Inference from Complex Samples

By LESLIE KISH and MARTIN RICHARD FRANKEL

The University of Michigan and The University of Chicago

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, October 17th, 1973, Professor J. GANI in the Chair]

SUMMARY

The design of complex samples induces correlations between element values. In stratification negative correlation reduces the variance; but that gain is less for subclass means, and even less for their differences and for complex statistics. Clustering induces larger and positive correlations between element values. The resulting increase in variance is measured by the ratio *deff*, and is often severe. This is reduced but persists for subclass means, their differences, and for analytical statistics. Three methods for computing variances are compared in a large empirical study. The results are encouraging and useful.

Keywords: CLUSTERS; COMPLEX SAMPLE; SAMPLING ERROR; DESIGN EFFECT; BRR; JACKKNIFE; INFERENCE; STANDARD ERROR; REPLICATION; INTRACLASS CORRELATION; SAMPLE DESIGN; SAMPLING VARIANCE; SUBCLASS ANALYSIS; STRATIFICATION; REPLICATION

1. INTRODUCTION

STANDARD statistical methods have been developed on the assumption of simple random sampling. The assumption of the independent selection of elements (hence independence of observations) greatly facilitates obtaining theoretical results of interest. It is essential for most measures of reliability used in probability statements, such as $\sigma/\sqrt{n}$, chi-squared contingency tests, analysis of variance, the nonparametric literature and standard errors for regression coefficients. Assumptions of independence yield the mathematical simplicity that becomes more desirable—and at present necessary—as we move from simple statistics such as means, to the complex statistics typified by regression analysis. Independence is often assumed automatically and needlessly, even when its relaxation would permit broader conclusions.

Although independence of sample elements is typically assumed, it is seldom realized in the procedures of practical survey work. Randomization of the sample would be unnecessary if the population itself were randomized, but "well-mixed urns" are seldom provided by nature or created by man. This uneasy situation exists widely in the social sciences. In the natural and physical sciences it is typically even more difficult to achieve complete randomization of the sample over the target population, but more often it may be somewhat reasonable to assume its existence in the population—although not entirely and not always.

Much research is actually and necessarily accomplished with complex sample designs, especially in social, health, economic and business studies. It is often economical to select existing clusters or natural groupings of elements. These are characterized by relative homogeneities within the clusters that negate the assumption of the independence of sample elements. The assumption may fail mildly or badly; hence standard statistical techniques result in mild or bad underestimates in reported probability intervals. Overestimates can seldom be severe.

Survey sampling was developed, mostly in the social sciences, censuses and agriculture during the past half century, to provide statistical techniques and theory for the complex selection methods needed for large-scale surveys. It was developed (1) for samples which were not simple random samples, and (2) for finite populations rather than hypothetical infinite populations. It was developed chiefly for descriptive statistics, that is for means, proportions and aggregates. Early landmarks in the literature of sampling were papers by Tchuprow (1923), Neyman (1934), Mahalanobis (1944) and Yates (1946). Five classics in five years—Yates (1949), Deming (1950), Cochran (1953), Hansen *et al.* (1953) and Sukhatme (1954)—outlined the boundaries that have largely defined and confined developments of methods in this subdiscipline of statistics.

The literature of survey sampling concentrates on providing estimates $\bar{y}$ and its standard error, ste$(\bar{y})$. The estimate $\bar{y}$ may be for an aggregate like $\hat{Y} = Fy$, where F is a constant and $y = \sum y_j$ is the sample sum over elements of a variable; or it may be a simple mean y/n of the sample elements, or a weighted mean, $\sum w_j y_j / \sum w_j$; or it may be a ratio, regression or difference estimator of the mean or aggregate. Further, ste$(\bar{y}) = \sqrt{\text{var}(\bar{y})}$ is the estimated standard error, computed from the sample data in accord with the complexities of the sample design. The function of these statistics is to provide statistical intervals of the type $\bar{y} \pm t_p$ ste$(\bar{y})$ for inference about population values of the estimated $\bar{Y}$.

We think it imperative and urgent to extend these statements to more complex statistics. More and more researchers are able to obtain data from complex samples, and to write computer programs for complex analytical statistics. We need methods for dealing properly with complex statistics from complex samples. We need statistics for probability statements. Such statements are symbolized here with $b \pm t_p$ ste(b), where b is some complex statistic, and ste(b) is its computed standard error. For example, b can be the difference of two subclass means or a regression coefficient. Inferences based on standard errors are acceptable on the assumption that survey samples are large enough to yield the needed approximate normality in spite of the nonindependence of the observations. Standard errors should be computed in accord with the complexity of the sample designs; neglect of that complexity is a common source of serious mistakes (Kish, 1957; Kish and Frankel, 1970). On the other hand, trying to obtain more exact but more complicated statistics than standard errors would become too difficult for complex selection designs.

When discussing the independence of observations, we deliberately neglect here the problems of sampling without replacement from finite populations. A learned literature devoted to these problems has recently arisen, but is limited essentially to the relatively simple problems of means and aggregates. We believe that the important theoretical issues of representing defined finite populations concern all statistical applications, not only survey sampling. We welcome similar recent views from C. R. Rao (1971): "Unfortunately the same situation prevails in other areas and considerable literature in statistics is devoted to an examination of the foundations of statistical methodology." The theoretical implications are important, pervasive and subtle, but practical effects are usually small, and we may safely ignore them in the present discussion.

For our discussions of inferential procedures, we propose dividing into three levels of complexity both the selection methods and the statistical estimates, as shown in Fig. 1. The divisions are arbitrary but useful. Among estimators beyond sample means, the analysis of subclasses is common, presents fewer problems than more

complex measures of relations, yet provides analogies and conjectures about them. Among selection methods, the stratification of elements generally has effects that are simpler and weaker than those of clustered sampling.

	STATISTICS		
SELECTION METHODS	1 Means and totals of entire samples	2 Subclass means and differences	3 Complex analytical statistics, e.g. coefficients in regression
A. Simple random selection of elements			
B. Stratified selection of elements		Available	Conjectured
C. Complex cluster sampling		Available	Difficult: *BRR, JRR, TAYLOR*

FIG. 1. The present status of sampling errors. Row 1 is the domain of standard statistical theory, and column 1 of survey sampling.

Standard statistical theory continues to supply new and improved inference procedures for row A, always assuming independent observations. In contrast, the literature of survey sampling is mostly confined to column 1, with theoretical discussions in cell A1 about its finite populations. In stratified element sampling, solutions are clear and simple for means and totals of entire samples (cell B1); they are also fairly simple for subclass means and their comparisons (cell B2, discussed in Section 2). For the more complicated analytical statistics used for relations between variables (cell B3), the solutions seem theoretically difficult and unclear, but rather simple conjectures appear reasonable, with some empirical justification. In clustered samples, for simple means and totals (cell C1), and for subclasses and their comparisons (cell C2), the answers are usually relatively simple and useful (discussed in Section 3). Our main concern (discussed in Sections 4 and 5) must be with complex analytical statistics from clustered samples (cell C3). We have some useful results, but we also have suggestions for further work. All four of the areas (cells B2, B3, C2 and C3), and cell C3 in particular, present challenging problems in need of both theoretical and empirical contributions. They are of utmost importance to statistical applications, and of great difficulty and variety.

2. STRATIFIED SAMPLES OF ELEMENTS (B2 AND B3)

The problems of subclasses are common and not difficult here. There are some useful and surprising results, especially for the kind of domains we shall call *cross-classes*—subclasses that cut across the strata used in selection. In crossclasses, the M_{ch} subclass members of the cth subclass among the N_h elements of the population in the hth stratum are distributed *roughly proportionately*; so that $\bar{M}_{ch} = M_{ch}/N_h$ in the stratum roughly equals $\bar{M}_c = M_c/N = \sum M_{ch}/\sum N_h$ in the population. This is

typical of most subclasses used in analyses of survey data. Conversely, the case when the subclass can be placed into separate strata before selection belongs to standard sampling theory B1. So does the situation when the data can be adjusted after selection with post-stratification weights M_{ch}/M from known values of M_{ch}. But for most subclasses the values of M_{ch} are unknown, and the sample of elements m_{ch} of subclass members among the n_h selected at random from the N_h in the hth stratum is a random variable. This common situation has drastic effects on the behaviour of the sample, as was first noted by Yates (1953).

