INTERNATIONAL STATISTICAL INSTITUTE

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

THE SURVEY STATISTICIAN

n° 17

SURVEY STATISTICIAN Nº 17

EDITORIAL

The last issue of Survey Statistician has been criticized. Because of an error, the chapter titles were not printed. In addition to being aesthetically displeasing, this may have caused problems for readers. The secretariat apologizes to the members for the error.

Some Anglophone members felt that the English translations of articles and comments written in French were of questionable quality. Dr FELLEGI, the President, has proposed that the task of French-to-English translation be given to Statistics Canada's translation service. We should like to thank him, we trust that, in future, everyone will be satisfied with Survey Statistician, whatever the language.

Moreover, the secretariat has examined the possibility of increasing the size of the print in Survey Statistician without significantly increasing the volume of the publication (and, therefore, the shipping costs). Many members are of the opinion that, in its present form, our journal can be read only by those with good eyesight. Once again, we hope that we have found a solution that will please everyone.

1. NEWS FROM THE ASSOCIATION

1.1. Tokyo Session

The following IASS meetings will be organised during the Tokyo session :

- Council meetings(September 10 th and 15 th)
- General Assembly (September 14 th)
- IASS Programme Committee (September 9 th)

Meetings of the other IASS Committees and of the Local Representatives with also be held, but the dates are not fixed yet.

1.2. Subscriptions to Survey Methodology

More than 100 members of the IASS have subscribed to the journal "Survey Methodology/Techniques d'Enquêtes", which is now proposed to the members at a special low rate by Statistics Canada. This is already a good success. The President and the Secretariate hope that more subscriptions will be received in the future and that this review, which is probably the only one dealing specially with surveys, will be a strong means of information for all the members of the IASS.

1.3. IASS Reduced Subscription Rate for International Statistical Review

All members of the International Association of Survey Statisticians are cordially reminded of their entitlement to subscribe to the ISI journal"International statistical Review." The cost to IASS members for a 1987 subscription is therefore only US $18 or Dutch Guilders 45 or UK£LL or French Francs 120 or Deutsche Marks 37.50 or Swiss Francs 29.50 (members can subscribe in whichever currency they wish).

Recent articles which may be of particular interest to IASS members include :

The machine editing of large sample surveys : The experience of the World Fertility Survey, T.W. Pullum, T. Harpham, N. Ozsever.

Parameters of superpopulation and survey population, V.P. Godambe and M.E. Thompson.

Survey non-response adjustments for estimates of means, R.J.A. Little.

Post-stratification in ratio estimation : Usages of auxiliary information in survey sampling and opinion polls, P. Jagers, A. Odén, L. Trulsson.

To place your order for the 1987 volume, simply forward your request and the required payment to the ISI Permanent Office. Your order will be immediately processed and copies despatched.

IASS orders at a reduced rate should be addressed to International Statistical Review (IASS Reductions), c/o ISI Permanent Office, 428 Prinses Beatrixlaan, P.O. Box 950, 2270 AZ Voorburg, Netherlands.

2. ANNOUNCEMENTS

2.1. Census Bureau : Fourth Annual Research Conference

The Bureau of the Census is planning its Fourth Annual Research Conference, to be held in March of 1988 in the Washington, DC area. As with ARC III, the conference will consist primarily of contributed papers, most of which will be reviewed in formal discussion at the conference. The conference will emphasize improvements in technology related to census and survey operations, but also seeks papers on other topics related to the broader range of Census Bureau research interests. Papers may address methodology, empirical studies, or relevant issues. Conference proceedings containing all papers and discussions will be published. Papers must be original and not previously published. Presenters will be reimbursed for transportation and per diem expenses and will receive a fee for manuscript preparation (expected range : $250-$450).

Topic areas include :

. new survey techniques
. census and survey automation
. automated cartography
. Processing longitudinal survey data bases
. methods and software for survey data analysis
. effects of data collection methodology on data quality
. mandatory versus voluntary compliance and other factors affecting survey reponse
. census undercount and adjustment
. microeconomic panel studies
. economic-geographic modeling
. use of economic models for imputation and related problems
. measurement problems with foreign trade statistics
. classification of economic activity (e.g., SIC code revision)
. other topics related to census or survey methodology

To submit a paper for consideration, send a 500-word abstract by June 1, 1987 to :

David F. Findley
Conference Chairman
Statistical Research Division
Bureau of the Census
Washington, DC 20233

To obtain registration information or to be icluded on the mailing list, contact :

Carolee Bush
Conference Coordinator
Office of the Director
Bureau of the Census
Washington, DC 20233

Please note, plans for ARC IV are dependent upon final approval and funding which are still pending.

