KEEPING SAMPLE SURVEY DESIGN AND ANALYSIS SIMPLE[1/]

Isidoro P. David[2/]

ABSTRACT

The experience with a number of analytic household sample surveys is recalled to explain why the analysis of such surveys often takes too long to complete, and with much higher sampling errors than expected. A major reason for this is hasty choice of sampling procedure with little or no consideration being given to ease of data processing and analysis. The result usually is a complex sample, the proper statistical analysis of which would have to be correspondingly complicated. The irony is that in the end the researcher often is compelled to proceed as if the sample were "simple random", an assumption which can potentially lead to serious inferential mistakes. Moreover, sampling procedures that look good in theory can perform disappointingly in developing countries. One example of this is the use of stratification on the basis of imperfect prior information which causes misclassification of units. Another is stratification in terms of a dynmamic variable in time series surveys in which units migrate in and out of strata. These errors and changes in classification complicate the analysis and render the estimates less precise.

KEYWORDS: analytic surveys, domain, misstratification, project benefit monitoring and evaluation (PBME), replicated sampling, replicated systematic sampling, stratification.

1. Introduction

Outside of a few market research firms, it used to be that government had a monopoly on censuses and surveys in developing countries (DCs) in the Asian region. Data users, particularly researchers, were content with secondary sources and thus - others argue - they had more time to reflect on their main problems.

The "survey bug" began spreading among DC scientists sometime in the late 1950s or early 1960s. It is likely that the germinal idea was picked up by the many students from DCs who went to western schools which tended to have graduate programs in the social sciences - particularly applied economics and sociology - that encouraged the use of survey data in research. These early graduates had high multiplier effects since many of them went into teaching, thereby replicating their training. Also, there was synergism between this orientation and the low cost of doing sample surveys in DCs. Further, as more resources became available for research and development projects directed towards low-income groups, e.g. small farm development integrated rural development, development of bypassed areas and the like, it became fashionable to work with data about the farmer and his household, the landless worker, the urban slum dweller, etc. Researchers who could not avail of detailed household level data for reason of confidentiality of individual government census and survey return began to

---

1/ Edited version for the Philippine Statistician of Statistical Report Series No. 5, Asian Development Bank.

2/ Head, Statistics Unit, Asian Development Bank.

collect their own data to suit their particular needs. Indeed, the demand for household data has so risen that - in the Philippines at least - traditional accounting firms and new outfits have joined the market by proposing to undertake research programs with surveys or by offering their services to do other researchers' surveys.

In recent years, government bodies charged with planning and implementing development projects, and external agencies which help fund these have done much to increase the demand for surveys. We refer to project or program monitoring and evaluation (M & E) which is increasingly being required by these agencies. Operationally, M & E, say of an integrated rural area development project, requires a schedule of data collection activities during the project cycle, often from a sample of household beneficiaries. Since projects of this type last anywhere from 5 to 15 years from start to full development, M & E, surveys overlap and grow in number with time. A quick count of multilateral and government agencies requiring M & E of the projects they fund should point to a snowballing effect in the number of M & E sample surveys in the years to come. Thus, in addition to surveys mandated to government statistical bodies, more have become obligatory (e.g. M & E surveys), or expected. In the Philippines, for example, there is an alarming trend among research proposals in agricultural economics toward basic data collection through rural household surveys lest by some convoluted reasoning those proposals that suggest utilizing data from secondary sources tend to be judged incomplete and ill-conceived - and denied funding.

The increased demand for surveys has not been matched by growth in survey sampling and data processing expertise. The researcher-statistician-data processor troika necessary to carry out a successful survey often is incomplete - if success means completing the data analysis on time, within the budget and with results meeting prescribed levels of accuracy. The researcher usually plays survey statistician in the latter's absence, or more truthfully, in the belief that he is a competent survey statistician; sometimes he also has to do his own data processing. Although such an arrangement can be educational, the outcome often is less than satisfactory: sampling design by rote; simplistic statistical analysis assuming "simple random sampling", thereby ignoring the complexities of the final sample; and protracted data processing. Mix these with a sample of moderate size and a propensity toward hefty questionnaires, and what we have is a recipe for massive amounts of data that are poorly processed, scarcely analyzed, or worse, hardly worth the trouble of analyzing at all.

This article was spurred by recent experience in trying to help analyze data from analytic household surveys, during which earlier suspicions that the same problems persist in this type of surveys were reinforced. There are two broad problems: (a) The sampling designs commonly used that look good and straightforward in theory can perform disappointingly in DCs for two reasons: (1) they are quite sensitive to errors in design (e.g. stratification) variables, which render them imprecise, and (2) they result in complex unbalanced samples which lead to

complications in data analysis. (b) The complex sample is often treated like a "simple random sample" during analysis, and oversimplication which can potentially lead to serious inferential mistakes.

The fact that these problems are perpetuated suggests that they have gone largely ignored or else they have been misunderstood by survey practitioners and data users. And one could not possibly begin to think of, much less appreciate, solutions to an unperceived problem. Hence, an ample part of the paper is an attempt to explain, through actual cases, the nature and consequences of these problems. After presenting illustrative sample surveys in section 2, section 3 concentrates on the stratification aspect of sampling design (for the reasons explained there), and discusses briefly why and how stratification is used, and what it is supposed to accomplish. In section 4, we go back to the actual surveys and point out the real results and consequences of the stratification procedures used. Section 5 presents an approach to the analysis of a sample survey complicated by misstratification and movement of units across strata. The resulting weighted estimates are compared with the unweighted estimates obtained under a simple random sample assumption.

2. The Surveys

Four sample surveys with varying degrees of detail will be used as case studies. All happen to be rural household surveys for no special reason other than that they form the set which the writer had been involved in during the last four years.

The Asian Development Bank encourages monitoring and evaluation of the likely benefits that Bank-assisted projects provide to intended beneficiaries. Guidelines (1980, 1984) have been prepared specifying how such M & E, called project benefit monitoring and evaluation (PBME), is to proceed. For agriculture and rural development projects, time series data are collected in and around the project area through sample surveys. The first of these, the benchmark survey, is conducted before project implementation, usually during project appraisal or feasibility studies. This is followed by monitoring surveys done every crop season during project implementation, which usually use a subsample from the benchmark sample of households. Next is a provisional evaluation survey done on completion of the project, which is followed years later (at full project development) by a final impact evaluation survey.

