is not EPSEM (with equal probabilities).

These minor differences and refinements should not be taken seriously here, because DEFFs should be viewed as rough measures for larger effects. Similarly we refrain from discussing here the factor (n-1)/n for computing s^2; especially when either pq/n or pq/(n-1) is used for s^2/n commonly for proportions $\bar{y}=p$. On the other hand, we should distinguish the population values (parameters) Deff and $Deft^2 = Var\ (\bar{y})/(S^2/n)$ from statistics based on sample results in (1.1). We also use $Ste(\bar{y}) = \sqrt{Var(\bar{y})}$ and $ste\sqrt{(\bar{y})} = \sqrt{var(\bar{y})}$. Then DEFF should refer to the term and concept of "design effects." These sometimes also bear other names, such as variance ratios or factors.

Design effects have also been formulated by analogy and computed for many other statistics: for aggregates $\hat{Y} = N\bar{y}$, for subclass means $\bar{y}_c$ and their differences $(\bar{y}_c - \bar{y}_b)$, for (partial) regression and correlation coefficients b_{yx}, r_{yx}, and other analytical statistics. Writing b for other statistics generally,

$$deff^2(b) = var(b)/srsvar(b). \qquad (1.2)$$

The simple random variances in the denominator are derived mathematically in classical statistics under assumptions of independent selections (I.I.D.). For the variances of the numerator several good methods are now available; chiefly a) linearization, or delta or Taylor methods; b) balanced repeated replications (BRR); c) jackknife repeated replications (JRR); d) and perhaps bootstraps.

The denominator is computed from the n sample cases as if they were selected with SRS (section 6). This is a relatively straightforward method for measuring the effects of clustering primarily and of stratification secondarily in EPSEM selection designs. The effects of weighting need separate attention (Section 7).

2. When DEFFS Are Unnecessary

First, and most important, note that DEFFs are needed only for "inferential" statistics like var(b), but not at all for "descriptive" statistics, like b or $\bar{y}$. Then for a better view of limits for using DEFFs, note four situations when DEFFs are not necessary. a) When selections are actually made with simple random selection (SRS), DEFF=1 by definition. However, judging what may be accepted as "approximately" SRS requires modelling and experience, to avoid pitfalls and bad surprises.

b) Variances can be used directly, without referring to DEFFs, for probability intervals, like $\bar{y} \pm t_p ste(\bar{y})$, when variances are self-sufficient, not requiring pooling or averaging. This can happen when : 1) the survey statistics $\bar{y}$ (or b) are few, so that all the $ste(\bar{y}) = \sqrt{var(\bar{y})}$ can be computed and presented; and 2) the values $ste(\bar{y})$ possess adequate precision to be useful: with sufficient degrees of freedom, numbers of primary sampling units (PSUs), "ultimate clusters."

c) Coefficients of variation $CV(\bar{y}) = Ste(\bar{y})/\bar{Y}$ can be useful for their direct relationship to probability statements. However, their "portability" for averaging is limited because they are inversely related to $\sqrt{n}$. These ratios do remove the units of measurement, which is their chief virtue. Thus, they are good for direct use for each statistic, but not for indirect uses for other statistics. Furthermore, their utility is limited to nonnegative variables. Moreover, many (or most?) survey results are based on a few, or a few dozen PSUs, and their variances lack adequate "measurability." Also, the large number of survey statistics impedes separate computation, presentation, and comprehension of standard error for all of them.

d) For periodic samples, based on the same design, variances (or coefficients of variation) for specific statistics may be averaged directly over periods, without going through DEFFs, and such averaging may overcome the lack of sufficient precision for each period (but high correlations between periodic samples may interfere).

3. When DEFFs Are Needed

Very often the selections for surveys are not SRS, and then DEFFs are needed for pooling or averaging sampling errors, for several reasons. We discuss here why they are necessary, and later (4.5) why they may not be sufficient for external uses. a) The precision of the variance estimates, $var(\bar{y})$ and var(b), may be too low in many clustered samples; the numbers of primary sampling units (PSU's, ultimate clusters, degrees of freedom) are too few for good "measurability." Sampling theory, by concentrating on "unbiased" and asymptotic variances, neglects this "dirty" topic. For example, a sample of 64 PSU's in 32 pairs yields roughly 32 degrees of freedom and even less [Kott 1989]. Hence the coefficient of variation of sampling errors and of deffs will be over $1/\sqrt{64} = 1/8$ or 12.5 percent; often not good enough alone.

b) There are too many statistics on most surveys to permit separate computations and presentations for the sampling errors of all those statistics. Most surveys are multipurpose for many survey variables. Furthermore for each of those survey variables, the statistics are needed not only from the entire sample for the overall target population, but also from subclasses for many kinds of domains, also for subclass comparisons, and often for other analytical statistics. Surveys are multipurpose in several dimensions.