The most drastic effect is on the variance of the simple crossclass aggregate $\hat{Y}_c = \sum y_{ch} N_h/n_h$, where $y_{ch} = \sum^{m_{ch}} y_{chj}$, which is the crossclass aggregate of the m_{ch} sample element values y_{chj} selected at random from the hth stratum. Here the effect of using $\hat{Y}_c$ for a subclass would be to increase the element variance approximately from σ^2_{ch}, the variance of element values of the cth subclass members around their mean $\bar{Y}_{ch} = Y_{ch}/M_{ch}$ in the hth stratum, to $\{\sigma^2_{ch} + (1-\bar{M}_{ch})\bar{Y}^2_{ch}\}$; the element relvariance is increased by $(1-\bar{M}_{ch})$. This drastic loss is well known and generally avoided in practice by using some other estimators such as $\sum M_{ch}\bar{y}_{ch}$.

Also well known is the variance for the mean $\bar{y}_c = \sum y_{ch}/\sum m_{ch}$ of a *proportionate* stratified sample of elements:

$$\text{var}(\bar{y}_c) = \frac{1-f}{fM_c}\sum_h W_{ch}\{\sigma^2_{ch} + (1-\bar{M}_{ch})(\bar{Y}_{ch} - \bar{Y}_c)^2\}, \tag{2.1}$$

where

$$f = n/N = n_h/N_h = f_h, \quad W_{ch} = M_{ch}/M_c, \quad \bar{Y}_{ch} = \sum_i Y_{chi}/M_{ch} \quad \text{and} \quad \bar{Y}_c = \sum_h \sum_i Y_{chi}/M_c.$$

fM_c is the expected value of the random sample size $m_c = \sum m_{ch}$. We can also express the variance in a slightly different form:

$$\text{var}(\bar{y}_c) = \frac{1-f}{fM_c}\{\sigma^2_c - \sum \bar{M}_{ch} W_{ch}(\bar{Y}_{ch} - \bar{Y}_c)^2\}, \tag{2.1'}$$

where

$$\sigma^2_c = \sum_h \sum_i (Y_{chi} - \bar{Y}_c)^2/M_c \quad (i = 1, \ldots, M_{ch}).$$

Notice that the element variance in brackets approximately takes the place of σ^2_c (or S^2_c) in the variance one would have from a simple random sample of m_c out of M_c elements in the subclass. On the other hand, a proportionate stratified sample with $m_{ch} = fM_{ch}$ in every stratum would have an element variance of

$$\sigma^2_c - \sum W_{ch}(\bar{Y}_{ch} - \bar{Y}_c)^2.$$

The last term is the between-stratum variance, to be gained from proportionate stratification for the subclass itself. Note that for *means of crossclasses the gains of proportionate stratification*, from the between-stratum components $(\bar{Y}_{ch} - \bar{Y}_c)$, *tend to vanish* in proportion to $\bar{M}_{ch} = M_{ch}/N_h$, *and the variance approaches that of simple random sampling*. Durbin (1958) wrote: "... if the proportion in the domain of study is small most of the advantage of stratification has been lost, while only if the proportion is close to unity has the advantage been retained." For example, a gain of 12 per cent for the entire sample would be reduced to 1·2 per cent for a crossclass of 10 per cent.

The simple formulas above neglect only the factors $N_h/(N_h-1)$ in the precise and general variance, without the assumption ($f_h = f$) for proportionate selection:

$$\text{var}(\bar{y}_c) \doteq \sum_h \frac{1-f_h}{f_h M_{ch}} \frac{N_h}{N_h-1} W_{ch}^2 \{T_{ch}^2 - \bar{M}_{ch}(\bar{Y}_{ch}-\bar{Y}_c)^2\} \tag{2.2}$$

$$= \sum_h \frac{1-f_h}{f_h M_{ch}} \frac{N_h}{N_h-1} W_{ch}^2 \{\sigma_{ch}^2 + (1-\bar{M}_{ch})(\bar{Y}_{ch}-\bar{Y}_h)^2\}, \tag{2.2'}$$

where

$$T_{ch}^2 = \sigma_{ch}^2 + (\bar{Y}_{ch}-\bar{Y}_c)^2 = \sum_i (Y_{chi}-\bar{Y}_c)^2/M_{ch}.$$

The approximation in the above is due only to the use of a ratio mean. For derivations see Durbin (1958), Hartley (1959) or Kish (1961, 1965).

The approach to simple random sampling signalled by the T_{ch}^2 terms is even faster for the difference of the means of two subclasses. Computing the difference of two subclass means, $\bar{y}_c$ and $\bar{y}_b$, from the same sample is a most common technique for measuring relationships. The variance of the difference may be written for proportionate sampling as

$$\text{var}(\bar{y}_c-\bar{y}_b) \doteq \frac{1-f}{fM_c}\sigma_c^2 + \frac{1-f}{fM_b}\sigma_b^2 - (1-f)\sum_h \frac{1}{n_h}\{W_{ch}(\bar{Y}_{ch}-\bar{Y}_c) - W_{bh}(\bar{Y}_{bh}-\bar{Y}_b)\}^2. \tag{2.3}$$

This neglects factors of $N_h/(N_h-1)$. More precisely and generally the above is

$$\text{var}(\bar{y}_c-\bar{y}_b) \doteq$$

$$\sum_h (1-f_h)\frac{N_h}{N_h-1}\left[\frac{W_{ch}^2 T_{ch}^2}{f_h M_{ch}} + \frac{W_{bh}^2 T_{bh}^2}{f_h M_{bh}} - \frac{1}{n_h}\{W_{ch}(\bar{Y}_{ch}-\bar{Y}_c) - W_{bh}(\bar{Y}_{bh}-\bar{Y}_b)\}^2\right]. \tag{2.4}$$

The third term will tend to become relatively small because $n_h = f_h N_h$ is large compared to $f_h M_{ch}$ for small subclasses. Furthermore, since strata typically tend to have "additive" effects, it tends to become negligible due to similar, and therefore cancelling, stratum differentials. Hence, *for the difference of two crossclasses, the gains of proportionate stratification tend to vanish.* The two variances may be computed as if for unstratified random samples, except for weighting for unequal sampling rates.

Controlling the sample sizes, m_{ch} and m_{bh}, for crossclasses is difficult in practical surveys; but where feasible, it suggests optimal allocation for the differences of cross-class means (Sedransk, 1957).

The formulas for computing sample variances reflect the above (2.2–2.4):

$$\text{var}(\bar{y}_c) = \sum (1-f_h)\frac{w_{ch}^2}{m'_{ch}}\{t_{ch}^2 - \bar{m}_{ch}(\bar{y}_{ch}-\bar{y}_c)^2\}, \tag{2.5}$$

where

$$w_{ch} = F_h m_{ch}/\sum F_h m_{ch}, \quad F_h = N_h/n_h, \quad \bar{m}_{ch} = m_{ch}/n_h,$$
$$m'_{ch} = m_{ch}(n_h-1)/n_h \quad \text{and} \quad t_{ch}^2 = \sum_j (y_{chj}-\bar{y}_c)^2/m_{ch}.$$

For the difference of two means approximately

$$\text{var}(\bar{y}_c-\bar{y}_b) \doteq \sum (1-f_h)\left(\frac{w_{ch}^2 t_{ch}^2}{m'_{ch}} + \frac{w_{bh}^2 t_{bh}^2}{m'_{bh}}\right). \tag{2.6}$$

For proportionate samples $f_h = f$ and (2.6) becomes

$$\text{var}(\bar{y}_c - \bar{y}_b) = \frac{1-f}{fm_c}\hat{\sigma}_c^2 + \frac{1-f}{fm_b}\hat{\sigma}_b^2, \tag{2.7}$$

where

$$m_c = \sum m_{ch} \quad \text{and} \quad \hat{\sigma}_c^2 \doteq \sum_h w_{ch} t_{ch}^2 n_h/(n_h - 1) = \sum_h n_h(n_h-1)^{-1} \sum_j (\bar{y}_{chj} - \bar{y}_c)^2/m_c.$$

Differences between means provide the most common bases for measuring relationships between variables in survey data. Furthermore, they also provide grounds for conjectures about the sampling fluctuations of other analytical statistics (cell B3 in Fig. 1) used for measuring relations between variables in stratified samples. These conjectures are necessary for the more complex statistics for which standard theoretical analysis cannot provide measures of sampling fluctuations in accord with stratified designs.