2.2. International telephone survey methodology conference (2nd announcement)

We recall that a conference, sponsored by the American Statistical Association (ASA) and IASS, on telephone surveys will take place in Charlotte (North Carolina, USA) in november, 1987.

The main topics will be :

- coverage of the household population
- sample designs
- administrative organization of data collection
- non responses
- measurement errors properties
- computer assisted telephone interviewing (CATI) systems.

For more details, please contact :

Robert GROVES
Survey Research Center
University of Michigan
Ann Arbor - MI 48109 - 1248
U.S.A.

2.3. More precisions on the symposium on statistical of Administrative Data (Statistics Canada)

For the fourth consecutive year, Statistics Canada is sponsoring an international symposium. This year the symposium, on the Statistical Uses of Administrative Data, seeks to bring together statisticians from the public and private sectors and the academic community for discussions on methodological issues arising out of the increased use of administrative data for statistical purposes.

Topics to be addressed at the symposium include policy issues and organizational experience (privacy, confidentiality and related legislation), data content and quality, record linkage and integrated approaches towards data developments. Both invited and contributed paper sessions are planned, followed by a panel discussion at the end. Presentations will be made by experts from Canada, the United States and three or four European countries.

The symposium will be held November 23-24, 1987, in the Main Hall of the Government Conferences Centre, Ottawa.

Those interested in contributing papers should submit an abstract as early as possible. A complimentary copy of the published proceedings will be included in the registration fee of C$80 or US$60.

For submission of abstracts, registration or further information please write to :

Dr. M.P. Singh
Methodology Branch, Statistics Canada
4th Floor C-2, Jean Talon Building
Ottawa, Ontario
Canada K1A 0T6
Tel. : (613) 990-9894

2.4. Statistical training at Bureau of the Census

Each year, the Bureau of the Census offers a program of statistical training, which is managed and organized by the International Statistical Programs Center. The training takes place over a year and comprises seven parallel courses, for which English is the language of instruction. The training periods for 1987-1988 are as follows :

1 - AUGUST 31 to OCTOBER 2, 1987
2 - OCTOBER 19 to DECEMBER 11, 1987
3 - JANUARY 4 to MARCH 11, 1988
4 - APRIL 4 to JUNE 10, 1988
5 - JUNE 20 to JULY 29, 1988

It is strongly recommended that registrants opt for a full year of training. However, partial training may be considered if need be. In particular, late registration implies that the registrant already has the training and experience that the missed courses would provide.

For further information on the objectives, schedules, diplomas offered, levels reguired and procedures to be followed, please contact the local representative of the United Nations or the FAO or write directly to :

Chief
Training Branch
ISPC
US Bureau of the Census
WASHINGTON, DC 20233 (USA)

2.5. Canadian Social Trends

The need for one report publishing careful enquiries into contemporary social issues, using data drawn from many sources, has grown remarkably in recent years. Canadian Social Trends has been created to meet that need.

Every quarter, Canadian Social Trends will analyse important issues and the influence they will have on general social conditions. In-depth analyses, by focussing on one specific topic and its related issues, will illuminate the forces that are altering our society. Probing feature articles will address such topics as the family, the work place, income and lifestyles.

Canadian Social Trends will also regularly publish trend reports outlining the direction in which Canadian society is heading ; articles about how to access electronic data sources in both the public and private sectors ; and brief summary results of Statistics Canada's social surveys.

In a journal-style report about 40 pages long, and with plenty of graphs and charts, Canadian Social Trends will discuss topical issues using a variety of formats.

Feature articles provide careful analysis, offering you - whether you're a business or academic researcher - a basis from which to expand your own studies.

Key social indicators covering 40 topics such as marriages, migration, housing starts and incomes provide important data, allowing you to take in current socio-economic conditions at a glance.

Trend reports pinpoint the strength and direction of emerging trends, giving you the insight you need to make the best business and policy decisions.

How-to articles about accessing public sector data sources describe where to find and how to retrieve electronically-stored data, opening new avenues of exploration and research to you.

Results of Statistics Canada surveys summarize the most recent statistical studies of social conditions, helping you to stay one research jump ahead of the competition.