Two of the four surveys discussed here are PBME benchmark surveys. One is for the Tulungagung Drainage Project in Indonesia, with a sample of 650 households out of more than 90,000 households spread in 150 villages over 30,000 hectares. Details of the Tulungagung survey are given in David (1982a) and the comprehensive report of the PBME research team (1981). The second survey, which is discussed in greater detail, is for the Davao III Irrigation Project in Southern Philippines, with a sample of 245 farm households out of 6,800 residing in the project area. A detailed account of the Davao III survey is given in David (1982b).

The other two are from the International Rice Research Institute's (IRRI) three-

country research project on the Consequences of Small Farm Mechanization, the primary objective of which was to assess the effect of rice farm mechanization on production, income and rural employment. The data was collected mainly through a series of crop season sample surveys. The Nueva Ecija province (Philippines) surveys will be used here. The three other sets of surveys, namely West Java (Indonesia), South Sulawesi (Indonesia), and Suphanburi Province (Thailand), have similar stratifications, both intended and real; hence only the first of these will be used. Details of the sampling procedures are documented in the research team's Operations Handbook (1982); the Nueva Ecija surveys are also described in Lim (1982).

3. On Sampling Designs for Analytic Surveys

3.1 Dichotomy of Surveys

Sample surveys may be classified into two types according to main purpose. Those which have as their main object the assessment of the characteristics of the sampled population or parts of it - usually through point estimates such as means and ratios and their sampling errors - are called descriptive or enumerative surveys. Many large surveys done by national statistics offices fit this category. Those that are used mainly to analyze relationships - for example, to compare means of subpopulations, to fit regression models and to test hypotheses - are called analytic or investigative surveys. The four surveys mentioned in the previous section fall in this latter group.

The choice of sampling design may vary with type of survey to be done. With few exceptions, the designs found in standard textbooks were developed for descriptive surveys the goal of which is, subject to budgetary and other resources constraints, to minimize the sampling variance of point estimates. Toward this end, auxiliary information is used in a variety of combinations - e.g. stratification, probability proportional to size (pps) selection, multi-stage and multi-phase sampling. However, with the possible exception of stratification, these variance- reducing techniques generally make sampling more complex in the sense that the resulting sample deviates further from a simple random sample (srs), thereby requiring the use of more complicated (e.g. weighted) estimates. Also, the theoretical underpinning of inference procedures based on more complex sampling designs and estimates have yet to be sufficiently developed, although research in this area has proceeded at a fast pace during the last 20 years. For example, the sampling distributions of non-linear statistics such as correlations, regression coefficients and t-ratios are extremely complex and still largely intractable. There are also empirical reports that demonstrate that the common practice of assuming complex samples to be srs or iid samples can lead to serious mistakes. (see e.g. Kish and Frankel, 1974).

An expert could take advantage of the full array of available sampling and estimation techniques and come up with a logical, sound and practical design for an analytic survey. However, generally, given the shortage of statistical know-how and computing facilities in DCs, prudent advice

would be to exploit stratification heavily, almost exclusively, to increase the likelihood of drawing a "preferred sample" (see next subsection) that is as close as possible to being a simple random sample. This is the reason why the ensuing discussions concentrate on stratification.

The same sort of advice holds for data analysis. A simple estimation procedure should be the goal in both types of surveys. A simple estimation procedure is also a necessary condition for expedient processing of massive data from descriptive surveys. It might sound surprising that many analytic surveys with samples between 200 and 600 households get bogged down at the data analysis stage. One reason for this is the sampling design, the choice of which more often is made without simplified data analysis in mind. Other reasons are unanticipated data collection problems and long questionnaires, which together can make data cleaning a long drawn-affair.

3.2 Stratification in Theory

Consider a hypothetical population of eight households: The number of possible simple random samples of size 4 is $\binom{8}{4} = 70$, and each has probability 1/70 of being chosen. But if the eight units are split into two equal-sized strata and two units are drawn from each, the number of possibilities is reduced to $\binom{4}{2} \cdot \binom{4}{2} = 36$, each with a selection probability of 1/36. Stratification gives zero selection probability to the 34 other samples. Presumably the sampler has reasons to prefer having one of the 36 samples, thus their selection probabilities are raised, and at the same time he makes certain that not one of the 34 he does not prefer will be drawn.

Note that although some combinations of units (samples) have zero selection probabilities, the population units individually still have positive probabilities of inclusion in the sample. Also, while stratification gives the sampler some control over the assignment of selection probabilities to the possible samples, it also allows him much flexibility. For example, various strata sizes and sample allocations give different partitions of the sample space into preferred and non-preferred subspaces. Additionally, the use of different sampling schemes among strata permits the setting of a preference scale (i.e. varying probabilities) within the subspace of preferred samples (although this last point is of more relevance to descriptive surveys).

Thus, stratification is a widely used technique for controlling the selection process in such a way that "preferred samples" are assigned larger and the rest smaller (sometimes zero) chances of being drawn. This, however, is still done within the scope of probability sampling, i.e. subject to the condition that every unit of the population has a positive probability of inclusion in the sample, and this probability is knowable beforehand.

The purpose of the survey should somehow determine which samples are to be preferred and which are not. Although conflicts may arise in multi-variable surveys in which some of the variables are not correlated or may even be inversely related, preferred samples in general are those that (a) give estimates with lower sampling errors, (b) are suitable for the proposed analytical

procedures, and (c) are amenable to easy statistical analysis.

Theoretically, stratification of the population - or more precisely, of the sampling frame - can effectively lower sampling errors, with the extent of reduction (relative to not stratifying) being dependent on how well we succeed in forming each stratum out of units that are as alike as possible, with likeness reckoned in terms of the observed values of stratification variables. Hence farms are routinely grouped by crop type (rice, corn, ...), size (small, medium, large), or water management (irrigated, not irrigated); non-farm households may be classified by level or source of income or by number of members. In principle, the internal variance of these strata will be small and the variance between them will be large, but the latter drops out of the sampling error of estimates since the strata are sampled independently.