Table 1 contains remarkable confirmation of these conjectures applied to a large and diverse group of chi-squared tests. Eight sets of data from stratified samples

TABLE 1

Ratios of three iterated chi-squared tests to SRS tests

Eight contingency tables based on proportionate stratified samples from Israel: Nos. 1–4 of savings, No. 5 of attitudes, No. 6 of hospital data, No. 7 of poultry medicament, No. 8 of perception experiments. Adapted from data of Nathan (1972)

Data set	*No. of strata*	*Row × columns*	*Sample size*	*Nathan's three tests*					
				First iteration			*Last iteration*		
				X^2	X_1^2	G	X^2	X_1^2	G
1	4	3×3	845	1·028	0·992	1·017	1·004	1·004	1·005
2	4	3×3	821	1·088	0·963	1·043	0·999	1·003	1·001
3	4	3×3	491	1·740	0·707	1·406	1·011	1·001	1·009
4	4	3×3	2581	1·095	0·959	1·049	1·003	1·005	1·003
5	6	2×4	500	1·079	0·967	1·040	1·004	1·003	1·003
6	3	2×2	120	1·013	0·967	1·009	1·008	0·969	1·007
7	5	2×2	269	1·076	0·989	1·043	1·011	1·015	1·011
8	2	2×4	81	1·368	0·889	1·186	1·029	1·037	1·029

were involved, and on each, three sophisticated iterated techniques (Nathan, 1972, 1973) were used to fit their stratified selections. Then Professor Nathan agreed to compute the same tests, but with "naive" *SRS* assumptions. Finally, we computed the ratios of the sophisticated to the naive results. Note that in the last iterations the ratios are all within 4, and mostly within 1 percent of 1·00. These values measure how close the naive estimates are to the last iterations, hence may slightly overestimate.

We conjecture that similar results usually will be obtained on other data, and also on other analytical statistics based on stratified random selections (case B3 in Fig. 1). It is a useful conjecture, because appropriate computations of sampling

errors for analytical statistics will be difficult for the foreseeable future. Clearly we need more research, both theoretical and empirical. The accumulation of empirical evidence will be most useful, but alone it is slow, subject to sampling fluctuations, and not completely convincing. Theoretical foundations would strengthen and hasten understanding, but alone they cannot suffice. The conjecture involves parameters with empirical content, and it can be contradicted in rare situations.

What attitudes should we adopt concerning the results on crossclass comparisons (2.6, 2.7), and the analogous conjectures about analytical statistics? On the one hand, if justified, we should welcome the convenience of the many formulas available on simple random assumptions. On the other hand, we may be surprised and annoyed that the effects of proportionate stratification tend rapidly to vanish altogether. It is not a result that would suggest itself to intuition.

We may conveniently summarize this section in terms of "design effects", a concept we shall use repeatedly. "The *design effect* or *Deff* is the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements" (Kish, 1965, p. 258); this may do here briefly for elements selected with equal probability. This concept, under diverse names, has been long and widely used. The design effect, in proportionate stratified element samples, (a) is less than 1 typically for means based on entire samples; (b) tends towards 1 from below for crossclasses, as these become small; (c) is close to 1 for differences between crossclass means; and (d), we conjecture, is close to 1 for analytical statistics in general.

3. Subclasses and Comparisons in Cluster Sampling (C2)

When diverse subclasses are completely segregated in separate clusters and strata (as for regional estimates in area samples) their treatment needs no new methods, although problems of multipurpose design and allocation arise (Kish, 1973). However, most frequently we must deal with new problems due to *crossclasses*, that is subclasses that cut across clusters; examples of these are age, sex and social classes in area samples. Dealing with crossclasses instead of the entire sample produces two principal effects on the sizes of the sample clusters, (1) a decrease in average size by the factor $\bar{M}_c$, and (2) an increase in the coefficient of variation CV (x).

Control of sample size, either with stratification or with PPS (probability proportional to size), is typically imperfect even for the entire sample. Unequal cluster sizes lead to common use of the combined ratio mean, $\bar{y}_c = r = y/x$, and this estimator also serves subclasses. This estimator is rather robust, but not when the denominator x is subject to wild variation. A sufficiently small coefficient of variation CV (x) assures a low bias for the ratio; this is also assumed for deriving its variance (Hansen *et al.*, 1953; Kish, 1965). For subclasses, if small or unevenly distributed, the loss of control over cluster sizes may permit CV (x) to become too large. Our computing programs (Kish *et al.*, 1972) have monitoring features to catch cases when CV (x) is too large for comfort. Actual situations are generally comfortable, because the bias ratio, Bias (r)/Ste $(r) = -\rho_{rx}$ CV (x), is small when either component is small; empirical investigations have been encouraging (Kish *et al.*, 1962).

It is useful to consider a simple model for the variance of crossclass means as a function of the proportion $\bar{M}_c$ in the crossclass. We begin with the well-known var $(\bar{y}) = \{1 + \text{Rho}\,(n/a - 1)\}\, S^2/n$, where brackets contain the design effect for clusters of size n/a. Then:

$$\text{var}\,(\bar{y}_c) = \{1 + \text{Rho}_c\,(\bar{M}_c\, n/a - 1)\}\, S_c^2/m_c. \tag{3.1}$$

Here

n = the number of elements in the entire sample,

a = the number of clusters in the sample,

$\bar{M}_c$ = the proportion of crossclass elements in the sample,

$m_c = n\bar{M}_c$, the number of crossclass members expected in the sample,

Rho_c = intraclass correlation for the crossclass,

S_c^2 = element variance of crossclass members.

The formula fails to make separate allowance for the effects of unequal sizes of sample clusters, and for the effects of stratification. We may consider these either as having been ignored, or as having been implicitly included in the definitions of the parameters of the equation. It obviously breaks down when $n\bar{M}_c$ approaches a, and should not be taken seriously for such small clusters of crossclass members. Subject to these limitations, the design effect in brackets is viewed as a function of $\bar{M}_c$, and of Rho_c and S_c^2. To the extent that the latter two are relatively constant for a group of similar variables, we see the increase over 1 of the design effect in relation to $\bar{M}_c$: the design effect tends toward 1 for decreasing crossclasses.

The estimation of the variance of $(\bar{y}_c)$ proceeds according to standard formulas for the ratio mean (y/x) of two random variables. However, small and fluctuating sizes of sample clusters cause problems; but these may be countered with two types of averaging procedures. First, with "combined strata" (Kish, 1965) we can combine primary selections, chosen at random across strata, to form larger units for computing the variance. The procedure introduces no bias into the estimated variance, but increases its variance.

Second, and much more important, are procedures for averaging computed variances for a group of means or other variates. Survey results are produced for so many variates—for different survey variables, and each of these for many subclasses—that the process of computing and presenting sampling errors for all of them usually becomes too costly and cumbersome. But we may compute them for a subset, and then make inferences from their average to the entire set. Furthermore, such averages may be computed for several meaningful sets of results. Another reason for averaging is to produce more stable estimates of variances than sample designs usually yield; averaging should increase the accuracy (lower mean-square error) in spite of introducing some "bias" for the individual variances. To control and reduce that bias, averaging should be confined to groups of similar variates, and should be performed with methods that promise stability within those groups.

A common method for averaging is to plot a graph of the computed design effects deff (or $\sqrt{\text{deff}}$) against subsample sizes m_c. Here Deff and Rho refer to population values and deff and rho to sample values. Using deffs removes two obvious sources of disturbing factors, S_c^2 and m_c, from the averaging of variances computed for different variates. This method assumes a common Deff and Rho_c for variates within a pooled set, as a function of m_c and due chiefly to the same design. The averaging may be done separately for more or less similar groups of the variates. If Rho is constant over values of m_c, then Deff approaches 1 linearly with decreasing m_c (Kalton and Blunden, 1973).

