FAO STATISTICAL DEVELOPMENT SERIES 3

SAMPLING METHODS FOR AGRICULTURAL SURVEYS

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
Rome, 1989

Reprinted 1998

M-70
ISBN 92-5-102748-X

FOREWORD

The Statistical Development Series is a sequence of comprehensive technical manuals on various aspects of the statistical programmes which make up a national information system for food and agriculture. The volumes **Food and Agricultural Statistics in the context of a National Information System. Programme for the 1990 World Census of Agriculture** and its **Supplement for Europe**, and **Microcomputer-based Data Processing** have already been published. An additional volume, **Guidelines on Socio-Economic Indicators for Monitoring and Evaluating Agrarian Reform and Rural Development** is in preparation. In the Statistical Development Series emphasis is placed on the need to conceptualize data sources within the framework of a national information system which requires standardized concepts and minimizes duplication of effort.

The publication, **Sampling Methods for Agricultural Surveys**, is intended to assist statisticians in their work on designing agricultural sample surveys. The manuscript was prepared for FAO by Professor Leslie Kish, Institute for Social Research, the University of Michigan.

Helmut Schumacher
Director, Statistics Division

TABLE OF CONTENTS

CHAPTER 1. SCOPE AND LIMITATIONS

1.1 SAMPLE DESIGN AS PART OF SURVEY DESIGN

This volume concentrates on sample design, which covers only part of the design of sample surveys. The field of survey design is broader because it also includes other aspects that precede, join and follow the design, selection and collection of samples. Those other aspects are listed in Table 1.1.1, with the understanding that the separation of its three classes is suggestive, not dogmatic. There should be interchange of ideas on all these aspects, but the group labeled sample design is mainly the responsibility of sampling statisticians and the chief concern of this book.

The class labeled "joint design" also concerns sampling, but these decisions must be shared with the subject matter experts (SME) directing the survey, who may be economists, sociologists, agronomists, biologists, etc. Choice of the population elements and its extent should begin with those SME, but the knowledge and advice of the samplers should be used, to restrict the population as necessary or to expand it as desirable and feasible, until a good design is agreed on. The estimation process (or estimator) is often included with the sample selection as joint aspects of the sample design; and unbiased estimators, for example, can only be defined by considering jointly the selection probabilities with the weights used in the estimator. However, the statistical analysis is typically combined with substantive analysis in practice; and they are both often separated by gaps of time from the selection process and from the efforts of the sampling statistician. Also, statistical analysis is the subject of most of the statistical literature and it would be futile to try to cover it in a sampling textbook, much less in this manual. Sampling errors also depend strongly on the selection design and sampling theory is needed for them, but the inputs of the SME, who are responsible for the presentation and utilization of data, are needed for the proper presentation of sampling errors. The allowed cost and the desired sizes and precisions for sample results generally come

from or through the SME, but the samplers may well influence them, because conflicts usually develop between desirable goals and the restrictions of resources. The skills and knowledge of samplers may also be utilized for dealing with problems of nonresponses and noncoverage, although these are largely functions of data collection in the field, which is largely beyond the samplers responsibilities.

Table 1.1.1. Relation of Sample Design to Survey Design

Sample design

Selection methods and procedures.
Sampling Units: choice, designation and identification.
Allocation of sample sizes to stages of sampling units.
Stratification and allocation (sizes,rates) to strata.

Joint design

Designation of population and elements.
Cost, size and desired precision.
Estimation and statistical analysis.
Computation and presentation of sampling errors.
Nonresponses, noncoverage, weights and imputation.

Survey design

Survey variables: choice, definition, measurements, observation.
Data collection, coding, processing, computing.
Substantive analyses: Methods and models.
Domains of analysis: choice, definition.
Response errors and biases.
Presentation of data, statistics, and results.
Utilization of results.

Because we concentrate on survey sampling, the related fields of experimental design, observational studies, census and registers are not covered. But many of the methods developed here are also useful in and relevant to those other methods of research [Kish 1987].

1.2 AGRICULTURAL SURVEYS

Agricultural surveys cover a great diversity of variables and a list may be useful even if incomplete.

1) Land areas and tenure.
2) Crop acreage and production, including pastures.
3) Vegetables, fruits, nuts.
4) Livestock, poultry, barns and pens.
5) Fishing, hunting and timber only as part of farming operations, not distinct industries.
6) Tool, machinery, fertilizers, pesticides, seeds and other inputs.
7) Irrigation, wells, drainage, fencing as parts of farms.
8) Income, marketing, expenses, savings and other economic data about agricultural production and population.
9) Population counts and characteristics; unpaid and hired labour.
10) Health, education, occupation, and social statistics of agricultural population.
11) Farm homes and buildings.
12) Transportation, communication of farm population.
13) Food sources and food consumption.
14) Attitude surveys about policies, methods, products, etc.

Other data of a great variety may also be collected as auxiliary variables related to the principal agricultural variables noted above. The sources of the data may be agricultural holders and operations, or they may be the agricultural households for other data. These are mostly small and numerous units. But sometimes large units (such as sugar mills, warehouses, grain elevators etc.) may also have to be included. These are unusual and very diverse, and they can only be covered generally and superficially. Chief attention must be placed on the many small farms and households.

Attention will also be focused on domesticated animals, fisheries and plants, as grown, raised or focused on farm operations. No separate attention is paid to maritime fishing, lumber and forestry, or hunting and trapping of wild animals.

Household surveys for agricultural income and food consumption may be included. Also the connection of agricultural surveys to integrated household survey operations must be noted also (Chapter 17). Most agricultural surveys are more restricted and omit some of the types listed above. On the other hand, some surveys may include other agricultural items of special interest in the country and attach non—agricultural items in integrated surveys. The list cannot and need not aim at being complete: Designating variables is not primarily a sampling task, as distinguished in Table 1.1.1.

1.3 MULTISUBJECT AND DIVERSE SURVEYS

Agricultural surveys are usually most difficult and complex because that single word covers a tremendous variety of activities and purposes in four ways.

First, they are *multisubject* (Ch. 9) in essential ways because "agriculture" covers a great variety of distinct "industries". Growing maize, wheat and irrigated rice differ greatly, and these grains are very different from vegetables and truck gardens, and from fruits and nuts. Then also raising sheep, goats, dairy cows, range cattle, pigs, poultry, rabbits and fish all describe different occupations. Then come the economics of buying seeds, fertilizer, tools and machinery, and of the even more varied activities of selling, including marketing in town. And so on down the list of 14 in the table above. Furthermore, each of these subjects also becomes *multipurpose*, because several variables are often required, and each of these for several domains (Ch. 9).

Second, often they must also be *multi—method* because different variables and subjects need drastically different methods of measurement. For example, consider the different skills needed to measure crop areas, crop yields (before and after harvest), counting animals, accounting for expenses, income, savings and loans. Then think of the household interview about age, sex, education, health items. And so on. Sometimes one set of field survey workers may be trained for all these skills, but in other situations two or more sets of

enumerators must be trained, employed and coordinated. The need for several languages may be mentioned here, but each village tends to use one language, which can often be ascertained.

Third, both *natural conditions and cultural norms* impose even greater variety than required by the several subjects and variables of each survey. The growing of any crop (e.g. rice) varies greatly between countries and between regions and even districts of the same country. Differences may be related to climate, moisture or irrigation, but also to tenure, economics, religions or ethnic cultures.

Fourth, *repeated or periodic surveys* (Ch. 16) are often needed for collecting agricultural data. Different crops mature in different seasons (including animal products) and often the same field may produce several crops of different or even of the same kind. In many cultures retrospective surveys over the whole year will not produce data of acceptable quality, because neither the records nor memory are good enough. Then repeated surveys may be planned to coincide perhaps with harvest (or birth) times, or sometimes with traditional marketing dates.

1.4 FOR DEVELOPING COUNTRIES (DC'S)

Although our treatment will be as general as feasible, chief attention must be paid to the hundreds of millions of peasants in the less developed countries, mostly in Asia, Africa and Latin America. First of all they are much more numerous, because the DC's have greater total population, and also greater proportions in agriculture, compared to the fast decreasing agricultural sectors in the industrialized countries of Europe, North America and the Western Pacific. Second, the wealthier, more developed countries are also likely to have already developed their own methods and offices for agricultural surveys and censuses. For these reasons their needs for basic instruction in sampling seems less pressing now.

The differences between the two situations are many and great, though many individual overlaps and exceptions can be found. There are great differences in income, expenses, lifestyle and type of farming. The differences with greatest impact on methods depend most probably on the prevalence of telephones, automobiles, and roads, a well developed transportation and marketing system. This last aspect also greatly affects home economics, food consumption and market reliance. The health and literacy status and social organization differ greatly. For all these reasons transfer of techniques of agricultural surveys from industrialized countries to DC's is risky and difficult.

There are great differences of agricultural practices between less developed countries as well as between regions and even districts within them. These concern nature of crops and growing practices, animal husbandry, type of settlement, such as villages, versus open country living, nomadism; etc. In many DC's there are also areas and groups with modern agricultural practice.

1.5 SIMPLE AND BRIEF

This manual is designed to be taught in a course of 2 or 3 weeks. It should also serve as a brief source of reference for practitioners of survey sampling. Also, both those uses should be available to agricultural officials who are not sampling experts, nor necessarily professional statisticians. However, it does assume, with its combination of brevity and depth, about 2 or 3 courses in statistics.

It must also assume knowledge of agricultural practices in order to be able to translate the general recommendations of this manual into the many specific variables and subjects referred to in 1.3 above, and in terms of specific situations and practices of their home countries. It aims to be entirely practical and concerned only with sampling aspects of agricultural surveys, as delimited in 1.1 and 1.2.

The style aims at being simple and direct so as to be translatable and comprehensible by readers of many countries. Because of its brevity it cannot be comprehensive, hence it attempts to concentrate on the most broadly useful methods. Thus it may bypass methods of theoretical interest and intellectually stimulating novelties. Even for the methods presented, often not all procedures can be developed in detail. The missing procedures can be developed in practice, or they may be found in the many references provided.

The manual is not mathematical and derivations of the formulas must be omitted for technical reasons as well as for brevity. These are provided in references to several books, but only one or two references for each formula. But most derivations can be found in several textbooks on sampling, and the readers should already have chosen one or two favorites among Cochran; Hansen, Hurwitz and Madow (HHM); Murthy; Sukhatme; Deming; Yates; and Kish.

CHAPTER 2. POPULATIONS AND ELEMENTS

2.1 DEFINITIONS FOR SURVEY UNITS

The *elements* of a population are the *elementary units* for which information is sought in a survey, and about which inferences are made. They are the *units of analysis* and their nature is determined by the survey objectives.

The *population* is the aggregate of the elements, defined jointly with the elements, and in terms of a) content, b) units, c) extent, and d) time. For example, the a) area in rice production of b) holdings in c) province P or country C in d) 1988. Most, surveys may yield information about several populations. For example, a) the content may include other crops, livestock, poultry, income etc.; b) the units may be subdivided or combined; c) the extent may be divided into subclasses, but also combined into national statistics; and d) data may be obtained for separate months and for several years.

For agricultural surveys the populations usually comprise one or more of *three kinds of elements*.

a) *Holdings*, or farms, or farm operations are the most common names used for the lands, crops, animals, buildings etc. involved in *agricultural operations*. The names differ between cultures, and holding is a neutral, international compromise.

b) The *holder* is a person (civil or juridical) who operates a holding, exercises management control over the holding operation, and makes major decisions regarding resource use. He may be called operator, or farmer; and peasant is still used widely in Asia, Africa and Latin America, less often in Europe.

c) A *dwelling*, and (its occupants) a household, with a homemaker, and a head of household, may be identifiable with each holding and holder. These may be basic units for food preparation and consumption, for income, and expenditures, and for social activities.

Identification of these three kinds of elements may be multiple, not necessarily one—to—one. For example, a holder may operate two holdings; and the holder may include two partners (brothers), living perhaps in separate dwellings. Definition of population elements is also a problem for censuses, and beyond the sampler's realm; for example, the definitions of holders, or holdings, or parcels. However, sampling considerations may indicate the use of dwellings for practical identification, although their occupants, the households, comprise the population.

Many *variables* can be usually measured on each element, and we can think of a *vector of p variables* (Y_{1i}, Y_{2i}, Y_{3i}, ... Y_{pi}) *associated with each element* i (i = 1,2,... N) in the population. Other subunits may be usefully noted, such as separate parcels in one holding, and animals, machines, buildings of the holding. These may denote the "contents" of each unit noted above. They may be considered as the elements on some surveys.

Observational units are sources for data about elements and variables and they are called *respondents* in interviews and questionnaires. For example the holder gives information about cattle as elements, and the homemaker informs about the children's health, vaccination, and education.

Sampling units contain the elements and they are used for sampling elements. Each sampling unit contains only one element (or none) in *element sampling* (Ch. 5), but the clusters used in *cluster sampling* (Ch. 6) may contain several, and often many, elements. *Listing units* are used to identify and select sampling units from lists or frames (Ch. 4).

Four Location types are important features of agricultural surveys, and these relate the locations of the holders' dwellings to the locations of their holdings. Usually either a) or b) predominates in any country or province, but the other three must also be covered.

a) In *open country* settlement most agricultural dwellings are located on the holding, as on the farms in the USA and in some African countries.

b) Most holders may live in *villages*, from which they travel daily to work their croplands. The animals may live mostly close to the dwellings in villages.

c) Living in *towns and cities* occurs in most countries, and is common in some.

d) *Nomadic* living may mix two of these types in seasonal migrations.

Definitions of what agricultural operations to be covered in any census or survey are varied between countries and sometimes even between surveys. Section 1.2 hints at some limits for what may be included in agriculture. Lower limits on sizes of operations to be covered in surveys of production are also needed, to exclude very small producers who would be too difficult and too costly to include for their negligible contribution to the total yield. Commonly small vegetable gardens and small pens for chicken, pigeons and rabbits in cities and towns must be neglected.

Agricultural surveys (and others also) must provide statistics not only for global population, but also for separate domains (subpopulations) of several kinds: provinces and other administrative divisions; types of farming, age of holder etc. (Ch. 8). For agricultural surveys separate field procedures may be devised for three levels of technological development: traditional, progressive, and modern. Furthermore some types of holders may operate large and widespread holdings, and these often require and benefit from special frames and procedures (Ch. 11).

2.2 FOUR LEVELS OF POPULATIONS

Sampling from every type of population — holdings, holders, dwellings etc. — is subject to several imperfections, such as nonresponse and noncoverage. Therefore, it is useful to describe four levels of populations — survey, frame, target and inferential — separated by three kinds of imperfections that are common to most surveys.

Samples give direct evidence about the *survey populations* they properly represent. *Sampling errors* indicate the fluctuations of sample statistics (e.g. the mean $\bar{y}$) around the population value $\bar{Y}$ that a complete census would have yielded with similar survey methods; these errors can be estimated from the data of the samples themselves, when these are properly designed to be "measurable" (Ch. 14). It is preferable to have the survey population so defined as to also carry the differences due to *item nonresponses*, which differ between variables; and these may be "adjusted" with imputations or weights (Ch. 15).

The *frame population* covers the elements from which the sample was actually selected, but it is larger than the survey population by the amount of *total nonresponses* due to not—at—homes, refusals etc.

The *target population* differs from the frame population by the amount of *coverage errors*: the frame population is smaller by the *noncoverage*, which may also be called missing units, or incomplete coverage. However, the frame may also suffer from overcoverage of units from outside the target population, and the algebraic difference of noncoverage minus overcoverage is the net undercoverage. It is worthwhile to distinguish these populations and these imperfections because of practical differences in their effects on statistics. The extent of noncoverage is difficult to measure; but for nonresponses the amounts may be counted, though their effects may be obscure; and effects from item nonresponses may often be better adjusted (Ch. 15). Another practical difference: developed countries may often have smaller noncoverage than DC's, but larger nonresponses, especially refusals. Noncoverage is due to

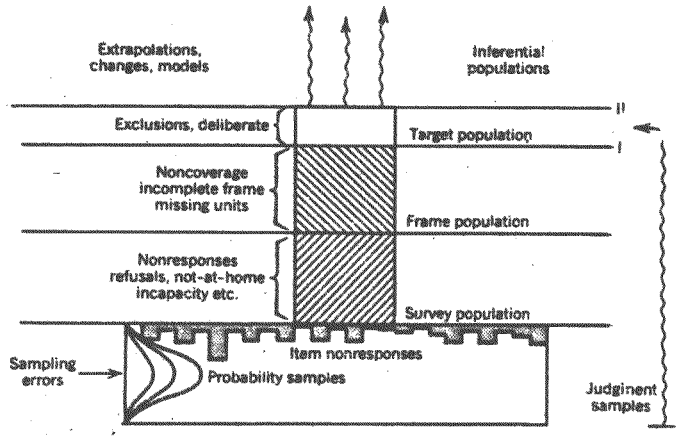


Figure 2.2.1 Discrepancies between four populations. [Kish 1987, 2.1]

Probability samples underlie the achieved survey population, but two discrepancies come even between them: sampling errors and item nonresponses. Both of these differ greatly among variables and the amount of item nonresponse is shown as differing greatly among variables. For both of these discrepancies the sample responses serve as bases; sampling errors are computed from them; and they are used for "imputing" or weighting for item nonresponses.

Thus probability samples are shown as a broad and solid foundation for the survey population, on which to build the structure of the inference above it. For the discrepancies beyond the survey population one must go beyond the sample data, with the help of implicit or explicit data. The span to the frame population is due to *total nonresponse* of diverse kinds (refusals, not–at–homes, etc.); the size of nonresponses may be estimated from sample records (with effort and care), but estimating their effects needs models and auxiliary data. The size of *noncoverage* can only be estimated with models or from checks with outside sources, yet this portion also belongs to the target population. This may also include a defined and deliberate *exclusion* from the coverage.

Furthermore, sample data are also used for inferences beyond the target populations, and these are many, various, and ill defined. "Superpopulations" of sampling theory are not only among these, but behind all these inferential populations. These model–dependent inferences (1.8) are too often merely implicit. Even more vagueness describes the path of judgment samples directly to the target population, and such vagueness is indicated by the thin, wavy, population line, as for the extrapolations to vague inferential populations.

faulty frames, and this imperfection should be distinguished from *deliberate exclusion* of part of the target population for practical, economic or tactical reasons; for example, a province though perhaps large in territory or even in total population may be excluded if it has too few holders, or is subject to rebellion, or too distant or may otherwise not be economically approachable.

Thus we ascend from the sample to the survey population, then to the frame and finally to the target population that is initially designated for the sample design, and these are three conceptually specified populations. But inferences must also be made from the target population to a wide variety of other populations. For example, from the statistics for one year, inferences are also made into the future and sometimes into the past. Also from statistics for one province, inferences may be made to national values; and often, on the contrary, inferences from national statistics are made to population values for provinces and other domains (Ch. 8).

Models are needed to span the gaps to inferential populations, though such models are usually only implicit and seldom explicit. Those leaps of inference differ greatly from the smaller and more explicit four steps needed to the target population. Designs for samples cannot possibly be planned for all the populations to which inferences will be made by all its users. But the four steps from the sample to the target population should be considered in the sample design. For example, an epsem overall sampling fraction should be designed for the planned sample size n in order to discount for anticipated proportions of noncoverage and nonresponse:

$$f = \frac{\text{n planned}}{\text{N (estimated)x(1}-\text{noncoverage)x(1}-\text{nonresponse)}}$$

2.3 ALTERNATIVE METHODS OF SAMPLING

Probability sampling requires *known nonzero* probabilities of selection ($P_i > 0$) for all elements ($i = 1,2,\ldots N$) listed in the specified frame population. Those probabilities must be assured with mechanical procedures of selection for all sampling units of the population, at each stage of selection. This manual is devoted to probability sampling and it is discussed in Section 3.1. Here we list some alternative methods that have been used for agricultural and other

sample surveys. The descriptions must be brief and incomplete, because they are beyond the scope of this manual, but also because the great variety of procedures prevents clear descriptions and definitions.

Quota sampling is widely known from political voting polls, but it has also been used in agricultural surveys. For example, from a (random?) selection of districts or census Enumeration Areas (EA's) the enumerators may be asked to fill quotas (numbers) of specified types and sizes of holdings. These numbers could be based on proportions from the last census, which could be obsolete and also based on discrepant definition of holdings. "Quota sampling is not one defined scientific method...Yet some general observations may enlighten readers ..." [Kish 1965, 13.7].

Judgment sampling denotes vaguely any method of selecting units – districts, EA's or holdings – based on *expert judgment.* These can be extremely varied, and thus they defy definitions and general descriptions. But even in specific situations they are difficult to evaluate scientifically.

Model dependent sampling has a larger place in theory than in the practice of agricultural or other surveys. It differs from judgment sampling in mathematical elaboration and in the more or less specified models for expert judgment (3.1).

Voluntary cooperators have been used for crop reporting systems to obtain rapid, economic reports with great geographic detail on the growth and yield of crops, on prices and other variables also. Although cooperating holders generally do not comprise a random (or representative) sample of the population, some hope that with ratios and adjustments their reports can reflect relative values of changes between reporting periods.

Fortuitous, accidental samples of holders may sometimes be persuaded to cooperate in a survey, after they are found as voluntary members of some association for marketing, religious practice etc. Expert opinions needed to

judge their representativeness link them with judgment sampling, or with voluntary cooperators, or with network sampling. These should not be considered equivalents of probability sampling.

Network or snowball sampling refers to procedures for using members of a sample to help identify associated, usually similar, members in the population. For example, growers of a rare crop, such as specific fruits or nuts, may identify similar growers in their own districts. Unfortunately, however, defining the probabilities of these identifications, hence of selections, remain unknown.

Sampling of time intervals has not advanced as far as sampling the space dimension, where probability sampling has been accepted as the standard. Sampling time is still mostly done by judgment selection of the best, or most representative, or some unique period. But sampling over time is possible in periodic samples and may be especially important for agricultural surveys (Ch. 16).

Restricted sites are still used when a widespread sample is not feasible for economic reasons. For example, instead of a sample spread over a national (or provincial) frame, one or a few "typical" districts may be chosen to "represent" the larger population. The inference from the few sites to the larger population must be based on expert judgment rather than on statistical inference, as we use that term [Kish, 1987, 3.1].

Telephone sampling and telephone surveys are two current methods that are often confused but should not be [Groves and Kahn, 1979; Groves et al., 1988]. Conducting agricultural and other interviews has long been done successfully by telephones in many countries, but interviewing is not the subject of this manual. The telephone numbers used may have come from area samples or from lists of special populations. On the other hand, random sampling of telephone numbers to reach households poses problems of lists and selections; it may be worthwhile in countries where 90 percent or more, but not

where less than 70 percent, have telephones. In any case to use telephone numbers to select agricultural holders does not seem practical, especially in the DC's. Not yet.

Mail Surveys from specially listed populations have had return rates that range from very bad to good, depending on several factors, such as the population (motivated, cooperative, literate?) the sponsor, the brevity of the questionnaire form, the number of repeat mailings, etc. [Dillman 1978, Kish 1965, 13.4]. "Mailbacks" of questionnaires left at dwellings on surveys and censuses can receive large returns from cooperative and literate populations. But sample selection of mailing addresses for agricultural surveys does not seem feasible, especially in DC's.

Mixed methods offer too many possibilities for a complete listing of all of them. Consider, for example, a judgment selection of four districts (or EA's), followed by good probability selections of holdings within each of the districts (or EA's). How would inferences be drawn from the sample of four districts to population values of the entire province or nation? On the contrary, sometimes districts (or EA's) may be selected with known probabilities, but selections within districts may be done with quota or other nonprobability methods.

2.4 POPULATION VALUES AND STATISTICS

Population values are expressions that summarize the values of some characteristics for all N elements of an entire population; they summarize some features in the defined population. The basic examples for survey sampling are the *population means* $\bar{Y} = \Sigma Y_i/N = Y/N$ and the population *totals* $Y = \Sigma Y_i$. Other examples are the element variance, either $\sigma_y^2 = \Sigma(Y_i - \bar{Y})^2/N$ or $S_y^2 = N\sigma_y^2/(N - 1)$, the covariance σ_{yx}, the correlation coefficient $R_{yx} = \sigma_{yx}/\sigma_x\sigma_y$, etc. Population values depend on four aspects of the population: 1) the specific target population, 2) the nature and distribution of the specified survey

variables, 3) the specified methods and procedures of measurement, and 4) the mathematical expression for summarizing the population value from individual element values.

True values denote numerical expressions that would be obtained from all element values in the population if these were free of errors of measurement. Thus the difference between the population mean and the true mean $\bar{Y} - \bar{Y}_{true}$ = Bias is the mean value of the errors of measurement. For unbiased measurements B=0 and $\bar{Y} = \bar{Y}_{true}$. The term *parameter* is often used for either of these, and to avoid confusion we may generally avoid it. The population value and the true value refer to concepts not observable in sample surveys, but they define the systematic Bias of measurement that is not affected by sample size. The population value would be obtained if the entire population of N elements, were designated for measurements under the same essential survey conditions as the sample of only n elements. Thus the sample statistics is linked through the population value and the Bias to the hypothetical true value being sought. This Bias is distinguished from the technical bias later.

The sample value, or *statistic* denotes a specific numerical *estimate* computed from the values of the *n* elements in a specific sample; for example, the sample mean $\bar{y} = \Sigma y_j/n$ of the n observed sample element values y_j. It is a *variate or random variable*, which depends on the sample design and on the specific elements that we selected into the sample. The particular estimate, or statistic, is only one among the many possible estimates that could have been obtained with the same sample design.

The *estimator* differs from the specific estimate from one sample: it refers to a defined method of sample selection and statistical estimation. The estimator applied hypothetically to a population would generate the *sampling distribution of the estimator* of all possible estimates, only one of which appears in the actual sample. The mean of this hypothetical distribution is the expected

value $E(\bar{y})$ of the estimator $\bar{y}$; and the standard deviation of this sampling distribution is the *standard error* $Ste(\bar{y}) = \sqrt{Var(\bar{y})}$ of the estimator $\bar{y}$, and $Var(\bar{y}) = Ste^2(y)$ is its variance.

The sampling distribution of an estimator $\bar{y}$ is the theoretical distribution of all possible values of the estimate ($\bar{y}_c$), each with its probability of occurrence (P_c). The possible values and their distribution depend on the sample design (size, selection, estimation) applied to the population distribution (which depends on the four population factors noted above). The mean of the sampling distribution is its expected value: $E(\bar{y}) = \Sigma_c P_c(\bar{y}_c)$ and the deviation of a specific estimate (statistics) $\bar{y}_c$ from its population value has two components:

$$\bar{y}_c - \bar{Y} = [\bar{y}_c - E(\bar{y})] + [E(\bar{y}) - \bar{Y}].$$

The component $[E(\bar{y}) - \bar{Y}]$ is the *sampling bias of the estimator* $\bar{y}$. When the mean of the sampling distribution equals the population mean, $E(\bar{y}) = \bar{Y}$ the sampling bias $[E(\bar{y}) - \bar{Y}] = 0$, and $\bar{y}$ becomes an *unbiased estimator*. We distinguish this technical statistical bias from those for the measurements $Bias = (\bar{Y} - \bar{Y}_{true})$, which are often more serious and more difficult to assess.

The *mean square error* of $\bar{y}$ defines the deviation of a specific estimate of $\bar{y}_c$ from the population value. The average of the squared error in the sampling distribution is

$$MSE(\bar{y}) = \Sigma_c P_c(\bar{y}_c - \bar{Y})^2 = \Sigma_c P_c[\bar{y}_c - E(\bar{y})]^2 + [E(\bar{y}) - \bar{Y}]^2 = Var(\bar{y}) + bias^2.$$

The variance $Var(\bar{y})$ depends on the entire sampling distribution and remains unknown. But from measurable samples we can compute estimates $var(\bar{y})$ for estimating $Var(\bar{y})$, as discussed in Ch. 13. Furthermore with the standard errors $ste(\bar{y}) = \sqrt{var(\bar{y})}$ we compute confidence intervals $\bar{y} \pm t_p ste(\bar{y})$ for inference from the statistics ($\bar{y}_c$) to the population value $\bar{Y}$ (3.5). As a basic example, in simple random sampling(*srs*) the value $var(\bar{y}) = (1 - f)s_y^2/n$ is computed from the n sample elements and is used for $ste(\bar{y}) = \sqrt{var(\bar{y})}$. This $var(\bar{y})$ is an estimate of $Var(\bar{y}) = (1 - f)S_y^2/n$ and this is a mathematical

expression for $\Sigma_c P_c(\bar{y}_c - \bar{Y})^2$. For srs both $\bar{y}$ and var($\bar{y}$) are unbiased estimators, but these unbiased qualities do not hold for most other epsem and probability samples (3.4–3.5).

In this simplified view of the sampling process the errors of measurement have been disregarded, except for the Bias noted above, but they are considered in Ch. 15.

CHAPTER 3. FOUNDATIONS

3.1 PROBABILITY SAMPLING WITH CHANCE PROCEDURES

Probability sampling assures for each element in the population (i = 1,2,...N) a *known positive* probability ($P_i > 0$) of selection. This assurance requires some mechanical procedure of chance selection, rather than only assumptions, beliefs, or models about probability distributions. The randomizing procedure requires a practical physical operation which is closely (or exactly) congruent with the probability model. The most common and best known chance procedure consists of the proper use of a good table of random numbers; but now computer programs can often replace selection by hand. The statistical inference in sample surveys depends on this chain of requirements: statistical inference → measurability → probability sampling → mechanical selection → lists or frames. Measurability will be touched on in 3.7 and based on sampling errors in CH. 13. The need for and use of frames are discussed in CH.4.

Simple random sampling provides a basic standard: From a table of random numbers select *n*(different) numbers independently from 1 to N. Each of the N population elements is identified in the list (frame) with one of those N numbers, and thus receives the equal probability $P_i = n/N = f$ (the sampling fraction) of being selected. The n "different" numbers produce srs *without*

replacement; otherwise, with repetitions permitted, srs *with* replacement results. This shows the survey sampling equivalent of *I.I.D*, the "identically and independently distributed" random variables of statistical theory.

Without independence of the n selections, there are other kinds of equal probabilities selection methods, or "*epsem*", with $P_i = f = n/N.$, For example stratified or systematic selection of elements (CH. 5) or clustered selections (CH. 6). When the sampling units can have several (or many) identifying numbers *unequal* selection probabilities results (P_i variable), instead of epsem. For example, if the ith unit receives m_i identifying numbers as its measure of size, with n selections it receives $P_i = m_i f$ selection probability, and "probabilities proportional to its measure of size, m_i, (or PPS).

This framework of survey sampling is often called "finite population theory". However, the finiteness of N elements in the population is not its crucial characteristic. For example, by sampling with replacement, the finiteness of N can be overcome in theory, though it seldom is in practice. The theory of survey sampling is distinctive because it is *population bound*: bound to the target population (2.2) and to mechanical selections from population frames (Ch. 4). Thus it differs from the *model–dependent* framework of some writings in the theory of sampling, which approach the theory of random variables [Kish 1987, 1.8]. In this framework the sample can be viewed as arising from a "superpopulation". The population–bound framework can and should admit a theoretical superpopulation that gives rise to the target population and also to the populations of inference. However, statistical inference in survey sampling links the sample only to the frame or to the target population, and model–dependent inference (or superpopulation theory) is needed only from then on (2.2).

The essential distinction of probability sampling lies in its insistence on mechanical selection of sampling units; and on basing the estimation of sampling errors on sample data in full accord with the sample design [Ch. 13]. However, models and expert judgments must be used to deal with designs and with imperfection.

3.2 MODELS FOR DESIGNS, BIASES AND INFERENCE

Models and the judgment of substantive experts are needed even for probability samples, and especially for two broad aspects: for designing samples at their start and for treating imperfections at the end.

The design of samples involves in all its many aspects population values (parameters) that the statistician cannot know, cannot even estimate well before completing the survey, and therefore must guess. For example, the desired overall *sampling fraction* (or rate) may be expressed as $f = n/N = D^2S^2/Ste^2(\bar{y})N$ for estimating the mean $\bar{y}$. The element variance S^2 and the design effect D^2 must be guessed, after determining the population size N and the desired standard error $Ste^2(\bar{y})$. On the other hand, the sample size n and fraction f are often based instead on the allowed total cost cn; then the $Ste^2(\bar{y})$ must be guessed from that (Ch. 9).

Many other aspects of the design process require guesses about unknown parameters and these will be mentioned in the appropriate chapters. The *stages of selection* used in the design must be determined, together with the nature and numbers of sampling units in each. The *numbers of selected units* at each stage are important design factors that must be determined also (Ch. 6). The design of *stratification* for each stage – the choice of variables, the number of strata, sampling fractions – also require judgment. Furthermore, it would be desirable to base all these choices on *multipurpose* considerations, not only for one single statistic, and their relative importance must be guessed. The multipurpose character of most surveys pervades all aspects of designs,

affecting all design parameters and this cannot be stressed too strongly, because it also multiplies the need for the use of models, judgment, guesses in the design of samples (1.3 and Ch. 9).

Errors in guessing design parameters are unavoidable, but fortunately their effects differ from mistakes in judgment in other sampling operations: *Errors in guessing sample design parameters reduce their efficiency but not the validity of sample statistics*. The statistics, including their sampling errors, are calculated from the sample results, hence they are not biased by mistakes and errors in the models. The design parameters can also be estimated from sample results to improve future designs. This situation differs from those created by biases due to errors of response, nonresponse and noncoverage, whose effects are difficult or impossible to assess well from the survey results.

Biases of nonresponse and noncoverage need models or judgment, first for the design stage before data collection, but also later for interpreting the research results and making inferences from them (CH. 15). The models needed differ with the amount, degree and nature of our ignorance. For each *item nonresponse* many other variables are available for the same individual that may be used for imputation of the missing item with models based on similar individuals. For *total nonresponses* the number of missing individuals is known plus perhaps some related variables by strata (subclasses) for reweighting. For *noncoverage* both the numbers and kinds of missing individuals are unknown without extraordinary and expensive efforts, and good models may be difficult to find and justify.

Measurement Biases lie mostly beyond the field of samplers and need the expertise of subject specialists, but statisticians may help with the construction of models. Statisticians can help even more with designs for measuring *variable errors of observation*, which increase in importance, relative to biases, with decreases in the sizes of subclasses (Ch. 15).

Methods of estimation and of statistical analysis may also need to be chosen from several possible alternatives. Also choice of auxiliary (ancillary) variables and of control variables, and of summary statistics all need decisions. Those decisions and choices may have to be based on imperfect information and models may be needed. Here the techniques of statistical analysis are available and must be utilized, together with expert knowledge of the substantive field.

A technical bias of some estimators occurs often. We distinguish technical biases denoted by $[E(\bar{y}) - Y]$, from measurement Biases denoted by $[\bar{Y} - \bar{Y}_{true}]$. Technical biases should be small and decreased by larger sample size for consistent estimates. Also models for their measurement and control are more readily available (Ch. 15).

Inferences from the target populations to other "inferential" populations need strong models. Statistical techniques may help, but broad knowledge of the subject matter is most crucial. Substantive, specific knowledge of the field should be combined with mathematical statistics to model causal systems that could produce the desired populations. "Superpopulation" is often used to denote a model population as the source (parent) of the separate populations involved. However, the approach in probability sampling insists on keeping distinct and separate from these models the inferences from samples to the target population (2.2).