Reviews of new books and publications report ongoing research, keeping you up-to-date with information available in a variety of social fields.

In future issues, we'll explore the stories behind the numbers on :

- changing family incomes
- native housing
- bilingualism
- pensions and retirement savings
- university grads in the job market
- the explosion of part-time work.

2.6. Kendall's Advanced Theory of Statistics : fifth edition of volume 1

In 1958, Sir Maurice Kendall wrote, in his Preface to the first three-volume edition of this work by Professor Alan Stuart and himself : "This book, in its original form, was planned before the outbreak of the Second World War and... since that time there have been many advances in statistical theory and an expansion in the use of statistical methods which is striking even by comparison with the development of scientific work generally... Owing to the growth of the subject, two further volumes are projected... nothing smaller would have covered the field after the manner of my original design".

Nearly thirty years later, the "advances in statistical theory" and "the growth of the subject" have continued without remission, and the three-volume edition has gone through four revised editions, the last beginning with volume one in 1977 and concluding with volume three in 1983, for which Professor Keith Ord joined the authors. Sir Maurice Kendall died soon after its publication.

It is, therefore, nearly ten years since the publication of the last revision of volume one, and the present authors, Professor Alan Stuart and Professor Keith Ord, have felt it desirable to extend their latest revision of the work into a projected four volumes, of which this is the first and is published in March 1987. Volume two will appear in 1989, and volumes three and four in 1991 and 1993 respectively.

It can be seen from this chronology alone that THE ADVANCED THEORY OF STATISTICS is the major sourcebook of statistical theory now in existence. The maintenance of its currency is an ongoing task of great complexity, requiring constant surveillance of the field and its professional literature on a scale which would make the creation ab initio of the work, had it not already existed, impossible today.

THE ADVANCED THEORY OF STATISTICS is unquestionably unique, and this new four-volume edition by Professors Stuart and Ord represents a complete reappraisal of the earlier work.

The entire text has been revised, word by word, and proofs modernised. There are many new examples in the text, and a new index to locate the examples, and many new exercices at the ends of chapters.

1 - Frequency Distributions
2 - Measures of Location and Dispersion
3 - Moments and Cumulants
4 - Characteristic Functions
5 - Standard Distributions
6 - Systems of Distributions
7 - The Theory of Probability
8 - Probability and Statistical Inference
9 - Random Sampling

10 - Standard Errors
11 - Exact Sampling Distributions
12 - Cumulants of Sampling Distributions (1)
13 - Cumulants of Sampling Distributions (2)
14 - Order-Statistics

3. LOCAL REPRESENTATIVES
Unified System of Household Surveys in Hungary

In Hungary the Unified Sys tem of Household Surveys /USHS/ was established in 1976 to meet the increasing demand for information on socioeconomic developments (1). USHS integrated the different, earlier separate surveys into an organizationally and methodologically unified frame with a longterm programme.

At the beginning the USHS was characterized by the seeking of methodological ways and means. On the basis of experiences gained in the first years essential changes took place in the system in 1983. USHS now consists of two subsystems and accordingly two independent, similarly designed subsamples / H and M/ with a unified network of interviewers.

The household budget survey goes back to 1949. Since 1983 it is carried out in every uneven year on the H subsemple which comprises 12 000 households.

---

(1) In addition to the USHS there are a few population surveys based on special samples. Probability sampling is also used in agricultural statistics, in contrast to the data of sociodemographic developments and economic statistics which come for the most part -- or rather exclusively in the latter case -- from complete or almost complete enumeration.

The other socio-statistical surveys are carried out in the even years on the M subsample covering 16 000 households. In this frame the survey of Basic Socio-demographic Data provides the most important demographic, socio-economic information, the so called core data.

The main objective of the surveys is the cross sectional analysis for the given period. Changes in time are followed up by regularly repeating surveys of 2-5-10 years periodicity on the aggregated level of major social-demographic groups.

The periodicity of surveys :

Household budget survey	2 years
Basic socio-demographic data survey	3 years
Income survey	5 years
Time budget survey	10 years
Social mobility and prestige survey	10 years

The nonrecurring or irregularly repeating surveys cover a great variety of topics : life and work conditions, housing, cultural habits, state of health, social conditions of women, unregistered incomes /second economy/, travelling and recreational habits, etc.

Organization : USHS surveys are carried out by the regional /county/ agencies of the CSO. There are 430 interviewers who are supervised by 120 specialized staff members at the CSO regional agencies.