Variances for differences between subclass means raise new issues. Formulas for computing $\text{var}(\bar{y}_c - \bar{y}_b) = \text{var}(\bar{y}_c) + \text{var}(\bar{y}_b) - 2\,\text{cov}(\bar{y}_c, \bar{y}_b)$ are merely extensions of variances for single means (Kish, 1965, Section 6.5). Because subclass comparisons are so basic and common in survey analysis, it is annoying to find computations of their variances so rarely even today. At the Survey Research Center they have been imbedded into our computing programs for variances since 1952, and from hundreds of computations we found bases for the inequalities:

$$\frac{S_c^2}{m_c} + \frac{S_b^2}{m_b} < \text{var}(\bar{y}_c - \bar{y}_b) < \text{var}(\bar{y}_c) + \text{var}(\bar{y}_b). \tag{3.2}$$

In words, the variance of the difference of two means from clustered samples shows the design effect of a positive intraclass correlation, but that effect is less than for the separate means. In other words, the covariance is positive, but not great enough to cancel the design effects of the separate means. See Kish (1965, Section 14.1), and Kalton and Blunden (1973).

This is an empirical statement about the additive nature of positive clustering effects in crossclasses. In actual computations subject to large sampling variations, it has often been contradicted, but in our experience these exceptions were negated when recomputed on similar data. Although it cannot be logically perfect, it is a dependable and useful empirical law. It is clearly preferable to the common practice of assuming equality at either extreme. Most commonly, equality is assumed on the left, as if the samples were simple random. Less commonly, the equality on the right is assumed, with the "conservative" estimate that disregards the covariance of crossclasses selected from the same sample of clusters.

Subclass comparisons represent a basic measure of relations between variables. Our findings about them lead to conjectures about design effects for other statistics that measure relations, such as regression coefficients. When techniques were unavailable for computing variances for them, we conjectured that design effects were greater than 1, but less than for the means of the variables involved (Kish, 1957). These conjectures have received empirical confirmations (Kish and Frankel, 1970; Frankel, 1971) as discussed in the following two sections.

4. COMPLEX STATISTICS FROM COMPLEX SAMPLES (C3)

Here we deal with clustered samples and with statistics more complex than the difference of subclass means. The following section describes how new techniques now make possible the computations of variances that incorporate the complexities of the sample. We shall justify the need for such computations with three broad propositions:

(1) Statistics (means, regression coefficients, etc.) approach their population values as the sample size increases.
(2) The approach is generally slowed by design effects.
(3) The design effects differ for different statistics, for different variables, and for different sample designs.

The three propositions presuppose that we are concerned with finite and real measurements and populations. This philosophy, which should be assumed by anyone involved in the application of statistics, pervades survey sampling theory.

Here we extend it from means to measures of relationships. Consider a realistic view of regression:

(I) There exists a finite population of N elements. Associated with each of these elements is a vector of $k+1$ values $Y_i, X_{1i}, X_{2i}, \ldots, X_{ki}$;

$$P = \{(Y_i, X_{1i}, X_{2i}, \ldots, X_{ki}) \,|\, i = 1, \ldots, N\}.$$

(II) Our parameters are numbers B_j such that $\sum_i^N (Y_i - \sum_j^k B_j X_{ji})^2$ is minimum subject to

$$\sum_i^N (Y_i - \sum_j^k B_j X_{ji}) = 0.$$

(III) Given a sample of n vectors from the population of N vectors our desire is to estimate the parameters B_j.

The regression model stated in (II) does not in practice correspond exactly (or even closely) to the complex relationship among the actual population of vectors. The error term measures (usually in a least-square sense) the extent to which the model departs from the actual complex relations among the population of vectors.

The statistical theory of regression begins at the other end—the theoretical end. It first assumes a basic structure of relationships. Letting $\mathbf{X}_i = (X_{1i}, \ldots, X_{ki})^{\mathrm{T}}$, and $\mathbf{B} = (B_1, \ldots, B_k)^{\mathrm{T}}$, it uses the model $Y_i = \mathbf{B}^{\mathrm{T}} \mathbf{X}_i + \varepsilon_i$. It then makes several strong assumptions:

(A) *linearity*: $E(\varepsilon_i | \mathbf{X}_i) = 0$, for all i;
(B) *homoscedasticity*: $\mathrm{var}(\varepsilon_i | \mathbf{X}_i) = \sigma^2$, for all i;
(C) *independence between observations*: $\mathrm{cov}(\varepsilon_i \varepsilon_j | \mathbf{X}_i, \mathbf{X}_j) = 0$, for all $i \neq j$;
(D) *normality* for the ε_i.

Assumptions (A), (B) and (D) concern the basic structure of the universe of the model, whereas (C) involves independent selections from it. This (or a similar) well-specified model yields several desirable results; the standard least-square estimates $\hat{b}$ are minimum variance, linear, unbiased, normal, etc. Literature and textbooks are written about this pretty model; this is what the research workers find in statistical textbooks, but they find very little about how to reconcile this model with the real population they are investigating.

Specifically, we need the real population model to describe the principal effect of a complex selection design: assumption (C) fails to hold. Clustered selection tends to introduce positive correlations between the errors of the model and, as will be shown later, these often have serious consequences.

The first proposition states that the correlation between elements *does not prevent* the approach of "first-order statistics" based on large samples to their respective population values (parameters). By *first-order statistics* we mean estimates of *parameters of the population distribution*; these parameters are (a) the substantive objectives of research, (b) based on *all* population elements taken *individually* and (c) unaffected by the sample design (for example, means, element variances, regression and correlation coefficients). On the other hand, by *second-order statistics* we mean *measures of variation* (variance, standard error, mean-square error) of first-order statistics. They estimate second order parameters, e.g. $E\{y - E(y)\}^2$, that are based on aggregate results of samples obtained under specific designs, and they are affected by the correlations between elements induced by the designs.

We need laws of large numbers and central limit theorems for samples in which elements are not independent. A difficult problem here is to specify what a "large" sample may be. It is not sufficient merely to say that it contains a large number of elements if these come highly correlated from a few clusters (i.e. the primary selections). To allow the clusters to become very numerous would suffice, but that would place unrealistic demands on many practical sample designs. Large numbers of primary selections are often neither possible nor necessary; a moderate number will suffice if the elements are numerous and the correlations between them not too great. We recognize that a rigorous theoretical formulation of the last statement, so needed for practical work, stands as a difficult challenge for theoretical statisticians. We shall merely sketch first what we know today about this topic in terms of the unbiased (or almost unbiased) nature of the expected values of the results of probability samples, without assuming independence of selections.

The weighted sum of sample observations is an unbiased estimator of the population aggregate: $E(\sum_j^n y_j/P_j) = \sum_i^N Y_i$, where P_j is the selection probability of the jth sample element. This is also true of vectors $\mathbf{y}_j$ and $\mathbf{Y}_i$ for several variables. (This expectation seems to be the basis for selections without replacement.) It is also true of moments of sample values, such as $\sum y_j^k/P_j$ or $\sum y_j^k x_j^m/P_j$, where k and m are real numbers, usually integers. Means based on these sums are also unbiased if the denominator is fixed; otherwise they are consistent and close (Kish, 1965, Section 2.8). Often the denominator is a correlated sum of sample observations as in the ratio mean $r = (\sum y_j/P_j)/(\sum x_j/P_j)$; empirical evidence is reassuring (Kish *et al.*, 1962). Many statistics are complex versions of functions of ratio means, e.g.,

$$b = (\sum y_j x_j/P_j)/(\sum x_j^2/P_j).$$

We conjecture that the biases in estimates of regression and correlation coefficients (as with ratio means) are functions of the departure of actual population relations from assumptions (A)–(D) above, and of sample sizes.

TABLE 2

Biases in complex samples of five types of estimators

Averages of relative biases = bias $(p_s)/P$, and of the bias ratios = bias (p_s)/ste (p_s). Data from Tables 5.1 and 5.2 of Frankel (1971)

Estimator	*No. in average*	6 *Strata design* 300 *samples*		12 *Strata design* 300 *samples*		30 *Strata design* 200 *samples*	
		Relative bias	*Bias ratio*	*Relative bias*	*Bias ratio*	*Relative bias*	*Bias ratio*
Ratio means	8	0·00425	0·04653	0·00216	0·03909	0·00295	0·08520
Simple correlations	12	0·06972	0·19847	0·05399	0·19013	0·01748	0·09896
Regression coefficients	8	0·04978	0·04429	0·03320	0·07108	0·02776	0·05558
Partial correlations	6	0·12333	0·21155	0·08365	0·20165	0·05863	0·17358
Multiple correlations	2	0·16002	0·72855	0·11115	0·52975	0·04670	0·29105

Analytical expressions would be complex and are not available now. Hence our inferences must rely heavily on accumulating empirical evidence. In Table 2 we

summarize results on the biases of five types of estimators from a complex sample (Frankel, 1971). The results from this large study, described in Section 5, are reassuring. Even for small sizes (12 units from 6 strata) the relative biases (ratio of bias to parameter) are small, and they decrease with the sample size for means and for three diverse types of regression coefficients. Because the standard errors also decrease the bias ratios (ratio of bias to standard error) fluctuate, but they remain small or moderate. The multiple correlation coefficient shows worse behaviour; but we think this is due to the basic defect of the estimator, rather than to the design's complexity.