Further reduction in sampling error is possible by allocating the total sample to the strata in some optimal manner, for example, proportionate to $N_i S_i$ where (N_i) and (S_i) are the strata sizes and standard deviations, respectively. Here again, the (S_i) are computed a priori from some design variable, possibly the stratification variable as well. A popular alternative is allocation in proportion to the (N_i) only, which in single-stage stratified srs leads to equal selection probabilities for all the population units. The sample becomes "self-weighing" and the estimation of means and totals proceeds as if one has a single srs. Also, the sample comes closest to being an independent, identically distributed (iid) sample, a key and often stringent assumption in many statistical techniques applied to survey data, e.g. ordinary least squares regression, analysis of variance and analysis of contingency tables.

What suits (a) usually serves (b) well also. For example, grouping households by income level or some proxy variable into low, medium, and high income strata and drawing a sample from each would lower sampling variances of estimates for variables substantially correlated with income. The procedure also guards against drawing, say a sample that has no high income households. The latter type of sample is non-preferred in many ways: it underestimates the medium income, it does not allow a comparison of mean income among groups, and a fitted regression from it with income as a dependent variable will be subject to truncation bias (see e.g. Amemiya, 1982).

If strata are composed of subpopulations for which separate estimates are desired, computations are simplified greatly by the fact that the subpopulations (also called domains) are sampled independently; hence their estimates are uncorrelated. This is an important practical consideration since, as will be shown in Section V, estimation and inference concerning domains that cut across strata boundaries can be a tedious affair.

4. Stratification in Practice

The message here is this: in DCs, stratification usually means grouping sampling units on the basis of imperfect prior information, the effect of which nullifies the purpose for stratifying in the first place.

4.1 Stratification of Primary Sampling Units

A typical household survey in the Asian region usually involves sampling in two, sometimes three, stages: district or town, village or enumration area, and finally, the household. The higher-stage units often are stratified based on geographic contiguity, and on information from past censuses or administrative records.

The Nueva Ecija surveys which were conducted in 1979-1980 employed a two-stage sampling procedure (not counting the towns) with stratification at both stages. In line with the objectives of the project, the surveys focused deliberately on two towns only, Cabanatuan and Guimba, which were found to have suitably large areas of rice under different levels of water management and methods of land preparation. Based on data from the Bureau of Agricultural Economics' (BAECON) 1976 barangay (village) census, the rice-growing villages of each town were divided into four strata according to main source of water (rainfed, irrigated) and number of tractors (less than five = low mechanization, five or more = high mechanization). The idea was that four sample villages per town - one chosen at random from each stratum - should be adequate to supply sample farm households which possess sufficient variations to allow comparison of the main effects of, as well as interactions between, mechanization and irrigation. The strata and sample villages are shown in the first two columns of Table 1.

The BAEcon data-based stratification was compared with the (factual) classification based on 1979 information obtained from a complete enumeration of households, shown in the last column of Table 1. The results revealed changes in classification in three of the eight villages: Lagare in Cabanatuan, and Burol and San Andres in Guimba. These recorded changes could have been due to genuine changes in the values of the stratification variables, but response errors in the 1976 data and differences in data definitions between the two sources are more likely explanations.

Table 1. Sample Villages, Nueva Ecija Survey

Town/Stratum	1976 BAEcon Census	1979 Household Census
Cabanatuan		
Rainfed, low mech.	Kalikid Sur	Kalikid Sur
Rainfed, high mech.	Lagare	-
Irrigated, low mech.	Caalibangbangan	Caalibangbangan
Irrigated, high mech.	San Isidro	San Isidro, Lagare
Guimba		
Rainfed, low mech.	Galvan	Galvan
Rainfed, high mech.	Burol	San Andres
Irrigated, low mech.	Narvacan	Narvacan
Irrigated, high mech.	San Andres	Burol

In either case, mis-stratification reduces the utility of data. It can spoil carefully conceived sample allocation plans both for the first and second stage units. For example, since Lagare village turned out to be irrigated instead of rainfed, a comparison in Cabanatuan between low-mechanized rainfed and high mechanized rainfed farms would be constrained by a smaller sample, or what amounts to the same thing, high sampling errors. Moreover, unless adjustments are made at the last stage of sampling, the final sample size for rainfed farms (n_r) would likely be much less than for irrigated farms (n_i), and this would affect the variance of the

difference between group means,

$$\sigma^2_{\bar{x}_r - \bar{x}_i} = \frac{\sigma_r^2}{n_r} + \frac{\sigma_i^2}{n_i}$$

where σ_r^2 and σ_i^2 are the variances of the rainfed and irrigated farms, respectively. When these two variances are the same, equal allocation ($n_r = n_i$) will minimize $\sigma^2_{\bar{x}_r - \bar{x}_i}$. An allocation in which n_r is much less than n_i is preferable only if σ_r^2 is proportionately much smaller than σ_i^2.

Notice from Table 1 that the two villages, Burol and San Andres in Quimba town switched strata. Outwardly, the effects of this "compensating" misclassification may not look as adverse as the one-sided case in Cabanatuan. However, both types make the analysis of the data more complicated and the sampling variances higher than if all units were correctly classified. Numerical illustrations of this are presented in section 5.

4.2 Stratification of Households

Household frames are harder to get, costlier to construct, and more subject to time-induced changes and errors than village lists. In two-stage sampling the usual procedure, as in the Nueva Ecija surveys, is to list all households in the sample first-stage units. This list serves as the frame from which the sample households are chosen.

The Tulungagung (Java) drainage project area covered 1,209 census blocks. For the first-stage sample of the BMBE benchmark survey an srs of 600 census blocks was chosen. The households in these selected census blocks were completely enumerated and asked a few questions, one of which was whether the household experienced flooding during the previous year. The result was a frame of 41,544 enumerated households, which was used to construct two strata: flooded (12,497) and not flooded (29,047) households. Six hundred and fifty srs households were drawn from these two strata respectively. About two months after the enumeration, in September-October 1980, the benchmark survey was conducted which allowed more detailed information on flooding incidence, including flooding experience in the previous year, to be obtained from the 641 sample households who responded. Table 2 compares the flooding incidence as reported during the household frame construction and during the benchmark survey. Note that only about half of those in the flooded stratum (308/592) were classified as flooded during the main survey; on the other hand, about one-fourth of those in the not flooded stratum (13/49) should have been labelled otherwise.