There is a central section which is responsable for the coordination and evaluation of USHS surveys and also for the development and promotion of suitable survey techniques and estimation methods. The subject matter statisticians are responsible for specific proposals, considerations and also for data analysis and dissemination of results.

There are quite a few methodological problems, arised during recent years : how to improve questionnaire design, the performance of interviewers, how to estimate the sampling and nonsampling error, how to make estimations for small areas or domains. In the following we briefly expose only one point, the design of the present general purpose sample, its advantages and disadvantages.

The samples or master sample is selected in two or three stages : the primary sampling units /PSU's/ are enumeration districts of the 1980 census. The ultimate sampling units are dwellings in all surveys.

The selection of enumeration districts constituting the sampling frame was carried out by stratification in the following way :

- all settlements with 25 000 or more inhabitants were included in the sample. Smaller settlements were grouped by size and selected in a way that for those with relatively more cooperative peasants the sampling fraction was higher than the average, and for those where the proportion of the old and inactive population was relatively high the sampling fraction was lower than the average ;

- in order to have more smaller settlements in the master sample, the sampling fraction in Budapest and in the large cities was lower than in other settlements.

Since the selection of enumeration districts is carried out with unequal probabilities, sample estimates need a weighting procedure. There are 270 different strata in the sample, and, consequently, 270 different weighting factors. The weighting factor for a given stratum is the ratio of the updated total number of dwellings and the number of dwellings in the stratum observed in a survey.

The sample is a critical point of the USHS. The sample size is fixed. One third of the selected dwellings are replaced yearly. The selected families are visited many times during the survey period. An interview takes sometimes one hour or even more time, which leads inevitably to diminishing tolerance on the respondents' side, and, consequently, to increasing response effect or even to possible refusals. The only feasible solution to this problem is a flexible system of subsamples which allocates the annual survey program among more respondents.

We suggest that the /sub/sample M is to be divided into two independent subsamples M_1 and M_2 of the same size (2). The surveys could then be distributed between the two parts, taking - if necessary- the possibility of matching also into account. Clearly it is useful to split M on the level of dwellings, since in that case the same enumeration districts are included both in M_1 and in M_2. If a survey needs a larger sample than M_1 /or M_2/, then additional dwellings used on one occasion only are to be selected. In this way a system of partially overlapping samples could be defined with occasional distribution in time. The flexible sampling design allows to ensure the desired level of reliability.

The principles of the new USHS sampling design -for the decade 1993-2002 - will have to be worked out in the near future. It is very likely that the sampling frame will consist of the census enumeration districts and not of the national population register. One of the important methodological questions to be settled is to check if the present system of unequal probabilites of selection is good enough or not in which case they have to be modified.

Adam Marton

---

(2) For the household budget survey the subsample H has its own problems and solutions which is not discussed here.

4. PAPERS ABSTRACTS

4.1. Title : Regression Analysis and Ratio Analysis for Domains : A Randomization - Theory Approach

by

Author : E. Elvers, C.E. Särndal, J.H. Wretman and G. Örnberg

Journal : The Canadian Journal of Statistics, vol. 13, no. 2, 1985, pages 185-199.

SUMMARY

The methods presented in this paper are of interest in analytic studies involving comparisons of (small) domains of a finite population from which a probability sample is drawn.

Consider a population of, say, N households, divided (on a geographical basis, say) into D subpopulations (domains). A probability sample is drawn from the total finite population. The number of sampling units falling in a given domain is random and perhaps small (especially if the domain is small).

We want to estimate, domain by domain, analytic parameters such as

$$R_d = \frac{\sum_{k \in U_d} y_k}{\sum_{k \in U_d} x_k} \quad \text{and} \quad B_d = \frac{\sum_{k \in U_d} (y_k - \bar{y}_d)(x_k - \bar{x}_d)}{\sum_{k \in U_d} (x_k - \bar{x}_d)^2}$$

Where U_d is the d : th domain ; $d = 1, \ldots, D$.

If x_k is household income, and y_k is household savings for the k : th household, both R_d and B_d can be viewed as measures of savings per unit of income for the d : th domain ; B_d is, of course, the standard regression coefficient. It is of interest to compare how B_d, for example, varies over the domains $d = 1, \ldots, D$. Thus estimates $\hat{B}_1, \ldots, \hat{B}_D$ are required.