In addition to small biases, in Table 3 the same study also gives us comforting news about the approaches to normality of complex statistics from complex samples.

TABLE 3

Approach to normality in complex samples of five types of estimators

Averages of relative frequencies within stated intervals of statistics $(p_s - \bar{p})/\text{ste}\,(p_s)$. Data from Table 5.3 of Frankel (1971)

Intervals	±2·576	±1·960	±1·645	±1·282	±1·000
P for normal deviate	0·9900	0·9500	0·9000	0·8000	0·6827
6-strata design (300 samples)					
8 Ratio means	0·9875	0·9533	0·9067	0·8075	0·6929
12 Simple correlations	0·9861	0·9533	0·9039	0·8061	0·6986
8 Regression coefficients	0·9742	0·9387	0·9067	0·8392	0·7425
6 Partial correlations	0·9822	0·9444	0·9050	0·8178	0·7039
2 Multiple correlations	0·9767	0·9467	0·9167	0·8417	0·7233
12-strata design (300 samples)					
8 Ratio means	0·9900	0·9550	0·8987	0·8804	0·6750
12 Simple correlations	0·9872	0·9461	0·9003	0·8064	0·6919
8 Regression coefficients	0·9808	0·9492	0·9092	0·8192	0·7079
6 Partial correlations	0·9872	0·9506	0·9050	0·8133	0·6861
2 Multiple correlations	0·9733	0·9383	0·9067	0·8350	0·7233
30-strata design (200 samples)					
8 Ratio means	0·9887	0·9544	0·9100	0·8069	0·6744
12 Simple correlations	0·9900	0·9567	0·9017	0·7929	0·6867
8 Regression coefficients	0·9869	0·9444	0·9019	0·8144	0·6912
6 Partial correlations	0·9875	0·9575	0·8958	0·8000	0·6942
2 Multiple correlations	0·9900	0·9525	0·9100	0·8275	0·6825

The approach seems to be good even for 12 units from 6 strata, and it improves markedly in moderate sample sizes (see also the tables in Section 5).

The approach to population values promised by the first proposition must be accompanied by the second proposition's warning about the slowing of the approach caused by positive correlations among selected elements. The extent of the slowing is simply expressed by the design effect on the variance of the mean: $\text{Deff} = \{1 + \rho(\bar{n} - 1)\}$. This expression is well known for random selections of equal clusters. It also has been extended for sample means obtained from probability samples in general, with ρ expressing the pairwise correlation of sampled elements, and $\bar{n}$ a parameter of the selection design (Tharakan, 1969). The same fruitful approach was used by "Student" (1909) in a paper that, surprisingly, has been neglected in the literature of sampling.

Similar analytical expressions, in a few useful parameters, are needed for the effects of design on the variances of complex statistics. The effective content of the expressions

must be statements about the structure of population variables, and the effect of the selection design on the variates studied. Meanwhile we must accumulate evidence about the magnitudes of these effects. Some of this empirical evidence is shown in Table 4, which summarizes results of Kish and Frankel (1970) and Frankel (1971).

TABLE 4

Values of $\sqrt{\text{Deff}}$ *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

We would like to know how the design effects tend to differ for different statistics obtained from complex selection designs. In addition to scientific curiosity, we have practical needs to discover reasonable regularities.

Often it is difficult to compute standard errors for all statistics, or to compute them with adequate precision. Hence reasonable conjectures would be most useful to researchers. Theory will help eventually, but it will need to be buttressed with empirical content. Our present conjectures have a light theoretical framework and some empirical background. They are phrased in terms of design effects $\text{Deff}(b)$ for complex statistics b from complex samples.

(i) $\text{Deff}(b) > 1$. 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) $\text{Deff}(b) < \text{Deff}(\bar{y})$. 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) $\text{Deff}(b)$ is related to $\text{Deff}(\bar{y})$. For variates with high $\text{Deff}(\bar{y})$, values of $\text{Deff}(b)$ tend also to be high. See Kish and Frankel (1970, Section 7) for a set of striking results.

(iv) $\text{Deff}(b)$ 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) $\text{Deff}(b)$ tends to have observable regularities for different statistics. This is a hope based on theoretical considerations; confirming results would help us make useful conjectures.

A simple model of the above would be

$$\text{Deff}(b_g) = 1 + f_g\{\text{Deff}(\bar{y}) - 1\}, \tag{4.1}$$

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

5. THREE METHODS FOR COMPUTING SAMPLING ERRORS

We shall compare here three basic methods for computing sampling errors from stratified clustered sample designs: The Taylor expansion method (*TAYLOR*), the method of balanced repeated replication (*BRR*) and the method of jackknife replication (*JRR*). These names are convenient, but not unique. The comparisons are based on a large-scale empirical study which contains a fuller discussion of all three (Frankel, 1971).

In that study for the sake of simplicity, we used sample designs with equal numbers A of primary units within all strata, and with two of those units selected from each stratum with random choice, without replacement and without subsampling. Thus, we have a clustered stratified sample design, where each population element has equal probability ($f = 2/A$) of appearing in the sample. However, all three methods for computing variances can deal with appropriate weighting to compensate for unequal probabilities of selection within and between strata. The extension to any number of primary selections per stratum is straightforward for the *TAYLOR*. With modifications and with the use of collapsed and combined strata techniques, methods *BRR* and *JRR* also can be applied to other sample designs (Kish and Frankel, 1970, Section 12).

5.1. *Taylor Expansion Method*

The use of the Taylor expansion for computing variances of ratio means has been described in textbooks. Deming (1960), Kish (1965) and Woodruff (1971) describe its use for estimating variance for other functions of the basic sample sums. The method is also known as the linearization or delta (δ) method. A detailed published extension of this method to more complex first-order estimates specific to survey sampling is due to Tepping (1968). This method produces an approximate estimate for the variance of a first-order statistic, based on variances of the linear terms of the Taylor expansion of the statistic (Brillinger and Tukey, 1964).

Let $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_k)^{\mathrm{T}}$ be a vector of sample totals; the sample total y_i is the aggregate of primary selection totals y_{iha}, where the indexes h and a denote strata and primary selections:

$$y_i = \sum_h \sum_a y_{iha}, \quad \text{where} \quad h = 1, \ldots, H \quad \text{and} \quad a = 1, 2. \tag{5.1}$$

The y_{iha} values are sums over primary selections of element values y_{ihaj}, weighted by the inverses of selection probabilities, so that $E(y_i) = KY_i$, the corresponding population value, with K some convenient constant.

The y_i's are chosen so that $g(Y)$ is the parameter we wish to estimate with the first-order statistic $g(y)$. Using the linear terms of the Taylor expansion $g(y)$ near $g(Y)$, the estimator of $\mathrm{var}\{g(y)\}$ is given by

$$\mathrm{var}\{g(y)\} = (1-f) \sum_h \left\{ \sum_i \frac{\partial g(Y)}{\partial Y_i} y_{ih1} - \sum_i \frac{\partial g(Y)}{\partial Y_i} y_{ih2} \right\}^2. \tag{5.2}$$

$\partial g(Y)/\partial Y_i$ is the partial derivative of $g(Y_i)$ with respect to the variable Y_i, and taken at the expected value Y_i; we must use $\partial g(y)/\partial y_i$ as sample estimators of $\partial g(Y)/\partial Y_i$.