3.3 LARGE, COMPLEX SAMPLES

Survey sampling is mostly concerned with large, complex samples from large, widespread populations. To cover large, complex populations a large survey organization is needed to design the sampling frame, select the sample and then to collect the data; and only large samples can justify the expense for engaging such large survey organizations. Fortunately those large samples permit reliance on simplifications based on asymptotic results which are needed for complex samples.

Simple selection methods may be used for sampling from small, compact populations. For example, for a population of 1,000 or 5,000 units (e.g. dwellings of a town, farmers of an area) a frame may be available or be prepared to facilitate both a probability selection and then the collection of data. Simple selection is also possible even for a large, widespread population if a good selection frame and easy collection methods are both available. For example, telephone sampling or mail surveys may be carried out for large populations if they are well listed and willing and able to become good respondents. This will seldom be true for agricultural surveys.

What are complex samples, and how and why are they complex? The nature and causes of complex designs are sketched in Table 3.4.1 and developed throughout this manual, but a brief summary is desirable here. a) *Stratified element sampling* is common, because stratification is usually preferred even when simple random sampling would be available, as for the simple populations cited above. Systematic sampling may also be used instead of stratified random. These complicate the analysis but often the effects may be mild enough to be disregarded (Ch. 5). b) *Cluster samples* pose the most important common problems in survey sampling. They are commonly needed for reasons of the costs of listing and of data collection. They also make for considerable increases in variances and in the costs of their computations. So they are troublesome but often unavoidable. Cluster samples can be selected in two or more stages, and stratification is commonly used in all stages of clustered and multistage samples. Variable sampling probabilities, and probabilities proportional to size (PPS) may be used in several stages (Ch. 6). c) *Two–phase* or multiphase sampling may be used for screening operations (Ch. 11). d) *Weighting* may often be used to compensate for unequal selection probabilities that may result either from frame imperfections or from deliberately designed allocations. *Weights can have drastic effects* on the estimates and on their variances.

What is large enough is even more difficult to specify, but the distinction must be made although no sharp boundaries separate large from small samples. A large number of elements alone n is not a sufficient guide because the *number of primary selections* must also be large enough for dependable results. For example, 3 or 4 districts will not provide a secure base for inference in survey samples even with large numbers of elements n, both because of large variances and because of the instability of variance estimates from few "degrees of freedom" (Ch. 14). *Outliers* may also cause problems, because even large numbers of elements may fail to include enough of these. These typically are caused by the few large elements on the extremes of skewed distributions; e.g. large farms, or large incomes.

The theory of survey sampling has been developed chiefly simple statistics like means $\bar{y}$ and aggregates $\hat{Y}$ and for inferences based on their standard errors ste ($\bar{y}$), as described briefly next (3.4).

3.4 MEANS AND STANDARD ERRORS

To cover large complex populations complex sample designs and operations are necessary. To justify those complex samples usually large samples are needed, and these are also needed for the accuracies and details required of the statistics yielded by those samples. Those design complexities prevent us from relying on the usual assumptions of I.I.D., and we must rely instead on the asymptotic assumptions for large samples. This background (often unstated) justifies the concentration of this manual, and of other textbooks on sampling, on estimates and their standard errors for inference.

The complex distribution of population elements interacts with the complex sample designs to produce complex samples. For example, agricultural area samples composed of selected districts and area segments show the clustering effects of soils, climate and farming practice. Such "design effects" have been found in thousands of surveys in agriculture, labor force, economic, social, health, education statistics etc. Social and other scientific

theories combined with mountains of evidence should persuade us that the world does not resemble the "well–mixed urn" of random, chance events assumed in probability and statistical literature.

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

Figure 3.4.1 The present status of sampling errors. Row 1 is the domain of standard statistical theory, and column 1 of survey sampling [Kish and Frankel 1974].

On the other hand, most statistical analysis techniques begin with "given n random variables" – either stated or implied. This assumption of the I.I.D. property facilitates the derivations for complex statistics, and that is the reason for the assumption of I.I.D., rather than any explicit belief in the randomness of either the selections or the populations of actual sample sets. Many derivations for small sample theory (like "Student's t") also assume normality of the population distribution. Nonparametric and robust statistics dispense with the normality but not with the I.I.D. assumptions. Those assumptions facilitate having one theory for small and large samples, whereas survey sampling must rely on asymptotic large sample theories, which are older [Yule and Kendall, 1965, Chapters 17–19].

Complex selection designs have quite different effects on the two classes of statistics that concern us: On descriptive or first–order' statistics like the mean $\bar{y}$ on one hand, and on inferential or second–order statistics, like their standard errors ste($\bar{y}$), on the other. For descriptive statistics the estimates in

probability sampling must represent the frame population. Hence the simple estimate of the population aggregate becomes $\hat{Y} = \Sigma_j y_j/p_j$ where y_j is the value and p_j is the probability of selection of the j—th sample element. For the mean this becomes $\bar{y} = \Sigma_j w_j y_j/\Sigma w_j$, where $w_j = 1/p_j$. But note that for the mean, and for other similar statistics, any convenient weights may be chosen, so long as they are inversely proportional to selection probabilities. For "self—weighting" samples from epsem selections where $p_j = n/N$ constant, we can use $\bar{y} = \Sigma y_j/n$, the simple, usual mean of sample cases.

Later we shall note modifications of these weights in order to make the estimates represent modifications of the frame populations (Ch. 12). Proportions and quantiles utilize similar weights. Furthermore, with these weights we may compute other descriptive statistics such as s_y^2, s_{yx}, r_{yx} etc., which are consistent estimates of similar population values S_y^2, S_{yx}, R_{yx} etc. Moreover, these simple estimates also hold for statistics for subclasses, which estimate similar values in domains (subpopulations) of the population. Thus *estimation for descriptive* (first—order) *statistics* seems relatively simple because it *may neglect the complexity of the methods used for selecting the sample.*

On the other hand, *estimates of inferential second—order statistics must reflect the methods of selection actually used.* The inferential statistics are confidence intervals and similar probability intervals such as Bayesian credible intervals, fiducial limits, tolerance limits), also tests of significance. These inferential statistics depend on the methods used for sample selection, because the sampling variability (the sampling distribution) of the statistics (like $\bar{y}$) depend on the selection design, which often has profound effects on that variability. That variability depends not only on the probabilities P_i of selection, but also on the joint probabilities P_{ij} of all pairs of elements in the population. These joint probabilities can vary greatly between the $N(N-1)/2$ possible pairs of population elements. For example, two elements (i and j) selected from the same *complete* clusters have joint probabilities $P_{ij} = P_i = P_j = P_\alpha$, (where P_α is the selection probability of the cluster),

instead of $P_{ij} \propto 2/N(N-1)$. *Computing inferential statistics must reflect the actual sample design.* They are based on sampling errors, whose computation for the many complex design is a most important function of survey sampling (CH. 13,14).

Standard statistical theory deals adequately with sampling errors for statistics from simple random selections as indicated on the top row of Table 3.4.1. Methods for computing sampling errors for complex selections tend to concentrate on relatively simple statistics, typically on the mean $\bar{y}$ and proportions p, which are also the most important for agricultural statistics; these means and the totals (aggregates) are in the first column of Table 3.4.1. Complexity of selections can take many forms, but we can conveniently separate stratified element sampling from all forms of clustered selections, because these have very different and often drastic effects on sampling errors.

The complexities of statistics are sorted into only three columns in Table 3.4.1. Subclass means and differences between them $(\bar{y}_c - \bar{y}_b)$ are used frequently in the critical analysis of survey data; sampling errors for the ste$(\bar{y}_c - y_b)$ are available, fortunately, as simple extensions of the methods for the ste$(\bar{y})$. The effects of stratification (cell 2B) and of cluster sampling (cell 2C) are often drastically different (and usually less) on ste$(\bar{y}_c - \bar{y}_b)$ than on ste$(\bar{y})$. The effects on sampling errors of more complex statistics, like ste(r_{yx}) and ste(b_{yx}) are more complex; and these are also more difficult to compute (Ch. 14).

3.5 CRITERIA FOR GOOD DESIGNS

Theoretical articles are often devoted to optimizing some single criterion. However, in practice sampling statisticians usually must balance the advantages and problems of several criteria, and that makes sample design less clearcut, more difficult, but also more interesting. Thus none of the five

criteria below has absolute dominance over the others and all five criteria must be balanced for each design. That balancing is art rather than science, because there is no criterion for choosing between the criteria.

1. *Probability sampling* to represent a defined frame population receives prime consideration in the manual (2.2,3.1), although it is circumvented in favor of other methods in some circumstances (2.3). The chosen frame population often falls short of the desired target population in order to better satisfy other criteria. For example, a strict probability sample of one district, or a few of them, may be less desirable overall than a somewhat weaker probability sample of the whole country.

2. *Measurability* refers to probability samples that permit the computation, from the sample data, of valid and close estimates of sampling errors. This is usually expressed in standard errors and in functions derived from them (CH. 14). These are the necessary bases for statistical inference from sample statistics to population values. Nonmeasurable samples, even if based on probability selections, cannot provide those objective measures of error. For example, selecting a single cluster (district, province) will not yield such measurability for national statistics; and two or four districts are not much better. A fairly large number of randomized replications identifiable in the sample are needed for measurability (Ch. 14). On the other hand, random replications of nonprobability samples (if ever properly designed) could yield sampling errors but about unspecified values only.

3. *Useful goals* are most difficult to write about, because they seem most obvious, but we must protect this criterion against possible neglect due to pedantic emphasis on other criteria. A national sample of holders may justify compromises in other criteria. These views openly admitted should apply not only to sample design, but also to other aspects of survey design (1.1). For example, problems of good measurement and data collection may require

restrictions on the spread of the sample. The *timeliness* of surveys must be an important aspect of the goals, especially for agricultural surveys. The *sustainability* of survey efforts should also be part of the *grand overall strategy*.

4. *Feasibility and practicality* refer to obstacles to achieving sampling methods and procedure as designed and intended. Probability sampling cannot be created by assumptions either about the selection design or about the population. "Go and get a random sample" is not a practical instruction to either field workers or statisticians. Selection models must be translated with care into detailed procedures that have been tried and found to work in past surveys or in pretests conducted under the actual field conditions that will be used in the sample survey. The field instructions must be simple, clear, and practical; also fairly complete, yet brief enough to be remembered and used. It cannot be entirely complete; judgment must be used about irregularities that must be either treated as recognizable exceptions for the attention of experts, or else to be tolerated as errors. The art of sampling involves making the practical design conform well, even if not perfectly, to the model for selection. It concerns especially the proper construction and use of frames for selecting units from the population into the sample (Ch. 4).

5. *Efficiency and economy* concern achieving the greatest accuracy for allowed cost; or (equally) achieving the surveys goals with minimum cost. Total survey cost is a broader concept than merely minimizing the number of elements n, the aim of "efficiency" within the *srs* concepts of standard statistical analysis. *Accuracy* is the inverse of the mean square error, including the Bias^2 term plus the variance, whose inverse alone defines the *precision*.

Achieving maximal accuracy (minimal MSE) for "allowed" total cost, or minimal cost for "required" accuracy (or MSE), are two ways of stating the aims of efficiency, depending on which of the two (cost or MSE) is "fixed." Minimizing the $\text{MSE} = \text{Variance} + \text{Bias}^2$ should be our aim, but often our knowledge of Bias^2 is so poor that we must be satisfied with trying to minimize the Variance alone. In this criterion of economy we should also include

measures for the *degrees* of achieving those aims, because these are never completely reached. The ideal optima (maxima or minima) can never be truly attained, especially for multipurpose designs (CH. 9). However, comparisons of relative economies for different designs can be done realistically and approximately, and for many survey objectives, (aims, purposes). Here the theoretical and practical aspects of the science and the "engineering" of sample designs can be practiced to good effect.

CHAPTER 4. SIMPLE LISTS AND COMPLEX FRAMES

4.1 SIMPLE LISTS AND COMPLEX FRAMES

Good procedures from adequate frames are necessary for probability selections (3.1). Finding or constructing and utilizing selection lists or other frames is basic to practical survey procedures. A listing of identifications (like numbers) of all N population elements may be the most desirable and simplest kind of frame, but often more complex frames must be found or constructed. Most lists and other frames have problems and practicing survey samplers must overcome them skillfully and efficiently. Failure to recognize or deal properly with such problems is a common cause of biased survey results. The ability to discover and overcome such frame problems and to accomplish these reasonably well and efficiently is perhaps the most important aspect of the "art" of survey sampling.

Population elements cannot be selected physically and directly like balls from an urn or cards from a deck. Instead we must select their identifying numbers or names from a list or frame; and numbers can be selected more conveniently from tables (or programs) of random numbers. Thus when we talk about a "list of elements" or a list of sampling units we refer to a listing of their identifying numbers. (We simply consider finite, countable numbers; sampling from continuous lines, areas, volumes, etc. can be reduced to finite counts for adequate approximations, which then permit convenient selections with random numbers.)

A simple list of all N elements, numbered 1 to N, is the simple ideal in most minds and books when people think of random sampling. On such a perfect list each element appears separately, once, only once, and nothing else appears on it. In such a list a population of 11,111 elements would be numbered 1 to 11,111, so that a five digit random number (10^5) would have 8/9ths blanks. However, the listed elements may come numbered 20,001 to

31,111; or X+1 to K+N; or they may have any of the 100,000 numbers from 00,000 to 99,999 and with the other 100,000−N numbers left blank, and these denote only minor changes.

However, most lists have worse problems, discussed in later sections. Selection procedures may be applied, instead of elements, to *sampling units* which may contain several, even many elements. For example, neat numerical lists do not exist for most populations, such as farms or dwellings or people in most countries or counties. They would be too expensive to construct and too expensive to cover in the field, and instead of lists we must use other frames for selection. Thus, often we must consider a multistage selection of farms from segments, segments from ED's (Enumeration Districts), ED's from counties (Ch. 6).

"Frame is a more general concept: it includes physical lists and also procedures that can account for all the sampling units without the physical effort of actually listing them. For example, in area sampling the frame may consist of maps, but the frame can be constructed without mapping the entire population." [Kish 1965, 2.8]. "The frame consists of previously available descriptions of the material in the form of maps, lists, directories, etc. from which sample units may be constructed and a set of units selected" [UN 1950]. "Frame: The materials or devices which delimit, identify, and allow access to the elements of the target population. In a sample survey, the units of the frame are the units to which the probability sampling scheme is applied. The frame also includes any auxiliary information (measures of size, demographic information) that is used for 1) special sampling techniques, such as stratification and probability selection proportional to size sample selections; or for 2) special estimation techniques, such as ratio or regression estimation." [Wright, 1987].

We note here three important examples of the use of frames. 1) *Clustered and multistage selections* are commonly used instead of direct selection of elements (Ch. 6). For example, adequate lists of farmers within each village

may exist or may be constructed for a small sample, but not for the entire country; then a sample of villages may be selected in two or three stages (districts and townships). Similarly, a frame for school children may consist of school districts containing schools, then classes, then children. Note that at each stage only a sample (usually only a small portion) of the sampling units needs to be prepared for the selection of the units of the next stage. Thus in a large country instead of trying to list millions of farms, the sampler may need to list only hundreds of units at each stage.

2) *Area sampling* is commonly based on area frames for farms, households, and other units that associate them uniquely, adequately and feasibly with area units like segments, villages, districts, counties etc. (Ch. 10). Stable and unique associations of farms, households etc. and of sampling units with clearly defined areas are the basis for area sampling, together with available records, measures of size, data for stratification etc. for the units. Thus area samples often serve as convenient means for clustered and multistage sampling. At each stage the entire population is divided into sampling units from which a sample is selected.

3) *Dual frame and multiframe selections* have been used sometimes in agricultural and other surveys when one frame does not appear as obviously both better and cheaper (Ch. 11). For example, a list of farms (or farmers) from an agricultural census may be a less expensive frame, but it needs to be supplemented by area samples for new and missing farms. Area frames can similarly supplement frames from other records, such as registered farmers, telephone subscribers, etc. Furthermore, the imperfect lists may also contain elements which are missed in area frames; for example, small farms in cities (poultry, eggs, vegetables) may be easily missed by area frames for farms. Or a third frame may also be used to find those missed by both frames.

Special frame problems are discussed in more detail in chapters 6, 7, 10 and 11, postponed in order to continue here with more urgent matters. There are other treatments of frame problems [Kish 1965, 2.7, 11.1 – 11.6; Hansen, Hurwitz, Madow I, Ch. 2; Wright and Tsao 1983]. However, in 4.2, 4.3, and 4.4 we call attention to solutions to three classes of problems.

4.2 FOUR FRAME PROBLEMS AND SOLUTIONS

We list in Table 4.2.1 the four possible contradictions of the rule of unique identification of sampling units (and elements) with frame (and listing) units, represented by L – U. Most (or all?) frames contain one or more, sometimes all four, of these imperfections; but in different quantities, as some frames are better (closer to the target population) than others. More complicated imperfections are possible as combinations of these four. For example, if several farmers work jointly several farms, this may appear as a combination of duplicate listing with a cluster of units. But by providing general solutions to the four basic problems, we expect to help also with more complicated cases.

For each of the four problems stated here, solutions are proposed to maintain or restore the intended probability, usually a constant rate f, that imperfections in the frame (listed below) would disturb. Two of the proposed remedies accept the inequality, but compensate for them with proper weights. Also note that the "common sense" remedies often used for all four problems would cause biases, which should be avoided.

1) *Blanks or foreign elements* occur in many frames: the listings contain no elements, because they expired or moved; or they were nonexistent or nonmembers of the survey population. Nonmembers are often numerous when subpopulations (by age, sex, occupation etc.) are excluded either from the selection or from the analysis.

Suppose the list contains N = M+B listings, M population members plus B that are blanks for various reasons. If the blanks can not be excluded before selection (as in 4.3), they *must be rejected* after the selection because they do

not contribute to the sample. Sometimes a first phase "screening" process is needed (Ch. 11). During subclass analysis, nonmembers are also excluded from the sample base.

If a sampling rate f is applied to all the N listings, the number n = fN of listings can be specified and all M members also receive the same probability f. The expected sample size is m = fM, but the actual sample size m' becomes a random variable. Were the selection process altered in order to fix an exact m, the sampling probability would become random when M is unknown (Ch. 11). It is usually better to fix the probability f and let m' vary a little, as it must in most cases.

Table 4.2.1. The Four Basic Deviations from Unique Identification (L − U) of Listing (Frame) Units (L) and Sampling Units (U)

L − U	Ideal lists one-to-one correspondence of units in frame and population
L − O	Blanks or foreign units; also extinction, emigration; also subclass analysis. No (0) units for listings.
L}U L}	Duplicate (replicate) listings; dual (multiple) frames
L{U {U	Clusters of units with single listing: small clusters
O − U	Missing units, noncoverage, incomplete frames. No (0) listing for units.

Avoid the common fallacy of accepting the next valid units as substitutes for selected blanks. This procedure increases the selection probabilities of all units in proportion to the number of blanks proceeding it. The "densities" of blanks often differ in parts of the frame and can be associated with different values for the member units.

2) *Duplicate (replicate) listings* would give sampling units selection probabilities proportional to the number P_i of listings. One may choose from three alternative ways to deal with this problem, depending on the situation.

a) *Unique identification of a single listing* for each element may be defined before selection. For example, define the first (or last) selector as the unique selector, particularly if all listings for each unit are clear and contiguous. Selections are confined to the specified unique listings; other elements become blanks and treated as above. If the ordering of listings is not simple and contiguous some unique feature may still be designated. For example, selections of farmers from listings of farm parcels may be accomplished by associating each farmer uniquely with his *largest* parcel. Sometimes random choices with probabilities $1/P_i$ may be substituted for unique choices, when the replicates may be found or at least the size of P_i ascertained. (This problem differs from those in 4.4, where equal probabilities for P_i elements in single units *must* be assured.)

If selections must be first completed. elimination of replicates from the entire population may have to be postponed. This modification may be needed when the replicates are scattered over the population, rather than contiguous. This also occurs when the replicates occur in other frames in multiframe selections (11.1). But even this procedure is less laborious than eliminating all replications from the entire list (4.3).

b) *Weighting* each selection by the inverse $1/P_i$ of its probability of selection should be used in cases when all P_i listings must be accepted because unique identification does not seem feasible. This may occur especially when the P_i are established only after expensive interviews or measurements. So,

although equal selection probabilities are abandoned, equal estimation weights are restored. Weighting has two disadvantages: greater complexity of analysis and usually increases in variances (Ch. 12).

Avoid the common fallacy that eliminating duplicate listings from the sample selections can deal with the problem of unequal probabilities. If the sampling rate is 1/1000, the chance that both listings of a duplicated element get selected is one in a million only (with srs assumptions).

3) *Clusters of elements* (or sampling units) can be associated with single listings. Suppose this problem is not common and the clusters are small. For example, listings of dwellings may contain occasional "duplex" dwellings; some households may have two holders or two women of childbearing age; a list of farms may contain a small proportion that have split into two farm operations. But if clusters are common and large it is best to resort to formal cluster sampling (Ch. 6).

Avoid the common belief that selecting at random one sampling unit maintains the equal probabilities for selecting the listings, after the selection factor $1/p_i$ is introduced (4.4). Choose the most feasible of three alternative ways of dealing with these occasional small clusters.

a) *Select one of the P_i elements at random* with $1/P_i$ but then compensate by weighting it with P_i. Weighting avoids the bias of the "common sense" fallacy. But this procedure has three disadvantages: the subselection may be difficult in the field; the analysis is complicated by weighting; and variances are also increased, especially if the weights and their proportions are not small (12.6).

b) *Include all elements identified with each listing*, when clusters are neither large nor common. This is usually true for the three examples above: dwellings, women and farms. This is often the best, most practical procedure, and it is the simplest because it preserves the probabilities, often a constant f,

originally assigned to the listings. Although variances of means per element may be increased due to the clustered selection, these increases will be slight when the clusters are small and not common.

c) *Relist a larger sample* and then subselect an *epsem* of elements. If small clusters are common this may produce a final epsem of elements for a modest expense. For example, list the adults from a preliminary *epsem* sample of dwellings to produce about 3n adults, and from this subselect with f = 1/3 to obtain about n adults.

4) *Missing elements*, also called *noncoverage, undercount and incomplete frames*, pose some of the most serious and difficult problems for many agricultural and other surveys. Though theoretically simple and similar to nonresponse, in practice they can be even more troublesome because even their magnitude may be unknown (15.3). The "common—sense solution" of taking larger or supplemental samples fails, because they only increase the already covered population, but these may differ significantly from the noncovered. There are three alternative remedies, but often none of them are feasible, unfortunately.

a) *Supplements with special procedures* may be added with smaller samples in separate strata (11.2). These procedures should be significantly better, but also more expensive than the main sample. Yet to be useful these subsamples should be affordable and also yield accurate results in order to either measure with comparisons the effect of noncoverage, or preferably to correct the sample results. Because such contrary and difficult conditions are seldom met, these desirable methods are only used on censuses and very large samples.

b) *Linking procedures* or *half—open intervals* offer appealing alternatives when the listing can be viewed and applied in linear fashion. In addition to the selected l—th listing also investigate and include any unlisted hence "missing" sampling units up to, but not including, the (l+1)th listing. This procedure assigns to those missing sampling units between the listings numbered l and

l+1 the same known selection probability assigned to the !—th listing. This procedure can be successful sometimes, under proper conditions. The linear order should be made feasible for the enumerators. The extra instruction should not be too cumbersome for them; or the burden may be reduced with sub—sampling. The missing sampling units should be well scattered, so that large clusters of them are not picked up at single locations.

c) *Estimation of the size and effects of noncoverage* may be useful. This requires skilled use of reliable and available auxiliary data in the estimation process. Actually using ratio, regression and post—stratified estimates may bring great benefits for reducing the effects of noncoverage (12.3).

4.3 AVOIDING FRAME PROBLEMS

There are also three general procedures for *avoiding* frame problems, and one of these may be useful in some situations.

1) *Ignore and disregard* the problem if we may be convinced that the effects are small compared to other errors and if the corrections would be too expensive compared to its results. For some small agricultural surveys, for example, it would be too expensive to search city areas for missing farms, especially for crops like grains. From other studies and from external evidence we may be convinced and also convince others that the problem is small enough, compared to other biases and to sampling variances, to be ignored. A statement about the probable magnitude of the problem should be added to the description of the sample (15.3).

2) *Redefine the population to fit the frame*, but only if the difference can be ignored (as in 1 above), or if the redefined population is actually preferred (as in examples 1,2,3 in 4.4). This should be avoided if the sample results would be seriously deflected from the aims of the study. We may accept target populations reduced by deliberate exclusions of regions small in population (through large in areas) when the bias caused by the exclusions would be

outweighed by the cost of covering those areas. We then deliberately exclude those areas, with explicit redefinition of the target population, and preferably with some estimates of the magnitude of the exclusion and of its effects.

3) *Correct the entire population list*, eliminating blanks and replicate listings, and splitting clusters. Finding missing units may be difficult but perhaps not necessary. Clerical correction may be less expensive, for even tens of thousands of simple records, than technical corrections would be; with machine treatment of tapes even populations in the millions may be treated. But hand treatment of lists running into the millions may be so costly that clerical routine must be replaced by skill. Such clerical labor may be reduced by introducing *multistage sampling or multiphase sampling* to reduce the size of the population being treated. It would be too difficult to explore here all the possible ramifications of these possibilities.

4.4 FRAMES WITH UNEQUAL PROBABILITIES

Eight situations are described below as distinct problems with separate treatments, but they have basic similarities; therefore this joint treatment would be heuristic and instructive. From the common basic principles the reader can more readily learn to treat other similar problems as well; and there are many others. These situations resemble partly those of "replicated listings" in 4.2.2, which we symbolized with (L – U – L) to show two (or more) listings for one sampling unit. But in the nine situations below both the frequency and the size of replicate listings can often be greater than conjectured in 4.2.2. Furthermore, quite often the listings may also be elements of meaningful populations so that (e – U – e) may better symbolize the situations below, with e for events and U for units.

1) *Sampling contacts with a facility*. These may refer to sampling the distinct visits to offices (e.g. agricultural agents), stores (e.g. fertilizer and seed stores), clinics, hospitals, libraries, doctors etc. Shares in companies, which can be used to sample shareholders with several or many shares, present similar

problems. Such samples result in biases because the contacts (visits, shares etc.) are not evenly distributed: units with more contacts (or shares) are overrepresented, whereas others receive reduced representation or none at all. Should all units (U_i) receive equal representation in the selection or in the estimation? Or should visits be accepted as the selection units, thus giving equal representation to contacts (e). This also gives representation to units proportional to the number of contacts (e_i) for each unit (U_i); and this amounts to a "redefinition of the population" (4.3.2).

2) *Size of family (or group) as selection factor.* Samples of families (U) have been drawn by selecting from lists of children (e) in schools or from lists of residents in cities. Larger families would be over-represented in those selections and estimates for families would be biased toward large families, unless corrected by weighting. [See Kish 1987, 7.4 for situations 1,2,5].

3) *Sampling parcels for holdings.* Selections from lists of farm parcels (e) would obtain samples of holdings (U) with overrepresentation of holdings with many parcels. This situation resembles the preceding two cases. For some purposes statistics based on parcels would be adequate, but for most purposes statistics based on holdings are needed. We need one of the solutions for replicate listings (4.2.2). See also 11.5.

4) *Selection grid with random points.* Selection grids with equidistant points can be placed on maps with random two–dimensional choices; thus every point on detailed maps has the same probability of selection. On the maps the location of each farm has been uniquely pinpointed (11.3). Procedure A: "Take the 4 (or m) farms nearest to each random point". Procedure B: "Take all farms within a radius of 2 km. from each random point". A is biased in favor of large farms, which have greater probabilities of having their identifying point near a random point. B is unbiased, though it permits variation in the sample size at each point. This is a two–dimensional extension of the problem of blanks in 4.2.1. It is also an example of problems arising from fixed sample sizes (7.7); and so are situations 6 and 7.

5) *Waiting times.* This is a familiar problem at all kinds of lines and queues for waiting. At busy airports, have you noticed that most check–in counters have only 1 or 2 persons, but you and we usually find ourselves in long queues? This is the natural result of unequal queues. Busses are scheduled for similar arrival times, but in big cities they arrive at unequal intervals due to traffic delays. The longer the delay the longer the queue and most people find themselves in the longer lines. The average waiting time for riders becomes much longer than the average and scheduled intervals between buses [Kish 1987, 7.4].

6) *Fixed sample sizes from unequal clusters.* When clusters are selected with equal probabilities f and then subsamples of fixed size b are selected from each, the probability of selection in the two stages becomes $f_1 x(b/N\alpha)$, inversely proportional to the variable cluster sizes $N\alpha$. Needless and inefficient deviations from equal probabilities occur in many situations. a) Fixed subsamples of farms from unequal segments; b) Fixed number of dwellings from blocks (or buildings) of unequal sizes; c) Fixed number of employees from firms of unequal sizes. Two stage (or multistage) selection with PPS (probability proportional to size) yields procedures to keep the sizes of subsamples approximately constant, yet they preserve the constant overall sampling rate f (7.7). Two stages of selection are represented by $(M_\alpha/bF) \times (b/M_\alpha) = 1/F = f$, where the M_α are "measures of size" for the clusters.

Selecting a single adult from the N_α adults in households that were selected with equal rates f, seems theoretically a similar problem: the selection probabilities of adults become f/N_α, inversely proportional to be number of adults N_α in the household. Weighting with N_α restores unbiasedness in the estimates. The practical results are not serious because N_α is 1, 2 or 3 and seldom larger (11.4).

7) *Telephone sampling with random digit dialing.* Telephone numbers of seven digits may be represented by AAA – BBrr, where the rr denote clusters 100 numbers, many of which may be blanks. Thus $100 = N_\alpha + B_\alpha$ digits

represent N_α occupied telephone numbers and $B\alpha$ represent unoccupied blanks. These could also represent pages of a register with 100 lines of which N_α are registered farms and B_α are blank lines. If we select random digits, the chances of hitting occupied number will be proportional to N_α. If then constant sizes of subsamples b are selected the overall selection rate becomes in two stages $(N_\alpha/bF) \times (b/N_\alpha) = 1/F = f$.

We may summarize briefly the problems of *weighting* that will be treated in 12.5. In situations 1,2,3,4 the selection of units was proportional to numbers of elements, or contacts as events symbolized with e_i for the ith unit value U_i. The simple mean would estimate the element weighted mean $\bar{Y} = \Sigma e_i U_i/\Sigma e_i$; but to get the mean of units $\bar{Y} = \Sigma U_i/N$ calls for weights proportional to $1/e_i$. On the contrary in 6,7,8 if the units are selected with equal probabilities f, to produce the element $\bar{Y}_w$ we need to use the weights e_i. When they must be used the weights: 1) must be known for all selected units; 2) tend to increase variances per selection; 3) tend to increase the complexities of analysis.

8) *Observational units of unequal sizes.* This problem represents situations contrary to those of 1 to 7: where it may be preferable to depart from simple equal probabilities for units for the sake of better representation and efficiency.

Large units of variable sizes can become observational units as well as sampling units when a single measurement Y_α is used to characterize it. Examples: a) Y_α measures the quality of drinking water, or climate, or available primary school etc. in village α; b) Y_α measures the quality of science of mathematics teaching in secondary school α; c) Y_α is the size, or altitude, or age of city α. In most of these cases it is likely that instead of the simple unit mean $\bar{Y}_u = \Sigma Y_\alpha/A$ of the A units, it would be better to estimate $\bar{Y}_w = \Sigma N_\alpha Y_\alpha/\Sigma N_\alpha$, the mean weighted by the numbers of elements in the units. The relative difference between the two means can be large: $(\bar{Y}_w - \bar{Y}_u)/\bar{Y}_u =$

$R_{ny}C_nC_y$, where R_{ny} is the correlation between sizes N_α and values Y_α, also C_n and C_y are coefficients of variation of the two variables. PPS selection of units is useful [Kish 1987, 7.5; Kish 1965, 11.6].

CHAPTER 5. ELEMENT SAMPLING

5.1 SELECTING ELEMENTS WITHOUT CLUSTERING

The populations and their elements should be defined by the aims of survey analysis. The next most basic question is: May the selection be made directly and confined to the elements, or must clusters of elements serve as sampling units? Because clustering increases both the complexities of analysis and the variances per element, it should be used only when needed, and it is often needed for agricultural surveys (Ch. 6). *For element sampling to be economical, we need two things*: first, adequate listings of the elements, fairly complete and up—to—date, must be available for selection. Second, locating the elements and collecting the data individually must be feasible and economical.

Later sections of this chapter describe four methods for selecting elements, but some other aspects of element sampling are left for later chapters. *Estimation methods* with ratio and regression estimates and with poststratification are treated in 12.3. *Two phase sampling* for screening and for rare items is discussed in 12.4. *Dual frame* selection may use element samples from one frame and supplement it with a clustered area sampling frame (11.1).

In this section we would like to describe some situations where element sampling *may* be feasible, and with special attention to agricultural surveys, farmers, and to households.

1. In some countries telephone ownership is over 90 percent (in 1987); also selecting telephone numbers, then identifying the defined population, and then obtaining "high enough" and "good enough" responses can be done fairly well. There exist many articles on this subject that it is growing and changing rapidly; situations differ greatly between populations and subjects, hence no review will be attempted here [Groves and Kahn 1979; Groves et al., 1988]. For DC's telephone sampling lies only in the future for general agricultural surveys, but they may be used for some special populations on special lists.

2. Some registers contain interesting data that can be sampled directly from the records without the need for obtaining data from the individuals. From data cumulated in registers even longitudinal statistics may be computed. Samples of population registers and health records have been used to produce interesting statistics.

3. Members of voluntary groups and organizations are often listed with clear and current addresses. They may be willing and able to answer mailed questionnaires. Or they may reside in a relatively small area so they may be interviewed with low location costs; for example, farmers in marketing or buying cooperatives may be sampled.

4. In some countries, mostly in Northern Europe, complete and up—to—date registers exist for the general population, with good, current addresses. On many mail surveys, with short questionnaires on non—sensitive subjects, good responses have been obtained from largely literate and cooperative populations, who are able and willing to respond in large proportions by mail. Also in these countries household surveys are expensive.

5. If a small country, or state, or province, or city, has a good population register, the sample may be drawn from it, and the sample (of households, persons, etc.) can be located and visited for interviews or observations without unduly great location costs.

6. Districts or cities that are not too large may be completely listed for element sampling for reasonable, though not negligible, cost. Suppose, for example, that a sample of n = 1000 households is wanted from a district of 10,000 households (or 50,000 persons); the sampling interval would be 10 and the cost of cheaply listing 10 households per interviewed household can be made reasonable. However, for a sample of 1000 from a state of 100,000 or 1,000,000 households, those ratios of sampling and of listing to interviews (100 or 1000) would become insupportable; hence the listings should be sampled also.

