ON ANALYTICAL STATISTICS FROM COMPLEX SAMPLES[1]

Leslie Kish[2]

I want to plead the case that an important and urgent task facing mathematical statistics consists of providing useful expressions for analytical statistics for complex sample designs. I should like to describe these problems to mathematical statisticians who should find them interesting because they meet the criteria of all good problems: they are important, unsolved and solvable.

The most important and difficult problems of survey sampling still await adequate mathematical treatment: the textbooks are aimed almost entirely at producing good estimates of aggregates, means and ratio means. One may also deal with the differences of two of these, but there is only fleeting and occasional reference to this problem. However, with that we come to the end of the statistical tools available for complex samples.

As sampling theory developed, probability sampling has been capturing the field of respectable sampling practice with sample designs, which are often simultaneously economical and complex. One result has been an increasing volume of sample survey data which is of high quality and which researchers wish to put to more involved analytical use. But the mathematical statistics for doing this validly are lacking. The available analytical statistics assume independence among the selected elements: but this independence is lacking in complex sample designs. Thus the researcher may be forced to forego the analysis which he considers desirable and valuable. But if he is too impatient or too ignorant for that act of self-denial, he may go ahead and use the *srs* formulas he finds in books on statistics, which often result in very serious errors.

I hope that mathematical statisticians will be impressed with the importance of the unsolved problems of analytical statistics for data arising from complex sample designs. The lack of these is a more frequent source of gross mistakes than any other kind of departure from the usual assumptions.

[1] Overview talk for the symposium.

[2] Leslie Kish, Institute for Social Research, The Univeristy of Michigan.

These problems are important, unsolved and interesting. You may ask: are they solvable now? Supporting my affirmative answer are three sources of justification. First, we observe the great recent advances in statistical theory. Secondly, the rapid increases in the quantity and quality of electronic computing machines make the time ripe for the solution of some of these problems. There is new interest in a general method which holds promise of rapid advance toward useful approximations. At the Survey Research Center we are now introducing this method for computing estimates of variances for regression coefficients and other statistics for which formulas are not now available.

It seems to me that this procedure resembles that of Alexander when he "solved" the Gordian knot. From a theoretical viewpoint I don't know whether it constitutes a solution of the problem or its avoidance. But insofar as it promises to give good approximations for much needed variances the practicing statistician will welcome its development with enthusiasm and interest. In this way one may obtain estimates of the confidence intervals of some analytical statistics for which specific formulas are not now available.

All of the above is verbatim from my talk to a joint session of the American Statistical Association and the Institute of Mathematical Statistics in 1957. Since then our situation has changed but little. Our 1957 hopes for that cut of the Gordian knot is now much used as BRR or Balanced Repeated Replications (Kish and Frankel 1970, 1974). But my moving plea for distribution theory for doubly complex analytical statistics did not move the mathematical statisticians. I know now why not, since I am sadder and wiser now. First, statisticians like other scientists work not on what solutions are needed but on those that seem feasible at the time. (Like nuclear bombs, for example.) Second, distribution theory for complicated statistics for complex samples seems too difficult to solve. Third, the solutions would have too many parameters to be useful. Thus my views in (Kish 1978) and today are more sober: "New computational methods can give us approximate variances that appear satisfactory for practical purposes. However, it would be more satisfying to have mathematical distribution theory for analytical statistics (e.g. regression coefficients) without the assumptions of independence, but with complex correlations between sample observations. We may hope for some progress, but not for generally useful results, because of mathematical complexities, and even more because the numbers of needed parameters will prove too great for practical utility."

Here follow seven important points about complex samples put boldly. They are not all widely known or believed, but I ask you to know, believe, use and teach them, as I do.

1. The effects of complex designs must be considered *separately* for point estimates and for probability statements, like confidence intervals or tests of hypotheses. For point estimates we have for all sample designs consistent approaches to parameters from similar probability-weighted (H-T) estimators. But the probability statements like confidence intervals are highly subject to design effects, especially in cluster sampling.

a) "Statistics (means, regression coefficient, etc.) approach their population values as the sample size increases.

b) The approach is generally slowed by design effects.

c) The design effects differ for different statistics, for different variables and different sample designs." (Kish and Frankel, 1974).

That paper also presents the most convincing evidence for these points: and evidence is widespread: e.g. (Verma et al 1980). Nevertheless two famous statisticians completely misstated our position in discussions of our paper: "Here the authors make the important observation that the confidence interval statements for the unknown parameter are numerically *not* much affected by the *lack of independence* of observations introduced by complex survey techniques such as stratified cluster sampling." Alas, that mistake gets quoted by other theoreticians who fail to read our answer of survey samplers: "They misunderstand completely our principal and repeated message: that confidence interval statements are numerically greatly affected by the lack of independence of observations introduced by complex survey techniques such as stratified cluster sampling." (Kish and Frankel, 1974).

This misunderstanding shared by naive non-statisticians with sampling theorists causes troubles for us survey samplers: hence we are working on a clearer statement.

2. Do we need sampling errors for analytical statistics for data from complex surveys? Or have a few of us been devoted to a negligible even trivial problem? I feel like a St. Sebastian, the target practice for the

slings and arrows of diverse outrageous heathen. (Mixed metaphors are better than fixed or random.) First come the market researchers and pollsters who ignore us, though some have learned to put a $\sqrt{\ }$ between a 2 and (pq/n). Second, some demographers write that with their large samples and larger measurement errors they have no time for sampling errors. Third come the mathematical psychologists, econometricians and biometricians who take their linear models straight from mathematical statistics, and that hurts. Fourth, even more hurtful are the mathematical statisticians themselves, who either forget that their n's do not justify their means, or they invoke IID, or they use some Bayesian exorcism against the spirits of the sample design. Fifth and worst are sampling theorists who display theorems to prove that, with completely specified models of arbitrary superpopulations, we need not worry about whence or how our elements were selected, nor weight them for unequal selection probabilities. They even convince a few survey samplers that they can dwell on some Olympus with their models and not come down to earth where the population lives.