5.2. *Balanced Repeated Replication (BRR) Methods*

The approach of repeated replications was developed at the U.S. Census Bureau (Deming, 1956) from basic replication concepts (Mahalanobis, 1946) and orthogonal

balancing was added later (McCarthy, 1966; Kish and Frankel, 1969). The *BRR* methods can be briefly described as follows. Assume that we have a stratified sample design with two primary selections from each stratum. Let S denote the entire sample; let H_i denote the ith half-sample formed by including one of the two primary selections from each of the strata; and let C_i denote the ith complement half-sample, formed by the primary selections in S not in H_i. The method we used for choosing the pattern of primary units that form the half-samples, H_i and C_i, is known as "full-orthogonal balance". If we form k half-samples $H_1, \ldots, H_k$, and corresponding complement half-samples $C_1, \ldots, C_k$, then we form *BRR* second-order estimators in one of two ways:

$$\text{var}_{BRR\text{-}S}\{g(S)\} = \frac{1-f}{2k}\sum_{i=1}^{k}[\{g(H_i)-g(S)\}^2+\{g(C_i)-g(S)\}^2], \tag{5.3}$$

or

$$\text{var}_{BRR\text{-}D}\{g(S)\} = \frac{1-f}{4k}\sum_{i=1}^{k}\{g(H_i)-g(C_i)\}^2. \tag{5.4}$$

Each of the two components in the *BRR-S* form also may be used separately for a less costly but less precise second-order estimator (Kish and Frankel, 1970; Frankel, 1971).

5.3. *Jackknife Repeated Replication (JRR) Methods*

The term *JRR* refers to a set of second-order estimation methods motivated by jackknife estimation procedures (Brillinger, 1964) and by *BRR*. With *BRR* methods, each of the k replications estimates the variance of the entire sample. With the *JRR* methods, each replication measures the variance contributed by a single stratum. The technique used to measure these contributions to the variance from the strata was suggested by the jackknife method for variances; it was formed by leaving out replicates from the sample. The specific procedures below were first used and described in Frankel (1971).

Assume again that we have two selections from each of H strata. Let S denote the entire sample; let J_i $(i = 1, \ldots, H)$ denote the replicate formed by removing from S one selection in the ith stratum, but including twice the other selection in that stratum. Let CJ_i $(i = 1, \ldots, H)$ denote the complement replicate formed from S by interchanging the eliminated and duplicated selections in the ith stratum.

Two *JRR* estimators of variance are defined as follows:

$$\text{var}_{JRR\text{-}S}\{g(S)\} = \frac{1-f}{2}\sum_{i=1}^{h}[\{g(J_i)-g(S)\}^2+\{g(CJ_i)-g(S)\}^2] \tag{5.5}$$

and

$$\text{var}_{JRR\text{-}D}\{g(S)\} = \frac{1-f}{4}\sum_{i=1}^{h}\{g(J_i)-g(CJ_i)\}^2. \tag{5.6}$$

5.4. *Accuracy of the Three Methods*

If we assume first-order estimates $g(y)$ or $g(S)$ that are linear functions of statistics, then a number of exact analytical results can be derived for all three variance estimation

methods: *TAYLOR*, *BRR* and *JRR*. However, when we consider the first-order estimates actually used by survey analysts (e.g. ratios, correlation and regression coefficients) we find that usable methods for exact (non-approximate, non-asymptotic) results evade us. Since estimators of sampling errors are needed now, we follow a tradition among statisticians that goes back at least as far as 1907, when "Student" (1908) selected 750 simple random samples to evaluate his theoretical derivation of the distribution of the sample mean divided by its estimated standard error.

We empirically evaluated and compared all three variance estimation methods, using three clustered and stratified sample designs. These called for paired selections (approximately 14 elements each) from 6 strata (approximately 170 elements), 12 strata (approximately 340 elements) and 30 strata (approximately 847 elements). The coefficients of variation $CV(x)$ of the sample sizes were 0·19, 0·13 and 0·074 respectively. Thus we imposed rather harsh, demanding tests on the empirical validity of these methods. For a more complete description of this study, which makes use of data from the *Current Population Survey* of the U.S. Bureau of the Census, the reader is directed to Frankel (1971).

The three methods were used to compute sampling errors of several statistics: ratio means, simple correlations and multiple regression coefficients. *BRR* and *JRR* methods also were used to compute sampling errors for partial and multiple correlation coefficients, but for the *TAYLOR* method we were unable to find tractable forms for the partial derivatives.

For standards of comparison we used robust fundamentals based on the definitions of means and variances. The bias of first-order statistics was judged against population parameters, based on the entire population of 45,737 households in 3,240 primary sampling units. The statistics were based on 300 independent drawings for the 6-strata and the 12-strata samples, and 200 drawings for the 30-strata sample. We used 8 variables in 2 multiple regression equations, each with 3 predictor variables. Thus the 8 coefficients of regression, 12 of simple correlation, and 6 of partial correlation, represent averaging $300 \times (8, 12 \text{ and } 6)$ statistics in the 6 and the 12 strata, and $200 \times (8, 12 \text{ and } 6)$ in the 30 strata. The total of about 400,000 complex computations made good use of modern computers.

We could not afford to compute *all* the possible combinations for the second-order statistics such as $E\{\bar{y}-E(\bar{y})\}^2$ for the variance, and $E\{\bar{y}-\bar{Y}\}^2$ for the mean-square error; these were averaged from the 300, 300 and 200 statistics. To these standards for second-order statistics were compared the corresponding results of the three methods.

Here we can only summarize a large set of results. In the original publication (Frankel, 1971), the large volume of details for distinct statistics provides firmer bases for the tables here and for the conclusions derived from them. For first-order estimators the biases on the average were relatively small; this was true both in terms of the relative bias (bias/estimate), and of the bias ratio (the ratio of bias to standard error). These were in the neighbourhoods of 0·05 for means and regression coefficients, and of 0·1–0·2 for simple and partial coefficients; the multiple correlation coefficient was around 0·3–0·7 and clearly presented problems as noted above (Table 2). As for the variability of first-order estimators, the strong design effects were shown in Table 4.

We are chiefly concerned here with the performances of the three methods of computing variances. These are summarized in Tables 5 and 6. Table 5 summarizes the averages of relative biases for the mse's (three mean-square errors), and the averages of their dispersions, measured as mean-squared errors of the mse's. The

corresponding results for the computed variances (Frankel, 1971, Tables 6.1 and 6.3) were close to these results shown for mse's, because the biases that would separate them (mse = var + bias^2) were small or negligible.

TABLE 5

Accuracy for mean-square errors (MSE) for three methods or error computations

Adapted from Tables 6.2 and 6.4 of Frankel (1971)

Relative bias of MSE Bias (MSE)/MSE				*Relative MSE of MSE* MSE(MSE)/(MSE)^2		
BRR	*JRR*	*TAYLOR*		*BRR*	*JRR*	*TAYLOR*
			6 Strata			
0·032	−0·019	−0·041	Means	0·543	0·501	0·483
0·188	−0·006	−0·075	Regression coefficients	4·207	2·803	2·437
−0·040	−0·163	−0·278	Simple correlations	0·772	0·678	0·431
0·029	−0·153	—	Partial correlations	0·989	0·852	—
−0·297	−0·426	—	Multiple correlations	1·168	1·079	—
			12 Strata			
0·064	0·035	0·022	Means	0·437	0·418	0·381
0·097	−0·010	−0·034	Regression coefficients	1·425	1·180	1·134
−0·072	−0·159	−0·243	Simple correlations	0·530	0·483	0·326
−0·013	−0·157	—	Partial correlations	0·686	0·603	—
−0·330	−0·439	—	Multiple correlations	0·993	0·906	—
			30 Strata			
0·004	−0·011	−0·014	Means	0·156	0·152	0·147
0·068	0·019	0·014	Regression coefficients	0·608	0·558	0·554
−0·036	−0·104	−0·159	Simple correlations	0·405	0·349	0·231
0·012	−0·101	—	Partial correlations	0·578	0·497	—
−0·161	−0·286	—	Multiple correlations	1·050	0·895	—