5.2 SIMPLE RANDOM SAMPLING (SRS)

The status of SRS in the field of survey sampling involves confusion and contradictions: SRS is basic in the theory of sampling, although it is and should be seldom used in practice for selecting samples (outside of introductory classrooms and simulation exercises). For the mathematics of statistics and of sampling the *independence between selected elements* is a powerful tool; it appears in assumption of "I.I.D., independently and identically distributed random variables"; also in "selections from a well mixed urn," etc. All classic statistics are based on these assumptions; and sampling theory also begins with them. But in practical survey samples various restrictions are used either to reduce variances (as with stratification) or to reduce costs (as with clustering). Nevertheless SRS serves as a basic standard, as in the denominator of "design effects" later. Furthermore, the concept of independent replications is basic to all measurements of sampling errors. However, natural populations do not exist in random states, nor can they be physically "well mixed" like the balls in the urns of the textbooks. Selection with SRS would yield the independence of elements in the sample that is needed and assumed (often unstated) by basic theory.

A. Procedures for SRS.

A simple operational procedure reads: From a *good* table of random digits select with *equal* probability n *different* selection numbers, corresponding to n of the N *listing numbers* of the population elements. The *n* listings selected from the list, on which each of the N elements appears once, uniquely identifies n different elements for the sample. At any $(k + 1)$th selection all the $(N - k)$ *unselected* elements have equal $1/(N - k)$ probabilities of selection, but all the k selected numbers receive zero probability (like "blanks," discussed in 4). The words *different* and *unselected* above mean that no element may be selected twice, and denote *SRS WITHOUT replacement.*

In SRS WITH replacement, n random numbers are drawn, each from 1 to N; all N elements receive the same probability 1/N on each of the *n* selections. The elements may be reselected on any draw and the words "different" and "unselected" are omitted from the definition above. Thus any element may appear not only once, but twice, seldom three or more times (though theoretically even *n* times) in the sample. Thus in SRS WITHOUT replacement, the selection receives some restriction; since SRS WITH replacement remains *unrestricted*, we may name it *unrestricted random sampling, or URS*. In SRS WITHOUT replacement sampling each of the $C_n^N = N!/(N - n)!n!$ possible combinations has the same equal probability of selection; but this theoretical definition does not lead to a practical procedure. In unrestricted URS sampling there exist N^n theoretically possible outcomes, all different permutations. URS has somewhat higher variance (for the same n/N) by the factor (1 − n/N) than SRS, because the samples may contain the same elements more than once. If the replicates are eliminated from the sample, the sample size becomes a random variable; but the fixed size n can be restored with a supplement, which can be designed before selection. [Kish 1965, Ch.2]

Bernoulli sampling is a name for selection procedures that now seem convenient with electronic computers: Each of the N listings of elements receives independently in its turn the same f = n/N probability of selection. This method would allow the sample size to become highly variable, but computing programs can now deal with this problem by changing f at each selection; or a rate f* > f can be set and the excess eliminated with equal probability f/f*. This yields n selections without duplicates and with f = n/N for each of the N elements; and the equivalent of an SRS.

Subclasses in the sample "inherit" the SRS properties of the entire sample. A subclass n_c of the sample is equivalent to having selected n_c from the N_c elements of the subpopulation (domain), *EXCEPT* that: 1) the

subsample size n_c becomes a random variable and 2) the population size N_c may remain unknown, as with sampling from a list with blanks (4.2). These theoretical problems may be neglected if n_c is not small.

Models for SRS are often assumed without actually selecting an SRS. The assumptions of I.I.D. are common in statistics and they also occur in the theory of sampling [Kish 1987, 1.8]. But actually creating a "well mixed urn" of listing numbers to identify elements is "never" practical and populations are "never" in truly random order. In general we should be cautious with assumptions of SRS. When we read "simple random sampling" for describing a survey sample we may well question either the description, or the wisdom of the procedure, or both. However, there exist situations where such assumptions may be good approximations, especially for small samples whose sampling errors probably dominate the small biases caused by SRS assumptions. A few examples can illustrate situations where the convenience of the assumption may overcome our caution. a) Day of birth *may* serve as a convenient identifier on some lists that are approximately random. b) Day of arrival at a facility *may* also select groups, which may be accepted as unbiased after an investigation. c) The last digits (2,3 or 4 of them) of social insurance numbers may be assigned (almost) at random; but the first digits may be (weak) stratifiers. But telephone numbers, auto licenses and such may be subject to individual choices, hence not acceptable as random digits. d) Members of small subclasses from stratified (PRES) selections approach SRS properties (5.5).

The population size N' is unknown in some situations with B' blanks among the N = N' + B' listings. Often with a fixed sampling rate (probability) of f, the sample size n' = fN' will be allowed to become a random variable; that retains the "known probability" f and the "unbiased estimator" E(n'/f) = N'. That variation often has probably no important practical consequences, except if n' or E(n') is allowed to become too small; and this is preferable to fixing n

and allowing f' to vary. But if n is fixed, f' = n/N' becomes a random variable and we lack a "known" probability, although we still have equal selection probabilities.

B. Descriptive Statistics from SRS.

SRS is an *epsem* selection and *self–weighting*: that is, the simple statistics based on equally weighted (i.e., unweighted) sample cases, have desirable properties. The simple sample total is the most basic statistic, and uses weights of 1:

$$y = \Sigma_j y_j \ .$$

From y with constant factors we obtain the sample statistics (estimators) for the mean and the total (aggregate), with weights of n and n/N:

$$\bar{y} = y/n = \Sigma_j y_j/n \quad \text{and} \quad \hat{Y} = y/f = Ny/n \ . \tag{5.2.1}$$

These estimators are simple and also "unbiased":

$$E(\bar{y}) = \bar{Y} \quad \text{and} \quad E(\hat{Y}) = Y \ .$$

These properties also hold for element variances and for covariances terms:

$$E(s_y^2) = S_y^2 \quad \text{and} \quad E(s_{yx}) = S_{yx} \ . \tag{5.2.2}$$

Here $s_y^2 = (\Sigma_j y_j^2 - y^2/n)/(n - 1)$ as usual, and S_y^2 is similar with N and Y instead of n and y; $s_{yx} = (\Sigma\, y_i x_j - yx/n)/(n - 1)$.

Some other descriptive statistics, such as the standard deviations s_y, and ratios of random variables are not technically "unbiased estimators" of their population values S_y, but they are technically "consistent" and they are good, and commonly used statistics. For example, the coefficients of correlation and of regression: $r_{yx} = s_{yx}/s_x s_y$ and $b_{yx} = s_{yx}/s_x^2$ (Ch. 12) are such consistent estimators of their population equivalents.

For subclasses the sample mean $\bar{y}_a = y_a/n_a$ is also clearly useful, unless it is unstable because n_a is too small, because it is a random variable. But for estimating subclass aggregates, instead of $\hat{Y}_a = y_a/f$ the ratio estimators $N_a\bar{y}_a$ maybe preferable, because these have lower variances. In general, *ratio estimators* may be used to improve SRS selections (Ch. 12).

C. Inferential Statistics from SRS.

The descriptive statistics in B, such as $\bar{y}$, $\hat{Y}$ and s^2_y, have general validity for many kinds of designs and depend only on the individual probabilities P_i of selecting the sample elements. In contrast, the inferential statistics in this part C depend on the joint selection probabilities P_{ij} (pairwise for population elements i and j). These formulas require independent selections and are strictly valid only for SRS and URS respectively.

The population value and its sample estimator for the variance for sample means ($\bar{y}$) are:

$$\text{Var}(\bar{y}) = (1 - f)S^2_y/n \quad \text{and} \quad \text{var}(\bar{y}) = (1 - f)s^2_y/n \,. \qquad (5.2.3)$$

This sample variance, like s^2_y, is also a technically unbiased estimator of the variance: $E[\text{var}(\bar{y})] = \text{Var}(\bar{y})$. This again, as for s_y, does not hold strictly, technically for the standard error, because $E[\text{ste}(\bar{y})] \neq \text{Ste}(\bar{y})$ where: $\text{ste}(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \sqrt{(1 - f)}s_y/\sqrt{n}$ and $\text{Ste}(\bar{y}) = \sqrt{\text{Var}(\bar{y})} = \sqrt{(1 - f)}S_y/\sqrt{n}$.

The theoretical value of $\text{ste}(\bar{y})$ is that, *with assumptions of SRS*, mathematical derivation can show [Kish 1965, 2.8B; Cochran 1977, 2.5, 2.9] that $\text{Var}(\bar{y})$ estimates the variance of the sampling distribution of $\bar{y}$:

$$E[(1 - f)S^2_y/n] = \Sigma_c P_c(\bar{y}_c - E(\bar{y}))^2 \,.$$

The important practical value of $\text{ste}(\bar{y})$ is that, because it is a good estimator of $\text{Ste}(\bar{y})$, it can be used (for SRS selections) to construct inferential statistics (probability statements, confidence intervals) like $\bar{y} \pm t_p\text{ste}(\bar{y})$.

For URS the population variance of the mean and its unbiased sample estimate are:

$$\text{Var}(\bar{y}) = \sigma_y^2/n \quad \text{and} \quad \text{var}(\bar{y}) = s_y^2/n \,. \tag{5.2.4}$$

The variance of SRS is less than these by the factor $(1 - f)$, where $f = n/N$, the "finite population correction" (FPC). This factor arises in the mathematical derivation from the lack of complete independence in ruling out replicate selections. It is usually disregarded either because it is small or for theoretical reasons. Other modifications of S^2/n, which we call "design effects" and express as D^2S^2/n, will be seen as much more important. Stratification will be seen with $D^2 < 1$ to reduce variances slightly in element sampling (5.5), and cluster sampling to introduce often large increases with $D^2 > 1$ (6.6).

Computing formulas for variances begin conveniently for the sample sum $y = \Sigma_j v_j$:

$$\text{var}(y) = (n \Sigma_j y_j^2 - y^2)/(n - 1) = ns_y^2 \,. \tag{5.2.5}$$

Then since $\text{var}(\bar{y}) = \text{var}(y/n) = \text{var}(y)/n^2$, we may use:

$$\text{var}(\bar{y}) = (n\Sigma_j y_j^2 - y^2)/n^2(n - 1) = s_y^2/n \,. \tag{5.2.6}$$

For the element variances we use $\text{var}(y)/n = s_y^2/n$. For the aggregate $\hat{Y} = N\bar{y}$ we can use:

$$\text{var}(\hat{Y}) = \text{var}(N\bar{y}) = N^2\text{var}(\bar{y}) = N^2s_y^2/n \,. \tag{5.2.7}$$

These serve for URS, however, for SRS with the FPC we may use $\text{var}(\bar{y}) = (1 - f)s_y^2/n$ as noted above.

Proportions $p = \bar{y}$ are used frequently as sample means to estimate population proportions (means) $P = \bar{Y}$, when the variables "y" are dichotomies and the Y_i takes only the values 0 or 1; and $\sigma^2 = PQ$. The computing formulas for the variances become simple because $y = np$; and then $\text{var}(np) = (n\Sigma y_j^2 - y^2)/(n - 1) = (n^2p - n^2p^2)/(n - 1) = n^2pq/(n - 1) = ns_y^2$. Then $\text{var}(np)/n^2 = \text{var}(p) = pq/(n - 1)$.

Coefficients of variation (c.v.) and *relvariances* are used sometimes as measures of *relative errors*: the units of measurement are removed by dividing by the mean $\bar{y}$. Thus for element values $c_y = s_y/\bar{y}$ is used to estimate $S_y/\bar{Y}$, and $c_y^2 = s_y^2/\bar{y}^2$ to estimate $C_y^2 = S_y^2/\bar{Y}^2$. For the distribution of the mean ($\bar{y}$) we have:

$$\text{c.v.}(\bar{y}) = \text{ste}(\bar{y})/\bar{y} \text{ to estimate C.V.}(\bar{y}) = \text{Ste}(\bar{y})/\bar{Y} ,$$

and

$$\text{c.v.}^2(\bar{y}) = \text{var}(\bar{y})/\bar{y}^2 \text{ to estimate C.V.}^2(\bar{y}) = \text{Var}(\bar{y})/\bar{Y}^2 . \qquad (5.2.8)$$

These relative errors are useful for chiefly positive quantities (like areas of holdings, yields, income), but not at all for variables which can be negative and have $\bar{y}$ near zero.

Subclasses of SRS are also SRS and the variance for the mean $\bar{y}_c$ of a subclass of size n_c has $\text{var}(\bar{y}_c) = (1 - f)s_c^2/n_c$; this formula neglects the difference between f and possible distinct f_c for subclasses. The variance for the difference of two means based on distinct (independent) subclasses n_c and n_b equals the sum of the two variances. Those variances are greater than for the entire sample by the factors n/n_c for a subclass and $(n/n_c + n/n_b)$ for a difference.

However, the mean for the difference $(\bar{y} - \bar{x})$ between two variables based on the same sample n (e.g., before/after differences) benefits from the effects of the covariance on the variance: $\text{var}(\bar{y} - \bar{x}) = s_y^2/n + s_x^2/n - 2s_{yx}/n = \text{var}(\bar{d})$, where $d_j = y_j - x_j$. This arithmetic identity also carries useful content.

D. The Design of Sample Sizes for SRS.

1. "Will a sample of five *percent* be large enough?"

2. "What sample *size* n should we take?"

3. "With a sample size n, how large will be the sampling errors?"

The ordering of these questions represent how commonly they are asked: the first is the most common, also the least sensible. First, the sampling error depends more on the sample *size*, and hardly at all on the sampling percentage or fraction of the population.

Second, the needed sample size depends on the required sampling errors; but most surveys are multipurpose and the precisions for the many aims usually differ greatly. For these reasons the design of sample sizes is mostly postponed to chapter 9 on "Multipurpose Designs" and only briefly treated here.

Third, the "permissible sampling errors" or the "required precisions" for the many survey aims are not usually available or realistically obtainable. Much more realistic and common is to have a permissible budget limit and from that to estimate a total field and processing cost. That may be stated as $C = \tilde{c}n$, then *from a reasonable and realistic cost per element* $\tilde{c}$, the permissible $C/\tilde{c} = n$ can be obtained. Then from the SRS variance we get $\tilde{V}ar(\bar{y}_g) = \tilde{S}_g^2/n_g$. The subscript g denotes that several (or many) aims should be represented for most surveys and that these can differ greatly. Especially, the n_g for subclasses may be 1/10 or 1/100 of the total sample.

The element variances $\tilde{S}_g^2$ in the design must be guessed, as noted by the tilda (~), with one of more of the several methods below. We first note with comfort that errors in guessing $\tilde{S}_g^2$ do not bias the sample estimates of the true values of S_g^2, because those are based on values of the s_g^2 computed from the actual values. Underestimating $\tilde{S}_g^2$ result in larger values of s_g^2 and of $ste(\bar{y}_g)$ than we hoped and designed for, but not in biases.

1) *Past surveys with similar variables* may be used either directly from publications and reports or from the advice of experts. 2) *Models* of the survey variables from experts in the subject matter can be most useful (Ch. 14). From reasonable guesses of the *coefficient of variation* $C_g = S_g/\bar{Y}_g$ and of $\bar{Y}_g$ one may guess $S_g = C_g\bar{Y}_g$ well enough. The values of C_g for domains can often be guessed reasonably well from the C for the entire population. 3) *Proportions*

are commonly used in surveys, and $S_g = \sqrt{P_g Q_g} = \sqrt{P_g(1 - P_g)}$ varies only slightly even for moderate variation in P_g, especially between 0.3 and 0.7 for P_g. Therefore, even mediocre guesses for P_g can yield useful values for S_g. 4) *Pilot studies* would seem like reasonable sources for S_g, but in practice most studies are too small and too hurried to support a large enough pilot study to yield useful estimates of S_g^2. Results from small pilots are almost useless, if they are less reliable than guesses from the first three alternatives.

With reasonable guesses about $\tilde{S}_g^2$ we can use these preliminary variances, n_g' and $f_g' = n_g'/N$:

$$\tilde{V}ar(\bar{y}_g) = S_g^2/n_g' \text{ and } n_g' = \tilde{S}_g^2/\tilde{V}ar(\bar{y}_g) . \qquad (5.2.9)$$

If the finite population correction FPC = (1 − f) must be taken into account, the preliminary n_g' can be corrected to the needed n_g and f_g:

$$\tilde{V}ar(\bar{y}_g) = (1 - f)S_g^2/n_g \text{ and } n_g = n_g'/(1 + n_g'/N) . \qquad (5.2.10)$$

The n_g here refers to desired sample sizes for the entire SRS sample, but those desired n_g will generally vary between variables, because of different guessed values $\tilde{S}_g^2$, but even more because of different "needed" variances $Var(\bar{y}_g)$ for different variables. Furthermore in complex designs, "design effects" $\tilde{D}_g^2$ should also be guessed, with $\tilde{D}^2 < 1$ slightly for stratified element samples, and $\tilde{D}^2 > 1$ for clustered samples. Even greater variations are introduced by designing for domains of the sample (Ch. 9).

In my view, the usual *increasing ordering of difficulties* in guessing design parameters is as follows:

1. $\tilde{c}$, the cost per element of collecting and processing data.
2. $\tilde{D}_g^2$, the "design effects", but this may be varying and difficult in clustering.
3. $\tilde{S}_g^2$, the element variances for diverse variables.

4. $\tilde{V}ar(\bar{y}_g)$ and $\tilde{S}te(\bar{y}_g)$, the "desired precision" or permissible sampling errors are most difficult to state with any reasonable dependability and especially for several (many) survey aims. These can vary a great deal especially when subclass sizes n_g vary greatly (Ch. 9).

5.3 STRATIFIED RANDOM ELEMENT SAMPLING

Stratification is treated in more detail in 6.2 under clustered sampling, because the gains from stratification are usually greater and more important there. But the general principles are similar for both element and cluster sampling. Briefly stated, stratification consists of four steps. 1) The *entire* population of sampling units is *divided into distinct* subpopulations called *strata.* 2) *Separate samples are selected independently* from each stratum. 3) The *separate statistics* (means, proportions, etc.,) from each stratum are *weighted and combined* into overall estimates. d) *Variances* for those estimates are weighted and added into overall variances. However, modified simplifications of steps 1, 2 and 3 are often feasible and convenient.

Four principal motivations for stratification, alone or together, account for the common use of stratification both for element sampling and for clustered, multistage samples. 1) Stratification *reduces variances* for given effort, measured either in the size of the sample or in costs. Variances may be reduced either with *proportionate* stratification (5.4 and 5.5), or contrariwise with deliberately disproportionate "*optimal*" allocation (5.6). These are the reasons justified with formulas in theory, but in practice the other three reasons may sometimes be even more important. 2) Stratification may be used for *safety, comfort, insurance* against suffering from distorted random selections. Because this aim is difficult to formalize, it has been ignored in theory, except for "balanced" or "controlled" selections (7.2). 3) Stratification *facilitates allocations to domains* of desired sample sizes, often proportionate, but especially when disproportionate allocations are desired. Domains are subpopulations that may contain several, even many, strata; for example, large

regions or provinces may be divided into many strata before selections (8.1). 4) Stratification facilitates *using different methods and procedures* for diverse portions of the sample. For example, for selecting farms in the metropolitan area the sampling methods should differ from those in the rural area. Also, farmers may live in villages in some provinces but in the open country in others and therefore different procedures may be suited to each.

Weighted means. Fundamentally, means (and other statistics) from stratified samples represent weighted combinations of separate means:

$$\bar{y}_w = \Sigma W_h \bar{y}_h = \Sigma N_h \bar{y}_h / N . \qquad (5.3.1)$$

Capital W_h represent relative weights, so that $\Sigma W_h = 1$, and the weights may be freely chosen to satisfy the needs of substantive analysis. The weights $W_h = N_h/N$ signify common situations where numbers N_h of elements in the strata are used for weights. Note that the sampling methods used to obtain the stratum means $\bar{y}_h$ are not specified, and even the selection methods may differ between strata. If the $\bar{y}_h$ are unbiased estimates of the stratum means $\bar{Y}_h$, then the combined sample mean $\Sigma W_h \bar{y}_h$ is also an unbiased estimate of $\Sigma W_h \bar{Y}_h$. With separate, *independent* selections from the strata, there are no covariances between strata, hence the variance for the combined mean is the simple sum of the stratum variances, $W_h^2 \text{var}(\bar{y}_h)$:

$$\text{var}(\Sigma W_h \bar{y}_h) = \Sigma W_h^2 \text{var}(\bar{y}_h) . \qquad (5.3.2)$$

This is merely a general expression for variances of weighted means. *For separate SRS selections within strata,* the combined variance becomes the weighted sum of the stratum variances $W_h^2(1 - f_h)s_h^2/n_n$:

$$\text{var}(\Sigma W_h \bar{y}_h) = \Sigma W_h^2 (1 - f_h) s_h^2 / n_n . \qquad (5.3.3)$$

The sizes of the n_h have not been specified and they can be arbitrary. They may be determined by availability of data, and also by size requirements for separate domains. We distinguish two special, important allocations: proportional selections (5.4) and disproportionate "optimal allocation" (5.6).

Often the means $\bar{y}_h$ are *proportions* p_h and then

$$p_w = \Sigma W_h p_h \text{ and } \text{var}(p_w) = \Sigma W_h^2(1 - f_h)p_h q_h/(n_h - 1)\,, \qquad (5.3.4)$$

and with the p_h available this formula saves having to compute the values of s_h^2.

For *estimating totals* (aggregates) with the weights N_h, we may use

$$\hat{Y} = \Sigma N_h \bar{y}_h \text{ or } \Sigma N_h p_h \text{ and } \text{var}(\Sigma N_h \bar{y}_h) = \Sigma N_h^2(1 - f_h)s_h^2/n_n\,. \qquad (5.3.5)$$

For simple expansion totals n_h/f_h may be substituted for the N_h in $\Sigma y_h/f_h$, but this estimator is usually not as good as $\Sigma N_h \bar{y}_h$ (12.3).

5.4 PROPORTIONATE STRATIFIED RANDOM ELEMENT SAMPLING (PRES)

Theoretically PRES may be viewed as merely one special kind of allocation (5.6) for stratified element sampling in general (5.3), but in practice this is probably the most commonly used among all methods of element sampling. It is also the most popularly known sampling method: it is probably what people mean when they think of "representative samples," which are "miniatures of the population" in which different portions of the population are "properly represented."

We can also assign to it three technical advantages in comparison with other methods of random element sampling. First, it is often *simple to select*, sometimes even simpler than SRS. It may be especially simple when approximated with systematic selection applied to an already stratified listing of population elements (5.5). Second, it helps to satisfy the *safety motivation* for stratification: to guard against unusual or extreme results from simple random selections. Third, PRES may often yield allocations that approximate (close enough) reasonable compromises between the conflicting "optimal" allocations of multipurpose designs (9.5)

We may take two distinct views of PRES which are mathematically equivalent. A) *In PRES the sampling fraction* f_h *within all strata are made equal*: $f_h = f$ for all strata (h), where $f_h = n_h/N_h = f = n/N$. Thus the overall uniform sampling rate $f = n/N$ is applied to each stratum size N_h to obtain the sample sizes in the strata: $n_h = fN_h = f_hN_h$. B) *In PRES the sample sizes* n_h *represent proportionately the population sizes*: $n_h/n = N_h/N$ for all strata (h). Thus the sample is made into a "miniature" representation of the population.

These proportionalities make some simplifications possible in formulas for PRES:

$$\bar{y}_{pres} = \Sigma\,(N_h/N)\bar{y}_h = \Sigma\,(n_h/n)y_h/n_h = \Sigma_h y_h/n = n\Sigma_j y_j/n\,, \qquad (5.4.1)$$

because $\Sigma_h y_h = \Sigma_h \Sigma_j y_j = \Sigma_j y_j$, the two step summation, within and over strata, is replaced by simple summation over all n cases. Thus PRES is "self-weighting", like other epsem samples, because $W_h = N_h/N = n_h/n$, population weights equal the sample weights. For proportions from PRES the self-weighting mean is simply p. But it is also possible to introduce other weights W_h for improved estimators with ratio and other methods (12.3).

Whereas the self-weighting means for PRES may be computed with a single summation, the variances must still be computed separately within strata:

(5.4.2)

$$\mathrm{var}(\bar{y}_{pres}) = \frac{1-f}{n}\Sigma W_h s_h^2 = \frac{1-f}{n^2}\Sigma_h n_h s_h^2 = \frac{1-f}{n^2}\,\Sigma_h \frac{n_h}{n_h-1}\left[\Sigma_j^{n_h} y_{hj}^2 - \frac{y_h^2}{n_h}\right].$$

For *proportions* p_h the variance becomes:

$$\mathrm{var}(p_{pres}) = \frac{1-f}{n^2}\,\Sigma_h n_h^2\, p_h q_h/(n_h - 1)\,. \qquad (5.4.3)$$

For *totals* (aggregates) we may use, from (5.4.2):

$$\hat{Y}_{pres} = N\bar{y} \text{ or } Np \text{ and } var(N\bar{y}_{pres}) = N^2 var(\bar{y}_{pres}) . \qquad (5.4.4)$$

Designs for PRES.

The element variances S^2 of SRS selections may be decomposed mathematically into two components [Kish 1965, 4,6A; Cochran 1977, 5.3–5.6]:

$$S^2 \simeq \Sigma W_h S_h^2 + \Sigma W_h (\bar{Y}_h - \bar{Y})^2 . \qquad (5.4.5)$$

The variance $Var(\bar{y}_{pres}) = (1-f)\ S_w^2/n$, based on the *within stratum* variance $S_w^2 = \Sigma W_h S_h^2$, is less than the total variance, because the *between stratum* variance $\Sigma W_h(\bar{Y}_h - \bar{Y})^2$ is eliminated by proportionate stratification in PRES. In the stratification process we should aim at increasing the between–stratum variance in order to decrease the within–stratum variance of PRES. The relative reduction by PRES may be called the design effect of the PRES = S_w^2/S^2, which are measured by

$$\frac{var(\bar{y}_{pres})}{var(\bar{y}_{srs})} = \frac{(1-f)s_w^2/n}{(1-f)s^2/n} = \frac{s_w^2}{s^2} . \qquad (5.4.6)$$

Both values can be computed from the PRES sample, where $s_w^2 = \Sigma n_h s_h^2/n = \Sigma W_h s_h^2$ measures the element variance (within strata) of PRES; and s^2 is the simple variance of sample cases. This ratio, $D_{pres}^2 = S_w^2/S^2$ is "always" less than 1.0 but seldom less than 0.95 or 0.9 or 0.8; thus the gains in variance of PRES are seldom greater than 5,10 or 20 percent for the mean $\bar{y}_{pres}$ of entire samples. For proportions these gains tend to be small, close to 5 percent, because the element variances $S_h^2 = P_h Q_h$ are not sensitive to changes in P_h (5.6).

Furthermore, even those *modest gains of PRES* tend to be further *reduced for subclasses* that are crossclasses (cut across strata), *reduced in the proportion* $\bar{M}_c$ *of the crossclass.* So that a gain of 20 percent for an entire sample (e.g., of all farmers) will be reduced to a gain of 2 percent for a crossclass of $\bar{M}_c = 1/10$ (e.g., a five year age group of all farmers). For

comparisons of two crossclass means ($\bar{y}_a - \bar{y}_b$) (e.g., difference of means for two age groups) the gains tend to be eliminated almost entirely. Thus for *crossclass analysis of PRES samples, the simple SRS formulas may yield* good approximations, only modest or negligible overestimates of variances. For small subclasses these may actually be preferable, when the sample sizes n_{ch} become too small and unstable for crossclasses (c) within strata (h) (8.3).

For PRES both the need and opportunities for highly efficient stratification are often less than either for "optimal" element selection (5.6) or for clustered samples (7.2). Therefore we may treat the procedures of stratification more briefly here and present only a few alternatives. 1) The listing of population elements may come already separated by strata (e.g., by provinces, districts, etc.) that may be adequate. Then a formal random selection can be applied with the same f to select $n_h = fN_h$ from each of the strata; or a simple approximation with the systematic interval $F = 1/f$ may be applied throughout (5.5). When selecting from several lists, the lists themselves can prcbably constitute strata also. 2) The N elements may all be sorted into the H strata and the sampling rate f applied to each. When several stratifying variables are used and each with several classes, then too many cells may result and "multiple stratification" may be needed. 3) When sorting all the N—elements seems too laborsome, a process of "random quotas" may be useful, if the stratum sizes (N_h or W_h) are known: select at random (SRS or an approximation) a sample with a preliminary f ' rate larger than the required f and then eliminate at random enough of the preliminary n'_h elements to get the sample size down to the required size $n_h = fN_h$.

5.5 SYSTEMATIC SAMPLING OF ELEMENTS (SYS).

Procedures for SYS are simple and they are widely known. The interval of selection is computed as $k' = N/n$, after the desired sample size n and the population size N are determined. Integers for k are preferable and an integral $k = N/n'$ can yield a reasonably close n; but techniques for using fractional k

are also easy [Kish 1965, 4.1B]. *Select a random start r from 1 to k and then apply the interval k to designate the selection numbers r, r+k, r+2k, etc.* This will select N/k elements, one element from each of the n' intervals of size k from the N listings. The last interval can be smaller than k and yield either 0 or 1 selection; thus the sample size n' can vary by one selection. But much greater problems of variation often occur due to blanks and other frame imperfections (4.2).

Stratification in the ordering of the N listings commonly exists in the composition of the list. Or it can be introduced by sorting on stratifying variables to yield designed strata, and these can be fitted to the selection interval k. Thus the strata can be of size k or 2k or ik, with i any integer. However, when we accept existing sortings within strata, it may still be worthwhile to apply the intervals to the strata linked in some meaningful order. The selection interval is then applied to strata after strata, with some intervals including fractions of two strata. This notion is the basis for the "serpentine" order of numbering area segments for agricultural and other samples.

Systematic selection is commonly used, especially as a *simple alternative to SRS and to PRES.* It is simpler to apply and to supervise than independent random selections, especially in the field work. For example, it is often too risky and difficult to trust the field workers to select a prefixed sample size n_i of farms or dwellings from the ith block. But they can be instructed to apply an interval k, after finding the random start r (from 1 to k) in sealed envelopes (10.4)

. *Unequal probabilities between strata* can be introduced with easy modifications. To increase the sampling rates in some strata from f to i × f, one may either a) use a smaller interval k/i, or b) use i distinct random starts $r_1, r_2,...,r_i$ and the interval k with each. It may even be better to select a complete sample with the shorter interval k/i (thus with a higher rate) and then subselect with the interval i in the strata that is to receive only f. Or to

reverse symbols, when k is an overall interval we can subsample in some strata with the interval i, in order to lower the sampling rates from f = 1/k to f/i = 1/ik.

Systematic samples selected with the uniform rate f = 1/k are epsem and the self–weighting sample mean and simple expansion totals are:

$$\bar{y} = \Sigma_j y_j/n = y/n \text{ and } \hat{Y} = y/f . \qquad (5.5.1)$$

For computing variances we must face a difficulty discussed below: because the entire sample depends on a single selection (r), computing variances must be based on one of several alternative models.

1) The population list may be divided into large strata and then the sample treated as if it were selected with *PRES*, with variance (5.4.2). This model pretends that within strata the n_h cases were selected with random rather than the actual systematic selection used.

2) One extreme of 1) would disregard stratification and use an *SRS* variance formula (5.2.6) When the ordering of the population list is a powerful stratifier this SRS formula overestimates the variance of the sample, because it neglects the reduction induced by the stratification. This upward bias would be less and perhaps negligible in PRES (1). It may be advisable to make some computations of the design effect by computing the ratios of PRES variance to SRS variance. This may also be done with the paired selection formula below to see if the effects of implicit stratification induced by the SYS selections are small or even negligible. Note that even moderate variance reductions by PRES tend to vanish toward $deft^2 = 1$ for SRS for small subclasses (8.3).

3) Other approaches would recognize the fine stratification achieved with systematic sampling. It assumes a model of n/2 "pseudo–strata" and random *paired selections* within those strata. However, instead of only the n/2 "even" pairs of contrasts, like $(y_1-y_2)^2 + (y_3-y_4)^2$ + etc, it is preferable (13.2) to use *all* the (n−1) pairs available from $(y_1-y_2)^2 + (y_2-y_3)^2 + (y_3-y_4)^2 + \ldots + (y_{n-1}-y_n)^2$ in:

$$\text{var}\,(\bar{y}_{sys}) = \frac{1-f}{2n(n-1)} \sum_{j}^{n-1} (y_j - y_{j+1})^2 \,. \qquad (5.5.2)$$

4) *Combining strata* has the same expected (average) value as the above, but it also avoids the problems of only two selections per "pseudo—stratum". Missing observations create blanks (zero) in the computations, these blanks would dominate in variances for subclass means. These may be treated with the PRES or SRS formulas of 2 or 1 above. But combined strata offer alternatives as well as briefer computations in some situations.

Suppose, for example, that about n = 800 selections were made with SYS. These can be divided into 40 replicates for computing variances, each replicate containing *about* 20 selections:

Replicate 1 contains selections 1 + 41 + 81 + ... + 761
" 2 2 + 42 + 82 + ... + 762

" 40 40 + 80 + 120 + ... + 800

Some of these selections may be missing, or blank, or nonmembers of the subclass. The number of elements per replicate may be larger (than 20 here) for stability. But the number of replicates may also be greater (than 40 here) for more "degrees of freedom" and stability of the variance estimation. The 40 replicates can be used either for 20 pairs or preferably for 39 successive differences as in 3 above.

Three theoretical problems of systematic sampling must be noted here, and because of them some statisticians prefer to avoid SYS altogether. Others would avoid SYS for primary sampling units in multistage selection, but use it within later stages where they are less risky and more useful (7.3). Those concerned with these problems may consult Cochran [1977, Ch. 8] or Kish [1965, 4.2].

1) First, SYS *are probability samples*, when each element receives the selection probability 1/k with random starts r from 1 to k. However, the single random selection (r) determines the entire sample, without independent

replicates for measuring sampling error. Thus, based on single selections, *SYS are not measurable samples* strictly speaking, and therefore they require *models for measuring sampling errors*, as we saw above. Because of these concerns, some statisticians prefer to take *separate random starts for each major stratum*; but this causes a variation of one selected unit in each stratum that is not exactly of size k; it also complicates the selection procedure." It is also possible to use i random starts, each with interval ik = i/f instead of k = 1/f. This would be a special case of *replicated selection*; but this has complications and problems also (13.5)

2) A strong, *consistent linear trend* could lead to biased estimates of the variance. But this is less likely than an irregular monotonic trend, or some mild, smooth trends that merely result in weak stratification, which the variance formulas can reflect adequately.

3) A *regular periodic fluctuation* with the period k or 2k or ik or k/i could be disastrous. But it seems hard to imagine a practical situation when this would occur and the practicing sampler would remain unaware of it.

5.6 OPTIMAL ALLOCATION

Assume the following situation: 1) Population elements are sorted into strata, so that $N = \Sigma N_h$. 2) The chief (or single) aim of the sample is to produce either the sample mean $(\bar{y}_w) = \Sigma W_h \bar{y}_h$ or the total $\hat{Y} = \Sigma N_h \bar{y}_h$. 3) Allocate the sample sizes n_h so that with either the variance $Var(\bar{y}_w)$ or the cost $\Sigma c_h n_h$ fixed, the other is minimized.

Under these conditions it may be shown [Cochran 1977, 5.5; Kish 1965, 4.6B] that the *optimal allocation* of the n_h would occur with

$$n_h \propto W_h S_h/\sqrt{c_h} \text{ or } f_h = \frac{n_h}{N_h} \propto S_h/\sqrt{c_h} \text{ or } n_h \propto W_h S_h \,. \tag{5.6.1}$$

The first of these in general for any weights W_h that are relative (i.e., $\Sigma W_h = 1$), whereas the second uses population sizes N_h for weights, so that $W_h = N_h/N$. In the third the cost factors $\sqrt{c_h}$ are neglected, because in many situations the element costs c_h do not differ enough between strata to make variations in $\sqrt{c_h}$ important.