Table 2. Frequency Distribution of Sample Households by Flooding Experience, Tulungagung Survey, 1980

From Household Frame	From Benchmark Survey: Flooded	Not Flooded	Total
Flooded	308	284	592
Not flooded	13	36	49
Total	321	320	641

Stratification by flooding incidence was resorted to (a) so that the chances of having a sample with both flooded and not flooded units was certain, (b) so that separate estimates for these two strata could be easily computed, and (c) since

these estimates would have lower sampling errors relative to non-stratified sampling estimates. That (b) is lost is obvious here: one has to cross over both strata in order to compute an estimate for either floode or not flooded households. Consequently, the sampling variance of a difference (in means, say,) between flooded and not flooded households involves a covariance term, making computations more complex than if the households were correctly classified. The problem is computationally equivalent to estimation methods for subclasses or domains whose elements cut across strata boundaries (see e.g. Kish, 1965, pp. 132-139). Item (c) is forgone as well, as the numerical examples of subsection 4.3 and section 5 will show.

What could have caused these many discrepancies between the frame and survey data? We cite two main reasons from experience. First, owing to their sizes, frames or lists are done in a hurry, and any member of the household who in the judgement of the enumerator is "knowledgeable" serves as the interviewee; this is in contrast to the survey proper in which more probing cross-validating questions are asked, and the head of the household is usually the designated respondent. Second, the survey teams regard sampling frame preparations as preliminaries, reserving more thorough and careful work for the main survey. This latter effect is stronger if the survey teams get the impression that sampling frame information is not used in computing estimates, which of course is not necessarily true: frame data become among other things the bases for stratum weights, selection probabilities and values for the auxilliary variables in ratio and regression estimation.

If the stratification variables change values quickly, the frame may be inaccurate the moment it is completed. This happened in the IRRI mechanization consequences surveys in which farm households were stratified by type of power used for land preparation (animal, tractor, or some combination of the two). These are classifications directly relevant to the objectives of the research project; however, the frequency and ease with which farmers switched from one type of power to another from one season to the next may have been unanticipated. The case of the Nueva Ecija surveys is discussed in secton 5 in conjunction with the discussion of the effects that changes in classification exert on data analysis. This case is very similar to that of the West Java machanization survey discussed below, and for that matter, to the other IRRI mechanization surveys in South Sulawasi and Thailand.

Table 3 presents information relating to the West Java mechanization consequences surveys. The first three columns show the strata, stratum sizes and sample allocation based on a household listing in early 1979. In brief, the listed households were grouped according to the type of power used for primary and secondary tillage: manual (M), animal (A), own tractor (OT), hire tractor (HT), and manual and animal (MA). A sixth group composed of landless workers (L), i.e. households that do not operate farms, but which earn a living as hired farm workers, was also included. Sixty households were selected from each group using srs, and the same sample was retained for the next three crop season surveys. However, the actual

sample distribution during the following three seasons not only differed widely from the initial equal allocation, but also from one another as well (see the last three columns of Table 3). For instance, the number of manually cultivated (i.e. man and hoe) rice farms increased from 60 to 217 during the 1979 dry season. When traced to the original sample composition this change is accounted for as follows:

217 new M = 60 original M
- 2 who shifted to HT
+ 57 who shifted from A
+ 5 who shifted from OT
+ 42 who shifted from HT
+ 55 who shifted from MA

In addition, new classifications emerged; namely, manual + tractor, animal + tractor, and manual + animal + tractor. This posed new problems in analysis. For example, had there been no movement of households, the simple average of the sixty sample observations (of a variate x) in stratum M would be an unbiased estimate of the mean of that stratum; the sampling variance is just as easily obtained. On the other hand, an unbiased estimate of the main of stratum M during the 1979 dry season is a weighted average of the 217 sample observations, the weight of each observation being the relative size of the stratum from which it originally came. Simplifying the analysis by applying a common weight to each observation (i.e. an unweighted estimate) biases the estimate because of the large differences in the stratum sizes. The problems become more complex when two strata are to be compared, or when a regression model is to be estimated. These complications figure prominently in delays in bringing surveys to successful conclusions.

Table 3. Household Stratification, Initial Sample Size, and Final Sample Sizes, West Java Surveys

Type of Power (Stratum)	Stratum Size	Initial Sample Size	Actual Sample Size, by Crop Season 1979 Dry	1979/80 Wet	1980 Dry
Manual (M)	584	60	217	55	161
Animal (A)	181	60	1	75	1
Own tractor (OT)	66	60	55	68	65
Hired tractor (HT)	254	60	26	49	38
Manual + animal (MA)	77	60	1	25	2
Manual + tractor[a/]	0	0	0	4	22
Animal + tractor[a/]	0	0	0	8	0
Manual+animal+tractor[a/]	0	0	0	1	0
Landless worker (L)	404[b/]	60	60	74	69
Total	1,565	360	360	359[c/]	358[d/]

a/ These cells were empty in the initial population and sample.

b/ These are households that do not operate farms, but which earn a living as hired farm workers.

c/ One landless worker respondent died.

d/ One respondent in stratum (A) moved out of the study area.

4.3 Stratification of Farm Households by Size of Landholding

This is commonly done to ensure representation of different farm sizes in the sample and to reduce the sampling error of production, area and yield estimates. The usual practice is to create strata - e.g. small, medium and large - with boundaries set more or less arbitrarily at integer values, based again on interview data from a list or census of households. There are two problems with this method of stratification.

One, which may be peculiar to the Philippines, has to do with the respondents' predilection to give landholdings in whole numbers (and to a lesser extent in multiples of one-fourth and one-third of a hectare).