Table 6 presents results for the criterion we consider most significant because it measures directly the inputs of the three methods into inference statements. Against the accepted standards of probability levels (for 6, 12 and 30 degrees of freedom), this table shows the levels actually attained on the average by the sample functions

$$t(s) = \frac{g(s) - E\{g(s)\}}{\text{ste}\{g(s)\}}. \tag{5.7}$$

The proportion of times that the ratio $t(s)$, computed for each sample, fell within fixed symmetric intervals t_p ($\pm 2{\cdot}576$, $\pm 1{\cdot}960$, $\pm 1{\cdot}645$, $\pm 1{\cdot}282$) are shown against the Student's P_t expected probabilities. Relative frequencies are shown for three methods: *BRR-S*, *JRR-S* and *TAYLOR* (from Frankel, 1971, Tables 7.4, 7.8 and 7.1 respectively). We omitted data for *BRR-D*, *BRR-H*, *BRR-C* and for *JRR-D*, *JRR-H*, *JRR-C*; the differences of these from the results shown for *BRR-S* and *JRR-S* are less important and are discussed elsewhere (Kish and Frankel, 1970; Frankel, 1971). The latter gives Tables (7.1–7.9) for all nine variations of the three methods and for asymmetric (one-sided) intervals, for which the performances were less satisfactory

TABLE 6

Relative frequencies of P_t intervals for three methods of error computations

Value of $t = \{g - E(g)\}/\{\text{ste}\,(g)\}$ computed, then for each type of (statistic × design × method) the proportions that fall within $\pm t$ intervals. Adapted from Tables 7.1, 7.4, 7.8 of Frankel (1971)

BRR	*JRR*	*TAYLOR*		*BRR*	*JRR*	*TAYLOR*
	$t = \pm 2{\cdot}576$				$t = \pm 1{\cdot}960$	
			6 Strata			
	$P_t = 0{\cdot}9580$				$P_t = 0{\cdot}9023$	
0·956	0·951	0·948	Means	0·904	0·894	0·888
0·966	0·952	0·942	Regression coefficients	0·915	0·883	0·873
0·948	0·931	0·916	Simple correlations	0·886	0·863	0·837
0·957	0·937	—	Partial correlations	0·908	0·868	—
0·935	0·912	—	Multiple correlations	0·895	0·840	—
			12 Strata			
	$P_t = 0{\cdot}9757$				$P_t = 0{\cdot}9264$	
0·972	0·971	0·971	Means	0·922	0·920	0·919
0·973	0·968	0·966	Regression coefficients	0·934	0·916	0·912
0·955	0·944	0·933	Simple correlations	0·897	0·875	0·859
0·966	0·949	—	Partial correlations	0·912	0·888	—
0·920	0·895	—	Multiple correlations	0·850	0·813	—
			30 Strata			
	$P_t = 0{\cdot}9848$				$P_t = 0{\cdot}9407$	
0·983	0·982	0·982	Means	0·944	0·943	0·943
0·983	0·980	0·979	Regression coefficients	0·938	0·933	0·932
0·973	0·966	0·965	Simple correlations	0·911	0·902	0·898
0·955	0·946	—	Partial correlations	0·897	0·879	—
0·913	0·895	—	Multiple correlations	0·825	0·793	—
	$t = \pm 1{\cdot}645$				$t = \pm 1{\cdot}282$	
			6 Strata			
	$P_t = 0{\cdot}8489$				$P_t = 0{\cdot}7529$	
0·845	0·836	0·833	Means	0·756	0·742	0·738
0·860	0·830	0·815	Regression coefficients	0·768	0·731	0·717
0·836	0·805	0·774	Simple correlations	0·739	0·699	0·671
0·855	0·810	—	Partial correlations	0·766	0·705	—
0·823	0·780	—	Multiple correlations	0·738	0·660	—
			12 Strata			
	$P_t = 0{\cdot}8741$				$P_t = 0{\cdot}7760$	
0·870	0·866	0·865	Means	0·769	0·765	0·763
0·875	0·854	0·850	Regression coefficients	0·773	0·750	0·744
0·844	0·826	0·803	Simple correlations	0·758	0·731	0·705
0·869	0·826	—	Partial correlations	0·754	0·711	—
0·790	0·738	—	Multiple correlations	0·677	0·633	—
			30 Strata			
	$P_t = 0{\cdot}8896$				$P_t = 0{\cdot}7903$	
0·891	0·889	0·888	Means	0·789	0·786	0·784
0·890	0·884	0·884	Regression coefficients	0·789	0·779	0·778
0·862	0·847	0·836	Simple correlations	0·753	0·735	0·723
0·844	0·819	—	Partial correlations	0·753	0·725	—
0·735	0·703	—	Multiple correlations	0·638	0·595	—

(because of skewed distributions) especially for the 6-strata design. Many tables of Frankel (1971, Appendices) give results for separate variables: 8 means, 8 regressions, 12 simple, 6 partial and 2 multiple correlation coefficients.

We derive from these tables several summary conclusions useful for survey sampling.

(i) All three methods gave good results for several statistics: means, coefficients of regression and of correlation, simple and partial. The mse values have small relative biases (Table 5), and the proportions of $t(s)$ values conform well to P_t expectations (Table 6). We now have three good methods for these difficult tasks.

(ii) The relative biases and the $t(s)$ proportions improve as expected for increasing sample size, from 6 to 12 to 30 strata.

(iii) The results for coefficients of multiple correlation are poor on all criteria, and they fail to improve for larger samples. We conjecture that this pathological behaviour does not result from the complexity of the selection design, but from more basic faults of the statistic.

(iv) The *BRR* method was consistently the best when judged by the criterion we believe most significant: the closeness to expected P_t of the actual proportions of $t(s)$ values. The *BRR* performed consistently better than *JRR*, and *JRR* performed better than *TAYLOR*. The *BRR*'s better performance is particularly noticeable for simple and partial correlation coefficients, where *JRR* and *TAYLOR* are less satisfactory.

The weaker performance of *JRR* and *TAYLOR* for correlation coefficients on the $t(s)$ criterion is probably associated with the negative relative biases of the mse measures of order $-0{\cdot}10$ to $-0{\cdot}16$ for *JRR* and $-0{\cdot}16$ to $-0{\cdot}28$ for *TAYLOR*. Moderate positive biases for betas with *BRR* may explain its "conservative" high proportion of $t(s)$ values for 6 and 12 strata.

(v) The variability, measured with mean-square errors in Table 5, shows interesting and surprising results. The values generally are greater than we should expect. Also, the decrease (consistency) for larger sample size is weaker than we expected. These results contrast sharply with the much better (and, we believe, more significant) results for proportions of $t(s)$ values in Table 6. Perhaps large numerators (deviations) and denominators tend to occur jointly with strong positive correlations. This possibility deserves further investigation.

The variability is consistently lowest for *TAYLOR* and highest for *BRR*. The differences are small, and apparently have less effect than the relative biases on the closeness of $t(s)$ values.

Clear differences in variabilities appear for the five kinds of statistics. Relative variation is least for means, and consistently decreases with larger sample sizes. For regression coefficients it is much greater, but also consistent. For correlation coefficients, both simple and partial, variability is somewhat greater than for means, and decreases for larger samples are rather weak.

(vi) When judged by several criteria, none of the three methods showed up strongly and consistently better or worse. The choice among methods may depend in most cases on relative costs and simplicity, and these will vary with the situation and with the statistics. *TAYLOR* methods may be best for simple statistics like ratio means, and *BRR* and *JRR* for complex statistics like coefficients in multiple regressions.

6. COMPUTING SAMPLING ERRORS

For complex samples (row C in Fig. 1), computing sampling errors seems both necessary and difficult. These computing methods are the necessary tools for inference, because the alternatives perform poorly in many practical situations. The difficulties must be great because actual computations still occur only as rare exceptions, rather than as the normal complement they should be for probability samples. The failure to compute sampling errors is a widely known scandal among practitioners. What difficulties cause this widespread evasion of an admitted duty? The list of difficulties to be overcome can also serve as *criteria for good practical programs.*

(i) *Complexity.* Computing second-order statistics is inherently more complicated than computing the first-order statistics they serve. This problem becomes more acute for complex multivariate statistics.

(ii) *Approximations.* Computations of variance typically involve approximations, and strategy involves a choice among them for validity and utility. We compared three methods; references contain further discussions (also see Kish, 1965; Sections 6.5, 8.6, 12.11, 14.1, 14.2).

(iii) *Data input.* This appears as the most important component in machine-time, because surveys are typically large-scale, involving thousands of cases. It weighs heavily in large-scale computations for multipurpose surveys, and especially for many subclasses.

(iv) *Multipurpose.* Surveys typically concern many variables, and these require many separate computations. The input for thousands of cases is multiplied by the number of distinct survey variables.