In the right (but rare) situations "optimal" allocation (OPT) can produce spectacular reductions of the variance for fixed allowed cost $\Sigma\, c_h n_h$ – or reductions of the cost for fixed "required" variance, $\mathrm{Var}(\bar{y}_w)$. The reductions can be estimated with the third term in

$$\frac{\mathrm{Var}(\bar{y}_{opt})}{\mathrm{Var}(\bar{y}_{srs})} = \left[1 - \frac{\Sigma\, W_h(\bar{Y}_h - \bar{Y})^2}{S^2}\right] - \frac{\Sigma\, W_h(S_h - \bar{S})^2}{S^2} \tag{5.6.2}$$

The third term measures the reduction of the variance due to OPT allocation $n_h \propto W_h S_h$, and it depends on large variations among the S_h around the mean $\bar{S} = \Sigma\, W_h S_h$. For an allocation $n_h \propto W_h S_h/\sqrt{c_h}$ use a cost–weighted mean $\bar{S}_h^{\,\cdot} = S_h\sqrt{(c_h n_h/\Sigma\, c_h n_h)}$. The second term measures the reduction due to PRES over SRS selection, and it depends on large variations among the $\bar{Y}_h$ around $\bar{Y}$.

The allocations are only proportionalities, but constants are available to yield fixed number, [Cochran 1977, 5.5; Kish 1965, 4.6B]. For example for fixed $\Sigma\, c_h n_h$ make $n_h = K W_h S_h/\sqrt{c_h}$, where $K = \Sigma\, c_h n_h/N\Sigma\, W_h S_h\sqrt{c_h}$. There is a similar constant for minimizing $\Sigma\, c_h n_h$ for fixed "required $\mathrm{Var}(\bar{y}_w)$, and two more when the cost factors c_h are disregarded and $n = \Sigma\, n_h$ is either fixed or minimized. However, even without these constants proportionality is sufficient because the n_h can be adjusted up or down to the allowed $\Sigma\, c_h n_h$, or to the

"required" $\mathrm{Var}(\bar{y}_w)$. Proportional adjustments can also take care of the problems due to the limits $n_h \leq N_h$, or $f_h \leq 1$; and larger allocations must be reduced and the surplus added to the other n_h.

Means based on only two strata $\bar{y}_w = W_1\bar{y}_1 + W_2\bar{y}_2$ present interesting special cases and the "optimal" allocation (5.6.1) becomes:

$$\frac{n_1}{n_2} = \frac{W_1S_1/\sqrt{c_1}}{W_2S_2/\sqrt{c_2}} \tag{5.6.3}$$

However, we may be also interested in the *difference* of the two domains means $(\bar{y}_1 - \bar{y}_2)$, when the "optimal" allocation should be

$$\frac{n_1}{n_2} = \frac{S_1/\sqrt{c_1}}{S_2/\sqrt{c_2}}. \tag{5.6.4}$$

This is a simple instance of conflict between two aims: the weighted sum is based on the weights W_h, whereas for the comparisons the weights are equal. The two allocations would be similar when $W_1 = W_2 = 0.5$, but if the two parts differ greatly in size there arises a conflict between allocation for the comparisons (differences) with (5.6.4) and allocation for the weighted sum with (5.6.3), which was the stated aim at the start of this section. These conflicts will be explored under multipurpose designs (9.5).

A fairly common situation for establishments (farms, stores, firms, etc.) is to have a highly skewed distribution of very different sizes, with a small portion of very large units accounting for a much larger portion of total outputs (production, sales, employment, etc.) and of total variability. In some cases it may be both possible and desirable to separate the large units into a "certainty" stratum of complete coverage ($f_1 = 1$) and with a selection of f_2 for the rest of the smaller units. The basic question of design is to find the boundary (approximately perhaps because of existing groupings) that minimizes the variance for a fixed total $n = n_1 + n_2$ or total cost =

$c_1n_1 + c_2n_2$. This can be done with trial and error or with available formulas [Hiridoglou 1981]. The first stratum with $f_1 = 1$ and $(1-f_1) = 0$ has no sampling error, hence $\text{var}(\bar{y}_w) = W_2^2\text{var}(\bar{y}_2)$. Sometimes two separate frames have been used, with a good (though not "perfect") list for the large units (farms, stores, etc.) and area segments for finding the smaller and the missed units [HHM, 12A.11 and 11.6]. Further modifications are feasible; for example $f_1 < 1$ though large; also instead of only one f_2, two (or more) strata with f_2 and f_3 may be used for medium and small units.

A different use of two strata occurs in cases of allocation for *nonresponses*: inexpensive methods (mail, telephone, registers) that obtain most of the responses with sampling rate f_1 (perhaps a census $f_1 = 1$), but a much lower rate f_2 is used for the stratum of nonresponses, with a more expensive ($c_2 > c_1$) method (15.4). Another extension of two strata uses *two–phase sampling with screening*, with a larger f_1 for the stratum of positives than the f_2 for the negatives, obtained on a preliminary, inexpensive and imperfect test for "susceptibles" (12.4).

Several lines of guidance may be stated simply and usefully in qualitative terms; these can be reasonably quantified with feasible guesses about the parameters (S_h,c_h especially) (9.4).

1. *"Optimal" allocation can yield large*, even spectacular, *gains in some "proper special" situations*, when the strata can be identified with large differences in values of S_h and/or c_h, and when they can be guessed well enough. These situations seldom arise for sampling persons or households. But they do occur for establishments with highly skewed distributions, when reliable, though imperfect, data are available for efficient stratification that distinguish well $S_h/\sqrt{c_h}$ for the survey variable Yi. Generally, for good reductions of variances, ratios of several fold are needed in the $S_h/\sqrt{c_h}$; that is ratios of 4 or 10 or more, which means ratios of 16 or 100 or more in the S_h^2/c_h.

2. *Approximate values only* of $S_h/\sqrt{c_h}$ are usually available and needed. If exact design parameters were available, we would not need the survey. But approximate values of $S_h/\sqrt{c_h}$ are also adequate. Modest or large mistakes (by factors less than 2 to 4 in $S_h/\sqrt{c_h}$ or 4 to 16 in S_h^2/c_h) only incur trivial or small departures (less than 4 to 10 percent) from optimal gains. For these reasons I have often written "optimal" in quotation marks.

3. *The multipurpose* nature of most surveys presents much more formidable problems (Ch. 9). a) Above we noted a conflict between designs for domains and comparisons and for overall means. b) Different survey variables may require conflicting allocations. c) Different statistics for the same variable may also have conflicting "optimals"; e.g., whereas means may require disproportionate allocations, "optima" for medians may be close to proportionate (9.1). d) In case of multipurpose conflicts between recognized aims (also some unrecognized) a simple PRES may be best. However, large establishments, for example, may have large values of $S_h/\sqrt{c_h}$ for most statistics, and some compromise "optimal" may be better.

4. *Proportionate* sampling is probably preferable for *small differences* between $S_h/\sqrt{c_h}$; e.g., not over 2 (i.e., not over 4 for S_h^2/c_h). Therefore, it seldom pays to depart from PRES for sampling for proportions, because $S_h = \sqrt{P_h Q_h}$ seldom varies by ratios over 2. Also, sampling for persons or families seldom justifies departing from PRES, because large departures from uniform $S_h/\sqrt{c_h}$ are seldom available for allocation.

5. *Simplicity* of analysis also favors PRES. If departurec for "optimal" allocation are needed, a *few strata* (2 to 5) may be adequate. The sampling rates may be *simple integral multiples*, so that $f_h = if$, or $f_h = f/i$, where f is a base rate and i some simple integers.

6. The concepts and formulas were developed for element sampling, but they are also applicable to cluster and multistage sample designs.

CHAPTER 6. CLUSTER AND MULTISTAGE SAMPLING

6.1 REASONS FOR CLUSTER SAMPLING

Chapter 6 and 7 on cluster sampling contain the most important aspects of this book, because cluster sampling is so frequently used in survey sampling, especially for agricultural surveys. Furthermore, clustering also represents the most basic departures, both theoretical and empirical, from most of the results of standard statistical analysis. Those mathematical results depend heavily on assumptions of independence (I.I.D.) between elements, which is contradicted by the correlations of elements within clusters in most surveys, and particularly in agriculture. That difference of independence, assumed in statistics and lacking in surveys, explains the principal justification for the field of sample surveys.

Instead of element sampling (Ch.5), the sampling units contain (often, typically) several or many elements. For example, samples of holdings (farms) and households can be based on selections of villages or segments. The population of elements (farms, or farmers, or households) are defined by survey objectives (1.1, 2.1). However, the sampling units used in the selection and for the collection of data depend on the situation of the survey and the resources of the survey organization. The sampling units are usually based on actual, often spatial, and sometimes social organization of the population; thus correlation of elements within clusters is usually inevitable, because elements of existing units tend to be similar – more or less, but almost invariably (6.6).

Typically, in cluster samples, as compared with element sampling we find that: 1) *The cost per element is lower*, due to the lower costs of both listing and data collection (locating). 2) *The element variance is higher* ($deft^2 > 1$), because of the usual, though irregular, average positive correlation (homogeneity) of elements within clusters (6.6). 3) *Statistical analysis is more*

complex, especially the inferential statistics, e.g., ste($\bar{y}$). Cluster sampling is used widely because the advantage of lower element costs more than compensates for the increased element variances and the increased complexity.

The nature of clusters and their use depend both on the cost of listing and on the costs of locating and collecting data. First, assume that a good, complete, up—to—date listing is not readily available on tapes or sheets. Then the listing costs depend on the size and the spread of the population, and also on the size and number of samples to be selected from a listing or frame. For example, it may be feasible to list a village with 500 farms or a district with N = 10,000, but too expensive for a province of 100,000 farms. However, even this may be feasible for a large sample of n = 10,000, because the ratio of listing to sample is 10:1. Furthermore the cost of the frame will seem less expensive if it can be used for several surveys without too many changes (11.6). Also 100,000 farms (or households) are more expensive to list if they are widely spread in a large province rather than concentrated in a fertile valley (or city).

However, even if a good frame exists (as in the population registers of Scandinavia), the costs of locating elements and collecting data may be much less expensive for clustered samples — unless mailed or telephone surveys are appropriate. Even good lists get quickly dated due to migration, births and deaths. Also clusters are often more convenient for field interviewers. The methods of data collection are important; and the need for callbacks may require a longer stay than a mere hour's interview. These considerations must also be related to the spread of the population, because a population spread over a large country or province needs clustered collection, whereas households of a city may be interviewed individually. Over a large country clustered samples can be operated and supervised more conveniently.

6.2 STRATIFIED SAMPLING OF UNEQUAL CLUSTERS

Much of sampling literature deals with random selections of equal sized clusters, because mathematical derivations and concepts can be more clearly and readily developed in that framework. Textbooks then proceed to develop formulas for selecting at random a sample of a clusters from the population of A clusters, then subsampling b from B elements from each of the a sample clusters. Thus from a population of $N = AB$ elements an *epsem* selection of $n = ab$ elements can be selected in two stages of unstratified random selection, with the uniform overall sampling rates of $f = (a/A) \times (b/B) = ab/AB = n/N$. We may consider that simple design as a special case of the more complex situations in practical samples of real populations.

However, in our brief treatment of cluster sampling we begin directly with unequal clusters, where the population consists of $N = \Sigma N_\alpha = (N_1 + N_2 + \ldots N_\alpha \ldots + N_A)$ elements. Selecting a sample of a from the population of A clusters with equal probability $f = a/A$ for all clusters would yield a sample of $n = \Sigma n_\alpha = (n_1 + n_2 + \ldots + n_a)$ elements. Note that: 1) All N population elements receive the same equal probability of selection $f = a/A$ that the clusters receive because the selection of any cluster results in the selection of the elements within it; 2) The size of the sample becomes a random variable, because the cluster sizes n_α of the sample vary. Because of that variation the sample mean $\bar{y} = y/n = \Sigma y_\alpha / \Sigma n_\alpha$ becomes a ratio of two random variables, which has some methodological consequences. Methods for controlling extreme variations in the cluster sizes n_α and in the total sample size $n = \Sigma n_\alpha$ are presented later, but without eliminating all variation (7.4).

Furthermore, we also begin directly with stratified selection of clusters, because stratification is most useful and most commonly used in cluster sampling, and unstratified random selection is rare in practice; it may be viewed as a special case when $H = 1$. In general the sample will be perceived as composed of $n = \Sigma_h n_h = \Sigma_h \Sigma_\alpha n_{h\alpha}$ sample elements, with $n_h = \Sigma_\alpha n_{h\alpha}$

elements from the h-th stratum (h = 1,2,...H). These come from $a = \Sigma_h a_h$ sample clusters randomly selected from the A_h population clusters in the h—th stratum.

Complete clusters is a name for the technique described above of including all n_α elements of the a selected clusters. This technique is simple both in conception and for practical field procedures of coverage (7.1). However, often the sizes of available and identifiable clusters are both too large and too variable for efficient sampling; for example, villages in some countries may vary from 10 to 1000 or even more farms or households, with an average of perhaps 100. *Subsampling* of the primary clusters then becomes necessary if a population listing of smaller units are not available. This leads to *two—stage sampling* of primary selections and of elements from them. More generally, we may need *multistage sampling*, first of *primary selections* and last of elements, but perhaps one or two (or more) stages in between (6.5). Note the emphasis on the n elements defined by analytical needs, and on the a primary selections that depend on the sampling resources and design. The importance of the kind and number of the a primary selections will be found both in practical considerations of field costs and in the formulas for computing variances.

Techniques based on *a primary selection* for sample designs and for variance computations are widely known and used under different names: *simple replication* [Kish 1965, Ch 6] and *ultimate clusters* [HHM 1953, 6.7; Kalton 1979]. Briefly and directly the ratio means can be written as

$$r = \frac{y}{n} = \frac{\Sigma_h y_h}{\Sigma_h n_h} = \frac{\Sigma_h \Sigma_\alpha y_{h\alpha}}{\Sigma_h \Sigma_\alpha N_{h\alpha}} . \qquad (6.2.1)$$

The variance of this ratio means may be computed as

$$\mathrm{var}\left(\frac{y}{n}\right) = \frac{1-f}{n^2}\left(\Sigma_h dy_h^2 + r^2 \Sigma_h dn_h^2 - 2r\Sigma_h dy_h dn_h\right) . \qquad (6.2.2)$$

Here

$dy_h^2 = (a_h \Sigma_\alpha y_{h\alpha}^2 - y_h^2)/(a_h - 1)$ and $dn_h^2 = (a_h \Sigma_\alpha n_{h\alpha}^2 - n_h^2)/(a_h - 1)$

also similarly $dy_h dn_h = (a_h \Sigma_\alpha y_{h\alpha} n_{h\alpha} - y_h n_h)/(a_h - 1)$.

These are general formulas that we shall use throughout for selection designs for cluster means and their variances. We shall see some simplifications and modifications for special cases, especially for paired selections. They are based on the general formula for ratio means $var(y/n) = [var(y) + r^2 var(n) - 2rcov(y,n)]n^2$. But note that all computations are based only on the $a = \Sigma_h a_h$ pairs of values for the primary computing units $y_{h\alpha}$ and $n_{h\alpha}$. The more complex and lengthy formulas for multistage sampling may be disregarded under this formulation of simple replication based on primary selections (ultimate clusters).

We assume equal probabilities of selection *f* for all N elements in the population, but we shall see that weighted elements can also be introduced into the computing units $y_{h\alpha}$ and $n_{h\alpha}$ (12.5). The factors $(1 - f)$ can be modified to $(1 - f_h)$ inside the summations for strata when different f_h are used within different strata. We shall also note that equal f for elements can be obtained in multistage sampling with probabilities proportional to measures of size (PPS) (7.4).

The presentation is organized around the ratio means $\bar{y} = y/n$. Other statistics will be discussed later. But we note now that estimates of the total (aggregate) Y are more commonly and preferably based on the ratio estimate $N\bar{y}$ rather than on the simple (unbiased) expression $\hat{Y} = \Sigma_j y_j/f_j$ (12.3).

6.3 STRATIFICATION FOR PRIMARY SELECTIONS

A. *Variance reductions from stratification are more feasible, frequent and larger for clustered and multistage sampling* than for PRES selections of elements. A chief reason is that in the gains $\Sigma W_h(\bar{Y}n_2 - \bar{Y})^2/S_u^2$ for stratified clusters, the denominator S_u^2 for variances between cluster means is much less than S^2 in $\Sigma W_h(\bar{Y}_h - \bar{Y})^2/S^2$ for element sampling; S_u^2 may be on the order of S^2/b,

and the relative gains from stratification b times greater for clustering. If stratification reduces the PSU (primary) variance by 1/k, the variance for a PSU's (districts or E.D.'s) stratified equates to ka unstratified PSU's; k varies between surveys and variables, but $k = 2$ is not surprising. Furthermore, much more *data are available for stratifying cluster means* than for elements, and this permits better sorting of clusters, especially PSU's, into strata with different means $\bar{Y}_h$.

B. *Other motives* also exist for introducing stratification into most clustered samples, stronger reasons than for element sampling (5.1). There are strong urges for better spread for *safety* across the many available stratifying variables, while the number of PSU's must often be limited by high costs. Linked to those desires for safety, there are also "public relations" aspects for making the sample of PSU's appear "representative" over the population.

Usually the PSU's must represent not only the entire population but also specified major *domains*, such as provinces. Stratification within these domains helps, especially because relatively few PSU's within domains are permitted by their costs. *Disproportionate allocation* between domains may result from concerns for domain estimates. *Different sampling procedures* may be introduced into different domains and strata. For example, procedures in urban and metropolitan areas for households, but especially for farms, may differ greatly from rural procedures; and sampling in rural areas with dense, irrigated, small farms may differ from sampling large, dry, livestock farms.

C. *Numbers of PSU's, of strata and of stratifying variables* become sources of conflict for reasons implied above. Economic reasons bring severe limits on the number of PSU's, say between $a = 20$ and $a = 200$ usually, seldom higher. But for surveys using mobile teams, sometimes more numerous (several hundred) small PSU's have been used. However, if we would use 6 stratifiers, each with only 3 classes we would need $3^6 = 729$ strata, thus $a = 2 \times 729 = 1458$ PSU's with only $a_h = 2$ PSU's per stratum. Or, for fine stratification with 27 classes for two variables one would also need $27^2 = 729$

strata. The need for several stratifying variables and the availability of data for them, tend to lead to only a few, broad classes for each variable, and to small number of PSU's per stratum, often $a_h = 2$. The need for several stratifiers is enhanced by needs for *multipurpose* design for most surveys (Ch. 9.3).

Theory can guide us, first by showing that a *few classes each on several stratifiers tend to yield higher gains* for most purposes, and especially for multipurpose designs, than many fine strata for one or two stratifiers [Kish and Anderson 1978]. Second, this arises chiefly because a few classes (3, 4 or 5) *yield most of the gains* from any stratifier. Third, it is best to use stratifiers that are not strongly related to each other, but strongly to the survey variables.

To compute variances, for measurability (3.5) we must select from each stratum at least $a_h = 2$ sample PSU's; thus the number of strata should be limited to $H = a/2$, as in paired selections below, where single selections, with $H=a$, are also discussed (6.4). However, some large samples use techniques of *multiple stratification* (or controlled selection or deep stratification or lattice designs) that permit the use of much greater numbers of cells than a [Kish 1965, 12.8; Hess, Riedel, Fitzpatrick 1975].

D. *Symmetry, regularity, and objectivity* are not necessary in the procedures for forming strata. This flexible, judgmental approach to stratification, and generally to design decisions, stands in contrast to the objectivity that we emphasize for the selection process. Flexibility in design helps to cut down on the number of strata needed. For example, most of the cells of the 729 strata we mentioned above may be empty (nearly or entirely) and these can be combined with others, so that we may have $H = 100$ or 200 strata instead of 729. These empty cells arise because correlations between stratifiers are unavoidable (despite our efforts).

E. *"Optimal stratification" of boundaries* for strata refers to techniques for choosing the best boundaries for strata. Quantitative stratifiers often seem to have skewed frequency distributions, with small portions of sampling units possessing large portions of the stratifier. For example, much of some important crop or farm product may come from only a small portion of districts. If the population were divided into equal sizes, with W_h constant, the range of y values $(y_h - y_{h+1})$ and S_h on the extreme(s) would become too wide. But if the ranges of the strata $(y_h - y_{h+1})$ and S_h are controlled then the stratum sizes W_h vary greatly. Theory shows that the compromise with $W_h S_h$ constant would yield optimal boundaries [Dalenius 1959]. This has been developed into a practical procedure of cumulating values of $\sqrt{f_y}$, where f_y is the frequency at the value y, and dividing that cumulation into equal parts [Cochran, 1977, 5A7]. The technique needs to be modified for common situations when (a) the stratifiers are qualitative, and (b) the data are only available or must be used within arbitrary class divisions.

F. *Equal allocation* denotes the selection of constant numbers of sampling units from strata. This approach can be convenient, and the special case of a_h constant with $a_h = 2$ receives much attention here and elsewhere (6.4). Furthermore, it also has theoretical basis in (a) optimal stratification with $W_h S_h$ constant, and (b) optimal allocation with sample size $a_h \propto W_h S_h$ (5.6); together they point to a_h constant.

G. *Stratification for later stages* usually receives less attention than for PSU's, though it is practiced. For example, a design for farms may use districts or counties for PSU's, but sampling farms directly from large PSU's may be too costly, hence E.D.'s and segments (or blocks) may be introduced as second and third stages. We may well introduce stratification into the selection of E.D.'s and segments, perhaps with simple systematic selection applied to ordered listings. More formal stratified selections in later stages are not likely for several reasons. First, there are too many populations (E.D.'s in each district, segments in each E.D.) to be treated at length. Second, there is probably less

to be gained from stratification in those later stages. Third, less data are available for stratifying those smaller and more numerous units. Nevertheless, a modest effect is probably worthwhile, though there exists little empirical evidence on these aspects.

6.4 PAIRED SELECTIONS

Paired Selections refer here to a special case of stratified cluster sampling, when $a_h=2$ in all strata. They serve here as basis for most clustered designs, described with several procedures in Chapter 7. Paired selections are used often both in the literature and in actual designs. They also appear often as useful models and approximations for designs, which are not strictly measurable, as we note below. Two principal reasons account for the popularity of paired selections in theory and practice.

First, selecting two replicates per stratum permits the *maximal* number of strata $H=a/2$ for a number a of PSU's, fixed and limited by costs, but also subject to the need for at least two random replicates per stratum for measurability. This also means "equal allocation" per stratum, which joins optimal allocation with optimal stratification (6.3). In practical terms this also means that if a stratum is so large that it warrants $a_h=4$ or 6 PSU's, we generally have enough stratifiers that we can use in order to split it into 2 or 3 strata with $a_h=2$ from each.

Second, variance calculations can be somewhat simplified with $a_h=2$, instead of a variable $a_h>2$. The computing units for var(r) in (6.2.2) are simplified by this identity:

$$dy_h^2 = (a_h \Sigma\, y_{h\alpha}^2 - y_h^2)/(a_h - 1) = (y_{ha} - y_{hb})^2 , \qquad (6.4.1)$$

$$dn_h^2 = (n_{ha} - n_{hb})^2, \text{ and } dy_h dn_h = (y_{ha} - y_{hb})(n_{ha}' - n_{hb}) .$$

With these we have:

$$\text{var}\left(\frac{y}{n}\right)=\frac{1-f}{n^2}\left[\Sigma_h(y_{ha}-y_{hb})^2 + r^2\Sigma_h(n_{ha}-n_{hb})^2-2r\Sigma_h(y_{ha}-y_{hb})(n_{ha}-n_{hb})\right] .$$

The ratio (6.2.1) also appears a little simpler with:

$$r = \frac{y}{n} = \Sigma_h\Sigma_\alpha y_{h\alpha}/\Sigma_h\Sigma_\alpha n_{h\alpha} = \Sigma_h(y_{ha} + y_{hb})/\Sigma_h(n_{ha} + n_{hb}) . \qquad (6.4.2)$$

In a similar manner, variances for more complex, analytical statistics are also simplified. This has first been pointed out by Keyfitz [1957], and paired selections are sometimes referred to as the "Keyfitz method" (but confusingly that name is also used for two other techniques). Though convenient, paired selections are not necessary, as some have suggested, for complex sampling. Also, for modern computing programs even that convenience becomes less important, though it still remains useful for hand calculators. Similarly paired selections remain useful for "half—sample replications," but the BRR method can be modified for $a_h > 2$ (13.5).

Paired selections per stratum are well suited to robust *combined* estimators, of which the ratio mean (6.4.2) is a prime example. This ratio of means differs greatly from $\Sigma_h W_h y_h/n_h = \Sigma_h W_h(y_{ha}+y_{hb})/(n_{ha} + n_{hb})$ called the separate ratio estimator (12.7). We should also warn that these separate ratios, if based on only two selections in the denominators, would be very unstable.

Paired selections also serve as models for variance computations for selection procedures with only one selection per stratum, which are commonly used for two reasons. First, more control with further stratification may be designed by dividing each stratum into two strata. Second, selection procedures can be simplified especially for PPS selection without replacement (7.4). For all designs with single selections, the variance computations are based on pairs of *collapsed* strata to simulate models of paired selections (13.2).

6.5 SUBSAMPLING: MULTISTAGE SELECTIONS

The first and major decision concerns giving up the simplicity of element sampling for the greater complexities and variances of cluster sampling. The simplest cluster samples would be based on finding or creating clusters of the right kind and size, to select a sample of these, and include all elements from these *complete clusters*. This can actually be done economically in some situations and we gain some advantages from such one—stage cluster samples. We could select a clusters from a population of A clusters as a/2 paired selections from H = a/2 strata. EPSEM selections with f = a/A would yield a simple (stratified) random sample of complete clusters (7.1).

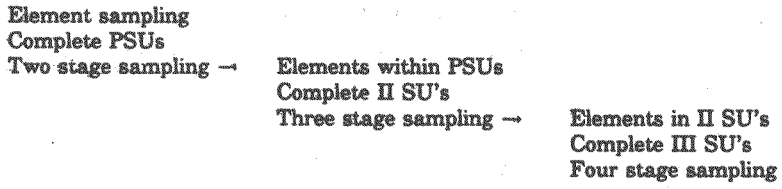
Second, we often cannot find nor create economically, clusters of the right size and kind for direct and complete selection. For example, when the sampling units are districts or villages their average sizes $\bar{B}$ may be too large and their individual sizes B_i too variable for complete coverage. The available clusters must be subjected to *subsampling from the primary selection of a clusters*.

Note two sources of cost reductions: only the fraction $f_1 = a/A$ needs to be listed; also the travel and locating costs for the subsamples $\bar{b}$ may be confined to the populations $\bar{B}$ of the secondary sampling unit (SSU's). From this we may deduce that usually we try to make the selection rate for the first stage, $f_1 = a/A$, small, in order to reduce the costs for the second stage; thus $f_1 = 1/100$ or $1/1000$ are common. On the other hand we prefer small $\bar{B}$ and large second stage selection rates $f_2 = \bar{b}/\bar{B}$, in order to keep low the cost of listing within clusters. This means that we try to find PSU's that are small and numerous; also preferably not too variable in sizes B_i.

Census enumeration districts (E.D.'s) are fairly small and numerous and often used for two—stage sampling. But where districts or villages (and other "natural" and administrative units) are used for PSU's, not only is the coverage size $\bar{B}$ often too large, but the variation in the cluster sizes B_i is too great. We should avoid great variations in the subsample sizes b_i, but we

should also avoid unequal sampling rates for elements from different PSU's. This dilemma is solved by *subsampling designs with probabilities proportional to size (PPS)* (7.4).

Figure 6.5.1 – Decisions leading to Multistage Sampling.



We described until now three alternative methods from which the basic design must be chosen: element sampling, or complete clusters, or *subsampling with two stages of selection*. The subsamples within PSU's may also be selected with one of three alternate methods. 1) List all the N_i elements in each of the selected a PSU's, and select a subsample n_i elements from them. · 2) Divide the PSU's into B_i clusters of secondary sampling units (SSU's or IISU's) and select a subsample of b_i *complete clusters* from them. Element sampling is a special case when $B_i=N_i$ and $b_i=n_i$. The n_i elements in each of the b_i selected clusters are covered completely. 3) When the sizes n_i in the subsampled clusters tend to be too large *subsampling with three stages of selection* may be introduced. Similarly the third stage of selection can also be accomplished in one of three alternative ways: listing and sampling elements directly, or creating and selecting third stage clusters; or subsampling those clusters in a fourth stage. And so on.

At each stage of selection we may also choose from several alternative sampling methods: simple random, stratified random, or systematic. We may also choose equal probabilities, or different probabilities for different strata or

units, or PPS selection. We may use different numbers of stages and different procedures in different strata. It is not feasible to describe all the alternatives, and we must rely on knowing the basic principles.

The most important choice concerns the numbers and kinds of PSU's. This importance is clear for the field collection of data. It is also true for the primary selections (ultimate clusters) used in variance computations (13.2). It is also the basis of simple and adequate cost factors and formulas (6.7). And the design effects of clustering and the roh values are also expressed approximately, simply and adequately in terms of ultimate clusters (6.6).

6.6 DESIGN EFFECTS OF CLUSTER SAMPLES. ROH

Clustered selections generally increase variances and the increases can be denoted generally by $Deft^2$=actual Variance/SRS Variance. For sample means this ratio, which shows the effect of clustering, can be measured by:

$$deft^2(\bar{y}) = \frac{var(\bar{y})}{SRS\ var(\bar{y})} = \frac{var(\bar{y})}{s_y^2/n}. \qquad (6.6.1)$$

The numerator $var(\bar{y})$ can be computed for a selections with the computing formulas we give for different designs. The denominator is also computed from the sample of n cases, but as if these had been selected with SRS (5.1): $s_y^2 = (\Sigma\, y_j^2 - y^2/n)/(n-1)$. The n used for s_y^2/n is valid for EPSEM, but for weighted samples it has to be modified (12.5). This s_y^2/n does not show effects of the clustering and therefore it would underestimate the actual variance by the ratio $deft^2(\bar{y})$, as shown in millions of calculations.

Both the numerator and the denominator are functions of S_y^2, the population variance and its unit of measurement. They also depend inversely on the sample size n. The denominator s_y^2/n determined by these completely,

except for sampling variations around its expected value $S_{\bar{y}}^2/n$. Taking the ratio removes these two parameters and leaves all the other design factors which affect only the numerator.

The numerator $var(\bar{y})$ is subject to sampling variation around its expected value $Var(\bar{y})$, and it can be highly variable, because it is based on only a selections, which is usually not large. Most importantly, the numerator is also a function of the kind and size of PSU's, also of the selection design; this variation between PSU's is reduced by the stratification, and that is why stratifying PSU's is important. (6.3)

Variance computations based on primary selections (or ultimate clusters) reflect the variations among the b_α values for the a primary values (more precisely among the a_h values of $b_{h\alpha}$ within each stratum). Whatever complexity (stratification, control of size, auxiliary variable) went into their selection should be reflected in the computed variances. Therefore, $deft^2(\bar{y})$ is a complex measure that summarizes many design features, and separating them would be too complex for most surveys (14.3).

However, it is frequently desirable to separate the effect of the average cluster size $\bar{b}$ in order to make the computations useful also for subclasses (crossclasses) and for other sample sizes. This is done by formulating the adequate approximate relationship $var(\bar{y}) = (s_{\bar{y}}^2/n)\ [1 + roh(\bar{b}-1)]$, and therefore:

$$deft^2(\bar{y}) = [1 + roh\ (\bar{b}-1)]\ . \qquad (6.6.2)$$

From this we compute $roh = (deft^2-1)/(\bar{b}-1)$, a good approximation if not taken close to or below $\bar{b}=1$. This roh, a "ratio of homogeneity," is a summary measure of several design factors, as we said, of which *homogeneity within the subsamples* $\bar{b}$ (really the b_α) is the most important. It is a substitute for the famous RHO (or ρ or δ) in the literature, *the coefficient of intraclass correlation* [Kish 1965, 5.4, 5.6; Cochran 1977, 8.3]. However, RHO is well defined only for unstratified, random selection of fixed size b in one or two stages from equal

sized clusters. We use *roh* as an adequate measure for the many multistage probability sample designs we actually need. Furthermore, using an average value $\bar{b}$, instead of a constant b, is necessary, and this practice has been justified in many computations for ratio means, with reasonable control over variations in b_α (which is necessary for practical and theoretical reasons) [HHMI, p. 608].

The population value Roh is almost always positive and $Deft^2 > 1$; although we do encounter some negative computed values of roh and of $deft^2$ under 1, but usually because of sampling variation due to small numbers a of PSU's. We may assume $deft^2 > 1$ and roh > 0 in our practical work, but the practical question is quantitative: By how much? There is a great deal of variation, because some of the effects of homogeneity are great indeed. Agricultural variables, in particular, can be subject to large factors of homogeneity due to soil, climate, cultural and other factors.

$Deft^2$ should be computed for many variables in order to measure increases in variances due to design effects, also deft for increases in standard errors. 1) They serve as warning devices because values of deft < 1 or deft > 10, though possible, can discover errors in computations. 2) They permit inferences to other statistics within the same survey. 3) They can be used for inferences to other and future surveys. 4) They can be used for improving designs for future surveys (14.2).

"The variance of cluster samples, particularly in social research, is typically greater than for a comparable sample of elements. This is not a logical necessity, but a generalization based on research with groups of many kinds. In most groups roh tends to be positive: the individuals within groups tend to resemble each other. The homogeneity of groups is greater than if individuals were assigned to them at random. The homogeneity may be due to selective factors in grouping, to joint exposure to similar influences, to the effects of mutual interaction, or to some combination of these three sources. Regardless of source, roh measures the homogeneity in terms of the *portion of*

the total element variance that is due to group membership. Sampling units employed as convenient clusters typically possess some group homogeneity" [Kish 1965, 5.4].

For reducing $deft^2 = 1 + roh(\bar{b} - 1)$, the emphasis here has been, as it is in practice, on controlling $\bar{b}$ and on reducing roh through subsampling. Reductions of the subsample size $\bar{b}$ must balance cost and variance factors (6.7). For reductions in *roh* we depend mainly on methods of subsampling for spreading the sample as widely as possible within the clusters. For example, when we use districts for PSU's, instead of selecting one complete segment or a few, for secondary units, we may select many segments within districts, then subsample each. Of course, once again one must balance the reduction of element variances against increases in element cost. These concepts lead to multistage sampling: each stage between PSU's and elements involves compromises between reducing variances and the increased costs for listing sampling units and for collecting data.

Subsampling is the principal means for spreading samples. Its foundation rests on assumptions that values for *roh* (clustering effects homogeneity) for elements are less in large units (like districts) than in small units (like small, complete segments); empirical evidence backs those assumptions. Models based on those assumptions explore this basic notion of decreasing roh with increasing area size [Cochran 1977, 9.5; Jessen 1979, 4.8; Murthy 1967, 8.,3]. These models attempt to present *roh* in terms of one or two parameters. But these tend to differ greatly between variables, therefore they result in conflicts for multipurpose designs. However, even for single variables, the simple uniform decrease of homogeneity for distance fails to stand up to reflection or to evidence [Yates 1981, Ch. 8; Kish 1961]. Nevertheless, we may base designs on the general concept of larger values of *roh* for elements nearer to each other.