The frequency distribution of landholdings from the Davao III project list of households shown in Table 4 is a typical example: Close to one-fifth of the responses are "one" hectare, and one-third of the responses are either "one," "two" or "three" hectares. This bunching causes a downward or upward bias on (stratified) estimates depending whether the integer boundary is counted in the previous stratum or the following one. For the same reason, the bias persists in (grouped) estimates computed from frequency distribution tables. As an illustration of this, consider estimating the mean and standard deviation (the true values of which calculated from the individual observations are $\mu = 2.56$ and $\sigma = 3.26$) from the two frequency distributions in Table 5 which differ only in the choice of class (stratum) boundaries. The left half shows the frequency distribution of the same landholding data using the class intervals 0.01 - 0.99, 1.00 - 1.99, ..., with the corresponding mid-points (x_i) and frequencies (f_i). These give the mean μ_ℓ (ℓ is for left),

$$\mu_\ell = \Sigma f_i x_i / \Sigma f_i = 2.82$$

and standard deviation

$$\sigma_\ell = \left\{\frac{\Sigma f_i x_i^2 - \Sigma f_i \mu_\ell^2}{\Sigma f_i}\right\}^{\frac{1}{2}} = 3.15$$

Note that μ_ℓ overshoots the actual mean by 10 per cent. The bias increases if more and finer class intervals are used and decreases if some classes are merged. On the other hand, the right half of Table 5 is different from the left half only in that the integer lower boundary of each class interval is made the upper boundary of the preceding one, i.e. 0.01 - 1.00, 1.01 - 2.00, ... Notice, however, the remarkable changes in the f_i. The mean $\mu_r = 2.33$ this time understates the actual mean by 9 per cent. The standard deviation is $\sigma = 3.12$. Both σ_ℓ and σ_r exhibit small but negative biases; in each case this is largely due to understatement by the last open interval of the contribution of large units.

The choice of the integer boundaries also affects the sample allocation among the strata. In the Davao III household frame, for example, all farm households with landholdings of less than three hectares were put into a small stratum ($N_1 = 4885$), those with three to seven hectares inclusive were classified as medium ($N_2 = 1502$) and those with over seven hectares belong to the large stratum ($N_3 = 418$). If, hypothetically, 245 sample farm households were allocated proportionately to the sizes of the strata, the allocation is (176, 54, 15).

Table 4. Frequency Distribution of Landholding Showing the Bunching of Responses at Integer Points, Davao III List of Households

Landholding (Hectares)	Number of Households	Landholding (Hectares)	No. of Households
0.01-0.24[a/]	85	1.00	1,225
0.25	347	1.01-1.99	927
0.26-0.49	63	2.00	701
0.50	771	2.01-2.99	323
0.51-0.74	44	3.00	379
0.75	363	Over 3.00	1,541
0.76-0.99	36	Total	6,805

a/ Households with landholdings of less than 0.01 hectare are classified as non-farm households.

If, however, the strata were redefined ever so slightly - by taking out "3" hectares from the medium stratum and putting it in the small stratum - the sizes are N_1 = 5264, N_i = 1123 and N_3 = 418 and the allocation is then (190, 40, 15). This is all right, arithmetically. But it grates against common sense to find that a decision to make the single point "3" the end-point of a stratum or not means gaining or losing 6 per cent of the total sample from that stratum.

Table 5. The Effect of Integer Boundary Placements on Frequency Distributions, Davao III

Landholding (Hectares)			Landholding (Hectares)		
0.01-0.99	0.505	1,709	0.01-1.00	0.505	2,934
1.00-1.99	1.495	2,152	1.01-2.00	1.505	1,628
2.00-2.99	2.495	1,024	1.01-3.00	2.505	702
3.00-3.99	3.495	539	3.01-4.00	3.505	404
4.00-4.99	4.495	350	4.01-5.00	4.505	426
5.00-6.99	5.995	525	5.01-7.00	6.005	293
7.00-8.99	7.995	184	7.01-9.00	8.005	136
9.00-11.99	10.495	146	9.01-12.00	10.505	163
12 and over	16.783[a/]	176	Over 12.00	19.074[a/]	119

a/ Actual mean of the individual observations in the class.

A more serious problem is mis-stratification due to errors in the frame data. To continue with the Davao III example, a comparison of the stratification based on the frame of households and on the classification from the benchmark survey data showed only a 78 per cent match (see Table 6). It is not difficult to imagine what effect the 53 misclassified sample units will have on the sampling error of estimates. Consider for a moment the 87 sample units stratified as small (second column of Table 6). Had all of the 87 sample units truly been of small size, the estimate of the variance of, say, average paddy production per farm household in the stratum would have been correspondingly small. However, the eight sample units that were, in fact, medium sized, and the one unit that was, in reality, large did not allow this to occur. This was in fact the main reason why the precision of estimates did not live up to the target levels. In general, the effect of errors such as this is much worse for estimates of totals such as aggregate production and area, but less so for ratios such as yields.

Table 6. Classification of Davao III Sample Households by Landholding Size from the List of Households and the Benchmark Survey

From Survey	From List			
	Under 3 ha.	3-7 ha	Over 7 ha.	Total
Under 3 ha.	78	31	0	109
3-7 ha.	8	47	4	59
Over 7 ha.	1	9	67	77
Total	87	87	71	245

Turning to specific variables, Table 7 gives the relative errors or coefficients of variation (CV) of some estimates from the Davao III benchmark survey. The sampling design involved several levels of stratification, including main crop grown, tenure and size of landholding. Details of sampling and estimation procedures are given in David (1982b). These CVs refer to estimates of totals for major geographic strata, i.e, the project area consists of two distinguishable units (I and II), each

of which is subdivided further into two sub-areas (upstream and downstream) according to the location of each relative to the proposed main water canals. The CVs can only described as high; hence it is doubtful whether the estimates can be useful as PBME benchmark indicators.

Table 7. CV of Selected Statistics, Davao III PBME Benchmark Survey, 1981 (Percent)

Stratum	Sample Size	Total Production Rice	Total Production Corn	Total Area Planted Rice	Total Area Planted Corn	Total Landholding
Unit 1 - Upstream	64	13	47	14	22	15
Unit I - Downstream	40	60	98	83	100	12
Unit II - Upstream	68	19	27	21	17	12
Unit II - Downstream	73	17	29	13	13	12
Unit I -	104	31	50	18	21	10
Unit II -	141	13	20	13	11	9
I + II (Project)	245	12	19	11	10	6

Thus, one could ask whether the extra effort and expense of the elaborate frame and use of the complicated sampling design were justified by the results, and whether more efficient approaches could be found and recommended for future surveys of a similar nature. A partial answer to the latter question is that there is much room for improvement via simpler but more efficient sampling designs. As for the first question, one could compare the CVs in Table 7 with the results if unstratified srs were used instead. In the latter, the CV of a total (or mean), ignoring finite correction factors, is $(\sigma/\mu)/\sqrt{n}$. For landholding, σ/μ = 3.26/2.56 = 1.27, hence for a sample of comparable size such as the Davao survey, the CV of an estimate of the total or mean is $1.27/\sqrt{245}$ = 0.08 or 8 percent. Notice that this is two percentage points higher than the sample estimate in Table 7; this gives us a rough idea that the rather complex sampling design used was not substantially more efficient than straightforward srs of farm households.