(v) *Subclasses.* Survey statistics, and errors, are needed not only for the entire sample, but typically also for many domains. This further increases and complicates the volume of computations.

(vi) *Interface.* For computing sampling errors we often need a triangular interface involving the researcher, the sampling statistician and the computer specialist. This is expensive, but omitting a side of this triangle without adequate planning can be dangerous.

We cannot present here a comprehensive treatment of these problems. In general we believe that for complex statistics (cell C3 in Fig. 1), the strongest emphasis should be placed on dealing with complexity and with valid approximations. Here the *TAYLOR* method becomes too complex for practical work and *BRR* or *JRR* is needed. But for simpler statistics (cells C1 and C2), we think that the *TAYLOR* method offers a better approach to dealing with the last four criteria.

These approaches are incorporated in a set of computing programs we have designed and used over the years; more recently they have been made available to others. *SEPP* (Sampling Error Program Package) is a set of three programs which we have used for routine computations of sampling errors. Manuals and descriptions appear in a book with that title (Kish *et al.*, 1972), and a *SEPP* package of tape plus manuals may also be purchased. For brief descriptions see Kish (1971).

7. DISCUSSION OF ALTERNATIVES

The approximations proposed here for sampling errors should be useful for research workers involved with applications. We are concerned here chiefly with analytical statistics (B3 and especially C3 in Fig. 1) and somewhat less with subclasses means and their differences (B2 and C2). We know that we have raised more questions

than we have answered. There are important contributions to be made by both theoretical and empirical investigations; we think it preferable that they be performed jointly. We urge the importance of the task by contrasting the proposed methods with the alternatives below.

(i) To restrain analysis of data to those statistics for which mathematics provides explicit distribution theory for complex samples. That poses difficult and distant goals. Meanwhile there exist irrepressible demands for the analysis of data provided by survey technology and facilitated by computer technology.

(ii) To restrain samples to independent selections for which distribution theory is adequate. This would be wise sometimes, but often it is not practical because it would be too expensive. Furthermore, analyses of relations are often secondary to the collection of descriptive data, for which complex selections are much more efficient.

(iii) To omit computing and presenting sampling errors. This is common practice. The "first-order statistics" seem to be reasonably well behaved, and rigorous proofs of that may be obtained easier and sooner than for "second-order statistics" of variability. We believe, however, that this proposal is less acceptable to most than (iv).

(iv) To compute sampling errors with the available formulas based on independent observations. This often gives bad underestimates; our evidence will be buttressed by many others. The magnitudes of these mistakes testify to the magnitude of this problem (in ironic contrast with many research papers).

(v) To select simple replicated (interpenetrating) samples, and to compute sampling errors using simple replications or jackknife modifications. This fundamental idea has much (and many) to recommend it, and it is useful sometimes. But more often in practice it is unsatisfactory for numerical reasons. If the replications are simple and few, estimates of error are poor (perhaps worse than those of (iv)). Even averaging may not rescue them sufficiently and practitioners have been disappointed. On the other hand, many replications sacrifice stratification, simplicity and perhaps validity. Here we think that *BRR* is a better answer (Kish and Frankel, 1971).

ACKNOWLEDGEMENTS

This research was supported by Grant 3191X from the National Science Foundation. The first author was visiting the Statistics Department of the London School of Economics in 1972–73.

REFERENCES

BRILLINGER, D. R. (1964). The asymptotic behavior of Tukey's general method of setting approximate confidence limits (the jack-knife) when applied to maximum likelihood estimates. *Rev. Int. Statist. Inst.*, **3**, 202–206.

BRILLINGER, D. R. and TUKEY, J. W. (1964). *Asymptotic Variances, Moments, Cumulants, and Other Average Values.* Princeton: Memorandum.

COCHRAN, W. G. (1953, 1960). *Sampling Techniques*, 1st and 2nd eds. London and New York: John Wiley & Sons.

DEMING, W. E. (1950). *Some Theory of Sampling.* London and New York: John Wiley & Sons.

—— (1956). On simplification of sampling design through replication with equal probabilities and without stages. *J. Amer. Statist. Ass.*, **51**, 24–53.

DURBIN, J. (1958). Sampling theory for estimates based on fewer individuals than the number selected. *Bull. Int. Statist. Inst.* **36**, 113–119.

FRANKEL, M. R. (1971). *Inference from Survey Samples.* Ann Arbor: Institute for Social Research, The University of Michigan.
HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory.* New York: John Wiley & Sons.
HARTLEY, H. O. (1959). *Analytic studies of survey data.* In *Volume in Onora di Corrado Gini.* Rome: Instituto di Statistica.
KALTON, G. and BLUNDEN, R. M. (1973). Sampling errors in the British General Household Survey. *Proc. Int. Statist. Inst. Vienna.*
KISH, L. (1957). Confidence intervals in clustered samples. *Amer. Soc. Rev.*, **22**, 154–165.
—— (1961). Efficient allocation of a multipurpose sample. *Econometrica*, **29**, 363–385.
—— (1965). *Survey Sampling.* New York: John Wiley & Sons.
—— (1971). Multipurpose programs for sampling errors. *Proc. Int. Statist. Inst.*, **2**, 216–221.
—— (1973). Optima and proxima in linear sample designs. Submitted to *J. R. Statist. Soc.* A.
KISH, L. and FRANKEL, M. R. (1970). Balanced repeated replication for standard errors. *J. Amer. Statist. Ass.*, **65**, 1071–1094.
KISH, L., FRANKEL, M. R. and VAN ECK, N. (1972). *SEPP: Sampling Error Program Package.* Ann Arbor: Institute for Social Research, The University of Michigan.
KISH, L., NAMBOODIRI, N. K. and PILLAI, R. K. (1962). The ratio bias in surveys. *J. Amer. Statist. Ass.*, **57**, 863–876.
MAHALANOBIS, P. C. (1944). On large-scale sample surveys. *Phil. Trans. R. Soc.*, **B231**, 329–451.
MCCARTHY, P. J. (1966). *Replication: an Approach to the Analysis of Data from Complex Surveys.* Washington: National Center for Health Statistics, Series 2, No. 14.
NATHAN, G. (1972). On asymptotic power of tests for independence in contingency tables from complex stratified samples. *J. Amer. Statist. Ass.*, **67**, 917–920.
—— (1973). Tests of independence in contingency tables from complex samples. *Proc. Int. Ass. Survey Statist.*
NEYMAN, J. (1934). On two different aspects of the representative method. *J. R. Statist. Soc.*, **97**, 558–625.
RAO, C. R. (1971). Inference in sampling from finite populations. In *Foundations of Statistical Inference* (Godambe, P. V. and Spratt, D. A. eds), p. 177. Toronto: Holt, Rinehart & Winston.
SEDRANSK, J. (1965). Designing some multi-factor studies. *J. Amer. Statist. Ass.*, **62**, 1121–1139.
"STUDENT" (1908). The probable error of the mean. *Biometrika*, **6**, 1–25.
—— (1909). The distribution of means of samples which are not drawn at random. *Biometrika*, **7**, 210–215.
SUKHATME, P. V. (1954). *Sampling Theory of Surveys with Application.* Ames, Iowa: Iowa State University Press.
TCHUPROW, A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, **2**, 646–680.
TEPPING, B. J. (1968). The estimation of variance in complex surveys. *Proc. Soc. Statist. Section Amer. Statist. Ass.*
THARAKAN, T. C. (1969). *Inference Based on Complex Samples from Finite Populations.* Ph.D. Thesis, The University of Michigan.
WOODRUFF, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *J. Amer. Statist. Ass.*, **66**, 411–414.
YATES, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *J. R. Statist. Soc.*, **109**, 12–43.
—— (1949, 1953, 1960). *Sampling Methods for Censuses and Surveys*, 1st, 2nd and 3rd eds. London: Charles Griffin & Co.

DISCUSSION OF THE PAPER BY PROFESSOR KISH AND DR FRANKEL

Professor G. KALTON (University of Southampton): It is a great pleasure for me to propose the vote of thanks to Professor Kish and Dr Frankel for a very stimulating paper, and also to see both of them here to present it. The paper usefully brings the results of recent research in survey sampling into a coherent framework, and in doing so it draws attention to the gaps in our knowledge. The framework also enables conjectures to be made about the nature of the unavailable results, and such conjectures appear in various places throughout the paper. Besides providing valuable guidance to survey researchers,