Sometimes, though not often, we can *create artificial clusters that have lower roh values* on the average than other clusters of the same size would have. For example: (a) When selecting clusters of numbers from telephone numbers, or from social insurance numbers, use of last digits can yield heterogeneous units, whereas the first digits would yield meaningfully similar units. (b) From listings of dwellings or farms from areas (blocks; E.D.'s, etc.) systematic selection defines less homogeneous "clusters" of elements than would compact segments (10.3). (c) For selecting segments (from E.D.'s, villages, etc.) a serpentine numbering before systematic selection, will tend to produce a better spread and less homogeneity than a contiguous cluster of segments would. Each of these exemplify the desirability of *spreading the subsamples* within sample units.

6.7 COSTS AND EFFICIENCIES IN CLUSTER SAMPLING

We wish to investigate how the average size $\bar{b}$ of sample clusters affects the efficiency of sampling. Most of our models on clustered selections suppose a sample of n elements selected from a clusters (PSU's), with an average of $n/a = \bar{b}$ elements per cluster. A simple and frequently used linear cost function states that:

$$C = cn + C_a a = cn[1 + C_a/c\bar{b}] . \qquad (6.7.1)$$

The basic cost cn, which is proportional to the sample size in elements, includes field collection (interviewing), coding, processing, also some sampling costs. The component $C_a a$ denotes costs proportional to the number a of clusters (PSU), and it can vary widely. It may be low when it refers merely to selecting E.D.'s and perhaps finding their boundaries; more if farms and households must be listed in the a sample clusters. C_a can be much higher if it refers to a districts, in each of which one or more enumerators must be hired, trained, and retained; or where teams of enumerators in vehicles must be sent for a week or two. Thus the cost factor C_a and the ratio C_a/c can vary widely; perhaps mostly in the range 1 to 100.

The ratio C_a/c is more relevant here than the absolute value C_a. The factors c and n concern scale factors of wages and of the scope of the survey. By removing them we can concentrate on the factor $C_a/c\bar{b}$ that denotes the increase due to costs of the clusters relative to the unit costs of interviewing. It is inversely proportional to the average cluster size $\bar{b}$: the added cost factor C_a/c is relatively easily borne for large sample sizes $\bar{b}$.

On the other hand, we must also consider the effects of the average cluster size $\bar{b}$ on variances. For the sample mean this can be expressed as:

$$\mathrm{Var}(\bar{y}) = \frac{S^2}{n}[1 + \mathrm{roh}(\bar{b}-1)] . \qquad (6.7.2)$$

The "design effect" of clustering over SRS increases variances by the relative factor $\mathrm{roh}(\bar{b}-1)$, which increases with the cluster size. We can assume roh to be positive, because it almost always is (6.6); but its value differs greatly between variables.

Thus we have a conflict between the two relative unit values per element: for larger cluster sizes $\bar{b}$, the unit cost decreases but the unit variance increases. To find an efficient compromise we express the product of the two unit costs as:

$$C \times \mathrm{Var}(\bar{y}) = cS^2[1 + C_a/c\bar{b}] \times [1 + \mathrm{roh}(\bar{b}-1)] . \qquad (6.7.3)$$

The sample sizes n cancel, so that this relation does not depend on n, within the limits that the parameters remain adequate. S^2 depends greatly on variables, but not at all on the design or on $\bar{b}$. An optimal compromise value can be found, for either fixed variance or for fixed cost, which is reasonable in practice. The most efficient value for $\bar{b}$ is found to be [Cochran 1977, 10.6; Kish 1965, 8.3b, 8.5]:

$$\text{optimal } \bar{b} = \sqrt{C_a/c}\ \sqrt{(1 - \text{roh})/\text{roh}}\ . \qquad (6.7.3)$$

Thus the optimal subsample size $\bar{b}$ increases with the ratio C_a/c and decreases with roh. But both of the changes are mitigated under the $\sqrt{}$ sign, and even poor guesses about C_a/c are not likely to lead to bad allocations.

On the other hand, roh can vary greatly, say from 0.001 to 0.200 or even higher for different variables, and multipurpose design becomes difficult indeed. For agricultural variables we may expect high values for some roh values, hence also high variability among roh values. Multipurpose design must be taken seriously (Ch. 9).

The two models (6.7.2) and (6.7.1) must be viewed with some caution. The linear cost function probably works well only within modest limits. At the lower end of element sampling, with $\bar{b}=1$ we would have $n(c + C_a)$ and other procedures are needed to obtain lower C_a. At the higher end of large values of $\bar{b}$, one may afford higher values of C_a for better procedures.

We must consider three other situations with drastic effects on these factors. First, when a sample of clusters gets used for several periodic surveys or as a master frame for several distinct surveys, the C_a cost factor should be divided between them. For example, for a listing used 9 times (if it does not deteriorate), the optimal $\bar{b}$ increases by $\sqrt{9} = 3$. Therefore, most good surveys come from continuing operations, which can afford high C_a to be split among surveys (Ch 16).

Second, when sampling for rare elements we may well design for large clusters of $\bar{b}$ in terms of a total population of householders or farms because only a portion $\bar{M}_r$ will come into the sample, so that an optimal $\bar{b}\bar{M}_r$ may be small even for large values of $\bar{b}$. That is, we can use large subsamples of total households because only a portion $\bar{M}_r$ incurs the costs c or is affected by values of roh (8.2).

Third, for subclasses (crossclasses), the design effect concerns the subclass members: $[1 + \text{roh } (\bar{b}\,\bar{M}_c - 1)]$. Therefore for crossclasses the optimal cluster size $\bar{b}$ is increased approximately by $\sqrt{\bar{M}c}$:

$$\text{optimal } \bar{b} = \sqrt{C_a/c}\,\sqrt{(1 - \text{roh}\bar{M}_c)/\text{roh } \bar{M}_C}\,. \qquad (6.7.4)$$

For both rare items and crossclases the sample sizes b_i become more variable, and these should raise doubts about the use of an average $\bar{b} = m/a$. Some variability of b_α must also be tolerated for the entire sample of $n = \Sigma\, b_\alpha$, but these can usually be contained within reasonable limits (Ch 7.4). Our models for both costs and design effects intend to include such variations of b_α in the definitions.

CHAPTER 7. PROCEDURES FOR CLUSTER SAMPLING

7.1 ONE–STAGE SELECTION OF COMPLETE CLUSTERS

This is the simplest form of cluster sampling and the simplest selection procedure when element sampling is not feasible. It requires partitioning the population into a large number of uniformly small, well–defined units with identifiable boundaries.

1. We must either find or create clusters which divide the entire population of N elements into a large number A of clusters, which are small on the average $\bar{N} = N/A$; also not highly variable: the N_a in $N = \sum N_a$ should not vary unduly. Such uniformity seldom occurs "naturally": villages and districts typically vary too much in size and often their average also is too large for selection as complete clusters.

2. The clusters must also be clearly and easily identifiable for data collection; this often implies that they must be simple for field enumerators. This requirement often conflicts with the preceding requirement of uniformly small units. Creating and delineating clear, identifiable boundaries for a large number of uniformly small units that partition the entire population is usually too expensive a task for large populations.

3. Enumeration Districts (ED's), sometimes called Areas (EA's) are created in most countries for census purposes. They *may* be available and adequate in some countries and some situations. They seem to vary from 50 to 300 households in average size, the variation in sizes perhaps tolerable for a few postcensal years, and boundaries perhaps available and adequate. If some of them are too large, perhaps they can be split before selection (see 7.2). Decennial censuses pay for the great cost of creating "equal" workload for enumerators. But seldom can one find organizations, like the military, which succeed in dividing their populations into small, equal branches.

4. Sometimes the sampler can *create* small and (appropriately) equal sized units, which are identifiable; and there also exist some favorable situations that the sampler may utilize. Alphabetical listings, when "complete" and when they can lead to feasible data collection (e.g., mail surveys) can be used. The number of initial letters (26?) are too few and create unequal clusters; but samplers have created "alphabits" of 3 or 4 starting letters, which are numerous enough (10,000?), also small and equal enough in size, in order to serve as identifiable clusters. Also the 365 days of the year have been used as clusters (e.g., for births, insurance policies, traffic surveys); the number of clusters can be increased by using either years for longitudinal samples, or hours for others. Agricultural samples of small populations (e.g., a single district) can be based on dividing the entire population into area segments with clear boundaries, each with a small number (4–10?) of farm dwellings.

5. The need for many and for small clusters are interrelated and both of them should be related to the selection rates. This can be seen in the simple relations $\Sigma N_\alpha = N = A\bar{N}$ and $A = Fa = a/f$. Small average cluster size $\bar{N}$ means many clusters A and this population number of clusters must be $F = 1/f$ times greater than the number a of sample clusters. Also for paired selections from strata, there will be $H = a/2 = Af/2$ strata formed.

6. Complete clusters, if they are clearly identified, can be less difficult for field enumerators to cover completely; and they are also easier to control and check than subsampling. Thus, they can lead to better coverage of clusters [Kish, 1965, 8.4B].

For sampling rare elements (such as female holders or unusual crops) complete clusters may be especially well adapted. The large total size of clusters (for example of E.D.'s) may only yield a small average number of the subpopulation of interest. Complete coverage allows for easier instructions, operations and controls than subsampling would. Even some variation in the size of the clusters is more tolerable when compared to the usual variations to be expected in the subpopulation in any case.

7. On the other hand, the homogeneity of elements (e.g., holdings, holders, farmers and households) is likely to be higher than for subsampled clusters. Subsamples can have lower homogeneity (roh), hence lower variances in two–stage (or multistage) samples of the same size $n = a\bar{b}$, than in complete clusters. But the complete clusters can be cheaper for the same total n. Comparisons for the same cost will differ between variables and situations.

8. The selection equation for paired selections of complete clusters is simple but instructive:

$$\frac{2}{2F} \times 1 = \frac{1}{F} = f \ . \qquad (7.1.1)$$

That is, first determine f = n/N, the desired EPSEM sampling rate from the desired sample size n and the estimated population size N. Then with F = 1/f for selecting clusters (and elements) *form strata of size 2F clusters for paired selections. From each stratum of 2F clusters select 2 clusters with EPSEM.* This selection gives each cluster the probabiiity of 2/2F = f and the same for each element, because the probability of selecting elements is 1 from the complete selected clusters. The selection equation specifies only EPSEM selections within strata: These may be true random selections, or one random from each of two half strata, systematic selection (6.4). The variance computations assume two random selections (13.2).

The paired selection design can be modified for other stratified design of complete clusters:

a) $\frac{1}{F} \times 1 = f$ for single selections from strata

b) $\frac{k}{kF} \times 1 = f$ for k selections from strata

c) $\frac{a_h}{A_h} \times 1 = f$ with $a_h = fA_h$ for varying sizes of strata and selections

d) $\frac{2}{A_h} \times 1 = f_h$ for varying selection rates between strata, but constant $a_h = 2$.

7.2 SIMPLE INTEGRAL SUBSAMPLING

Often the average size $\overline{N}$ of available clusters, with clearly identifiable boundaries is too large. If the average $\overline{N}$ is much greater than the desired optimal $\overline{b}$ (6.5) then subsampling of the clusters is indicated. For example, if the E.D.'s average 200 households and optimal $\overline{b} = 5$ or 10 is indicated, then we may resort to subsampling. However, the discrepancy between the $\overline{N}$ and the optimal $\overline{b}$ should be by a factor of 4 or more ($F_b > 4$) before abandoning the simplicity of complete clusters; smaller departures from optimal $\overline{b}$ incur only moderate or trivial losses; and realistic multipurpose design should also indicate more flexibility in choosing $\overline{b}$ (9.3). However, sometimes subsampling intervals $F_b < 4$ have also been used; also too large F_b may increase unduly the segmenting work. Of the many possible forms of subsampling (6.5), the simplest is two-stage sampling, and each stage selected with integrals. For paired selection this appears as:

$$\frac{2}{2F_a} \times \frac{1}{F_b} = \frac{1}{F_a F_b} = \frac{1}{F} = f \ . \qquad (7.2.1)$$

For example, we first find that $n/N = f = 1/f = 1/900$ approximately, also that $\overline{N} = 90$ farms per E.D. roughly; and that the optimal $\overline{b} = 10$ roughly. Then $F_b = 90/10 = 9$ and $F_a = F/F_b = 900/9 = 100$. These numbers can be adjusted flexibly, to get two convenient numbers so that $F_a \times F_b = F$. It is preferable to have F_b an integer for easier field work, but a fractional F_a can be applied in the office (5.5). Usually F_a should be large in order to reduce the sample clusters from a large number A to a much smaller number a for easier field work; but a should be large enough to reduce variances and to permit stable computed estimates (14.3). On the other hand, F_b should be small enough so that not too much preparatory work on the $\overline{B}$ units (e.g. listing farms or dwellings) is needed to obtain the desired sample of $\overline{b}$. These numbers are

chosen after the overall sampling rate is determined tentatively as $f = 1/F = n/N$; then from $n = a\bar{b}$, a desirable pair of a and $\bar{b}$ are chosen; then from $\bar{N} = N/A$ the pair of rates in $F_a \times F_b$ are chosen, and these may result in a readjusted F.

Note several important aspects of (7.2.1), which may also be written in sampling rates $f_a f_b$, which are also selection probabilities:

$$f_a \times f_b = f = 1/F \quad . \qquad (7.2.2)$$

1. The f_a and f_b refer not only to sampling fractions of population units included in the sample, but also to true *probabilities that must be made operational*; e.g., with tables of random numbers.

2. The equal probability f for all N population elements results from two probability operations: second stage selection with f_b within selected clusters, *conditional* on the cluster's selection with f_a in the first stage.

3. The probability $f_a = 1/F_a$ in the first stage can be applied in a variety of ways described in 7.1, and with any k = 1, 2, or more in k/kF_a.

4. Subsampling with f_b can be applied in many ways and stages, as described in 6.5.

5. When the variation of cluster sizes is too large this simple technique becomes unsatisfactory because the subsample sizes $f_b \times N_\alpha$ vary too much and are too unstable for both practical and theoretical purposes (13.3). Just what variations in N_α should be judged "too large' to be tolerated depends on several factors, but the tolerable upper limits for the largest N_α are bound to be somewhere less than 6 times the average $\bar{N}$; however, with a large number A of clusters and for very rare and hard to find large values of N_α we may tolerate even higher extremes. If variations in size N_α are too large, we must resort to one of several techniques for controlling subsample sizes with PPS (7.3 – 7.6). As we

proceed the techniques become more general, so that each earlier method can be viewed as a specialized case of 7.4. But it is of practical use to get to know the details of these simpler techniques.

7.3 SYSTEMATIC SELECTION Of INTEGRAL PORTIONS

This is the simplest technique for introducing measures of size, and the crudest form of PPS sampling. Suppose you find or create a consecutively ordered list of clusters with measures of size. These may be identified cluster numbers on sheets of paper, with measures of size indicated for each. Or they may be only areas (blocks or segments) on maps or aerial photographs, with dwellings or farms or parcels identifiable on them. Next a decision about a basic average size $\bar{N}$ must be made in light of the procedures to follow. Then the clusters are numbered with consecutive integers, from 1 to ΣI_α. Each cluster can be assigned $I_\alpha = 1,2,...F_a$ integers; thus a block 2 (or 3 or I_α) times larger than $\bar{N}$ is assigned 2 (or 3 or I_α) integers as measures of size. The interval of selection F_a is applied to the cumulated (consecutive) integers successively, after a random start from 1 to F_a. For example, this serves to create I_α segments in blocks of unequal size, and the numbers of segments will vary approximately with the block sizes. The selection probabilities in the two stages are:

$$\frac{I_\alpha}{F_a} \times \frac{1}{I_\alpha F_b} = \frac{1}{F_a F_b} = \frac{1}{F} = f \ . \qquad (7.3.1)$$

In the second stage, the probability $1/I_\alpha F_b$ is applied to the elements of the α–th cluster, conditional on this cluster having been selected in the first stage with probability I_α/F_a. Thus, the arbitrary integral measure of size I_α cancels, because it assigns direct probabilities in the first stage and indirect probabilities in the second. Thus, no bias results from its consistent and contrary uses in the two stages. To the degree that $I_\alpha\bar{N}$ comes close to the actual size N_α, the variations in subsample sizes will be diminished, and to that degree also the expected sample size $N_\alpha/I_\alpha F_b$ will be close to the designed subsample size $\bar{b}$.

Furthermore, in the proportion that clusters with $I_\alpha = 1$ can be assigned good boundaries the design approximates 7.2. And when $F_b = 1$ hence $\bar{N} = \bar{b}$ seem justifiable, the design can resemble the complete clusters of 7.1.

The preparation of the frame is complex and it consists of four tasks: 1) identifying the cluster boundaries which involves joining small parts and splitting large ones (both in numbers of elements); 2) assigning measures of size in consecutive integers; 3) ordering the clusters into strata; 4) numbering the units to establish the listing for selection. These tasks must be done simultaneously. They can be better described in detail in connection with area sampling (10.2).

The measures I_α can be assigned arbitrarily and usefully, as will be noted in 7.4. It may be convenient to make the total $\Sigma I_\alpha = aF_a$, so that a is also an integer, thus the interval F_a will yield exactly a selections. Note that each implied stratum F_a yields one selection, and $2F_a$ yields two selections for the variance computations. Hence, $I_\alpha = F_a$ can serve as a maximal size and a "self–representing" PSU, where the rate $1/F_aF_b = f$ is applied within the cluster.

7.4 SELECTION WITH PPS: PROBABILITIES PROPORTIONAL TO MEASURES OF SIZE

Some statisticians distinguish probabilities proportional to size from probabilities proportional to *measures* of size, and use PPMS as symbol for the latter. But this distinction is not necessary in practice, where exact sizes are never available, only more or less imperfect measures. Sampling theory should accept exact sizes as a limiting case without errors when the measures equal the "true" sizes. Section 7.3 concerned PPS with crude integral measures and 7.5 also concerns PPS, but with explicit strata, whereas here we describe "implicit strata" of fixed sizes, also called "zones."

The selection formula can be written as:

$$\frac{Mos_\alpha}{Fb^*} \times \frac{b^*}{Mos_\alpha} = \frac{1}{F} = f \text{ , or for paired selections:} \qquad (7.4.1)$$

$$\frac{2Mos_\alpha}{2Fb^*} \times \frac{b^*}{Mos_\alpha} = \frac{1}{F} = f \text{ .} \qquad (7.4.2)$$

1. The *measures of size* Mos_α for the primary clusters may be available as numbers of farms, households or persons in the clusters in the last census. Or these numbers may have been adjusted for changes or because another population is sought. For example, the only available data may be 1980 Census persons whereas we seek 1988 actual holders. In 7.3 the Mos_α were simple integers $I\alpha$, and in 7.2 they were all 1. The numbers Mos_α can be arbitrary, they are "never" perfect "true" sizes of the population elements, but we must try to make them roughly proportional to the true size, in order to reduce at least the extreme variations in the actually obtained subsamples b_α. For the desired a of PSU's we compute $Fb^* = \Sigma\, Mos_\alpha/a$.

2. In the second stage the *same* Mos_α *must be used inversely*: the probability b^*/Mos_α must be applied to all $N\alpha$ elements in the selected α—th cluster for an expected subsample of size $N_\alpha b^*/Mos_\alpha$ and an actual subsample b_α. Many different ways of subsampling with the rates b^*/Mos_α exist in one, two or more stages, and we shall describe some later (7.5). Note that we had to use three symbols for the subsamples: b^* for the fixed, designed, intended subsample; $N_\alpha b^*/Mos_\alpha$ for the expected size directly proportional to the size N_α of the cluster and inversely to the measure Mos_α; and the actual subsample b_α a random variable around the expected size, resulting from the subsampling process.

3. Through the two selection procedures the constant probability f is assured to all the $N = \Sigma\, N\alpha$ elements in the population, because the same measure Mos_α is applied directly in the first stage and inversely in the second; otherwise the equal probabilities would not be maintained (7.7).

4. Selecting a single cluster with Mos_α from a stratum of size Fb* is not difficult: select a random number from 1 to Fb* and apply it to a cumulated list of the Mos_α, where $Fb^* = \Sigma\, Mos_\alpha$ for the stratum. Problems arise because of a conflict : in order to compute variances we need 2 (or more) selections per stratum. However, two selections per stratum, with varying Mos_α and without replacement is difficult. But one of the following six may be used, preferably a,b,c, or d.

a) Systematic selection with random start 1 to Fb* is easiest; but it faces technical objections (5.5).

b) Systematic selection with new random starts 1 to Fb* for every odd (1,3,5...) interval, in order to avoid the theoretical problems of a single random start; other similar schemes can be used.

c) Divide each stratum of 2Fb* into two half—strata of size Fb* and take a random start 1 to Fb* for each, a total of a selections for the a half—strata.

d) Select two random numbers from 1 to 2Fb*. If the same cluster is selected twice, take two samples from it, but without duplication of elements. For this method, clusters of $Mos_\alpha \geq 2b^*$ must exist or be created, so that two subsamples can be selected with b^*/Mos_α. This has been called "simple replicated subsampling" [Kish 1965, 8.6A], and "graduate variable probabilities" [Sanchez-Crespo 1977].

e) Select two random numbers from 1 to 2Fb*; if the same cluster is selected the second time, select again until a different cluster shows up. Accept the selection bias against larger clusters as "negligible"(?).

f) There are many methods in a large literature for selecting two (or more?) units (from 1 to 2Fb*) without replacement with varying probabilities. They are all difficult. [Cochran 1977, 9A.1–9A.12; Murthy 1967, 6.10–6.11; Brewer and Hauif 1983; Kish 1965, 7.4]

5. The next section, 7.5, deals with explicit strata of different sizes $M_h = \Sigma\, Mos_{h\alpha}$ created by cumulating *complete* PSU's. In this section we accepted a fixed constant Fb* to be the size of the half strata for methods a, b, c, and 2Fb* in d, e, f. If we simply cumulate the values of Mos_α this leads to a conflict because most of the stratum boundaries will cut across PSU values. This problem of *implicit strata*, or "zones" [Deming 1960] may not be of great

practical importance, and can be dealt with one of several alternative procedures. a) With systematic selection (as in 4a above) two selections do not occur, unless $Mos_\alpha > Fb^*$. b) With some labor one or a few values of Mos_α may be arbitrarily adjusted in each stratum to create explicit strata of complete measures $\Sigma_\alpha Mos_{h\alpha} = Fb^*$ (or $2Fb^*$). No bias result if the same adjusted Mos_α is used in both stages. c) Accept the implicit strata (zones) and if a value Mos_α is selected from two strata, then separate subsamples are selected with b^*/Mos_α for both strata. Here again we need $Mos_\alpha \geq 2b^*$ to permit two separate selections [Kish 1965, 7.5 for procedural details].

6. Adjustments of Mos_α, if they cancel in the two stages, do not destroy the desired EPSEM f. To several uses of flexibility, already noted, we should add deliberately low assigned values of Mos_α in order to decrease the selection probability Mos_α of some units and increase the size of the subsamples of those units when they are selected. For example, clusters that are distant, or expensive to reach, or costly to list, may all receive low values of Mos_α, with the expectation of larger subsamples Mos_α when selected. On the contrary we may deliberately increase Mos_α for units we prefer to include even if with correspondingly smaller subsamples b_α.

7. The above flexibility may also be had by altering the designed subsample size b^*, but that may be better done in separate strata, with different b^* in different strata. This in turn leads to possibly using different f_h in different strata, with inverse weighting in the analysis.

7.5 PPS IN EXPLICIT STRATA; SUBSAMPLING

In some situations more formal selections of the PSU's seems preferable and these from clearly defined strata in the first stage. Thus, denote the selection probability of a PSU by $P_h = Mos_\alpha/M_h$, with $M_h = \Sigma_\alpha Mos_{h\alpha}$, the sum of the measures of the PSU's in the first stage. In that case we have for the two stages

$$P_h \times \frac{f}{P_h} = \frac{Mos_\alpha}{M_h} \times \frac{M_h f}{Mos_\alpha} = f \ . \qquad (7.5.1)$$

The second stage probabilities must be inversely proportional to the first stage probabilities, and the factors in the second stage are worth perusing for their content. Note that $M_h f$ represents the designed size of the subsample from the hth stratum. This technique can be utilized when the inconveniences of implicit strata seem troublesome. It is used commonly when a selection of PSU's (counties, districts) is used over longer periods (e.g., between decennial censuses) for selecting many samples from a master sample. These have been described several times [USCB 1963, 1980; Kish 1965, Ch. 10; Hess 1985].

We may note several modifications of the basic (7.5.1). To compute variances it is necessary to collapse similar strata, perhaps in pairs (13.2). Or it may instead be preferable to double the stratum sizes and then select one PSU from each half of stratum:

$$\frac{2Mos_\alpha}{Mos_h} \times \frac{M_h f}{2Mos_\alpha} = f \ . \qquad (7.5.2)$$

Selections within PSU's can take many forms (7.6), and frequently the same PSU's can be used for many samples with different selection methods, sometimes in two or more stages. For example, one may select a small fixed number (e.g. 2) in the second stage and then use the last stage of selection to balance the equation (e.g. applied systematically to last stage listings of segments or elements):

$$\frac{Mos_\alpha}{M_h} \times \frac{2}{B_{h\alpha}} \times \frac{M_h B_{h\alpha} f}{2\ Mos_\alpha} = f \ . \qquad (7.5.3)$$

Yet another stage may be introduced; for example, b secondary units can be selected in the second stage, and from each of these c third stage units, and the entire operation balanced in the fourth stage with variable sizes:

$$\frac{Mos_\alpha}{M_h} \times \frac{b}{B_{h\alpha}} \times \frac{c}{C_{h\alpha\beta}} \times \frac{M_h B_{h\alpha} C_{h\alpha\beta} f}{bcMos_\alpha} = f \ . \qquad (7.5.4)$$

Here we assume that the numbers of sampling units in the second and third stages are variable $B_{h\alpha}$ and $C_{h\alpha\beta}$. They typically are variable in both numbers and in sizes (measured in numbers of elements. They may possess accepted (administrative) definitions and boundaries that the sampler merely recognizes and accepts. However, the sampler can and should use flexibility to "split and combine" them in order to form units that will be more equal in size (numbers of elements). The closer the sampling units are in size the less size variation is introduced by the ratios $b/B_{h\alpha}$ and $c/C_{h\alpha\beta}$. However, even more control of size can be introduced with PPS if measures of size can be obtained; then (7.5.3) may be reformulated:

$$\frac{Mos_\alpha}{M_h} \times \frac{bMos_{\alpha\beta}}{M_{h\alpha}} \times \frac{M_h M_{h\alpha} f}{bMos_{\alpha\beta} Mos_\alpha} = f \ . \qquad (7.5.5)$$

Here $M_{h\alpha} = \Sigma_\beta Mos_{\alpha\beta}$, the sum of measures $Mos_{\alpha\beta}$ for secondary (β) sampling units within the selected primary α sampling unit from the stratum (h); b sampling units are selected. In (7.5.3) b = 2, each selected with equal $1/B_{h\alpha}$. However (7.5.5) may serve more often to control variations in size, because assigning the measures $Mos_{\alpha\beta}$ may be feasible more often than creating secondary units with equal sizes. In both cases the last selection rate applied to numerous small units (elements or segments) is used to maintain the overall sampling rate and probability f; but this must permit variations in the number of selected last stage units. See also subsampling in 6.5, and in chapter 10 on area sampling, and the three earlier references.

7.6 SIMPLE TECHNIQUES FOR CONTROLLING SIZE

Control of b_a, the size of subsamples is valuable and useful in most surveys. Selections with PPS are powerful and flexible techniques for that purpose, but in the next two sections we describe some other techniques also useful sometimes.

"*Split and combine*" is a technique for forming artificial "pseudo" clusters that are less unequal than the natural clusters from which they are formed. This may be adequate when a large majority of natural clusters can be accepted unchanged, or can be combined into convenient (contiguous, neighboring) units of acceptable sizes; small clusters can be combined into one or they can be attached to acceptable clusters. A small proportion of large units, perhaps in separate strata, can be split into smaller units. For example, a list of "pseudo–E.D.'s" may be created from an obsolete census list.

Stratification by size can be a convenient method for reducing variation in size. With 3, 4 or 5 strata for size, the variation in size can be greatly reduced within those strata. On the other hand, we may need strata for other variables, when the numbers of strata we can use is narrowly limited; in those situations we should use PPS for control of cluster sizes, and thus leave stratification for other variables. However, sometimes sizes of units can also serve as domains of interest in the analysis, hence also as strata.

Size–stratified subsampling appears as a reasonable design: (1) clusters are stratified by size; (2) the sizes of the strata are made roughly equal in the population and in the sample; (3) the subsampling of clusters is designed for an EPSEM of elements. Then we can select 2 clusters (or other a_h) from each stratum; some strata can have few, large clusters while others have many, small clusters; yet from *roughly* equal sized clusters we can obtain roughly equal–sized subsamples from all strata. The selection rates can be written as:

$$\frac{1}{F_{ha}} \times \frac{1}{F_{hb}} = \frac{1}{F} = f, = f_{ha} \times f_{hb}, \text{ with } F_{ha} = \frac{F}{F_{hb}} \cdot \qquad (7.6.1)$$

Usually $f = n/N$ would be computed first as an overall sampling rate. Then probably the different F_{bh} can be computed to yield roughly the desired subsample sizes b_α from different cluster sizes $B_{h\alpha}$, since $B_\alpha/F_{bh} = b_\alpha$. Suppose that the cluster sizes run from $B_\alpha = 1000$ down to $B_\alpha = 10$ and even below, and $b_\alpha = 10$ seems about right. Then from $F_{bh} = 60$ to $F_{bh} = 1$ can give us most of the control we need, and in only 4 size strata:

$\frac{1}{F_{ha}} \times \frac{1}{F_{hb}}$	$\frac{1}{60} \times 1$	$\frac{1}{15} \times \frac{1}{4}$	$\frac{1}{5} \times \frac{1}{12}$	$1 \times \frac{1}{60}$
Range of B_α	5 to 20	15 to 70	60 to 300	250 plus
range of b_α	5 to 20	4 to 18	5 to 25	4 plus

Note that:

(1) We cut a 200 fold variation in B_α from 5 to 1000 down to a 4 or 5 fold variation within 4 strata. The standard deviation in size may be much less.

(2) This even allows for overlaps in stratum boundaries for cases when guessing exact values for B_α is too difficult or expensive.

(3) Any units with $B_\alpha < 5$ may be combined into "pseudo" clusters.

(4) The selection may be made with one cluster selected from the stratum of size F_{ha}; the size boundaries of strata may be shifted since we allowed for flexibility already. Otherwise we must tolerate variations in numbers selected.

(5) All the above was written with single selections per stratum. But double selections from strata of size $2F_{ha}$ would be advisable to prepare for variance computations.

(6) If some of the strata have $2F_{ha}$ plus clusters split the strata into two or more.

(7) The number 60 used above had convenient factors. In most cases $F = 1/f$ can be adjusted somewhat in order to yield convenient factors.

7.7 EXACT SUBSAMPLE SIZES

Throughout this chapter we faced the common problem that measures of size for clusters Mos_α differ from the actual sizes N_α found for subsampling elements; differences can be due to changes (obsolescence), imperfect measures, and especially different populations (e.g., Mos_α of persons and N_α of holdings). The preceding sections describe techniques for controlling and reducing extreme variations, but allowing some variations in the achieved subsample sizes b_α in order to maintain controlled probabilities f, usually EPSEM, for all population elements N. This practice is preferred and common in survey sampling. In contrast, in this section we describe procedures for exact subsample sizes b_f, but allowing selection probabilities f_α to vary between clusters.

The simplest situation would be for equal selection probabilities in the first stage (because measures of size for clusters were lacking, mistrusted, or ignored), followed by *fixed* subsample sizes b_f from the selected clusters:

$$\frac{a}{A} \times \frac{b_f}{N_\alpha} = \frac{ab_f}{AN_\alpha} = f_\alpha \text{ variable} \quad . \tag{7.7.1}$$

This procedure is not uncommon, and some naively believe that because each stage was EPSEM, so is the combined two—stage procedure. However, selection probabilities are inversely proportional to the cluster sizes N_α. Elements (holdings) from large clusters are underrepresented compared to elements from small clusters. Exact subsample sizes b_f have several disadvantages.

a) If the unequal probabilities are ignored in selfweighting estimates, the selection bias can produce biased results to the degree that it is correlated with survey results.

b) To eliminate that bias needs weighting proportional to N_α, but such weighting can often increase variances, sometimes considerably (12.5). Furthermore, obtaining good measures of N_α may be costly.

c) Selecting fixed numbers b_f of elements may be expensive, compared to selecting with fixed rates or intervals (7.1–7.6). The N_α elements must be obtained and usually listed. SRS selections are usually difficult, whereas systematic sampling with intervals N_α/f_f (or integers based on them) are easier. Usually the listing requires separate trips to the clusters, whereas fixed intervals may be applied on the first visit.

d) The symmetries of equal, fixed clusters of b_f are often destroyed in any case, by nonresponses unless substitutions are adopted. For subclasses the inequalities are even greater and practically irremediable.

Fixed sample sizes are assumed in standard statistical literature, and the prejudice seems difficult to overcome. But in the practice of survey sampling usually variable sample sizes and fixed probabilities are preferred, and fixed sample sizes are usually mistaken. [Kish, 1977]. The problems noted below are similar in principle, though perhaps different in quantity and in the types of biases incurred, in several other situations.

Suppose that a multistage situation is treated as a "hierarchical" situation would be treated in experimental design; equal sample sizes are selected at each stage and the probabilities of selection become

$$\frac{a}{A} \times \frac{b_f}{B_\alpha} \times \frac{c_f}{C_{\alpha\beta}} \times \frac{d_f}{D_{\alpha\beta\gamma}} = \frac{ab_fc_fd_f}{aB_\alpha C_{\alpha\beta} D_{\alpha\beta\gamma}} . \tag{7.7.2}$$

The overall probabilities would be difficult to ascertain. If ignored, probability sampling is abandoned for "model sampling" (3.2).

Suppose measures of size Mos_α are used in the first stage of selection, but in the second stage fixed sized subsamples b_f are selected:

$$\frac{Mos_\alpha}{M_h} \times \frac{b_f}{N_\alpha} = \frac{b_f}{M_h} \cdot \frac{Mos_\alpha}{N_\alpha} . \tag{7.7.3}$$

Perhaps b_f/M_h can be fixed and the chief variable becomes the ratio of Mos_α/N_α. To the degree that measures are proportional (or equal) to actual sizes, the ratios Mos_α/N_α are constant (or 1); otherwise selection biases can result similar to those of (7.7.1). If measures are "good," the disadvantages a and b become less important, but procedural problems for fixed b_f in c and d may remain, and the advantages of fixed b_f remain doubtful.

CHAPTER 8. DOMAINS AND SUBCLASSES

8.1. TYPES AND CLASSES

Domains can denote divisions, usually partitions, of the population, and *subclasses* the corresponding divisions of the sample. The sampling literature has been careless and confused in terminology, and it is useful to become clearer in the future. "Subpopulations" and "subnational domains" have also been used for population divisions and "subsamples" for subclasses. Strata and clusters also denote partitions of the sample, but they are used for improved sample design and are often numerous. The purpose of domains and subclasses is to serve the substantive analyses of data.