From these necessarily brief remarks you notice that I am an extremist for several reasons: a) Design effects for analytical statistics provide common evidence for imperfectly specified models for the best stratified samples; b) We frequently find the effects of selection weights on samples; c) Relations between predictor and predictand variables exist in actual individuals, and they in real populations, and these interact with sample designs. (I am developing these points in a book on Statistical Design for Social Research for Wiley, 1985.)

My philosophy is consistent, but in practice I am less dogmatic. I recognize that in practice: a) it is never possible to cover completely our target populations, hence we must always resort to models for inference; b) probability sampling is too costly and not feasible for most experiments; c) despite lack of randomization either in selection or in treatments, we often blunder our way to reliable results with care, replication, design, additivity and a little bit of luck.

3. Analytical statistics begin with subclasses and with their comparisons. In the last three decades much useful material has been published about variances and design effects for subclasses. There are masses of empirical results and several useful guiding rules based on them (Kish 1980, Kish et

al. 1976, Verma et al. 1980), also some recent theory (Rust 1984, Chapter 6).

a) Distinguish between proper domains and the more common crossclasses, on which we focus here.

b) Selection probabilities are preserved for crossclasses but sample sizes become highly variable.

c) Estimates of totals and means from complex samples are retained in ratio and conditional forms.

d) Design effects for crossclasses tend to approach to almost 1 proportionately as the subclass sizes per primary cluster approach 1. This approximate model needs care and qualification but it is preferable to all venerable alternatives about design effects: that it is simply 1, or some other constant, or the same as for the entire sample. The pooled model may be often better than separate and highly variable computations.

4. Comparisons of paired means tend to have design effects greater than 1 but considerably less than the sum of the two variances. These reductions due to positive covariances (hence to a kind of additivity) have been found widely and regularly for comparisons both of crossclasses and of periodic surveys (Kish 1965, 14.1, also the above).

5. For complex analytical statistics several methods exploit the potentialities of electronic computing: Taylor linearized (delta) methods, including machine differentiation, Balanced Repeated Replications and Jackknife Repeated Replications, all have been shown to yield useful estimates of variance and design effects for complex samples (Kish and Frankel 1970 and 1974; Woodruff and Causey 1978), Bootstrapping may also be added in the future (Rao 1984).

Analytical statistics consistently show design effects greater than 1, significantly greater in every sense, but also lower than design effects for means. The relations of design effects between diverse coefficients and comparisons with those for means show some regularities.

For useful guidance we need not only more empirical work but also more results from sampling theory and model building. I am disappointed frankly that since our early work we have not seen more publications in theory and models that would be directly useful for guiding inference for actual data. The empirical bases of design effects are necessary, but to satisfy our intellectual needs for understanding we need more theory and better models.

Furthermore, even our practical needs remain unsatisfied with merely empirical design effects, because they are functions jointly of the variables, of the type of estimates, of the sample design used and of the population basis for the data. That four-dimensional source of variation is too complex and we need theory to construct models for greater simplicity.

6. Categorical data analysis is an important area, rapidly developing, and several contributions have been made to apply these methods to complex survey data (Fay 1982; Landis et al. 1982: Koch et al. 1975). These also have implications for analysis of variance where some of the earliest models were started, but not followed (Kempthorne and Wilk, 1955; Tukey and Cornfield).

7. As for the future I am hopeful about contributions from theory to applications but for two exceptions. 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. They will however guide us toward better and more comprehensive inference. Also toward better utilization and presentation of analytical statistics from complex surveys.

REFERENCES

[1] Fay, R. (1982). Contingency table analysis for complex sample designs: CPLX. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 44-53.

[2] Kish, L. (1957). Some unsolved problems of complex samples. Paper for Joint Meeting of the American Statistical Association and Institute for Mathematical Statistics.

[3] Kish, L., and Frankel, M.R. (1970). Balanced repeated replications for standard errors. JASA, 65, pp. 1071-94.

[4] Kish, L., and Frankel, M.R. (1974). Inference from complex samples. JRSS (B), 36, pp. 1-74.

[5] Kish, L. (1980). Design and estimation for domains. *The Statistician* (London), 29, pp. 209-22.

[6] Kish, L., Groves R.M., and Krotki (1976). *Sampling errors for fertility surveys.* Occasional paper 17, London: World Fertility Surveys, 61 pages.

[7] Koch, G., Freeman, D., and Freeman, J. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, pp. 53-59.

[8] Landis, J.R., Lepkowski, J., Eklund, S., and Stehouwer, S. (1982). A statistical methodology for analyzing data from a complex sample survey. *Vital and Health Statistics*, Series 2 - No. 92. DHHS Publ. No. 82-1366. Public Health Service, Washington, U.S. Government Printig Office.

[9] Rao, J.N.K. (1984). Bootstrap inference with stratified samples. (Submitted for Publication).

[10] Rust, K.F. (1984). *Techniques for Estimating Variances for Sample Surveys.* The University of Michigan, Ph.D. dissertation.

[11] Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *JRSS* (a) 143, pp. 431-73.