When a domain happens to contain very few observations, it is highly desirable to strengthen the standard sample-weighted estimators of R_d and B_d by incorporating available auxiliary variables in the estimators $\hat{R}_d$ and $\hat{B}_d$.

Assuming that there are one or more auxiliary variables (z-variables) about which information is available (from census or other sources) beyond the set of sample units, we construct estimators of R_d and B_d (and, more generally, of multiple regression coefficients) that depend explicitly on the z-variables.

The paper shows that substantial efficiency gains can be realized from the presence of the z-variables. It is shown that the conditions under which the gains are particularly high have to do with the capacity of the z-variables to "explain" patterns in the residuals, $y_k - R_d x_k$ in the case of R_d and $y_k - \bar{y}_d - B_d (x_k - \bar{x}_d)$ in the case of B_d.

4.2. Title : Efficiency of Combining two complex sample Surveys

Author : Donna Brogan and Richard Porter, Emory University

ABSTRACT

In planning the 1983 Géorgia High Blood Pressure Survey, a choice of two sample designs was available. One design, referred to as the single design, was to allocate a total of 3000 sample housing units (SHU's) so as to achieve an equal probability sample of HU's throughout Georgia. In the second design, referred to as the dual design, an additional 960 SHU's could be added to the 3000 SHU's, but these additional SHU's needed to be blocated in six specific small counties in south Georgia. Then, the other 3000 SHU's would be allocated to the remainder of Georgia in an equal probability sample.

Although it might seem that increasing the total sample size by almost one-third,from 3000 SHU's to 3960 SHU's, would substantially increase the precision of estimates for Georgia, the increase was expected to be counterbalanced to some extent because of widely differential weights inherent in the dual design. Theoretical comparisons of the two potential designs indicated that the precision for statewide estimates would be about the same under each design. The dual design was used to do the survey. Statistical analysis of the actual survey data from the dual design were in agreement with the theoretical comparisons done prior to designing the survey.

4.3. Title : Effects of Rotation Group Bias on Estimation of Unemployment
Author : Gary Solon
Journal : Journal of Business and Economic Statistics. 4, 105-109

Summary for IASS

Many repeated surveys are designed so that respondents are retained in the sample for several occasions and then rotated out of the sample to be replaced by a fresh group of respondents. On each occasion a proportion (say 20 %) are in the survey for the first time, a similar proportion for the second time and so on. These groups are called rotation groups. There is plenty of evidence that responses vary with the length of time that respondents have been in the sample so that, for example, those responding for the first time give systematically different responses to those reponding for the second or third time. This systematic difference is called rotation group bias.

Previous analysis of rotation group bias in the Current Population Survey of the United States has concluded that if the biases are additive, the ratio and composite estimators of month-to-month change are unbiased. This article shows that if the biases contain a multiplicative aspect, both estimators of change are biased, and the composite estimator's bias depends on the entire history of the true series. Empirical tests produce evidence that the size of discrepancies between the unemployment reported by certain rotation groups and that reported by the full sample is related to the level of unemployment. This finding casts serious doubt on the additive model of rotation group bias.

Taken together, the theoretical and empirical results suggest that rotation group bias may lead to some bias in the estimate of change as well as the level of unemployment in the Current Population Survey. Moreover the form this bias may take in the case of the composite estimator raises the question of whether this estimator is preferable to the ratio estimator.

5. QUESTIONS/ANSWERS

Conducted by Leslie Kish. Please send Questions to him (ISR - The U of Michigan, Ann Arbor, 48106, USA), or, to IASS, Paris. Please indicate wheter or not you want your name given with the question. This has become an open forum, and we shall gladly print (after refereeing) additions, modifications, discussions of past published answers. Contributors to answers will be acknowledged with their agreement.

SS 17,No. 1

Question : In the Q/A of S.S. 16 on "Sampling errors" you described the design effect = $deft^2$ = actuel variance/srs variance, with s^2/n for srs variance of the sample mean.

You also referred to a synthetic roh = $(deft^2 - 1)/(\bar{b} - 1)$, where $\bar{b} = n/a$, the average sample size per primary selection. How do we apply these forms to our samples which are heavily weighted to counter unequal selection probabilities, with the weights $K_j \propto 1/p_j$?

These unequal P_j were used to obtain appproximately equal sample sizes from unequal clusters. Or are $deft^2$ and roh only applicable to essentially epsem and self-weighted samples ?