The practical treatment of sample designs and analysis is also confused by common presentations for subclasses of different kinds, which can have very diverse practical effects. We develop *three types* of domains and subclasses and *four size classes*. Realistically the arbitrary boundaries of these types and classes are not sharp and categorical, but gradual and also conditional on situations. But naming these $3x4 = 12$ kinds should serve to emphasize that not all subclass problems are similar.

This topic deserves the emphasis that this entire and early chapter presents, because most survey analysis concerns not only the overall estimates, like $\bar{y}$, but also similar subclasses $\bar{y}_c$ and $\bar{y}_d$ and their differences $(\bar{y}_c - \bar{y}_d)$. It is difficult to interrupt technical presentations at every point, which are mostly given in terms of $\bar{y}$, but the subclass statistics should be always in our minds. They will be noted in multipurpose designs (9.3). First, let us distinguish three *types of* crossclasses and domains.

a) *Designed subclasses* are those for which separate samples have been planned, designed, and selected, usually in separate strata. They are combined to form the entire sample, usually as (weighted) sums of independent samples. For example: major regions or provinces, or urban and rural portions, where each of these is composed of entire strata of PSU's.

b) *Crossclasses* are at the other extreme, because they cut across sample designs, strata, and sampling units. These are the most commonly used in subclass analysis; e.g., age, sex, education and income classes; kinds or sizes of agricultural holdings; behavioral and attitudinal subclasses, etc. Usually they cannot be separated by design before selection, because individual information is not available for their separation before the survey. Their spread in the population is never entirely even or random, such as for social and economic factors, but they can be relatively so.

Types of Classes	Sizes of Classes		
	Major	Minor	Mini
Design classes	Major regions, provinces	50 states of United States	3000 counties of United States
Mixed classes	Partial segregation: natural resources, cultural, ethnic, or mixed types: regions × age		
Crossclasses	Five-year age groups Major occupations	Single years of age Occupation × education	Years of age × education Age × education × income

Figure 8.1.1 Classification of Domains and Subclasses (with Examples)

c) *Mixed classes* are between the two extreme types and less commonly used; they have not been separated by design before selection like design domains, but not evenly spread like crossclasses, and they tend to concentrate unevenly and irregularly in strata and in sampling units. They are common in agriculture where soil and climate can create relative concentrations like rice in some regions, wheat in others, and grazing cattle in still others. Farming specialties and fisheries, as well as lumber and mining, are of this type.

For separating *sizes of subclasses* into four classes we must mention boundaries that are arbitrary and subject to changes. However, it would be more misleading to lump together all different sizes of subclasses, when there are such wide discrepancies between them, both as regards methods for sampling them and in survey results.

1. *Major subclasses* comprise between 1/4 and 1/20 of the sample and we shall name 1/10 for a usual boundary. For example, 4 to 6 major regions or provinces for major design classes, also 5 to 10 year age groups for major crossclasses. Most samples are large enough to yield reasonably good estimates for major subclasses also, though with standard errors increased by factors of 2 or 3. For crossclasses the number of elements is the key factor, but for design classes numbers of PSU's can also matter (14.3).

2. *Minor classes* contain perhaps from 1/10 to 1/100 of the sample. These may refer to 10—50 provinces of different sizes for design classes, and single years of age for crossclasses, or five—year age groups with four education classes in a two—fold classification. Separate statistics for minor classes are often sought for minor classes, and sometimes satisfied by large samples or by cumulations (8.6). They generally are not adequately represented by the sample design, especially by numbers of PSU's for crossclasses.

3. *Mini domains* may contain from 1/100 to 1/1000 or less of the population and may not even be separately and individually covered by the sample though they are represented by it. Requests for separate statistics (e.g., for the 3000 counties of the USA) arrive these years from administrative sources, who need statistics more current than the last census. Samples are too small to provide these alone and other methods are needed (8.5).

4. *Rare types* that occur in less than 1/1000 of the population and of the sample present problems for which the entire sample is useless. Separate techniques and lists are needed because survey sampling does not provide proper tools (8.6).

Domains and strata have different meanings and uses, but they are often confused, and both represent partitions of the population. Domains are used in the analysis, whereas strata serve the sample design. It is helpful to use strata to create design domains when one can, but strata can be made much more numerous than design domains, and should be so made so as to provide stable bases for them. Furthermore, most subclasses are crossclasses that represent crossdomains which cut acres strata and across sampling units because we cannot control them in the selection.

8.2 COMMON EFFECTS ON SUBCLASS STATISTICS

Subclass means $\bar{y}_c = y_c/n_c$ are the statistics most frequently used and also most developed, but much of the discussion is also relevant for other statistics based on subclasses.

Differences (comparisons) of pairs of subclass means, such as $\bar{y}_c - \bar{y}_d = y_c/n_c - y_d/n_d$, occur frequently also in the analyses of sample surveys. The *denominators* n_c and n_d of subclasses in practice tend to have several common characteristics.

a) They generally represent count variables of the sample sizes of subclasses: 0,1 variables denoting (non)membership in a specified subclass (and domain). They may be unweighted counts, but if weighted the same weights should enter the numerators as the bases. But sometimes the bases may also be other (continuous) variables x_c and x_d (e.g., hectares of a crop, when y_c/x_c is yield per hectare); and most statistics for ratio means (means, variances) are adequate for them also (13.1).

b) Subclass differences usually compare partitions (non—overlapping) of the sample (and of the population), based on *categories of the same variable*. This feature is assumed for simplifying some of the variance formulas and computations; for other comparisons (for overlaps, inclusions, different bases for subclasses) special care and notation are needed for variance computations. For example, farmers under and over 40 years old are non—overlapping

partitions; but rice farmers and wheat farmers can overlap, because some farmers may raise both. They may be exhaustive partitions (e.g., under versus over 40 years old) or only two of k partitions (e.g., under 25, 25–30, 30–35, etc.).

c) The denominators are usually random variables, because the total sample size was also, but especially because subclass sizes (n_c and n_d) can seldom be controlled. This feature has consequences that are handled in variance computations.

d) The *numerators* y_c and y_d represent the same variable from different bases for the compared pair. This obvious feature simplifies some computing aspects.

e) The above describe what are commonly understood by subclass means and differences. They should not be confused with other statistics: proportions such as y_c/y and y_d/y are better called "shares" of the entire sample, and the difference $(y_c/y - y_d/y)$ is a comparison of shares.

We describe now some general effects on subclass statistics, which arise when statistics are based on the values y_c and n_c of subclass members, and nonmembers are assigned zero values (blanks).

A. *Selection probabilities* p_j and weights $w_j \propto 1/p_j$ for individual sample elements *are preserved* for (inherited by) subclass members; they are not affected by the zero values assigned to nonmembers. *Sample means, totals and other descriptive statistics retain their forms for subclasses.* The unbiasedness of the simple expansion total $\hat{Y}_c = \Sigma\, y_j/p_j$ is retained. Ratio means $\bar{y}_c = y_c/n_c$ retain their sturdy consistency until the variability of n_c becomes too high for small subclasses, n_c (13.3).

B. *Sample sizes become smaller; also highly variable for crossclasses.* We shall later (8.3–8.4) compare subclass means with specifically designed entire sample means of similar sizes. Here we emphasize the drastic reductions of sample sizes and consequent increases of variances even for major subclasses.

For SRS the variances for subclasses of relative size $\overline{M}_c = M_c/N$ increase by that factor: var $(\bar{y}_c)$ = var$(\bar{y}_c)/\overline{M}_c$. For complex designs the effects are modified around those basic, major increases. We must consider those increases in multipurpose design (9.3).

C. *Design effects* of stratification and of clustering behave in opposite ways to each other: in PRES the $Deft^2 < 1$; but in clustering $Deft^2 > 1$, often considerably. However, *both of those effects tend to be reduced considerably*, proportionately as crossclass sizes $\overline{M}_c$ decrease. We may generalize specific findings (8.3, 8.4): both stratified and clustered designs represent controls (over the joint selection probabilities) and in crossclasses those controls are reduced, and tend to be lost (Fig. 8.2.1).

D. *Designs for domains induce conflicts with each other and with the entire sample.* Conflicts can arise in allocating sample sizes and rates to provide adequate sample sizes for subclass statistics. Optimal allocations to strata contain sources for further conflicts, greatly magnified if screening procedures are also needed (9.3).

E. *Variance estimators become unstable* as sample sizes decrease. This becomes particularly troublesome for design subclasses in cluster sampling, when small numbers a_c of PSU's can seriously challenge the utility of variance computations (8.4).

8.3 EFFECTS OF STRATIFICATION (PRES) ON SUBCLASSES

The principal effect of dealing with subclasses is to reduce sample sizes from n to $n_c = \overline{M}_c n$, with corresponding increases in the basic SRS variances by the factor $\overline{M}_c = M_c/N$, the proportion of domain c in the population. Here we investigate further effects, sc that we may compare subclasses of size n_c with specifically designed samples of similar sizes. These comparisons are useful for designing samples, and also for interpreting sampling error functions, especially $deft^2$ computed for the entire sample.

For *design subclasses* the variance estimators have *on the average the same deft*^{2} as the entire sample. This conjecture can be used: a) to save separate computations; also b) because separate computations may be too unstable, because the number of PSU's a_c are too few for adequate separate bases (14.3). However, we must be cautious with that average, because design subclasses may differ greatly. For example, the deft^2 between provinces may differ greatly concerning rice or wheat yields or production; on the other hand deft^2 may be similar for poultry and pig production, and permit averaging and "pooling" or "borrowing" deft^2 (14.3). Furthermore, if the sampling fractions are varied between design domains, then they must be separately considered.

For *crossclasses* the situation is quite different, because *design effects deft*^{2} *are reduced in proportion to decreases* in $\bar{M}_c = M_c/N$. "Exact" mathematical statements need more definitions and development than the subject warrants here (it involves factors like $(n_h-1)/n_h$); and they may be found elsewhere [Kish and Frankel, 1974; Kish, 1965, 4.5; Kish 1980; Cochran, 1977, 5A.14]. The principal results are interesting and useful. The mean for a crossclass can be computed simply as

$$\bar{y}_c = y_c/n_c = \Sigma_h y_{hc}/\Sigma_h n_{hc},$$

the simple self–weighting mean for EPSEM selections, assuming that PRES was used. However, if sampling rates differ between strata, then weighted statistics should be used. The variance of $\bar{y}_c$ may be viewed approximately as:

$$\mathrm{Var}(\bar{y}_w) = [\, S_w^2 + (1-\bar{M}_c)S_b^2] \,/\, n_c = [S^2 - \bar{M}_c S_b^2] \,/\, n_c\,, \qquad (8.3.1)$$

where $S^2 = S_w^2 + S_b^2$. The total element variance is composed of S_w^2, the within stratum variance, and $S_b^2 = \Sigma\, W_h(\bar{Y}_h - \bar{Y})^2$, the between stratum variance. For entire samples $\bar{M}_c = 1$, the second term in (8.3.1) vanishes and the variance is S_w^2/n_c. For small crossclasses, $\bar{M}_c$ is small, the second term of (8.3.2) tends to vanish, and the variance approaches S^2/n_c. Thus *the gains of*

PRES tend to vanish for small crossclasses in proportion to $\bar{M}_c$. For example, a "large" gain of 20 percent becomes merely 2 percent for a crossclass of 10 percent, because $1-\bar{M}_c \cdot 20 = 0.98$.

These approximations serve adequately for understanding and for designs. More precise formulations (as found in the above references) require lengthy descriptions. They are needed for extensions to stratified samples beyond PRES: also for computing variances. The tendency toward SRS is even stronger for differences of means of crossclasses, where a good approximations is:

$$\mathrm{Var}(\bar{y}_c - \bar{y}_d) = S_c^2/n_c + S_d^2/n_d \,. \qquad (8.3.2)$$

That is, *variances for differences of crossclass means for PRES tend to those for SRS*, losing all the gains of stratification. Furthermore, this also happens to ordinary kxm Chi–square tests, which can also be viewed as combinations of pairwise differences. We may conjecture similar tendencies for other analytical statistics from PRES [Kish and Frankel, 1974].

The most *drastic effect is on variances of simple expansion totals of crossclasses*,

$$\hat{Y}_c = \Sigma_h y_{ch}/f_h = \Sigma_h N_h y_{ch}/n_h:$$

$$\mathrm{Var}(\hat{Y}_c) = \Sigma_h N_h^2 \bar{M}_c [S_{ch}^2 + (1-\bar{M}_{ch})\bar{Y}_{ch}^2] / n_h \,. \qquad (8.3.3)$$

Because of the loss of control over the sample size n_c of the crossclass (c) the element variance becomes inflated by the second term, which adds the squared mean to the within stratum variance. On the other hand if the domain sizes M_{ch} are available for a ratio mean $\Sigma_h M_{ch} y_{ch} / n_{ch}$, then this term can be avoided (12.3).

8.4 EFFECTS OF CLUSTERING ON SUBCLASSES

Here we describe the effects of clustering on subclasses of sizes n_c in comparison with complete samples of similar sizes. These effects look beyond the principal effect of variance increases due to decreases of sample sizes from n to $n_c = n\bar{M}_c = nM_c/N$, where $\bar{M}_c$ denotes proportion of the population in subclass c.

In order to be generally meaningful, yet brief and simple, we present this problem entirely within a framework of ultimate clusters: samples of $n = \Sigma b_\alpha$ elements in a clusters with the average of $\bar{b} = n/a$ elements per cluster. The b_α vary some but not too much, so that the $\bar{b}$ may be useful in judging design effects. The sample of a clusters denote primary selections of PSU's, usually stratified. The subsamples b_α denote elements in ultimate clusters selected with any EPSEM design. Unequal selections with weighted estimates are discussed later.

Clustering generally increases the variances of cluster means, as measured by values of $deft^2 > 1$. The increases vary a great deal, and these variations are also reflectedby subclass means, but in modified forms. We should first distinguish the effects $deft^2 = [1 + roh\ (\bar{b}\text{-}1)\ ]$ on design subclasses from those on crossclasses.

Design subclasses (and domains) are based on separate and independent strata (by definition). That forms the basis for imputing to them variances and $deft^2$ from values computed for the overall sample mean $\bar{y}$. Conjectures from $deft^2(\bar{y})$ to $deft^2(\bar{y}_c)$ for design subclasses (c) seem reasonable when similar designs prevail over all domains; that is, when the design and $deft^2$ in the $a_c = \bar{M}_c a$ subclass clusters are similar to those in the entire sample of a clusters. For example, a sample of holdings may have similar subsample sizes $\bar{b}$ and designs for the separate provinces (domains). Then assuming similar values of $deft^2$ for subclasses, and for the entire sample we may impute that

$$\text{var}(\bar{y}_c) = \text{var}(\bar{y})/\bar{M}_c\,, \qquad (8.4.1)$$

because $\text{deft}^2(\bar{y}_c) = \text{deft}^2(\bar{y})$; and $\bar{M}_c = n_c/n$ may be assumed. Such imputations are more difficult and complex when the designs differ for subclasses; for example, designs for metropolitan areas, urban areas, and rural areas may all differ basically in the nature and degree of clustering (sizes of $\bar{b}$). Even then we may remember that the overall $\text{var}(\bar{y})$ and $\text{deft}^2(\bar{y})$ are the weighted averages of similar functions for the separate design subclasses (14.3).

For *crossclasses the situation is quite different*, imputation a bit more complex but also more stable, and backed by a great deal of empirical evidence. Crossclasses (by definition) tend to cut across all strata and all clusters, hence to reduce design effects in drastic and fairly predictable manner. The reduction of the sample size from n to $n_c = \bar{M}_c n$, and retention of all (or most) of the a clusters, also means that the average subsample size gets reduced from $\bar{b}$ to $\bar{b}_c = \bar{M}_c \bar{b}$; this reduction is assumed to happen in all the a sample clusters on the average. But we also expect some variation in the actual subsample sizes $b_{c\alpha}$ even for demographic variables like age, children, births; and even more variability for socio—economic classes like income and occupation, because of their partially clustered distributions. Some variables, like type of farming, that depend on soil and climate, can be even more clustered and may be considered neither crossclasses nor design classes, but mixed types.

The design effects on the entire sample have been denoted as $\text{deft}^2 = [1 + \text{roh}(\bar{b} - 1)]$: ratios of increases in element variances, where the *roh* summarizes the effect of clustering (with stratification) and averages them over the diverse parts of the sample. These synthetic roh values vary greatly between variables and designs, and are specific for them. The design effects for crossclasses can be similarly denoted as $\text{deft}_c^2 = [1 + \text{roh}_c(\bar{b}_c - 1)]$, but it would be laborious to compute these for all kinds of crossclasses. Fortunately, it seems generally true that for crossclasses, approximately $\text{roh}_c = \text{roh}$ for the entire sample. Therefore, we may use as an approximation:

$$deft_c^2 = 1 + roh_c(\bar{b}_c - 1) = 1 + roh(n_c/a - 1) . \qquad (8.4.2)$$

$Deft_c^2$ tends to be reduced toward 1 as the crossclass proportions $\bar{M}_c$ decreases, because roh_c for crossclasses tends to be similar to roh for the entire sample, for specific variables and designs. This relationship has been shown in many computations [Verma et al., 1980; Kish et al., 1976], but with some modifications. First, only $n_c/a = \bar{b}_c > 1$ should be considered, because the relationship tends to break down near $\bar{b}_c = 1$, where $deft_c^2$ approaches 1 in any case. Second, $roh_c > roh$ slightly even for "true" crossclasses, and somewhat more for socio—economic classes. But these slight increases (perhaps $roh_c =$

1.2 roh as a rough tendency) pale to insignificance compared to differences of roh between survey variables (which may be ten— or hundred fold); also to differences between subclass sizes $\bar{b}_c$, especially when compared to $\bar{b}$. The relationships seem to stand up best, fortunately, where they are most needed: for larger values of $deft_c^2$, due to larger values of roh_c and $\bar{b}_c$. They should not be used for values of $deft_c^2$ and of $\bar{b}_c$ near 1, where they are less needed. In a word: *roh_c* is much more "portable" for different crossclasses of specific variables than is $deft_c^2$, because it excludes the common, wide variations in $\bar{b}_c$.

Differences between crossclass means show further reductions of clustering effects: covariance terms, $2\,cov(\bar{y}_c, \bar{y}_d)$ tend to reduce the clustering effects, so that $var(\bar{y}_c - \bar{y}_d)$ comes closer to SRS variances but still remains higher:

$$S_c^2/n_c + S_d^2/n_d < Var(\bar{y}_c - \bar{y}_d) < Var(\bar{y}_c) + Var(\bar{y}_d) . \qquad (8.4.3)$$

These relationships have been found in many computations over the years [Kish, 1965, 14.1; Kish et al., 1976; Verma et al., 1980]. Though subject to sampling fluctuations, results tend to fall nearer the lower (SRS) limit than the higher limit (without covariances). Because of those sampling fluctuations in computed results, we prefer to use here the expected population values, symbolized with capitals. We may symbolize usefully and jointly the relationships of (8.4.2) and (8.4.3) with:

$$1 < Deft^2(\bar{y}_c - \bar{y}_d) < Ave[Deft^2(\bar{y}_c) + Deft^2(\bar{y}_d)] < Deft^2(\bar{y}) . \qquad (8.4.4)$$

If $\text{Deft}^2(\bar{y})$ is not much greater than 1, then some reasonable conjectures about $\text{Deft}^2(\bar{y}_c - \bar{y}_d)$ may be made between those limits. The third term denotes an average of its two components.

For simplicity, we had to assume for this section EPSEM selections with self—weighted results, also rather uniform design over the entire sample. For different designs, with differing probabilities or different designs over portions, the above may still be useful, but more complex analysis is needed. These would need developing at greater length than would be feasible here. We note here also that increases in variances due to "random" weighting persist also in subclasses (12.6).

8.5 SMALL DOMAIN STATISTICS: TECHNIQUES AND CUMULATIONS

Complete censuses and administrative records yield data for small domains, whether crossdomains or design domains like small local administrative areas. But data from decennial censuses are not timely and those from registers are not "rich" in depth and spread (17.1). Sample surveys alone cannot provide sample bases in the detail needed for small domains, because of small sizes in numbers of elements, and especially in numbers (a) of primary selections.

Demand for statistics for small domains, especially local administrative and geographical areas, has been growing in all fields. It is especially needed for agricultural statistics where crops, yields, practices and economics depend so much on local conditions. Statistics are needed not only for understanding and planning, but also for administrative action that should be specific to small domains.

Growth in supplies of statistical capabilities may have been even more crucial than the demands in the recent growth of techniques for statistics of small domains and local areas. Recent growth of capabilities has come in three basic fields. First, there are more and better data from three sources: from censuses, from administrative records, and from sample surveys. Second, we

witness an explosion in the capabilities of statistical computing machines and programs. Third, statistical estimating techniques have been designed to take advantage of the above two advances.

It would be difficult to describe briefly the diverse methods developed for population counts and for vital and health statistics. Two recent reviews with many references are by Platek et al. [1985] and Purcell and Kish [1980]. The techniques depend on different combinations of the three sources of data: censuses, registers, and samples. And each of eight major techniques now available also uses different statistical methods for combining those sources of data. The central concept consists of combining the strength of each source to overcome the weaknesses of the other(s).

"*Cumulating cases and combining statistics* from different samples is becoming more feasible with the growth of repeated and periodic surveys. Cumulating cases can refer to aggregating, summing, amassing individual elements from repeated surveys. Combining statistics from published results can be done for surveys and for experiments carried out in diverse places and at different times." [Kish, 1987, 6.6] Larger samples for domains, especially for minor domains, are needed for most surveys: sample sizes are limited, whereas interest in details is unlimited. Cumulating cases over time from periodic samples seems particularly attractive. For example, a national sample of farms, taken yearly or quarterly, may also yield reliable data for major domains but not for minor domains. However, cumulations over, say, five periods, may yield adequate data for minor domains, though less frequently. For minidomains some combined small area technique may be feasible.

8.6 SAMPLING FOR RARE ITEMS

Suppose that only a small proportion $\bar{N}_c = N_c/N$ of the total population possesses a variable (trait, item) denoted by $Y_i \neq 0$, and for the vast majority $(N - N_c)$ the variable $Y_i = 0$. For example, only a small proportion $\bar{N}_c$ of all N holders may grow Y_i kgs of walnuts, or raise Y_i turkeys, however the other

$(N - N_c)$ farmers have none, $Y_i = 0$. The rare item to be estimated may be the number N_c or the proportion $\bar{N}_c = N_c/N$ of the rare subclass; or it may be the mean $\bar{Y}_c = Y_c/N_c$ of the variable within the subclass, or the mean $Y_c/N = \bar{Y}_c\bar{N}_c$ over the entire population. Any or all of these may hold substantive interest, and they present related problems of finding a sample n_c to estimate N_c, for whom $Y_i \neq 0$.

Several alternative methods are presented below for dealing with these problems. The choice between them depends on the rarity $\bar{N}_c$ of the variable and on the costs of measuring Y_i; also on the available resources, special lists and the costs and biases of such special lists and resources. It may not be too difficult to find crops that are grown by 5 percent of holders, or only in a few districts of one province. But if only 1/1000 of the holders grow it and they are scattered over most of the country, the problem becomes difficult, unless a good list is available. For very rare, scattered and unidentifiable items, survey sampling may not be feasible at all. Detailed treatments or alternatives and further references may be found elsewhere [Kalton and Anderson 1986; Kish 1965, 11.4]. These methods are presented not only for the very "rare items," but also for minidomains and even minor domains, as defined in (8.1).

a) *Cumulation* of rare populations may be the best strategy for continuing or periodic surveys with changing samples (16.3). For example, from quarterly surveys of agriculture, four quarters over a year, or twelve over three years, may yield adequate samples. The cumulated statistics must be regarded as averages over the periods covered by the surveys, and such averaging has definite advantages, especially in agriculture, which is subject to seasonal and even yearly fluctuations.

b) *Multipurpose samples* cover several or many rare variables on the same survey(s) and divide the costs among them. They can also enrich analysis with statistics on relationships between these variables. Surveys for market research commonly cover several products, each relatively rare. In a sense every survey of agriculture is a multipurpose survey of distinct crops and

animals, most of them more or less rare. Similarly health surveys cover distinct diseases, most of them rare, and a health survey can cumulate cases from 52 weekly samples [NCHS 1958]. Thus cumulations go well with multipurpose surveys.

c) *Large clusters* can decrease drastically the costs of locating and screening for rare populations. For example, a sample of complete E.D.'s, or of villages, can be searched and screened for holdings or households with the needed trait(s). Even large clusters will yield only small clusters of the rare population: from B_α elements in the cluster we expect $\bar{N}_c B_\alpha$ members. The actual numbers will vary around this, hence one can also tolerate greater variation in the B_α (size of E.D. or village).

d) *Disproportionate sampling* with "optimal" allocation may be used if the rare population tends to be concentrated in some clusters, and especially some domains, which can be *identified before selection*. As an extreme example, suppose a crop is known to grow only in one province or valley, where a high sampling ratio f_1 may be used; the rest of the country may "safely" receive $f_2 = 0$ with this "cut–off" method that would omit no more than negligible amounts [HHM, 1953, 11.6]. But concentration is not so extreme for most variables, and "optimal" allocation with two of a few strata must be used, when the areas and levels of concentration can be identified in advance.

e) *Two–phase sampling* (double sampling) may be used when the clusters and strata with concentrations of the needed rare population cannot be readily identified in advance of the selection. The first phase consists of a screening operation for identifying members of the rare population. If that would be too expensive, screening may sometimes be used to identify clusters with high concentration. For example, districts or villages with many of the needed farmers may be identified from lists of inquiries on a sampling basis in the first phase and then "optimal" allocation applied in the second.

The screening operation must be cheap and therefore usually has errors of inclusion and exclusion. Screening is useful when it is cheap per element, yet has very few "false negatives" (erroneous exclusions), and not too many "false positives" (erroneous inclusions). The costs and ratios of both kinds of errors are important, but false exclusions are much worse. The allocation of effort between the first and the second phase should be determined by the efficiency of the screening and the ratio of costs in the two phases [Kish 1965, 12.1; Cochran 1977, 12.1–12.5].

f) *Special lists* may provide the most efficient (or only feasible) technique for populations that are: very rare, widespread and difficult to identify in the field; but available (mostly) on a good list or lists. For example, it is possible that (almost) all holders who raise a rare animal belong to an association (for legal or economic reasons). Most lists are far from perfect, and if a large proportion (or not small) of the rare population ($1-\bar{N}_c$) are unlisted, and if the unlisted differ in kind, samples from the list would give biased results. When finding and identifying a probability sample of a rare population just is not feasible, lists of such elements have been used. Means from such lists may or may not be badly biased. Analytical statistics of relations must also be viewed with caution (3.2). And estimates of aggregates must depend on outside data, models, and conjectures.

g) *Dual or multiple frames* can be useful in agriculture. A basic example uses a special list of a population, as above, but adds an area supplement for elements missing from the list. For example, a rare population may be identified on the list of the last census (or some other list), and an area supplement can be added to find new and missing elements. The sampling fraction for the list could be much larger than for the supplement ($f_1/f_2 > 1$) because the cost per element is much less for the listed holders ($c_1/c_2 < 1$). The area sample could exclude those on the list, or include them because that may be cheaper (11.2).

h) *Other methods* are available for finding rare elements but we may leave those to the references above, because they are not generally useful in agricultural surveys. *Batch testing* can be used to find rare elements, if part of the material, when tested in batches (as in blood tests) can reveal the presence of even one rare element, which may then be identified with further testing. *Snowball sampling* is a colorful name borrowed for various techniques of building up lists for special populations by using an initial set of the members as informants. *Multiplicity sampling* refers to using several informants to find, identify (and perhaps collect data about) each element of a rare population.

CHAPTER 9. MULTIPURPOSE SAMPLE DESIGN

9.1 UNIPURPOSE DESIGN

Most surveys involve several purposes during the planning stages. Typically many more purposes emerge later during the analyses of the data and during their interpretation and utilization. Why then is the true multipurpose nature of surveys neglected in sampling methods? These usually present a unipurpose orientation for economical sample designs. The reason for that oversimplification is that sampling theory is rather complex already, and multipurpose design would make it even more complex and difficult. For the same reason we also must begin with simple unipurpose designs, before we leave that simple base for multipurpose designs in later sections. We had already presented (5.2) the commonly known

$$\tilde{V}ar(\bar{y}_{srs}) = (1 - f)\,\tilde{S}^2/n \qquad (9.1.1)$$

for the designed variance of means ($\bar{y}$) with SRS of n elements. This result should be viewed as the $\bar{V}ar\,(\bar{y})$ obtained for a fixed total cost of C = cn; with c as the estimated cost per element, C buys n = C/c. The designed $\tilde{V}ar(\bar{y})$ depends on the guessed element variance $\tilde{S}^2$; both are estimated and that is why they carry the notation (~). From the data themselves valid estimates $var\,(\bar{y}) = (1 - f)s^2/n$ can be computed; and too low guesses for S^2 will result in actual $var(\bar{y})$ which then would be larger than the designed $\tilde{V}ar(\bar{y})$; but the wrong guesses for parameters would not bias the sample statistics. For this formulation we needed: the sample design (SRS); the sample size n, from n = C/c, hence C and c; estimate or guess for S^2; also f = n/N for (1 − f), but this is seldom important. If c is underestimated C = cn will be higher than planned, because adjusting n is seldom practical. The most important and difficult task is the decision to choose ($\bar{y}$) as the "only" or the most important "statistics," for which $\tilde{V}ar(\bar{y})$ is designed and $\tilde{S}^2$ guessed. This unique choice is avoided in later sections for multipurpose designs.

If the Var($\bar{y}$) seems higher than "required," the sample size n and cost C = cn may be scaled up in order to reduce it. Less often Var($\bar{y}$) may be lower than "required" and then costs may be reduced.

If the "required" precision Var ($\bar{y}$) is relatively fixed and the allowed sample size and cost C = cn can be suited to that need, then the above procedure may be reversed to proceed from Var ($\bar{y}$) to the n needed to attain it:

$$n' = S^2/\text{Var}(\bar{y}) \text{ and } n = n'/(1 + n'/N) . \qquad (9.1.2)$$

The size n' neglects the factor (1 − f), but dividing by (1 + n'/N) yields the final n corrected for this factor − which is usually negligible, and may even be inappropriate.

If the initial guess about $\tilde{S}^2$ was an underestimate and the sample results show $s^2 > \tilde{S}^2$, then $\text{var}(\bar{y}) > \tilde{\text{V}}\text{ar}(\bar{y})$. But the required $\tilde{\text{V}}\text{ar}(\bar{y})$ may be obtained with a supplemental simple n" so that $(n + n")/n \doteq s^2/\tilde{S}^2$. For practical reasons of data collection it may be better to design initially for a sufficiently large sample n(max), then select first a small sample n (min); on the basis of the simple estimates s^2, enlarge the sample to attain var($\bar{y}$) close to the "required" $\tilde{\text{V}}\text{ar}(\bar{y})$. This two−phase procedure assumes flexibility in the timing and the cost cn of data collection; that flexibility may not be feasible.

Two other sources of uncertainty, hence two other needs for size adjustments, concern nonresponse and design effects. Nonresponse and noncoverage must be anticipated in planning any design. These can vary a great deal and any planned value is subject to errors, but a simple procedure may suffice here. Express the coverage rate with the proportion p_c and the response rate with p_r and then use these to inflate the planned sample size: $n_t = n/(p_r \cdot p_c)$. Thus if one wants n = 1000 and expects $p_c = 0.96$ for coverage and $p_r = 0.93$ for response, $n_c = 1000/0.93 \cdot 0.96 = 1120$ should be the initial target. One may also view these factors as factors that decrease the effective population size from which the sampling fraction f = n/N is selected.

The effects of departures of the sample design from SRS can most simply dealt with estimates of design effects, but of course always with errors in these advance estimates based on extraneous sources (14.1). Remember that typically for stratified PRES samples $Deft^2 = S_w^2/S^2 < 1$ (slightly) (5.4). $Deft^2 = [1 + roh(\bar{b} - 1)] > 1$ for clustered samples and these can vary from mild to severe effects and deserve detailed investigations (6.6). In either case one may view the effects as $D^2S^2/n = S^2/(n/D^2)$: either directly on the effective element variance D^2S^2, or inversely on the effective sample size n/D^2.

Estimating the element variance S^2 is not difficult for proportions, when P can be estimated well enough to guess $S^2 = PQ$, which is not highly sensitive to moderate errors in P. When one can guess both $\bar{Y}$ and the coefficient of variation $C = S/\bar{Y}$ well enough, then $S^2 = C^2\bar{Y}^2$ may also serve us. Although never easy or precise, reasonable guesses about S^2 are often feasible, and using models can be helpful [Kish 1965, Fig. 8.2.II].

9.2 PURPOSES

The term "multipurpose" has been used with several meanings and especially for levels 3 and 4 below; and four other levels need recognition as well. Thus the twenty or so distinct meanings for the concept of "multipurpose" are presented in a hierarchy of six levels, rising from the least to the most complex type of survey operations. All those meanings need attention, because any of them can lead to the kind of conflicts listed later that we should become aware of during the design of surveys. We shall list ten areas of conflicts and most of these can occur in connection with most of the kinds of purposes distinguished here.

1. *Diverse statistics from the same variables* should be noted and treated separately. For example, hectares of wheat holdings can be presented not only with the mean of holdings, but also with the median (which can be much less) and other quantiles: and also as the percentage with less than some fixed number of hectares. Optimal allocation for the mean will point to selecting more large farms, whereas the median, the quantiles, and the percentages will all tend to point to PRES as nearly optimal (5.6).

Complex analytical statistics, such as regressions, categorical data analysis, and other methods for the analyses of relationships would, if pursued, also tend to lead to different allocations.

The time aspects of periodic studies often represent a variety of purposes, and these can lead to conflicts when designs for individual (micro—) changes and of aggregate (macro—) changes are compared to designs for cumulated data and for current (static) levels (9.3, 16.3). For example, periodic agricultural surveys should have maximal overlaps (panels?) of

holdings to measure changes in plantings (of wheat, rice, vegetables, in response to changes in prices or policies); but to cumulate for enlarged samples it is best to design for minimal (zero) overlaps (16.3).

2. *Statistics for domains* present most often the greatest problems and sources of conflict. Differences between sizes of domains and of the entire sample can be very large. Strong needs and reasonable requests for better provincial data are commonly expressed. The nature of different domains is described elsewhere (8.1), also the areas of conflict and of useful compromises (9.5). We merely repeat that the primary effects of domains that cover the proportion $\overline{M}_c = M_c/N$ of the population are to reduce sample sizes and increase variances by $\overline{M}_c$. Thus, even for major domains that cover 0.10 of the population and sample, variances increase by factors of 10. Furthermore, for subclass analysis based on comparisons of pairs the variances will tend to be 20 times greater than the overall sample variance.

3. *Multiple variables on single subjects* are very common and they occur in several forms. First, alternative measures may be taken on single variables. For example, the time, labor, and fertilizers spent on the farm can each be measured in alternative ways. Crop yields can be measured in wet or dry weights, and in other ways. The crops growing on specified fields may need several measures if they are interplanted, or if they are rotated between seasons. Second, different periods of coverage — per day, per week, per month, during the entire year — will each yield different measures of the same basic variable.