5. Inference Concerning Domains that Cut Across Strata

5.1 Introduction

The word domain is used to mean a subpopulation for which separate estimates are to be computed. Ideally, domains should be made equivalent to the strata so that, as mentioned previously, data analysis becomes simple and the estimates have small sampling errors. However, there are many reasons why domains and strata boundaries might fail to coincide. To begin with, domains may not be designated as strata, as when the sex of respondents cannot be known beforehand but sex-disaggregated estimates are desired. Secondly, sampling units may be misstratified, as in the Tulungagung and Davao III surveys. Thirdly, sampling units may move into other strata between the time of stratification and the actual survey, as in the West Java mechanization consequences survey. Occasionally, combinations of these three occur simultaneously. But their effect, individually or collectively, on the data analysis and estimates is more or less the same: the former becomes more complicated and the latter less precise. We saw in subsection 4.3 what misclassification did to the estimates of the Davao III survey.

Here we use the Nueva Ecija survey to illustrate a method of analysis and the effect on the sampling error of estimates when sampling units move across strata.

The procedure for choosing the eight sample villages in the Nueva Ecija survey was described in subsection 4.1. A complete census of households among them were grouped into nine strata according to water source and power source for land preparation as shown in Table 8, first column. A sample of 337 rice farm households, with allocation as shown in the third column, was drawn using srs in each stratum. This same sample was to be used for a series of surveys covering two crop years. Migration, changes in occupation and a few non-responses reduced the sample by one to three units per stratum by the time of the first (1979 wet season) survey, to a total of 320 (see last column). Since this attrition was more or less evenly distributed among the strata, this low attrition rate (5 per cent) does not warrant any change in the approach to the analysis of the data.

What should cause greater concern, however, is the significant shifting of the remaining units across strata. Compare the original classifications of the sample units according to the household census (last column of Table 8 which is reproduced as the column labeled "Before" in Table 9) with the actual classification discovered during the 1979 wet season survey interviews (labelled "After" in Table 9). For example, rainfed animal-powered farms increased from 43 to 97, with the latter number of farms coming from six different strata, including 38 of the 43 original rainfed animal-powered farms. While these classification changes provide useful information, at the same time they laid to waste the work that went into stratification and sample allocation. Intuitively, these changes also ought to be reflected in, and force a logically correct approach to, the statistical analysis of the data.

Table 8. Rice Farm Households Stratification and Sample Sample Allocation, Nueva Ecija Survey

Water Source-Power Source	Stratum Size	Initial Sample Allocation	1979 Wet Season Survey Sample Size
(1) Rainfed - animal power	257	46	43
(2) " - 2-wheel tractor	82	37	36
(3) " - 4-wheel tractor	54	35	32
(4) Irrig., 1 crop - animal power	25	21	20
(5) " - 2-wheel tractor	50	32	29
(6) " - 4-wheel tractor	24	19	18
(7) Irrig., 2 crops- animal power	98	42	41
(8) " - 2-wheel tractor	239	66	64
(9) " - 4-wheel tractor	64	39	37
Total	893	337	320

Table 9. Before and After Classification of Sample Farm Households, Nueva Ecija Survey

Before ↓ \ After →		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	43	38	2	3	0	0	0	0	0	0
(2)	36	23	10	3	0	0	0	0	0	0
(3)	32	21	5	5	1	0	0	0	0	0
(4)	20	7	0	0	6	0	0	6	0	1
(5)	29	0	0	1	0	1	0	3	16	8
(6)	18	2	0	0	2	0	1	3	6	4
(7)	41	6	1	1	1	0	0	24	1	7
(8)	64	0	0	0	0	0	0	11	44	9
(9)	37	0	0	0	1	0	0	9	7	20
320		97	18	13	11	1	1	56	74	49

5.2 Superpopulation Concept for Analytic Surveys

In order not to obscure the main object here - which is to point to an approach for analyzing data when units have moved across strata - with undue mathematical complexities, we ignore the fact that the sample was drawn in two stages and assume that:

(a) The sample in Table 8 (last column) is a stratified single-stage srs from the finite population described in the same table (second column).

(b) The sampled population in Table 8 is in turn a stratified srs from an infinite (super) population whose units are distributed to the nine strata in the same proportions as Table 8, i.e., .288 (= 257/893) of the infinitely many units are rainfed animal-powered, etc.

The concept of a finite population itself being a sample from an infinite superpopulation is an old idea (Cochran, 1939; Deming and Stephan, 1941) which has found increasingly wide acceptance (see e.g. Hartley and Sielken, 1975) because, among other things, it is a simplifying assumption. Applied to the Nueva Ecija survey, it means accepting that (i) the target of inferences in the mechanization consequences study was all the while some conceptual infinite population, and (ii) the intricate selection procedure that led to the sample rice farm households of Table 8 can be assumed to be equivalent to srs. It would not be surprising if (i) fits the research proponents' thinking, but (ii) could draw mixed reactions, particularly since some non-probability-based judgment went into the choice of the study areas from which the sample units came (for details, see P.L. Lim, 1982).

With a superpopulation as the object of inferences, finite population correction factors drop out of estimators, allowing for somewhat simpler computation. Also, if the sample is srs from the finite population, which in turn is srs from the superpopulation, then estimates that are unbiased for the first are also unbiased for the latter population. However, complications arise; for example:

a) Non-srs used to draw the sample from the finite population. This is the case with Table 8 which is a stratified sample and, although srs was used in each stratum, the sampling rates among strata (which are also the inclusion probabilities) are very unequal. Hence the total sample is non-srs.

b) The sample units change classification, as in Table 9.

Presumably, the superpopulation undergoes changes also. The problem is how to proceed with the statistical analysis in the face of these complications.