Answer : You direct our attention to a problem that has been mostly overlooked or avoided, because it is difficult. But it is common and it can be serious, hence it deserves a brief and simple treatment promptly. My treatment may be incomplete and imperfect and we may hope hear from others in later issues. Let us note at the outset that both $deft^2$ and roh are designed to yield rough, simple, single statistics for complex designs. They may cover several parameters, such as unequal population cluster sizes B_i and sample sizes b_i, also diverse selection methods for several stages involving stratification. Yet we may use these single, simple statistics like $deft^2$ and roh in order to compute and present them for many survey statistics and variables, because the variations between variables may be vastly greater than the neglected finer points. But should the effects of weighting be added to the complexities already included in these summary measures ?

I propose that we distinguish between internal (immediate, narrow) uses and external uses of design effects. The design effect $deft^2$ is the ratio the actual variance to the variance of an srs of the same sample size n, hence it is adequate for internal uses to measure the distortion of intervals and of tests of significance caused by naive formulas, like s^2/n. For these purposes the single overall statistic $deft^2$, which includes the effects of weighting along with these of complex clustered selections, is useful. It yields the combined effects on variances of all complexities in the designs, but that $deft^2$ is relevant only to the specific statistics (e.g. variance of a specific mean) for which it was computed.

However, it is mistaken and misleading to make external uses of those values of $deft^2$ for other statistics to which they are not relevant. Those mistakes are very common but they can be avoided with the cautions mentioned below. It is particularly misleading to attribute all $deft^2$ mechanically to clustering in the presence of the effects of weighting.

The basic strict definition $Deft^2 = 1 + \rho(B - 1)$ holds strictly for equal sizes (B) clusters [e.g Kisk 1965, 6.5B]. The concept has been stretched to apply to $deft^2 = 1 + roh(\bar{b} - 1)$ for average sizes $\bar{b}$ of unequal subsamples from primary selections (ultimate cluster) of epsem selections (equal overall selection rates f) [Kish 1955, 8.2].

This approximation $roh = (deft^2 - 1)/(\bar{b} - 1)$ applied well to means $\bar{y} = y/n$ of epsem samples from unequal clusters : This roh is a close estimate of p, a good measure of the effects of clustering, and fairly robust for changes in the average cluster size $\bar{b}$. But this formula does not hold well generally for the simple expansion or "unbiased estimator" $\hat{Y} = y/f$ of the total Y. The factor $\hat{V}^2/V^2$ has been introduced for this purpose by Hansen, Hurwitz and Madow (Vol. I,6.8) and they write [p. 608] "Table D-13 shows the ratios $\hat{V}^2_x / V_x^2$, the factor by which the relvariance of the unbiased estimate is increased... because of sampling psu's that vary in size. For the ratio estimates (means), the factors $\hat{V}^2 / V^2$ are all very nearly equal to 1, because the ratio estimate tends to remove the effects of the variation in size of the psu... with almost all the factors in the range. 99-1.03". Since the "unbiased estimator" $\hat{Y}$ is seldom used in practice (as against theory), we shall corcentrate on the ratio mean $r = \bar{y}$, and on separating the effects of weighting from those of clustering. Rather than foring a single answer, let us consider several alternatives (1 to 5).

1. Computing $deft^2 = var(r)/(s_y^2/n)$ does not describe accurately the effects of clustering alone, because this ratio also includes the effects of weighting. A basic alternative to s_y^2/n is to assume "on unrestricted random sample of n elements from the same population and then give different weights k_j to randomly selected subsets of elements, so that" the proportion with weights k_j is the same as in the actual sample [Kish 1965, 11.7]. With

this model we get $\dfrac{S^2}{n} \cdot \dfrac{n\sum k_j^2}{(\sum k_j)^2}$ for the variance of the mean. The element variance S_y^2 may

be estimated by $s_y^2 = \dfrac{\sum k_j \sum k_j y_j^2 - (\sum k_j y_j)^2}{(\sum k_j)^2} \left(\dfrac{n}{n-1}\right)$, and this last correction may often be neglected.