Third, surveys often cover diverse aspects with different variables of the same subjects. For example, production of a single crop involves the areas of planted plots, fertilizers, labor for planting and for harvesting, then the transport and economics of selling of the crop. These, like all the alternatives noted under the first three points above, are examples of multipurpose needs common in most surveys even for single subjects.

4. *Multisubject surveys* are, however, conducted very frequently. For example, agricultural surveys may fruitfully combine in the same schedule and interview some items about the production of some crop (like rice), together with its own food consumption and with related financial items. Furthermore, agricultural surveys often deal with several, often many, crops; and production of *every crop and domestic animal is a distinct subject* that requires separate attention, planning, hence design. Multicrop agricultural surveys are multisubject surveys.

In other fields also, multisubject surveys are common. Surveys of store inventories for market research combine several clients on single surveys. Health surveys often combine many diseases because each of them would be a rare item that would be both expensive and unstable (8.5). Socio—economic surveys also uncover more and richer relationships by combining several subjects in a single schedule, interview and field operation.

5. *Continuing and integrated survey operations* raise survey complexities to an even higher level. Furthermore, their influence is broader than has been commonly appreciated: perhaps most good surveys in many countries depend on the operations of some continuing, integrated organization (office, institute, etc.), often in the national statistical office. Setting up a high quality, widespread survey operation may just be too difficult and expensive for single surveys.

Survey operations of the U.S. Census Bureau (USCB) and its Current Population Surveys are well known examples. "The advantages and savings are substantial; they may be properly allocated with cost accounting, similarly to investments in large machine tools for mass production" [Kish 1965, 12.6, Ch. 10; USCB 1978]. The National Sample Survey of India is an early and outstanding example [Murthy 1967, Ch. 15]. The United Nations' National Household Survey Capability Program is successfully promoting integrated survey programs [UN 1980]. The survey operations of Statistics Sweden and Statistics Canada also cover different fields and methods, different technicians

and different field staffs, all from single coordinated offices for sample surveys. This also occurs outside of the central statistical offices, as in the Institute for Social Research of The University of Michigan [Hess 1985].

6. *Master frames and master samples* must involve at least some loose level of coordination needed for the combined utilization of common design aspects. Joint use implies some design constraints as well as benefits from common utilization. The two terms in the title have been used without clear definitions or distinctions, and perhaps some examples will suffice here.

An extreme of a master sample comes the Current Population Surveys of the USCB where sample areas are divided into 8 segments to be selected for 8 identical periodic surveys [USCB 1978]. Master listings of sample blocks, areas, E.D.'s are also used for selecting unspecified, diverse samples for data collection by a single organization. However, a sample of listings can be used also by several organizations; in The Federal Republic of Germany one firm sells addresses for most samples. A single selection of PSU's, staffed with skilled interviewers. may be used by several government bureaus for distinct sample surveys; agriculture, employment, health, education, etc.

9.3 TEN AREAS OF CONFLICTS BETWEEN PURPOSES

These areas of conflicts do not have one—to—one relationships with the twenty or so purposes listed under six levels (9.2). Any of the ten areas may occur with most of the purposes, and we may think of 60 or 200 kinds of conflicts, though some should be more common and important than others. References indicate where these conflicts are discussed in more detail.

a) *The overall size* of the sample size in numbers of elements n (holdings, households, interviews) should be a source of concern and conflict. For multistage samples the numbers of other units, especially of PSU's, must also concern the designer of samples. But it would be too complex to discuss all selection stages here, and we may here use D_g^2, the design effects, as surrogate

measures for the omitted complexities. For brevity, D_g^2 also includes the fpc = $(1 - f_g)$. The index g denotes a specific statistic, which has complex sources, because it depends on the variables, on the subclass base, and on other factors.

A sample of size n yields for a mean $\bar{y}_g$ the $\text{Var}(\bar{y}_g) = S_g^2 D_g^2/n$, and the sample size needed for a required $\text{Var}(\bar{y}_g)$ is $n_g = S_g^2 D_g^2/\text{Var}(\bar{y}_g)$. These can differ considerably between statistics g, because of variations in all its factors. Especially when dealing with subclasses the variations between sizes can be great. We may express the subclass size as $m_g = n_g P_g$, yielded by the subclass proportion from an overall sample size n_g. Then the sampling fraction $f_g = n_g/N = m_g/P_g N$ needed to yield subclass size $m_g = S_g^2 D_g^2/\text{Var}(\bar{y}_g)$ is

$$f_g = S_g^2 D_g^2/\text{Var}(\bar{y}_g) P_g N. \qquad (9.3.1)$$

Thus, the overall sampling fraction f_g may have to be increased greatly to satisfy the $\text{Var}(\bar{y}_g)$ specified for a small subclass P_g.

If it were not for cost constraints, it would be possible to satisfy all the required f_g by taking the largest f_g; as a result of such generous design the f_g, n_g and m_g for other statistics would be larger, hence $\text{Var}(\bar{y}_g)$ smaller, than needed. But restraints on total costs seldom permit such generosity (9.4).

b) *Allocation of the n_g among domains* presents other problems, though related to those of (a). For example, the overall sample sizes of (a) concern us most when designing for crossclass sizes, such as age groups. But, suppose that we want to: (1) satisfy requests for specified values of $\text{Var}(\bar{y}_g)$ for provinces g; or (2) distribute a fixed total sample size $n_t = \Sigma n_g$, so as to have the same $\text{Var}(\bar{y}_g)$ for all provinces. Of course, the provinces "always" differ greatly in size; thus to satisfy these provincial requests would result in widely different sampling fractions f_g for the various domains. These differences result in conflicts of allocation between provincial and other design domains, and also for the needs of overall national and crossclass data. Joint solutions and adequate compromises are presented later (9.4–9.5).

c) *Allocation among strata* should not be confused with allocation for domains, although they may be related. For example, production of a crop (say rice) may differ widely between provinces; thus for estimating total, national rice production some widely disproportionate sampling rates f_h may be "optimal," with greater f_h to provinces with greater and more variable production rates. In general, this concerns the allocation of a fixed total sample size $n = \Sigma n_h$ among the strata (h); or total fixed cost $\Sigma c_h n_h$ if the costs vary; or domain sizes $m_g = \Sigma m_{gh}$ for domain totals (9.4–9.5). Clearly the best allocations will differ between crops (rice, wheat, cattle, etc.). For overall agricultural surveys it may be best to have perhaps a smaller f_h for urban and metropolitan areas.

d) *Choice of stratifying variables* poses problems, especially for multisubject surveys and for integrated survey operations. If, for example, the same survey or operation covers both agricultural production and employment statistics, the best stratifiers will differ for the two. However, for such problems of optimal efficiency, adequate solutions may be found with partial use of both (or all) sets of stratifiers and with compromises (6.3).

e) *Optimal cluster sizes* differ between variables and especially between crossclasses and the entire samples. For simplicity, we use again the overall design effects $D_g^2 = [1 + roh_g(\bar{b}_g - 1)]$ to measure the effects of clustering (6.6). This neglects the separate effects for stages of selection and stratification. It also uses an overall average $\bar{b} = n_g/a$, the members of elements per PSU over the sample; but it may be desirable to design smaller $\bar{b}_g$ for blocks in metropolitan areas than for large outlying areas (6.7).

The variation in the D_g^2 may not be as great as for some of the conflicts in a,b, and c; thus compromises may be less difficult; the range of usual variation in most cases may be $1 < D_g^2 < 10$ perhaps. The relative range in roh may be great, say $0.001 < roh < 0.200$; yet it can have only mild effects on the

optimal $\bar{b} = \sqrt{[(C_g/c)(1-roh)/roh]}$ (6.6–6.7). The design must also consider divergent values of $m_g = P_g n_g/a$, and for small crossclass proportions P_g the optimal values of the total cluster size $\bar{b}_g$ may be large (8.4, 8.6).

f) *Measures for cluster sizes* may also differ. For example, whereas we may neglect the differences between total persons and total dwellings and occupied dwellings for measures of cluster sizes, numbers of holdings and of dwellings generally differ widely. This raises problems for integrated surveys that try to satisfy both holdingss and all households.

g) *Retaining sampling units (PSU's)* provides techniques for adequate compromises between divergent measures of size for different subjects and objectives. The techniques also serve (and were designed for) measures that change over time, e.g., between decennial censuses. They can also provide compromises for differences in strata between conflicting designs, as in d above (11.7).

h) *Designs over time* for periodic surveys must involve several decisions. a) How much overlap to include: total for measuring changes, or none for cumulating, or partial ($0<P<1$, but how large) for static measures? b) What units in the overlaps: elements, or PSU's only, or some intermediate units? c) What periods to overlap: monthly, or yearly, or quarterly? (16.3).

i) *Relations of biases to variable errors* in RSME $= \sqrt{(\sigma^2 + B^2)}$ provide interesting contrasts between variables, and particularly between statistics for small domains and for overall statistics (15.2). In summary: biases and potential biases may dominate for some overall statistics (means and totals), though not necessarily for all. However, for small domains, and especially for comparisons, the variable errors, especially sampling errors, increase faster relatively, and they tend to become larger than biases.

j) *Computing and presenting sampling errors* also require decisions and choices, because usually it is not feasible to compute, and even less to present, sampling errors for all of the many statistics that survey reports commonly offer in large numbers. This is especially true of sampling errors for all domain statistics. Thus generalized tables are often devised and presented (14.3).

As for measuring and presenting not only sampling errors, but also all sources of biases, that is but an idle dream stimulated by articles about "total survey errors."

9.4 COMBINED OPTIMA

Two distinct technical methods exist for the joint solution of conflicts in allocation, and particularly for the first three conflicts noted above: a) overall sample size, b) allocations among domains, and c) allocations to strata. One method designs weighted compromises among variances for fixed total cost (9.5). Another approach, merely outlined here, uses iterative, nonlinear programming in order to *satisfy for minimal cost the specified variances* (or required precisions) *jointly for all stated purposes*. These elegant solutions utilize the capacities of modern computers and have appeared in many articles since 1963 [Cochran 1977, 5.3–5.4, Bean and Burmeister 1978, Rodriquez–Vera 1982, Kokan 1963]. They deal chiefly with allocations among domains, and with allocations to strata.

The costs needed to satisfy the required precisions often turn out to be much too high for the sponsors, because the "required precisions" were unrealistic. It would be possible to scale the entire sample down to an allowable cost level; but such rescaling would give all the required precisions the same lack of importance, and this would not be as realistic as the fixed cost approach of 9.5. Such rescaling would expose the lack of realism of this elegant approach.

This situation perhaps points to a fundamental fault in an approach that is based on fixed specified variances. This implies assigning arbitrarily fixed constant values to any variance below the "required" Var($\bar{y}$) and zero values to variances above it. Instead of such dichotomous step functions, it is more realistic to postulate smoother functions of increasing worth for decreasing variances. The compromises of the next section attempt to provide designs for them.

9.5 COMPROMISE ALLOCATIONS

A potentially useful approach calls for a compromise by *averaging all the "optimal" allocations for various purposes, by minimizing the combined weighted variances for fixed cost*, or fixed sample size. Let us first note the four steps involved in this method and then critically evaluate them. "Potentially" above will be justified by examples of the method's applicability. But it also signifies that no formal application of these methods has been found in practice. Therefore, the reader may be justified in passing over these necessarily technical pages.

1. Denote with $\Sigma_i V_{gi}^2/n_i$ the variance attainable for a statistic g, with the allocations of sample sizes n_i for the i—th component of variation. The index g may refer to variables, domains, formulation (e.g. $\bar{y}$). The index i may refer to domains, strata, stages; for all of these components and statistics, linear summations of quadratic terms is assumed, as in sampling theory generally.

2. Denote with $V_g^2(\min)$ the minimal variance obtainable for the statistic g for a fixed cost $C = \Sigma c_i n_i$. This would be for the "optimal" allocation of the sample sizes n_i to domains, strata, stages; such optimal allocations are mentioned in these separate sections.

3. Now let

$$1 + L_g(n_i) = (\Sigma_i V_{gi}^2/n_i)/V_g(\min) = \Sigma_i C_{gi}^2/n_i \qquad (9.5.1)$$

denote the ratio of increase in the variance, due to any allocation n_i for the statistic g, over its own minimal variance. *Thus $L_g(n_i)$ measures the relative loss for the statistic g over its own minimal (optimal) value of 1.* Accepting these *relative* variances C_{gi}^2/n_i and relative losses $L_g(n_i)$ to be minimized represent a critical decision. They seem more reasonable functions to be combined in (9.5.2) than the V_{gi}^2, because these depend on arbitrary units of measurement, which are removed by the $V_g(\min)$ from the relative measures C_{gi}^2/n_i.

4. Next the separate relative losses for the various statistics g must be combined into a joint loss function. For this combination some *relative* weights of importance

I_g(with $\Sigma_g I_g = 1$) are assigned to the statistics (g) for any set of allocations (n_i) of the sample sizes:

$$1 + L(n_i) = \Sigma_g I_g(1 + L(n_i)) = \Sigma_g I_g \Sigma_i C_{gi}^2/n_i \qquad (9.5.2)$$

$$= \Sigma_i \Sigma_g I_g C_{gi}^2/n_i = \Sigma_i Z_i^2/n_i \,.$$

First, the order of summation was merely changed and then the new variables $Z_i^2 = \Sigma_g I_g C_{gi}^2$ were created. These Z_i^2 can be computed after the relative measures I_g are assigned and the C_{gi}^2/n_i computed.

5. Finally the function $1 + L(n_i) = \Sigma_i Z_i^2/n_i$ can be "optimized" for the allocations n_i, in order to yield *the combined compromise solution of the minimal weighted loss* $\Sigma_g I_g L_g(n_i)$ for the fixed total cost $\Sigma C_i n_i$. This solution is similar to "optimal" allocation of the n_i to strata in the simple univariate case (5.6):

$$n_i \propto Z_i/\sqrt{c_i} \,. \qquad (9.5.3)$$

For fuller justifications the reader may look at some references [Kish 1976; 1988; 1965, 8.5; Cochran 1977 5A.3–4]. However, some assumptions of the model may be briefly discussed here. a) The variances $\Sigma_i C_{gi}^2/n_i$ relative to a minimal $V_g^2(\min)$ are used deliberately, though other functions of variances may also be substituted. They seem to provide the best bases to standardize

the variances for units of measurement before combining them; they also seem to provide good bases for assigning the weights I_g for relative importance. But there may arise rare cases when very small values $V_g^2(\min)$ in the denominator would cause some rating to become wildly large and unstable; then those ratios or their I_g should be assigned arbitrary values or removed. b) Assigning relative values I_g may seem both arbitrary and difficult; but, compare that to the difficulties of its two alternatives. Univariate allocation amounts to assigning $I_g = 1$ to the single "principal" purpose and $I_g = 0$ to all other purposes; our method includes that as a special case, which would seldom be chosen willingly. On the other hand, the "combined optima" (9.4) assigns arbitrarily equal weights of importance to all purposes ($I_g = 1/G$) and then would try to satisfy all of them, *if it could*; this method also demands specified values of $\mathrm{Var}(\bar{y}_g)$ for each purpose, a difficult and unrealistic task, indeed. c) The method also needs estimators or guesses about parameters like $S_g^2 D_g^2$, as in the unipurpose case (9.1), but many more. In a realistic situation this may be done for several critical and contrasting types, chosen so as to "represent" the spectrum of all important purposes. d) The method relies on linear combinations of sums of squares of variance components. It would be difficult to circumvent this model, so common and useful in statistics.

Compromises can be shown to be generally feasible and worthwhile, because allocations and losses are insensitive to moderate changes in the weights I_g. Changing the relative weights by ratios of 2 or even 5 is less drastic than by the infinite ratios implied by I_g of 1 or 0 in the "combined optima"; insensitivity to weights is common in statistics, e.g. in regression.

Compromise solutions are applied in Table 9.5.1 to two numerical illustrations of allocations of sample sizes between domains. It is assumed for simplicity that S_g^2, D_g^2, and c_g are constant between domains; this holds approximately and on the average for many variables. In part A the two domain sizes stand in the ratio $W_1/W_2 = 0.8/0.2 = 4$. For the overall mean the optimal allocation would be $\propto mW_i$, proportional to the W_i, (thus $f_1 = f_2$)

yielding the optimal $\Sigma W_i \bar{y}_i = 1$, but incurring the increase to 1.56 for domains and comparisons (line 1). These are all in relative terms $1 + L(n_i)$. On the other hand, equal allocation ($m_1 = m_2 = m/2$) incurs $\Sigma W_i \bar{y}_i = 1.36$, but yields the optimal 1 for domains and comparisons (line 2). The optimal compromise for these cases is $m_i \propto \sqrt{(W_i^2 + 1/H^2)}$ and the increases are reduced drastically to only 1.116 and 1.080 respectively (line 4).

A more spectacular case of 133 countries, ranging from 0.2 million to over 100 millions is investigated in B. This 500 fold range in relative sizes is closer to the ranges (50 or 100) often found among provinces of countries. The loss ratios, of 6.86 for separate means and 3.34 for the combined mean (on lines 1 and 2), are shown (on line 4) to be reduced to 1.28 and 1.31 respectively by the compromise allocation proportional to $\sqrt{(W_i^2 + 1/H^2)}$. Lines 5, 6 and 7 show how well the compromise works even with different I_g [For more details see Kish 1987, 7.3; 1976; 1988].

9.6 FEASIBILITY AND PRACTICALITY

It is difficult to write about practicality without becoming banal, or to generalize about feasibility and yet be relevant in specific situations. Nevertheless some brief reminders may help some readers to learn from the mistakes of others rather than from their own mistakes [Kish 1965, 8.4; 1977].

a. *Simple designs*, when feasible, should be preferred over complex designs. *Element sampling*, when good lists and low costs of collection allow, facilitates simpler analysis than cluster sampling. *Complete clusters* from a single stage selection is simpler than multistage, when element sampling is not feasible. *EPSEM*, equal overall probabilities f for all N elements, allows *self–weighting* analyses, without the complexities of weighting. When constant f for all elements must be abandoned, try next for *simple ratios* such as ixf or f/i, where the i are integers devoting simple ratios of selection probabilities, which simplify weighting or imputation. Reasons for these guidelines appear in

appropriate sections. However, we must emphasize that simplicity is most needed in the field procedures, less in office routines, and least in the work of skilled statisticians.

Simple, visible boundaries for segments are stressed in area sampling. Nevertheless, recognizing these from sketches, maps or aerial photos becomes "simple" only relatively and only after adequate training and experience.

b. *Sturdiness or robustness* may be other words for simplicity above, but they are meant to emphasize lack of sensitivity to moderate deviations in actually achieved conditions from those envisioned in the design. Thus, sturdiness is useful especially for multipurpose designs, plus others that arise only later, during the analysis. To the suggestions under simplicity we may add paired selections of PSU's and simple replicated samples.

Different selection procedures may be used in different domains to better fit them to local situations and resources (although the measurements must be standardized). For example, for selecting dwellings in the central city element sampling (PRES) may be used, whereas cluster (multistage?) sampling may be needed in remote rural areas. Nevertheless, in general a *single procedure* for a team of enumerators at one collection time may be safer than several, each of which may be better in separate domains. Perhaps two or three procedures may be taught to a well trained field team.

c. *Practical field instructions* are essential tools for making the actual sample to approximate the sample design. Procedures should be simpler, and instructions more extensive and detailed, and the training more thorough, for field enumerators who are not expert, have little training or experience, and who work far removed from central control and guidance of experts. On the other hand less training time and shorter instructions may suffice for experts and also for enumerators working close to expert guidance.

Instructions cannot be perfect or complete. *Maximize the amount learned* rather than the amount "taught" (presented); the longest, most complete books of procedures do not lead to best performance, within training periods limited by budget. Train them to handle well the 95 or 99 percent of ordinary cases and to know enough about the exceptions to consult the instructions, or the supervisor, or the central office. "In writing instructions use the imperative tense; it is clearer. Put the instructions in simple outline form, not in long narratives. Underline or capitalize the essential points. Provide headings and stubs that facilitate rapid references to the needed sections. Write for the field workers and not for fellow statisticians. In providing ideas, arrangement, and language, put yourself in the field workers' place and see them with their eyes. Perhaps ask one of them to help rewrite the instructions. If practicable, pretest your instructions; or borrow them from good sources. Whenever feasible, give the purpose of each step in simple terms, to motivate the interviewer and to provide needed flexibility through understanding" [Kish 1965, 8.4B].

Complex field procedures may often be *divided into separated tasks*, so that the field worker can concentrate on each in turn. Specifically, field listing (of dwellings, holdings) may be separated form the main, intensive interviewing. Those two tasks may be done by either the same enumerators or by two separate, specialized crews. The increased costs of those two separations must both be taken into account.

You may also include *negative instructions about what should not be done* in spite of temptations for the naive. Example were shown for the four frame problems (4.2). We also end here with two negative instructions about instructions. Do not just send out interviewers with the instruction: "Go and find a random sample!" That violates all rules, being neither simple, nor clear, nor adequate. Also don't imitate the statisticians who said:'"I wrote a perfect set of instructions, but they were widely misunderstood by the ignorant enumerators." Instructions are not good or bad in the abstract, but only in relation to the available human and material resources.

d. *Random numbers for field work* need careful preparation unless a) statisticians, or skilled professionals select in the field, or b) all selection takes place under close supervision, in the office. Both of these procedures can often be too costly, but we must warn that selections from tables of random numbers by enumerators in the field are difficult to trust and check; and some biased results have occurred from unjustified trust.

Random numbers should be selected in the office and sent to the field in a simple form, easy to supervise and check. Sometimes it may be feasible to establish a fixed order for listing the units that leaves little room for doubt or personal choice and is easy to check. Otherwise conceal the selection numbers until after the list has been prepared, so that the selections will be "blind" (10.5). Selection numbers may be concealed behind tapes, or they can be sealed in envelopes, kept closed until the listing is completed.

Systematic selection is easier to apply and to check than random choice. After the random start and the selection interval are removed from the sealed envelope, the enumerator can be trained and trusted to apply them properly — usually, but this too must be checked.

e. *Fix sampling rates f, not sample sizes n.* Easier selection procedures, with systematic sampling above, is one reason for the preference. But others have also been noted elsewhere in connection with unknown population sizes (7.7) [Kish 1977].

f. *Pretest all procedures*: in the field: this is desirable and even necessary, unless the procedures have been used before in similar situations and with similar field staff. This need is widely recognized for questionnaires, but sampling procedures may also need them if they are new or changed. Furthermore, a double need may also arise. The statisticians in the office may need field experience with new procedures. However, that experience cannot substitute for the perception and application of those procedures by the nonprofessional field enumerators, who will actually collect the data.

g. *Sequential control of sample size* often may not prove practical or economical in practice. It may seem desirable to be flexible with sample sizes in the face of unknowns during the design stage: size of the population N, the variance factors S^2 and D^2, the unit cost c, the rates of nonresponse and noncoverage.

Sample size should not be controlled by limiting the response rates, because every reasonable effort should be made to keep this as high as possible. Cutting off responses will often result in bias, because early responses do not constitute random samples. Also, it is usually not economical to randomize the order of field collection.

Sometimes it may be possible to design a maximal sample size, subselect a proper (random) sample of minimal size, and from the difference, n(max) − n(min), send out a proper (random) *supplemental sample*. However, the two samples will represent separate times and efforts. This may be undesirable or too costly.

CHAPTER 10. AREA SAMPLING

10.1 AREAL FRAMES FOR DWELLINGS, HOLDINGS, PLOTS

Area sampling is commonly applied in many situations because it uses practical frames and listing procedures for dwellings, holdings (farms) and parcels in agricultural surveys; also for other populations, e.g., stores, traffic, trash, pollution, noise, rocks, bacteria, etc. Populations of dwellings serve directly for variables like home gardens, rooms, furnaces, toilets, kitchens, stoves, TV sets, etc. But, more often other populations can also be associated with dwellings: holders, families, persons (adults, children), domesticated animals (cats, dogs), flies, private autos. Dwellings have also served often to identify agricultural holdings.

Area sampling can also be used for sampling directly holdings, farms, and plots, also the crops, orchards and timber that grow on them; also the domesticated animals and poultry located in barns, stables, pens, and fields on them; and fisheries in ponds and tanks; even wild flora (mushrooms, herbs) on them. However wild fauna (deer, rabbits) cause problems because of their mobility.

Area sampling depends on "unique, adequate, easy" identification of specified populations with small area segments of the earth's surface; the (" ") quotations around the above adjectives refer to common imperfections in those relative terms, which must be restrained. However, such imperfections can overwhelm the utility of area sampling in some types of situations. a) *Three dimensional mobility* can be difficult to overcome: fish and aquatic animals in oceans, lakes and streams, and flying birds and insects may be too difficult to identify with ordinary area sampling. b) *Extreme mobility* makes it difficult to cover wild animals (deer, wolves), nomads, homeless persons, and frequent travelers. c) *Panel studies* must overcome relative mobility over the longer periods. d) *Multiple identifications* can cause problems: two parcels of one holding, two homes, extended vacations, seasonal laborers who live both on distant farms and at home, are all examples of potential problems.

Farm holdings and operations are particularly subject to double or multiple identification. Some holders may have parcels and holdings in different segments, perhaps widely separated, even in different districts or provinces. Also two or more partners living apart may operate one holding (11.5)

Area sampling is based on frames for farms, holdings, dwellings and persons, which are relatively *convenient and effective for several reasons.* 1) With office mapping procedures the entire population of ultimate units (dwellings or holdings) can be readily identified with defined listings of blocks and segments. 2) These small areal units can usually be linked into an entire hierarchy of geographical/administrative units, which possess and provide identification, stability, and useful auxiliary data: E.D.'s, districts, provinces. Such data from administrative records and from censuses are useful for measures of size, stratification, etc. 3) The identifications persist from the listing time through the survey's collection period. 4) Field enumerators can identify "clearly and readily" block and segment boundaries and the units (dwellings, holdings) within them. 5) The units serve as a convenient link for sampling persons, because they are readily identifiable, relatively stable, and usually contain few persons, each of which can be identified uniquely with one and only one dwelling. Dwellings often serve as small clusters of persons, and similar unique identification can also be made for many other populations. For populations that lack these conditions area sampling is less useful. The following sections provide only brief guidance and further details appear widely scattered [Kish 1965, Ch. 9].

Area frames and sampling have two aspects: First the frame provided by the available hierarchical identification of administrative boundaries: e.g., a country divided into "provinces," these into "districts," these into "villages" or "E.D.'s" and so on; second, the field identification of the survey elements (dwellings, holdings, parcels, people) with area segments. Surveys commonly use both aspects, but they may use only one. For example, the administrative

frame may be applied to an available list of specialized holdings. But a valley containing a crop, for example, may be segmented from aerial photographs without use of administrative boundaries. Both aspects are commonly applied in area sampling.

10.2 PREPARING MAPS

In segmenting and numbering maps (or aerial photos) several necessary tasks must be done together. Just what those tasks are depends on resources and situations, because sometimes the maps and materials may already be prepared. For example the E.D.'s of a recent Census may be adequate, supplied, accepted, and used with their boundaries clearly marked and with measures of size provided also. Then perhaps the survey office needs only to prepare stratification and to update and to correct some measures of size. If the E.D.'s are too large, they may be segmented either in the office or in the field. Instead of E.D.'s, the Census may provide city "blocks," or political/administrative subdivisions. However, below we assume only basic situations where one must begin only with detailed, local maps that show many features, including probably farm houses. We hope that more up—to—date, and detailed aerial photos are also available to help to identify boundaries and to help in counting holdings.

Boundaries for the population area must first be defined and this seldom comprises a single, entire, contiguous surface. It may consist of many separated islands (e.g., in the Philippines, Indonesia). It may exclude many large lakes, inaccessible regions, military reservations. *It may also exclude areas* where the experts can confidently expect only negligible fraction of the survey population (holdings, or specific crops): urban areas, deserts, or wrong kind of soil or climate. In those areas great cost would be expanded for negligible yields and a "cut—off" exclusion can be used instead. These exclusions may be extensive and important in agricultural surveys; for some special crops most of the national territory or most of the population (e.g.,

urban) may be confidently excluded. The population area in the following is assumed to compromise a "primary enumeration area" to be covered by a team of enumerators, perhaps residing within its boundaries, but not necessarily. This area may be only a primary sampling unit selected from a larger population.

Segmenting the entire area into blocks or segments must concentrate on two primary tasks: defining good boundaries, but for small and roughly equal populations. This must be done quickly and with incomplete knowledge; those tasks pose difficulties. The terms "blocks" and "segments" are used here to help deal with difficult contradictions: blocks refer to areas with good, identifiable boundaries, even if the areas are larger and more unequal than desirable for final selection. Segments should be small and similar enough in size to serve as complete clusters for final selection, even if their boundaries need more care and skill for field identification. Thus it may be necessary to define only blocks in the office and leave the final segmenting to the field enumerators.

Boundaries should be lines rather than areas in the sense that they must not contain elements (dwellings, holdings). Thus, streets, roads, rivers and lakes make good boundaries if there they contain no elements; but not where people live on houseboats or on the streets. Where people live only in villages and cities even the open country between them can be regarded as empty boundaries (like the ocean).

Use existing identifiable "permanent" physical landmarks, either natural or artificial, for good landmarks: streets, roads, rivers, irrigation ditches, power lines. The enumerator cannot identify a long, arbitrary, imaginary line drawn on the map. However, the inhabitants of dwellings in many places know to what administrative area their own dwelling belongs. The boundaries must be drawn except where they are obvious.

Numbering and stratifying may be done at the same time when the stratification is geographical. Serial numbers within the blocks can accomplish three tasks simultaneously: identify them, establish their list, assign them measures of size (MOS), and stratify them. The numbering should anticipate the method of selection, especially when systematic selection follows. Selecting with the interval F selects a pair of numbers from every *implicit* stratum of 2F numbers. Furthermore, a serpentine order of numbering may yield some desirable stratification within that.

Measures of size and numbering may be assigned jointly if the measures (MOS) are simple and small integers, perhaps single digits, 1 to 9. If recently prepared, detailed maps and area photos identify the elements (dwellings and holdings) reasonably well, this may be fairly well accomplished. But not in other situations: where maps are too old or fail to mark the elements, when dwellings are crowded in villages or cities, etc.

In those situations assigning MOS must be separated and done after the blocks have been numbered and listed. The MOS may be assigned from records (census or administrative), or with field work. This will be costly, and some rapid methods, such as driving, cruising around the blocks, are often used.

10.3 COMPACT SEGMENTS VERSUS LISTED ELEMENTS

For subsampling elements (dwellings, holdings) within selected blocks, two alternative procedures are often used: compact segments or listed elements (10.4, and 10.5). The choice depends on eight factors of which the first three tend to favor compact segments, the next three listed elements (dwellings), and the last two may favor either. Consider these remarks as tentative, clearly variable between different local conditions, but useful as reminders of factors to be considered.

1. *Coverage* tends to be more complete with compact segments because listings, prepared hastily from the outside, tend to miss elements (dwellings, holdings), which are difficult to recover during interviews.

2. *Stability* favors compact segments, which continue to reflect changes within stable, identifiable boundaries, whereas listed dwellings or holdings change with time, perhaps even between listing and interviewing.

3. *Simplicity* tends to favor compact segments, where it is easier to train enumerators to cover completely the defined segments or holdings. To create segments in the field may require skilled training, but perhaps this may be done separately by specially trained teams.

4. The *homogeneity* (roh) of elements within compact segments may be greater than among listed elements selected around larger blocks. Thus for the same average size sample cluster $\bar{b}$, increases of the variance, denoted by $deft^2 = [1 + roh(\bar{b} - 1)]$, tend to be greater in compact segments with the greater roh. This factor becomes less important for crossclasses and their comparisons, because of smaller $\bar{b}$.

5. *Variations in sample size* with compact clusters has two sources. The segments vary in size due to searches for good boundaries and to imperfect maps. Even greater may be the effects of using random numbers of segments per block. On the other hand, a systematic selection can reduce the variation to single elements per block.

6. *Screening operations*, which can be done from the outside, may be less expensive if connected with the listing operations around the block. However, the operation and the comparison of costs become more complex if they involve interviews within the dwellings.

7. *Costs per element* may be less for complete segments, because preparing complete lists for entire blocks may be expensive. However, this comparison varies between situations. Listing costs are relatively less where sampling rates within blocks are large (shorter intervals) and where the listings get reused for several samples.

8. *Social interaction* among neighboring people *may* be greater within compact segments. Some surveyors fear higher refusal rates and also the "contamination" of responses. But clear evidence for those conjectures is lacking and in some situation the data collection may be facilitated by interaction: first, information from neighbors may help in planning callbacks and also in screening; second, favorable interaction may even help the response rates.

10.4 INSTRUCTIONS FOR COMPACT SEGMENTS

Three criteria should guide the creation of compact segments. Compromises among these criteria are needed, because they conflict in actual practice. Instructions are needed for these compromises for the workers in the office and in the field. With good maps and aerial photos, most of the segmenting may be done in the office, but otherwise and elsewhere, trained workers must divide the blocks or E.D.'s in the field. This should be done preferably before the main interviewing in order to facilitate selection in the office, but cost consideration may force combined operations for segmenting, selecting, and interviewing. But this would require skilled and trusted field workers.

1. Determine the *desired average size* of segments as a compromise between increased variance ($deft^2$) and decreased costs of larger segments. But remember that statistics for crossclasses suffer proportionately less from the variance increases (8.4). Consider also the possibly *lower coverage biases from larger segments* both because of clearer boundaries and because of lower ratios of marginal cases (11.3).

2. Create *segments of roughly equal size*, in numbers of dwellings, or other elements. This task is more difficult for holdings and other elements than for dwellings, when identified and counted on maps or from the outside. Size variations become even greater when a screening operation within households is needed to find subclasses.

3. *Use clear, identifiable, stable boundaries*: roads, streets, rivers, irrigation ditches, tree lines, utility lines. Sometimes arbitrary straight lines must be used between two well—fixed points, where lines are short and through thinly settled areas. If clear boundaries cannot be found, create segments of double or triple (or k—tuple) sizes; these may have to be listed and subsampled with rates of 1/2 or 1/3 (or 1/k). Thus, compact segments may need to be modified to occasional listing, especially in the presence of dense settlements and multi—dwelling buildings. These have been called "take—part" segments.

Most often the segments must be created in the field, then brought into the office for checking and for selecting the sample. Irregularities and mistakes are corrected; oversized segments and listings are divided; undersized segments may be combined (7.6). From the numbered lists of segments the sample segments can be selected with the within—block intervals, and sent out to the field for interviewing.

"On the other hand, if the travel to the block is relatively expensive, we may want to combine the three distinct tasks of segmenting, selecting, and interviewing (or at least the first call) into one step. This requires proper training of field workers for the three tasks, with emphasis on preventing unconscious biases in selection. The interviewers must assign a specified order when numbering the segments they create, or segment numbers selected in the office must be hidden — either in envelopes or behind black tape — and revealed only after the interviewers have assigned their numbers to the segments. The selection numbers must run comfortably beyond the expected

average number to allow for reasonable variations in numbers of segments found in the blocks. Extraordinary large blocks, however, are best reported to the office for checking and handling" [Kish 1965, 9.5A].