5.3 Inference Concerning Proportions

A tempting and, unfortunately, commonly used procedure in analyzing survey data is to assign equal weights to the units in Table 9 regardless where they fall, and to ignore the fact that some were originally drawn from other strata. For example, from Table 9,

$$\tilde{p} = 97/320 = 0.303$$

is used to estimate the proportion of rainfed animal-powered farms, with standard error,

$$\text{s.e.}(\tilde{p}) = [\tilde{p}(1-\tilde{p})/319]^{\frac{1}{2}}$$

$$= 0.026$$

Estimates for the other domains are computed

similarly. However, this ignores whatever effect the sampling design has on the estimates, such as the fact that stratum weights and sampling rates vary across strata. These unweighted estimates are therefore biased even for the finite population of Table 8.

On the other hand, the usual stratified (weighted) estimator, $\bar{y} = \sum w_i \bar{y}_i$, is design-unbiased, where $w_i = N_i/N$ is the stratum weight and $\bar{y}_i$ is the simple average of the n_i sample observations in stratum i. When applied to the proportions, say, of rainfed animal-powered units, $\bar{y}_i$ is replaced by $p_i' = n_i'/n_i$ where n_i' denotes the number among the original n_i sample elements in stratum i that were found to be rainfed animal-powered. (These are 38/43, 23/36, ..., 0/37, as can be verified from Table 8 and 9). Hence,

$$\bar{p} = \sum w_i p_i' = 0.382$$

is a design-unbiased estimate of the proportion of rainfed animal-powered farms. Its sampling error estimate is

$$\text{s.e.}(\bar{p}) = \{\sum w_i^2 p_i' (1-p_i')/(n_i-1)\}^{1/2} = 0.018$$

For comparison purposes, estimates of $\tilde{p}$, $\bar{p}$ and standard errors for the other strata (also considered here as domains) are given on Table 10. In exchange for their easy computability, the cost of using the unweighted estimates can be severe:

(a) The biases cannot be assumed away as negligible. The average value of the relative biases $|\bar{p}_i - \tilde{p}_i| / \text{s.e.}(\bar{p}_i)$ in the nine strata is 1.66.

(b) The standard errors of the unweighted estimates tend to be higher than those of the weighted estimates. The former do not truly reflect the survey design's effects, i.e. they would be statistically appropriate had the observed frequencies in Table 9 come from a single srs of size 320 from the 893 population units. A measure of the soundness of this simplication is given by the ratio s.e. $(\bar{p}_i)$/s.e. $(\tilde{p}_i)$, with values close to unity being supportive of the adoption of unweighted sampling errors. This ratio of the standard error computed in accordance with the survey design to the standard error under srs is called the design factor (deft) by some authors (see e.g. Verma, 1982); its square, or the ratio of the corresponding sampling variances, is called deff, an acronym for design effect, due to Kish (1965). For details on further uses of deff, see Kish and Frankel (1974). The ratios range from 0.33 to 1.18 for the nine strata (domains), with a mean of 0.75. Thus, on average, s.e. $(\tilde{p}_i)$ exceeded the statistically correct s.e. $(\bar{p}_i)$ by a non-negligible margin.

(c) The use of unweighted estimates in this particular case therefore can be misleading. For example, "large sample" 95 per cent confidence intervals about the proportion of rainfed animal-powered farms are given by

$$(0.303 \pm 1.96 \times 0.026) = (0.25, 0.35)$$

and

$$(0.382 \pm 1.96 \times 0.018) = (0.35, 0.42)$$

from the unweighted $(\tilde{p})$ and weighted $(\bar{p})$ estimates respectively. Thus, if it were of interest to speculate whether there was a change in the proportion from the initial 257/893 = 0.29, the first interval would support a status quo conclusion, but the second would point to an increase.

5.4 Inference Concerning Continuous Variables

A problem that needs settling first concerns the small frequencies in some domains of interest (strata in this case), 2 of which have 1 observation each and another three of which have less than 20 (see last row of Table 9. The usual solution is to collapse these in some meaningful way which would result in larger samples, e.g.

	Animal	Tractor	Total
One-crop	108	33	141
Two-crop	56	123	179
Total	164	156	320

The one-crop domain includes rainfed and irrigated one-crop categories. Again, researchers occasionally assume that these four frequencies represent simple random samples and proceed with an unweighted analysis.

Table 10. Unweighted and Weight Estimates of Domains and their Sampling Errors, Nueva Ecija Survey

Domain[a/]	$\tilde{p}$	s.e. ($\tilde{p}$)	$\bar{p}$	s.e. ($\bar{p}$)
(1)	0.303	(0.026)	0.382	(0.018)
(2)	0.056	(0.013)	0.051	(0.013)
(3)	0.041	(0.011)	0.042	(0.013)
(4)	0.034	(0.010)	0.018	(0.005)
(5)	0.003	(0.003)	0.002	(0.002)
(6)	0.003	(0.003)	0.001	(0.001)
(7)	0.175	(0.021)	0.146	(0.017)
(8)	0.231	(0.023)	0.240	(0.018)
(9)	0.153	(0.020)	0.118	(0.016)
Total	1.000	-	1.000	-

a/ For illustration purposes, the domains were chosen to coincide with the stratum classifications (see Table 8 for descriptions).

Thus, if x denotes net farm income,

$$\tilde{x}_{oa} = \Sigma x_k/108 \text{ and } \left[\Sigma(x_k - \tilde{x}_{oa})^2/107\right]^{\frac{1}{2}}$$

are estimates of the average net farm earnings and its standard error respectively of the one-crop animal-powered (oa) farms. These and similar estimates for the other domains are given in Table 11.

One can think of the actual net income among the oa farms as X_{oa}/N_{oa}, where N_{oa} is the number of oa farms (in the finite population) and X_{oa} is the sum of their net incomes. Under assumption (b), section 5.2, this ratio is a consistent estimate of the corresponding parameter of the superpopulation. However, both numerator and denominator are unknown and need to be estimated from the sample. If n_i' denotes the number of oa farms among the n_i sample units in the original i^{th} stratum, the quantity N_i (n_i'/n_i) estimates the number of oa farms among the original N_i units, and the sum across strata

$$\hat{N}_{oa} = \Sigma_i N_i \ (n_i'/n_i)$$

is unbiased for N_{oa}. It can be verified that $\hat{N}_{oa} = 356$, which is almost 20 percent higher than the unweighted srs estimate, 893(108/320) = 301.