The factor $n\sum k_j^2 / (\sum k_j)^2 = (1 + L)$ represents the ratio of increase in variance due to random weighting. we may view $n_e = n/(1 + L)$ as the effective sample size reduced by the factor (1 + L) due to random weighting. Then $s_y^2/n_e = (1 + L)s_y^2/n$ represents srs variance increased by the same weighting factor as the numerator in $deft^2$, and

$$deft_u^2 = \frac{var(r)}{s_y^2/n_e} = \frac{def^2}{(1 + L)}$$

represents the increase in variance due to clustering, without the effects of weighting. Thus $deft^2 = deft_u^2 (1 + L)$ is composed of two factors : $deft_u^2$ due to unweighted clustering and (1 + L) due to weighting. Therefore it is misleading to neglect large factors (1 + L) and use the large $deft^2$ as if entirely due to clustering ; e.g. to compute $roh = (deft^2 - 1)/(\bar{b} - 1)$.

On the other hand we may now use the design effect corrected for (1 + L) and estimate $roh = (deft_u^2 - (\bar{b} - 1)$. This will need some thought and care as shown below (7).

The ratio of increase attributed to random weighting may be computed several ways, depending on data and on needs :

$$(1 + L) = \frac{n\sum k_j^2}{(\sum k_j)^2} = \frac{n\sum n_h k_h^2}{(\sum n_h k_h)^2} = (\sum W_h k_h)(\sum W_h / k_h),$$

where n_h cases have weights k_h in a class or stratum h in the sample ; or $W_h = n_h k_h / \sum n_h k_h$ represents the size of stratum h in the population, with $\sum W_h = 1$. The class weights k_h and the element weights k_j are any positive weights desired. Note that each of the forms becomes correctly $(1 + L) = 1$ when the weights become any constant $k_j = k_h = c > 0$.

The factor $(1 + L)$ is rather robust, insensitive under moderate changes in the sizes and distributions of the weights, as may be seen in Table below. We shall note several consequences of that robustness ; for example usually a few computations from a sample may well serve for many defts computed.

2. We may neglect the effects if the <u>relative</u> range of weights k_j is within the bounds 1:2. This should be true, for example, if the weights are only for differences in nonresponses and for post-stratification. Within those bounds the increase due to weighting is usually within the ratio 1.04 for the variance and deft^2 and necessarily within 1.125 (see Table). These would not interfere greatly with the precisions needed and available, but occasional computations may be used for safety.

3. We may compute estimates of $\text{deft}_u^2 = (\text{var}(r)/(s^2/n)$ without weights in both var(r) and s^2/n to estimate the other effects of the design. Compared with (1), this has the disadvantage of estimating parameters in a population without arbitrary proportions. However, these results may not differ violently from results that (1) with proper proportions would give.

4. We may make separate computations and presentations for two or three domains with greatly differing sampling rates or different methods, but with epsem selections within each. For example, in a survey of a rurel country the sample of urban areas had essentially element sampling of its dwellings and deft^2 of 1.0, whereas the mostly rural residual areas had large area clusters and deft^2 averaging 4. But subclasses combined accross the two domains would need some care [Kish, Groves and Krotki, 1976, Section 3.4, Malaysia].

5. We may insist that deft^2 and roh are valid only for epsem samples, hence weighted samples need the full complexity of detailed computations of all the separate components. This may seem advisable in situations wher the sampling weights differ greatly so that we cannot neglect them as in 2), and we connot separate them simply as in 4), but we do not trust the simple computations of 1) or 3).

6. The simple formulas of (1) for $(1 + L)$ have been often used successfully and some examples may suffice. An interesting extreme case occurs in the design effects averaging 1.05 found for element sampling of telephone numbers, without any clustering effects. The source of these curious design effects is the weighting introduced to compensate for selecting a single adult from the unequal numbers associated with each telephone. These numbers and weights k_j are integers (1,2,3,4) concentrated around 2 ; the effects are small and conform to the model in (1). [Groves and Kahn, Table 6.5, "stratified"].

Examples of $(1 + L)$ ratios are found in section 6.3 and Table 9 of Verma, O'Muircheartaigh and Scott [1980], and these show rather small weighting effects. But elsewhere we have seen large effects published and attributed to clustering, whereas much or most of them were probably due to large relative weights. See the Table below for relation of L to weights k_h and proportions W_h in population strata. Only large differences in the k_h, combined with substantial relative weights W_h yield large ratios $(1 + L)$ and losses L, and experierces with computations are instructive. The robustness of these ratios also means that values from this Table or a few computations for any sample may suffice for many defts computed.