For all survey variables it is necessary to compute their distinct values of $deff^2$ from the entire sample because they can differ greatly. But these overall $deff^2$ will not be sufficient, because $deff^2$ can be greatly different and lower for subclasses and analytical statistics.

c) Beyond those two needs "internal" to the surveys, sampling errors are also used "externally," in four ways. First, defts can be readily used for periodic surveys. Second, the sampling office may "borrow" for the other surveys the Deffs or the derived Rohs (as we see in Section 5), in order to save computations. Third, Deffs and Rohs are also needed for designing future samples by the same office. And fourth, the "borrowing" of values of DEFFs and Rohs by other institutions is more common than proper and admitted.

Thus DEFFs and Rohs serve our needs for averaging either a) for greater precision, or b) for economizing on computations or presentation, or c) for borrowing for other survey samples.

4. Where DEFFs Are Not Sufficient

Having already noted in two sections the situations when DEFFs are necessary and when they are not, we now list four situations when they are not sufficient and must therefore be modified.

a) For any specified survey variable the DEFFs will be different for subclasses. For "crossclasses" the DEFFs tend to approach 1 asymptotically, and even faster for differences of subclasses (section 5). For analytical statistics, like regressions, the situations are more complicated.

b) For weighted samples, ordinary computations of $Deff^2$ would combine the effects of weighting with design effects due to clustering and stratification. These would often be confusing, and we shall note methods for disentangling the effects of weighting from the other design effects in section 7.

c) The values of $deff^2$ combine the several variance components of clustering and stratification that can arise in multistage sampling. These overall, rough $deff^2$ values yield the convenience and liberty needed for sampling errors of multipurpose surveys. The price of that liberty is eternal vigilance in the form of methodological checks of those components when occasions permit.

d) Those rough, approximate values of $deff^2$ and roh must also neglect some technical, theoretical factors - such as (n-1)/n, unequal clusters, sampling with/without replacement, etc. Such neglect may be necessary for their simplicity, portability, volume, presentation.

5. Subclasses: from DEFFs to ROHs

Portability was our chief reason for moving from variances to DEFFs. But we need even more portability, because $deff^2$ combine two distinct factors, as seen in $deff^2$ for clustered samples: $deff^2 = 1 + roh\ (\bar{b}-1)$. From this we may separate the effects of the average cluster size b from the average homogeneity of elements within primary clusters: $roh = (deff^2-1)/(b-1)$. I coined the name ROH (ratio of homogeneity) instead of the classical rho (ρ) for this rough measure of homogeneity within primary (alias "ultimate") clusters, which are usually stratified and of unequal sizes. Thus b denotes average sizes $b = n/a$ of n elements in a primary ("ultimate") cluster; and moderate variations of size have been shown to have only trivial effects.

Values of roh yield the portability needed often, and especially for three important tasks. First, we need these for "crossclasses," when the average cluster sizes $b_c = n_c/a$ are often much smaller than b. We then may use $deff_t$ from the overall mean for $deff_c$ for the subclass: $deff_c = 1 + roh_t(b_c-1)$. This may be improved perhaps with $deff_c \approx 1 + k_g\ roh_t\ (b_c-1)$, where k_g is slightly greater than 1, say 1.2, for subclasses that tend to be unevenly distributed in clusters, such as socioeconomic subclasses ( ).

Second, we may also use the roh_t values to design future samples from the same clusters, but with different average sizes $b_d = n_d/a$.

Third, we have often used the relationship for differences of subclass means $(\bar{y}_c - \bar{y}_b)$:

$$s_c^2/n_c + s_b^2/n_b < var(\bar{y}_c - \bar{y}_b) < var(\bar{y}_c) + var(\bar{y}_b).$$

Thousands of computations have shown that the subtraction of positive covariances tends to reduce the effects of clustering, but not to eliminate them altogether.

6. Computing the Values of deff and s^2/n

For the estimates $deff^2(\bar{y}) = var(\bar{y})/(s^2/n)$ there exists a formidable literature for computing $var(\bar{y})$, too large and varied to be summarized here. On the other hand, values of $s^2/n = (\Sigma y_j^2 - \bar{y}^2)/(n-1)n$ are computed from the n cases from complex clustered and stratified samples, with little justification either in the literature or in the minds of the computers. Fortunately, justification seems both simple and ample.