Similarly, with x_{ik} as the net farm income of the ik^{th} sample unit, where i denotes the strata and k the units within the strata, define

$$x_{ik}' = x_{ik} \text{ if the unit is oa,}$$
$$= 0 \text{ if not.}$$

The sample mean, $\bar{x}_i' = \Sigma x_{ik}'/n_i$, is an unbiased estimate of the average of the x_{ik}' among the original N_i units in the i^{th} stratum, $N_i\bar{x}_i'$ estimates the total, and the sum

$$\hat{X}_{oa} = \Sigma_i N_i \bar{x}_i'$$

is unbiased for X_{oa}. Hence,

$$\bar{x}_{oa} = \hat{X}_{oa}/N_{oa}$$

is a consistent ratio estimate of X_{oa}/N_{oa}. It is well-known that the variance of $\bar{x}_{oa}$ is more complex and involves a non-zero covariance term (see e.g. Cochran, 1977, Chapter 6):

$$\text{var}(\bar{x}_{oa}) = \hat{N}_{oa}^{-2} \Sigma(N_i^2/n_i)\left[s_i'^2 + \bar{x}_{oa}^2 p_i'(1-p_i') - 2\bar{x}_{oa}\bar{x}_i(1-p_i')n_i/n_i-1)\right]$$

where $p_i' = n_i'/n_i$ and $s_i'^2 = \Sigma(x_{ik} - \bar{x}_i')^2/(n_i-1)$.

Numerical results for the four domain are given in Table 12. Notice the differences between these and the unweighted estimates in Table 11 although the two sets lead to overlapping confidence intervals (or to the same conclusions on tests of hypotheses about means) owing to the large standard errors.

Inferences concerning differences between domain means, say, between one-crop animal-powered (oa) and one-crop tractor-power (ot) farms, require the variance of the difference, var $(\bar{x}_{oa} - \bar{x}_{ot})$. Since both types of farms are found together in some of the original strata, the covariance between $\bar{x}_{oa}$ and $\bar{x}_{ot}$ will not be zero; thus var $(\bar{x}_{oa} - \bar{x}_{ot})$ will have a forbiddingly complex form. Fortunately, the covariance term tends to be relatively small because it involves the product $p_{i(oa)}' p_{i(ot)}'$ (Kish, 1965, p. 135); hence,

$$\text{var}(\bar{x}_{oa} - \bar{x}_{ot}) = \text{var}(\bar{x}_{oa}) + \text{var}(\bar{x}_{ot})$$

usually is a reasonable approximation.

Table 11. Unweighted Estimates of Average Net Farm Income and Standard Error (Pesos)

	Animal	Tractor
One-crop	1218 (207)[a]	1102 (553)
Two-crop	2110 (376)	3279 (374)

a/ Standard error.

Table 12. Ratio Estimates of Average Net Farm Income and Standard Errors (Pesos)

	Animal	Tractor
One-crop	1126 (250)a/	827 (285)
Two-crop	2110 (422)	3397 (305)

a/ Standard error.

6. Summary

In summary, cross-classification causes complications in estimation, but neglect of this complexity in favor of simplifying assumptions such as srs-based estimation can lead to serious mistakes. (Kish, 1965, Chapter 14 and Kish and Frankel, 1974 present numerous examples). Cross-classification also leads to higher sampling errors, resulting in loss of power or sensitivity of significance tests. For example, the standard errors in Table 12 average to roughly 300 pesos; hence an observed difference between two domain means has to exceed $(1.96)(300)\sqrt{2} = 832$ pesos before the true means can be declared "significantly

different." Note, however, that this threshold value already exceeds one of the means in Table 12. (Likewise, earlier results from the Davao III Survey for rice area and production show a 30 to 100 per cent increase in sampling variance due to cross-classification compared to when the domains coincide with the strata (David, 1982b)).

Exceptions to the above-mentioned general conclusions are: (a) when the domain frequencies (say, M_i) are known, in which case the straightforward stratified estimates with N_i replaced by M_i can be used with little or no loss in precision, or (b) when the proportion of the domain units in the population is high. Durbin (1958) observed that most of the advantage of stratification will be lost if this proportion is small, while only if the proportion is close to unity will the advantage be retained.

7. References

Amemiya, T. 1982 Tobit Models: A Survey. Rhodes Associates Criminal Justice Research Series 4. (Palo Alto, California)

Asian Development Bank. 1980, 1984 Guidelines on Logical Framework Planning (LFP) and Project Benefit Monitoring and Evaluation (PBME). Manila

Casley, D. J. and D. A. Lury. 1981. Data Collection in Developing Countries. Clarendon Press, Oxford.

Cochran, W. G. 1939. The Use of Analysis of Variance in Enumeration by Sampling. Jour. Amer. Statist. Assoc. 34, 492-510.

_____ 1977. Sampling Techniques. (Third Edition), John Wiley, New York.

_____ 1982a. Two-Stage Sampling with Composite-Like Selection of Second Stage Units. Survey Statistician. No. 7, April, 11-13.

_____ 1982b. A Critique of Survey Sampling Practice and Use of Survey Data in Social Science Research. The Philippine Statistician. XXXI, 3-25.

Deming, W. E. 1960. Sample Design in Business Research. John Wiley.

_____ and F. Stephan, 1941. On the Interpretation of Censuses and Samples. Jour. Amer. Statist. Assoc. 36, 45-49.

Durbin, J. 1958. Sampling Theory for Estimates Based on Fewer Individuals than the Number Selected. Bull. Intern. Statist. Institute 36, 113-119.

Hartley, H. O. and R. L. Sielken, Jr. 1975. A Super-Population Viewpoint for Finite Population Sampling. Biometrics. 31, 423-447.

International Rice Research Institute. 1982. Consequences of Small Farm Mechanization Project Operations Handbook: A Documentation of Sampling and Post-Sampling Procedures and Data Processing. Los Banos, Philippines.

Kish, L. 1965. Survey Sampling. John Wiley, New York.

_____ and M. R. Frankel. 1974. Inference from Complex Samples. Jour. Royal Statist. Soc. Series B, 36, 1-37.