7. Both the care needed and the possibilities of using the factors $(1 + L)$ described in (1) can be illustrated with design effects for "crossclasses". Values of $roh = (deft^2 - 1)/(\bar{b}_t - 1)$ computed for an entire epsem sample can also be used for crossclasses such as age, sex, classes, and many others. This can be done approximately with $deft_s^2 = 1 + roh\,(\bar{b}_t - 1)$, where $\bar{b}_s = n_s/a$, the average size of the crossclass per primary selection area. This relationship is shown to hold fairly well when the b_i are not too small, nor too unequal (see the last two references). We have been using and testing these relations for years [Kish, Groves and Krotki, 1976] and often "variance estimates based on an assumption of constant roh across crossclasses are likely to be at least equal in quality (as measured by the mean square error) to directly estimated variances [Rust 1984].

On the contrary, the increase $(1 + L)$ due to weighting can be assumed to be constant (invariant, inherited) over crossclasses. Therefore, for inference from $deft^2$ for the total sample to $deft_s^2$ for a subclass we should separate the effects of clustering $deft_u^2$ from those for weighting $(1 + L)$. Thus the computation can proceed in the following steps :

$$deftu^2 = deft^2/(1 + L)$$
$$roh_t = (deftu_s^2 - 1)/(\bar{b}_t - 1)$$
$$deftu_s^2 = 1 + roh_t\,(\bar{b}_s - 1)$$
$$deft_s^2 = deftu_s^2\,(1 + L).$$

Of course, for inferences to other samples from the same clustered design but with different weighting system, the effect $1 + L$ will have to be changed. The formulas in (1) and the Table below should help those designs.

TABLE 4.5.1. Relative Losses (L) for Six Models of Population Weights (U_i) ; for Discrete (L_d) and Continuous (L_c) Weights : for Relative Departures (K_i) in the Range from 1 to K[a,b]

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

[a] From Kish, 1987

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

References

Groves R M and Kahn R L [1979]. Surveys by Telephone, NY : Academic Press, Table 6.5.

Hansen M H, Hurwitz W N and Madow W G [1953]. Sample Survey Methods and Theory, Vol I, 6.8, 12. DII.

Kish L [1987]. Statistical Design for Research, New York : Wiley, Table 4.5.1.

Kish L [1965]. Survey Sampling, New York : Wiley, Section 11.7.

Kish L, Groves R M and Krothi, K P [1976]. Sampling errors for Fertility surveys, The Hague : ISI, WFS, Occasional Paper, No. 17, Section 3.4.

Rust, K [1984]. Techniques for Estimating Variance for Sample Surveys, Ann Arbor : The U. of Michigan, Ch. 6.

Verma V, Scott C and O'Muircheartaigh C [1980]. Sample designs and sampling errors for the World Fertility Survey, JRSS(A), pp. 460-61.

SS 17, No. 2

Question : To compute $deft^2 = var(\bar{y})/(s^2/n)$ for design effects of means $\bar{y}$ from complex (stratified clustered) samples we need to estimate s^2/n. How can these be estimated simply, without computing the components for several stages and their strata ?

Answer : I take this opportunity to answer this older question together with the answer to no. 1 above. It can be shown that "descriptive" statistics are good, consistent estimates of their respective parameters. This holds for the sample mean $\bar{y} = y/n$ and for the sample element variance $s_y^2 = (\sum y_j^2 - n\bar{y}^2)/(n-1)$ as estimates of $\bar{Y} = Y/N$ and $S^2 = (\sum Y_i^2 - N\bar{Y}^2)/(N-1)$. Thus $\bar{y}$ and s_y^2 are good, consistent estimates of $\bar{Y}$ and S^2, regardless of the complexities of the sample. However these statistics only hold for self-weighted estimates.

For weighted estimates we need weighted means $\bar{y}_w = \sum k_j y_j / \sum k_j$ and element variances $s_y^2 = [\sum k_j \sum k_j y_j^2 - (\sum k_j y_j)^2]/(\sum k_j)^2 \, [n/(n-1)]$.

A somewhat improved (unbiased) estimator, instead of s_y^2, can be shown to be $\tilde{\sigma}^2 = s_y^2\,[1 + (deft^2-1)/n]$. But this correction is of small consequence unless $deft^2$ is large and n small [Kish 1965, 2.8C].

A more important issue arises when we compute the srs variance for the denominator of $deft^2$. That will be s_y^2/n for self weighted samples - or pq/(n-1) for proportions p. However, for weighted samples, $s_y^2/n_e = (1 + L)s_y^2/n$ the reasons noted above in answer no. 1. The factor (1 + L) may be considerable, as in the Table above.

IMP. MARIE - HONFLEUR