Define $\hat{s}^2 = \Sigma y_j^2/n - \bar{y}^2 = s^2(1-1/n)$, and then note that

$$\begin{aligned} Exp(\hat{s}^2) &= Exp[\Sigma y_j^2/n - \bar{y}^2] \\ &= \Sigma_i Y_i^2/N - Exp[\bar{y}^2] \\ &= [\Sigma_i Y_i^2/N - \bar{Y}^2] - [Exp(\bar{y}^2)] + \bar{Y}^2 \\ &= \sigma^2 - Var(\bar{y}). \end{aligned} \qquad (6.1)$$

This justification relies on $Exp(\bar{y}) = \bar{Y}$, $Exp(\Sigma y_j^2/n) = \Sigma_i Y_i^2/N$, and $Exp(\bar{y}^2 - \bar{Y})^2 = Var(\bar{y})$. These expectations are unbiased for any selection method that uses fixed sample sizes n and equal probabilities of selection n/N. Usually sample sizes n are not fixed, and often the sample selections are unequal and weighted. However the biased estimates from large samples are consistent enough for practical purposes.

From the above we deduce that $\tilde{s}^2 = \hat{s}^2 + var(\bar{y})$ or $\tilde{s}^2 = \hat{s}^2(1+deff^2/n)$ will give adequate estimates; and even $s^2 \approx \hat{s}^2(1+1/n)$ will suffice when $deff^2$ is near 1.

For analytical statistics there exist analogous but more complex methods. The values of var(b) can be computed with either Taylor methods or with repeated replications. The srs variances for the denominator of $deff^2(b)$ are often yielded by canned computer programs.

7. DEFFs for Weighted Means

This problem, mostly overlooked and neglected, because it is difficult to treat neatly, deserves at least

brief attention even here. How should $deft^2(\bar{y})$ be computed for weighted means $\bar{y} = \Sigma k_i y_i / \Sigma k_i$? The weights may represent unequal selection probabilities, response rates, or adjustments.

a) Sometimes $deft_a^2(\bar{y}) = var(\bar{y})/(s_y^2/n)$ has been computed, with both $var(\bar{y})$ and s_y^2 based on proper weights. This has several deficiencies, because both $var(\bar{y})$ and $deft_a^2(\bar{y})$ confuse the effects of weighting with those of clustering and stratification, whereas n does not: here s_y^2/n estimates the variance of a self-weighting srs sample of n cases. This is not highly portable. For example, increases due to weighting tend to persist undiminished, whereas effects due to clustering tend to disappear from small crossclasses. Also I have seen two papers with roh = $(deft^2-1)/(b-1) > 1$, absurd results due to weighting causing too large $deft_a^2(\bar{y})$.

b) It would be possible to compute $deft_b^2(\bar{y}) = var_u(\bar{y})/(s_u^2/n)$ with both $var_u(\bar{y})$ and s_u^2/n unweighted, and thus the effects of weighting excluded from the ratio. However, this would yield $deft_b^2(\bar{y})$ for an artificially distorted population, which could differ from DEFF for the target population. It would also add the burden of computing the unweighted $var_u(\bar{y})$.

c) A preferred procedure may be to compute $deft_c^2(\bar{y}) = deft_a^2(\bar{y})/(1+L)$, when the effects of "random" weighting can be expressed as the increase by (1+L) of the variances - in the decrease by (1+L) of the effective sample size. The effects of weights on DEFFs clearly deserve more thorough treatment.

References

Kalton G and Blunden RM [1973], Sampling errors in the British General Household Survey, Bulletin of the Int. Statistical Inst, 45(3), 83-97.

Kish L [1965], Survey Sampling, New York: John Wiley and Sons, Section 5.4, 8.2, 8.3, 14.1.

Kish L [1987], Statistical Design for Research, New York: John Wiley and Sons, Sections 2.2, 2.6, 7.1, 7.2.

Kish [1989], Sampling Methods for Agricultural Surveys, Rome: FAO, Stat. Div. Series No 3, Sections 14.1-14.4.

Kish L and Frankel MR [1947], Inference from complex samples, Jour. Roy. Statl. Soc (B), 36, 1-37.

Kish L, Groves RM, and Krotki KP [1976], Sampling Errors for Fertility Surveys, The Hague: ISI.

Kott PS [1989], Assessing linearization variance estimators, Proceedings of Survey Research Methods, Am. Statistical Assoc..

Lepkowski JM and Landis JR [1986], Design effects for linear contrasts of proportions and logits, Proceedings of the Survey Research Methods, Amer. Statistical Assoc.

Rao, JNK and Wu CFJ [1985] Inference from stratified samples, Jour. of the Amer. Statistical Assoc., 80, 620-30.

Rust KF [1984], Techniques for Estimating Variances for Sampling Surveys, Ann Arbor: U of Mich, Ph.D. dissertation.