Control of sample size may be obtained by separating the two procedures. The first concerns assigning measures of size (MOS) to the segments together with clear boundaries. In the second, the segments are combined to form "pseudo—segments," not necessarily contiguous, so as to reduce variations in measures between them. Furthermore, the numbers of these created units may be exactly kF_b, so that exactly k units will be selected with the interval F_b, thus the rate $1/F_b$.

Quicker, cheaper procedures of segmenting are often needed, when the procedures above for blocks or E.D.'s are too expensive. a) *Villages* may be divided into preassigned number of "segments" along identifiable streets, on sketches prepared in the field. b) *Creating segments from buildings* (e.g., using floors of high rises) has been used and described [Kish 1965, 9.7]. c) *Segmented listings* from dwelling listings of blocks has been used also [Kish 1965, 9.5D]. d) *Creating segments from listed buildings* may be used when buildings may be readily listed without entering them, and only a sample of them needs to be segmented [Kish 1965, 9.7]. e) Creating "segments" from alphabetical and similar registers has been described earlier (7.3).

10.5 INSTRUCTIONS FOR LISTING BLOCKS OR E.D.'S

Instead of creating compact segments, a sample of blocks or E.D.'s or villages or other identifiable sampling units, may be assigned for complete listings of its dwellings or holdings or other elements. Here we refer to blocks, and assume clear and identifiable boundaries for them; also that the elements are dwellings that may be distinguished and identified (numbered or described) from the exterior, without entering them for interviews. For other elements, such as holdings or families or persons, a brief interview may be needed for screening and listing [Kish 1965, 9.6].

1. Begin listing at the place on the block marked with an X on the Sketch Sheet, usually the NW corner. *Proceed in the direction marked* with → on the Sketch Sheet; usually clockwise, so that the sample block is on your right side as you walk around it. You must cover the entire block inside the boundaries and nothing outside it. Be sure to check all roads, streets and alleys on the block for possible dwellings, going in and out of them and listing them as you come to them. Also look for dwellings away from the streets but inside the block. Please explain unusual locations, alleys and dwellings on the Sketch Sheet.

2. *List the address or description* of each dwelling on a separate line of the Listing Sheet, proceeding around the block in the specified order. If there are more dwellings on the block than lines on the Listing Sheet, use Continuation Listing Sheets. Watch for obscured dwellings behind shops or above stores, in a barn or garage in the yard, etc., and list them. Watch for dwellings away from streets, in the middle of your block. Look for and list multiple dwellings in houses, searching for clues: entrances, doorbells, mailboxes, etc.; sometimes inquire briefly.

3. Local and specific *instructions are needed for apartment houses, hotels, rooming houses*, trailer camps, irregular slum settlements, etc. In buildings use their systematic numberings of dwellings, if these exist. Otherwise: a) List bottom first and work up. b) List dwellings on right first then left. c) List front dwellings first, then the rear. d) The ground (street) floor is called "first," then the second, etc.; (this differs in some countries) e.g., a dwelling above a store would be "second floor above store." In case of doubt, include a sketch with explanation.

4. *Write descriptions of simple, visible, stable features* when street numbers and dwelling (apartment) numbers are lacking.

5. *List vacant dwellings, doubtful dwellings, and dwellings under construction.* Vacant addresses cause little harm, but missing dwellings is very harmful.

CHAPTER 11 SELECTION PROBLEMS AND METHODS

11.1 DUAL (MULTIPLE) FRAMES

"In sample survey methodology one often finds that a frame known to cover approximately *all* units in the population is one in which sampling is costly, while other frames (e.g., special lists of units) are available for cheaper sampling methods. However, the latter usually only covers an unknown or only approximately known fraction of the population...For example, the 1960 Survey of Agriculture of the Bureau of the Census uses two frames, mainly (A) a frame based on conventional 'area sampling' approach; (B) a frame of farms conceptually and operationally 'associated' with the A—1 listings of the last (1959) Census of Agriculture...the combined use of these frames proved a successful combination for simulating screening and providing coverage" [Hartley 1962].

"Some units would have a chance of entering the final sample through both the list frame and the more complete frame, while others could enter the sample only through the complete frame. The optimal allocation of sample size to the frames must usually be made in a situation where the frames cover different portions of the population, where they entail different data collection costs, where the element variance of the survey variables might differ across frames, and where the efficiencies of possible sample designs vary across frames" [Groves and Lepkowski 1985].

"The Sample Survey of Retail Stores" in the USA used lists for large retail stores, supplemented by area samples with lower sampling rates for the smaller stores [Hansen, Hurwitz and Madow, 12.A]. This survey also had two other features, described in detail: (a) it was confined to 68 PSU's to reduce costs of data collection; (b) it was essentially a multipurpose sample of 9 kinds of business groups, such as automobiles, food stores, and gasoline stations filling stations.

It is convenient to concentrate here on dual frames of a special list (L) plus an area frame (A). This is more common, but the method can be generalized to several frames; for example, to several lists (L_1, L_2, L_3...) plus another frame that is fairly complete, which is often an area frame, but not necessarily. Some alternatives about the frames and the samples may be answered by considering the four subsets: AL, the overlapping elements present in both frames; $A\bar{L}$, present in A but absent from L; $\bar{A}L$, present only in L; $\bar{A}\bar{L}$, absent from both frames. (a) From the L frame one may select a "complete" census L or only a sample *l*. (b) Similarly, either the entire A frame or only a sample *a* may be used. (c) The sample may either include the overlap *al* in the sample or exclude it and use only $a\bar{l} + \bar{a}l$.

The list L may have been obtained from a previous census or sample; or from a list of telephone owners; or from a widespread farm cooperative; etc. The formulas for efficient ("optimal") allocations among the three sets ($A\bar{L}$, $\bar{A}L$ and AL) would be too time consuming here and they may be found in references, but we must distinguish three different approaches to the stratum AL of possible duplications. (1) They may be excluded altogether. In supplements for nonresponse (b below) they would be excluded, of course. But large units selected with f_l may also be excluded, with some screening expense, from area samples of small units selected with rates f_a. (2) On the other hand, it may be less expensive to permit the units from the L sample, selected with f_l from the list, to also appear in the A sample selected with f_a. Thus the combined probability of the L units becomes $f_l + f_a = f_l(1 + f_a/f_l)$, with $f_l > f_a$. For example, if $f_l = 0.2$ and $f_a = 0.01$ then $f_l + f_a = 0.2(1 + .05) = 0.21$; and this may be more practical than trying to screen out the L units from the A frame. (3) The above allows for duplicate selections of the same units from both frames; a rare event when both ratios are much less than 1, and independent. This problem can be easily treated by duplicating the data for these rare events. But they can be eliminated from the sample, and the units of the L strata are then selected with probability

$f_l + f_a - f_a f_l$, assuming independence, and appropriately weighted in the estimates. In the above example this would be .2 + .01 − .002 = .208 for the L units.

We concentrated here on the use of an area sample for the A frame for being more complete, but also more expensive per element for the list frame L. However, the method also resembles other combined samples treated elsehwere. (a) For example, *l* may be responses on a telephone or mail surveys, and *a* may refer to efforts to find nonresponses or the nonresponses plus noncoverage of telephones (15.4). (b) Or *l* may represent an imperfect frame and *a* the efforts to supplement it (11.2). (c) *L* may refer to a census and *a* to the post—enumeration—survey (PES) to correct for responses and perhaps also for noncoverage (17.4). (d) *l* may represent a survey of crop yields, and *a* some smaller sample from the *l* sample of more expensive crop cuttings to calibrate *l*, which may or may not attempt a better coverage (12.3). (e) The chief purpose may be to estimate those $\overline{AL}$ missing from sample data by using two "independent" frames *a* plus *l* and their overlap *al* (11.9). This use of the dual frames, often called the C—D or *ChandraSekar—Deming method*, estimates the quadrant missing from both frames with $\overline{AL} = A\overline{L} * \overline{A}L/AL$, using sample estimates of the three observed quadrants and assuming independence between the two frames [El—Khorazaty et all, 1977]. (f) Similarly, independence of dual captures is often assumed in *capture—recapture* methods for estimating fish and other mobile populations [Darroch 1958]. Attempts have been made for both of these methods to modify the independence assumptions and to apply them to human populations. (g) Some survey population may be defined as the sum (union) of elements on any of several *overlapping lists* but without the replications. For example, farmers who appear on the lists of any of H farm cooperatives may define a combined farm population; or "social scientists" may be defined as members of one of five national associations. Replicates can be removed with sampling methods used in some efficient order [Kish 1965, 11.2D). (h) *Replicate listings* in different forms have been covered earlier (4.4). (i) *Multiplicity sampling* is the name

used for techniques to increase probabilities of selection by defining several identifying informants for rare populations. For example, annual new births may be reported not only by the mother, but also by her sisters (or siblings) [Sirken 1970].

11.2 SUPPLEMENTS FOR THE MISSED, NEW, UNUSUAL

This topic is related to dual frames, but with a more specific remedial view. It is also treated briefly earlier (4.2), and then later under Post Enumeration Surveys as adjuncts to censuses (17.4); there are also many different sources of reference [Madow et al 1983; Kish 1965, 2.7A, 9.4C, 11.5, 12.6C, 13.3]. The scale and nature of the remedial procedures to be used should depend on several aspects, each of which presents alternatives. (a) The principal aim may be either to supplement the principal sample, or to measure the portion missed, or to improve future surveys. The PES attached to censuses aim chiefly at measuring errors, whereas dual frames and this section aim at supplementing the main sample. (b) Both here and in dual frames we aim chiefly at cumulating (aggregating) cases rather than combining the final statistics from the two samples. (c) In some situations the survey procedures in the supplement may be similar to procedures in the principal sample; but in others some distinct procedures may be more appropriate for the supplement, for reasons of either cost or feasibility. Perhaps even a separate team of enumerators may be used for the supplement. Also mail or telephone interviews in the main frame, and doorstep interviews in the supplement can be combined. (d) Simultaneous collection may be used for both samples, but often the supplement is collected later than the main sample. (e) The main sampling rate may be used also for the supplement, but often a smaller rate may be set for supplements. (f) The supplemental samples usually exclude those in the main frame; whereas in dual frame designs the second (area) frame may include those on the list frame.

We must assume that the main frame has covered most elements, and that the supplement tries to cover a small but nonnegligible noncoverage p_c, say perhaps $.02 < p_c < .20$, of the population. If $p_c > 0.20$ the main procedures need improving; but if $p_c < .02$, it may be futile or too expensive for most surveys to search for them; but those arbitrary limits are only illustrative.

Four important matters related to supplements must be mentioned, although they are treated elsewhere. (a) Nonresponses can also be treated with techniques that resemble supplements (15.4); noncoverage by the main frame is the objective of the supplements in this section. (b) "Overcoverage" may also occur in some situations (15.3). (c) "Surprises" from sampling units that are discovered in the field to be "too large" cause problems that may also be treated with a stratum or supplement [Kish 1965, 13.4].

(d) *Linkage procedures* may be used to supplement listed units (e.g. dwellings) with units that have been missed or newly added (built). These procedures, also called "half–open intervals," specify that in addition to the selected units, the intervals (spaces) following selected units *up to but not including the next listed* units be searched for adding any units thus discovered [Kish 1965, 9.6I). Thus, the selection probability of the ith listed unit f_i, is also given to any unit found between unit–i and unit (i + 1), and the missed, and new units would be added at the time of enumeration, without the need for special supplements. For the procedures to work well, three conditions are needed that are lacking in many practical situations. (1) The missing or new elements must appear alone or in small numbers, and not as large multiple dwellings, for example. (2) The units (dwellings) should be located in linear order, so that "next" and "up to" have definite field identification, and not spread out in two (or more) dimensions. (3) The enumerators must remain alert to discover those missing units, which appear rarely, perhaps only once or twice in a hundred; but enumerators are often too absorbed in other tasks. For

all these three reasons the linking procedures have disappointed some who used them without actual field trials; but they seemed to have worked well in other situations.

Procedures for constructing supplements must be suited to specific situations, which may vary widely, as shown by a few examples. (a) Samples of holders (peasants, farmers) from villages may be supplemented with (1) holders located in the open country between villages; (2) holders living in cities; (3) nomads; (4) special samples in deserts or jungles or distant islands. In all those places the procedures for the main frame may be unsuitable and the supplements need different procedures. (b) Samples of persons from dwellings may be supplemented with persons living in institutions, in prisons, in school and university dormitories, in the military [USCB 1978, Appendix H]. (c) Samples of dwellings may also need supplements with special procedures for mobile homes, trailer camps, boat houses. Some of these procedures may be too expensive for small samples, and the institutional population, for example, may have to be excluded from the survey population. (d) Samples of youth selected from schools may need supplements for students in special schools (private, religious, remedial) and for youth not in schools (sick, working, delinquent).

11.3 SIZES OF SELECTIONS OF BLOCKS AND SEGMENTS

When applying selection rates within PSU's (e.g. districts, cities, etc.) two kinds of random variations in size occur commonly: in the numbers of units selected and in the sizes of those units. By units we mean clusters like blocks or segments; where the elements (e.g. holdings or dwellings) are listed for entire PSU's the variations in size of the actual sample b_i around its expected size should be relatively small.

However if, for example, a selection rate of $f_i = 1/20$ is applied to 54 blocks, either 2 or 3 blocks will be selected, a large variation of 1 around the expected size $54/20 = 2.7$ blocks. Furthermore, the block sizes also vary in size, and techniques for dealing with these problems exist (7.4).

When designing PPS subsampling within blocks with the variable rates b^*/Mos_α(intervals of Mos_α/b^*) some difficulties are often encountered [Kish 1965, 7.5]. Let us suppose that the designed sample size is $b^* = 8$ dwellings from the blocks. First, some of the blocks may be undersized, *insufficient*, when sizes $Mos_\alpha < b^*$, (measures less than 8) are assigned to the block. In those blocks the sampling rate would be $b^*/Mos_\alpha > 1$, which are not feasible. One may choose from several alternative *remedies before selection*: create separate strata and procedures for all blocks with $Mos_\alpha < b^*$ and select *all* those blocks; or assign arbitrarily $Mos_\alpha = b^*$ for all block units with $Mos_\alpha \leq b^*$ and accept this added variation in size of subsamples (both methods increase the number of blocks in the sample); or link all these blocks with $Mos_\alpha < b^*$ with others before selection so as to create combined blocks, all with $Mos_\alpha \geq b^*$.

When these steps before selection are too cumbersome or for too few blocks, two remedies remain *after selection*: create linkings with an *unbiased* procedure after selection; or assign weights b^*/Mos_α to all B_i elements in the block (if block has 6 dwellings, either all dwellings get weights 8/6, or 4 dwellings get weights of 1 and a random 2 dwellings get weights of 2 for total estimation weight of 8) [Kish 1965, 7.5E]. Blocks with $Mos_\alpha = 0$ would be embarassing unless they can safely be excluded and not searched. Note also that the minimum block size may be set not at b^* but at kb^*, if the selected blocks must serve k samples of b^* each.

On the contrary, blocks that have measures much larger than the needed b^* (or kb^*) could require too much work in listing and a two—stage procedure may be introduced to reduce that work. Furthermore, there is no need to allow measures greater than the stratum size M_h, because with $Mos_\alpha = M_h$ the

block can be selected automatically as a "self—selecting stratum" and the two stages of selection can be compressed into one; that is, $(Mos_\alpha/M_h) \times b^*/Mos_\alpha = 1 \times b^*/M_h$ for those large blocks.

Different problems arise when blocks are discovered (after) selection to be too large; that is, selection with rates b^*/Mos_α would yield too large samples, because the actual size N_α is much larger than Mos_α. From periodic or repeated samples one may "borrow strength" by combining their surprises into a "surprise stratum" and select diluted samples from them [Kish 1965, 12.6C].

11.4 SELECTING ADULTS FROM DWELLINGS

Dwellings are commonly selected with EPSEM rates f, and if the population elements are uniquely identifiable with the dwellings they also receive the same probability; for example "head" or "homemaker" of the household. Furthermore, if the small cluster of elements are all included in the survey, they all also receive the same EPSEM f. For example, in labor force surveys information about all persons over 16 years (in some countries) in the household is obtained from one "responsible adult" (the observational unit). Similarly the homemaker may give information (about health, nutrition, education) for all children between specified ages. Most households contain either 0 or 1, seldom 2 (rarely more) women of child—bearing age, and fertility studies usually include both women in those few dwellings. Including all elements from small clusters (not only from households) seems generally practical and efficient if *any* one of these conditions hold: (1) Only a small portion of clusters have more than one element. (2) Information about all elements can be obtained simultaneously, reasonably cheaply, and "uncontaminated." (3) There is no large positive intraclass correlation between elements in the clusters. However, selecting all elements should be avoided when none of these conditions hold.

For many surveys of adults, a single adult from each household is commonly selected, especially for variables that are highly correlated and/or "contaminated" in the interview process. Such procedures depart from EPSEM, because the probability for adults becomes f/P_i, where the P_i are the numbers of adults in the households. However, these departures from EPSEM increase variances only little (by factors of 1.05 or 1.1), because for most dwellings $P_i = 2$, for most others $P_i = 1$ or 3, and seldom more for adults. (Selection of one person from all persons, children plus adults, is not reasonable.) It is important to operationalize simple and adequate field procedures of selection. Selection tables are available that have been used frequently, also for telephone samples, and also for other situations [Kish 1965, 11.3).

11.5 IDENTIFYING HOLDINGS, HOLDERS AND HOUSEHOLDS

Suppose that area segments are used for sampling farm holdings and to obtain interviews with the holders to collect data. A specific holder, for example, may be identified with one of the following: (a) a parcel located in a segment *a* that fell in the sample, (b) or another plot located in another segment *b* that is not in the sample, (c) or a household in a village, not located in either segment, (d) or in a city, where he is working or vacationing during the survey period.

The above is not an extreme case of the difficulties in identifying sample holdings in agricultural surveys. The first plot may be located only partly in the sample segment, and most of it may lie in nonsample segments. There may be two or more holders (brothers, father and son, partners) who live in separate households. The several plots of the holding may be located in different districts, even different states. The presence of the holder during the survey collection period is important for collecting the interview without prohibitive costs. These problems of observation and identification also occur in complete censuses, but many of them can be resolved by using the holders'

names for identification. In sample surveys they must be resolved with rules for identification, and they mostly can be, but only with good rules, care, and training.

Dissimilar situations lead to different methods for different surveys and in different countries, perhaps also in different provinces. *Household surveys* are often used for collecting data on food and agriculture, and especially in the presence of "integration of agricultural surveys in national household survey programmes" [FAO 1978b, Ch II and III]. Identification of area segments or dwelling listings with dwellings, dwellings with households, households with holders, and holders with farm holdings describes the chain whose links must be connected. Dwellings are needed because they can be identified by enumerators from the exterior, and *households define the occupants of dwellings*; empty dwellings are vacant, or unoccupied, or converted to other uses. Households identify farm holders, or homemakers for surveys of food consumption; but many (or most) households in the segments may have no holders, because they are non-farm households. Identifying all the holdings is the object of careful interviews with the holders. Households may be identified not only with area segments, but also with village listings and with listings in towns and cities. In some countries or provinces the villages may serve as preferred clusters of households. Sampling towns and cities may allow for the coverage of home gardens, fruit trees, and of small scale livestock farming (poultry, pigs, milk cows), but may be excluded in many countries.

Sampling land areas in the open country has led to publications on the advantages of open segments versus closed segments. *Closed segments* refers to methods for including in the sample only agricultural land and activities that are contained within the sample segments, but all of that. For example, in the example above only the crops and animals located on the portion of the plot within the sample segment a would be covered in the enumeration of that

segment. Closed segments may be convenient for observing growing crops. But in interview surveys it means that activities for holdings are divided into two or more portions that must be separately accounted.

Open segments methods identify the headquarters of holdings uniquely with only one segment. First, the household (dwelling) location of the holder within any segment associates the entire holding with that one segment, all its parts and parcels, whether they are in the sample or not. If no segment contains the holders household (because they live in a town, city or in a province outside the survey population), the location of stables, barns, and tool buildings locates the holding either in or out of the segment. If those do not exist in any segment, the northwestern corner (for example) of all the parcels in the holdings constitutes the unique identification. Thus all holdings are defined to belong to one and only one segment, whether in or out of the sample. The order of the rule for associations is aimed to facilitate identification and location in unique segments for the large majority of holdings.

11.6 REPEATED SELECTIONS FROM LISTINGS AND FRAMES

Chapter 16 discusses methods and designs for periodic surveys for some defined set of objectives. This section concentrates on problems and opportunities when a set of listings or a sampling frame is used for several surveys. Savings of costs and the increased opportunities are so great, both on the modest level of listings and on the greater scope of master frames and operations, that they more than overcome the problems and difficulties that must be solved from their repeated and joint uses.

At the simplest level consider a set of listings or segments for dwellings or holdings prepared for a sample of blocks; the sample of blocks may be only for a district or a city, or they may belong to a national sample. In the simplest case the listings or segments are prepared for k samples selected with the same sampling rate f for k periodic surveys spread over 1 or 2 or a few years. The sample can be first designed for a rate of kf and then divided into k

parts of f, each of which benefits from the greater care and better spread of the larger sample, yet carries only 1/k of the costs of preparation. The attained sample sizes of each of the k samples should represent the changes and growth of the k periods. Area segments reflect those changes and they can last for a few years, during which boundary changes will be few. On the other hand, listings of dwellings or farms should be prepared only for a year or two, and also supplemented for the new or missed. A well described example is available from Current Population Surveys of the USA, with listings prepared for 8 rotations [USCB 1978].

A set of listings or segments may also be used for different selection rates that are not foreseen at the time of their preparation, but this situation requires more care. If an EPSEM selection f_s is removed from a set of EPSEM listings f_l, the residual is also EPSEM, $f_r = f_l - f_s$, from which future samples can be removed. It may be even easier to select from the entire original listing of f_l; if a new selection hits a selected line, merely substitute the next one, and so on, because selected lines and successor lines are both EPSEM. It is important that the removed sample be EPSEM, because if an unequal sample were removed, it would leave remainders biased in the reverse direction [Kish 1965, 9.4D].

Unequal selection probabilities for clusters introduce problems for repeated selection. For example, if blocks are chosen with PPS, blocks with smaller Mos_α get exhausted first, leaving unselected, unused listings in the larger Mos_α. Before discarding these unused listings in favor of an entire new selection, there are several alternative remedies. First, one may resample and reinterview some households in the smaller blocks, especially after a few years. Second, we may use an unbiased procedure for forming larger "sufficient size" combined blocks (11.3). Third, this procedure may be even better when used with foresight to form combined blocks for listing, say, kb* instead of only b*

as the needed subsample size. Another problem with PPS and other unequal probabilities concerns the selection from strata of two or more units with PPS and without replacement; this was touched upon and referenced before (7.4).

Sufficient size PSU's may appear as another form of the same basic problem of "sufficient size": in a small city (district, county) a survey organization in a few years (5 or 10) may have visited most of the dwellings. The possible solutions resemble those above, but on a different scale perhaps. For example, revisiting dwellings after 10 years may not be troublesome, especially since a good proportion would have different occupants. Nevertheless, the formation of combined units can yield a different solution.

11.7 CONTINUING AND INTEGRATED SURVEY ORGANIZATIONS

The combined listing for several samples described above may be called a *master sample*, but that term is also used more broadly. As we move to broader levels perhaps *master frame* may be more appropriate, although the two terms have not been defined and distinguished. Even broader and higher levels may refer to *integrated survey operations* [UNSO 1980]. Some of the issues and advantages have been briefly noted earlier (9.2) and discussed under "continuing operations" [Kish 1965, 10.4, 12.6]. Descriptions of such integrated operations are illuminating [USCB 1978; Hess 1985; Murthy 1967, Ch 15, 16].

Master frames may serve integration and cooperation between separate survey organization. However, here we need to outline the several ways in which one integrated survey organization may undertake to serve the diverse needs for sample surveys of a nation, state or other community.

a) *Field staff of enumerators*: either a single permanent, staff may be used, or different teams may be hired for different subjects; e.g. agronomists for agriculture, women for fertility surveys, etc.

b) *The location of enumerators* may be "permanent" in the sampling areas, or travelling teams may be sent from the central office(s) to the selected points.

c) *One set of PSU's* may be used for all surveys, or different sets may be selected to fit better the diverging needs of different subjects. Samples of agriculture, industry and labor force may need different PSU's as we note below. This conflict also links with the possible need for different types for teams of enumerators.

d) *Scope and nature of populations and subjects* covered by the survey organization are closely connected to the above. These in turn must influence the *measurement methods* used for data collection.

e) *Sampling methods* are related to the above and to the *frames* and other resources available to the survey organization.

f) *Methods of analysis* can differ widely: even more than statistical estimation, the depth and variety of substantive analysis must depend greatly on the experts (economists, demographers, agronomists, etc.) within and around the survey organization.

Retaining PSU's for changed needs. The best measures of size and the best variables for stratification may be rather dissimilar for different subjects, such as agriculture and labor force or total population. If separate designs are used to "optimize" for each, the separate sets of PSU's would need separate sets of enumerators and preparatory work. These problems are closely related to *methods for retaining units after changed strata and probabilities.*

The needs for changed designs based on new data from decennial censuses resembles the needs for different designs for different subjects. "After the initial selection the units may be used for many surveys over several years. But as time passes, the needs of new surveys may be better served by new strata and new probabilities, based on new data, than those used for the initial selection. The difference between initial and new data may be due to changes

in survey objectives and populations; for example, a sample initially designed for households and persons may be later required to serve a survey of farmers, or college students. *Obviously, our methods are also applicable to designing simultaneously a related group of samples with differing objectives*" [Kish and Scott 1971; Kish 1965, 12.7]. Those methods allow for using the "best" measures, for size and for strata, separately for each sample purpose, and also for maximizing the retention of the overlap between sampling units (PSU's) between the samples for separate purposes. As an alternative it would be possible to design a compromise that would average the measures in order to achieve a complete overlap of all units, but sacrificing some efficiency for both purposes (9.3). A compromise between those two may be even better than either: increase the overlap with some small sacrifices of separate efficiencies by recognizing only differences of measures that surpass some arbitrary minimal criteria.

CHAPTER 12. ESTIMATION, WEIGHTING, ANALYSIS

12.1 STATISTICAL ESTIMATION AND ANALYSIS

That sample design combines selection and estimation is often stated in the sampling literature and those statements have some validity. For example, neither a selection of n cases nor the estimator $\hat{Y} = Ny/n$ is either biased or unbiased by itself, but only in combination: $\hat{Y} = \Sigma y_j/p_j$ is unbiased if the elements j are selected with probabilities p_j, and $\hat{Y} = Ny/n$ for EPSEM selections only. Much of sampling theory concerns these simple expansion estimators, but in practical survey work they are seldom used; most survey analyses use more complex estimators. It would be difficult to draw a sharp boundary between estimation and statistical analysis; but it is also difficult and unwise to separate statistical analysis from substantive analysis. Statistical and substantive analyses are jointly needed, even for differences of subclass means $(\bar{y}_c - \bar{y}_b)$, but even more for more complex analysis, such as multivariate regression analyses; also for presenting sampling errors.

However, the chain linking selection to estimation, then to statistical analysis and to substantive analysis would be too long. In most practical work the sampling statistician who selects the sample is not in control of the analysis (agricultural, economic, social, epidemiological, etc) of the survey statistics. Thus the separation of the functions and operations, both in time and in personnel, of the selection of the sample from the analysis is often desirable or even necessary. But that very separation should alert us to the need for relating the selection design to the aims of the survey analyses (Ch. 9).

A good deal of estimation depends on appropriate weighting for validity and for efficiency. Weighting is used not only to balance for unequal selection probabilities, but also to compensate for nonresponse, for noncoverage; also for improved ratio estimation. Weighting thus is closely connected to estimation and to analysis, hence the triple name for the joint aims of this chapter.

Statistical analysis fills most volumes on statistics and mathematical statistics and that vast subject cannot be condensed into books on sampling. (On the other hand, those volumes largely avoid the problems and methods of sample selection.) It would be even less feasible to touch on all the many aspects of substantive analysis used on survey data.

12.2 SIMPLE AND COMPLEX MEANS AND RATIOS

We may regard the sample sum $y = \Sigma y_j$, or its weighted form $y = \Sigma w_j y_j$, as the basic computing unit for most survey statistics. Other forms used for survey analysis are also sums of moments, especially the components of the covariance matrix: $\Sigma w_j y_j^2$, $\Sigma w_j y_j x_j$. The simplest functions involve only constant factors: $\hat{Y} = Fy = y/f$, simple expansion totals; or $\hat{Y} = Ny/n$, which may be ratio estimates (12.3.4); or $\bar{y} = y/n$ simple means; here F,N and n are all fixed constants. But those simple forms seldom suffice for analyses and we shall begin with the ratio means of elements:

$$\bar{y} = y/n = \Sigma y_j/\Sigma c_j, \text{ or}$$

$$\bar{y} = \Sigma w_j y_j/\Sigma w_j c_j = \Sigma w_j y_j/\Sigma w_j, \qquad (12.2.1)$$

where $c_j = 1$, the simple count variable for elements and the w_j are the weights for the elements, which can be regarded as 1 (or 1/n) for "self–weighting" estimates. The denominators are random variables, because either the sample size n, or the weights, or both, are random variables.

Two modifications are now introduced that readily increase the generality already implicit in ratio means. First, let x denote the base of the ratio means, $r = y/x$, to signify that it stands for a generalized random variable: for example, the ratio of two survey variables, such as yield/acre, or weight/height, proteins/calories. Most often x represents a simple count n or a weighted count Σw_j of elements, but it is a random variable, not a fixed sample

size, because of nonresponses, unequal clusters, and variable subclass sizes. Second, let $y_j = w_j y_j'$ and $x_j = w_j x_j'$ represent already weighted variables, where needed, to simplify the formulas. Then the ratio mean is:

$$r = \frac{y}{x} = \frac{\Sigma_j y_j}{\Sigma_j x_j} = \frac{\Sigma_h y_h}{\Sigma_h x_h} = \frac{\Sigma_h \Sigma_i y_{hi}}{\Sigma_h \Sigma_i x_{hi}} = \frac{\Sigma_h (y_{ha} + y_{hb})}{\Sigma_h (x_{ha} + x_{hb})} . \qquad (12.2.2)$$

The r denotes the ratio y/x of two sample sums of element aggregates (weighted if necessary), where y and x also denote the sum of stratum totals y_h and x_h, each the sum of elements within the h—th stratum. These in turn represent the sums of totals for the ith primary selections, y_{hi} and x_{hi} (or "ultimate clusters") from the h—th stratum. Often there are only *two paired selections* (i = a,b) taken from each stratum h. These forms are needed for computing variances for ratio means based on stratified sums of clusters of variable sizes, $n_{hi} = x_{hi}$ (13.1).

Other functions of random variables may also be used; for example products yx, linear forms $\Sigma_i W_i y_i$, etc. A great variety of commonly used functions are based on ratio means, and the difference of pairs of ratio means is probably the most common:

$$r_1 - r_2 = y_1/x_1 - y_2/x_2 . \qquad (12.2.3)$$

Such differences may denote several kinds of comparisons, with different effects on variances. a) For comparisons of design classes (e.g., differences of crop yields between two provinces) the two samples are independent and $\text{var}(r_1 - r_2) = \text{var}(r_1) + \text{var}(r_2)$. b) For comparisons of crossclasses (e.g., crop yields for two age groups of holders) the two samples come from the same sampling units and $\text{var}(r_1 - r_2) =$

$\text{var}(r_1) + \text{var}(r_2) - 2\ \text{cov}(r_1,r_2)$. c) For comparisons of two time periods of similar sample bases, there also are covariances when the same sampling units are used for both occasions; perhaps high correlations if the same elements (e.g., households) are used, but lower if only the same PSU's are used.

Differences are simple forms of the linear combinations of ratio means, and $\Sigma_i k_i r_i$ can represent the more general form. For example, $\Sigma_t W_t r_t$ could denote a weighted mean of crop yields over several periods (t) of similar surveys. In addition to the sum of variances, there will also be covariances when using the same sampling units.

Ratios of ratio means (or double ratios) are used sometimes with $r_1/r_2 = (y_1/x_1)/(y_2/x_2)$; for example, the ratio of mean crop yields for two styles of farming may be used, instead of the difference $(r_1 - r_2)$ of the two means. Then those ratios of yields may also be compared for two periods $(r_1/r_2)_t - (r_1/r_2)_u$ for two periods (t,u). Furthermore, an index may represent the weighted sum of several double ratios: $\Sigma_i W_i r_{1i}/r_{2i}$, with i denoting different items in the index. Medians and quantiles are also used and their variances need special methods [Kish 1965, 12.9–12.11].

12.3 RATIO AND REGRESSION ESTIMATORS FOR MEANS AND TOTALS. POSTSTRATIFICATION.

In many situations we can find auxiliary or *ancillary data* for improving the estimates, and those data and estimates may have different forms. In basic, common form the sample statistic $\bar{y}$ may be improved with some available ancillary population value $\bar{X}$, together with the ancillary statistic $\bar{x}$, which is taken from the same source as $\bar{X}$ but confined to the same sample as $\bar{y}$. For example, $\hat{Y} = \bar{y}X/\bar{x}$ may estimate aggregate production of a crop based on a mean sample yield $\bar{y}$, multiplied by the total yield X from a census and the mean yield $\bar{x}$ from the same census, but only for the same sample base as $\bar{y}$. Any "bias" from the earlier census affects both X and $\bar{x}$ similarly, thus cancels out. The ratio $X/\bar{x}$ may be viewed as improving the estimator $\bar{y}$, especially because the simple expansion total y/f can have much higher variance (12.3.4). Or the ratio $\bar{y}/\bar{x}$ may be viewed as "calibrating" the earlier census total X, which may be obsolete or of low quality, or both, because:

$$\hat{Y} = (\bar{y}/\bar{x})X = (X/\bar{x})\bar{y} = (y/x)X = (X/x)y \ .$$

The sample pairs, $\bar{y}$ and $\bar{x}$ or y and x, have the same standardizing factors whether 1/n, or 1/f or 1.

In addition to reducing variances, ratio estimators also serve extensively as adjustments for reducing the biases of noncoverage and nonresponse. There they merge with methods known as post—stratification, noted later. This great flexibility of ratio means accounts for its wide use in practice, and should be remembered during comparisons with two other estimators that follow. We can construct for any constant k

the *difference mean* $(\bar{y}_{diff}) = \bar{y} + k(\bar{x} - \bar{x}) \ .$ (12.3.1)

For example, $\bar{x}$ and $\bar{x}$ may be population and sample means from a previous census, and $\bar{y}$ the mean from a sample survey, with $\bar{x}$ based on the same sample as $\bar{y}$. The adjustment factor may be simply k = 1. When for the adjustment factor k the linear regression coefficient B is used, k = B and

the *regression mean* $(\bar{y}_{reg}) = \bar{y} + B(\bar{x} - \bar{x}) \ .$ (12.3.2)

When for the adjustment factor the ratio of means is used, $k = \bar{y}/\bar{x}$ and

the *ratio mean* $(\bar{y}_r) = \bar{y} + (\bar{y}/\bar{x})(\bar{x} - \bar{x}) = (\bar{y}/\bar{x})\bar{x} \ .$ (12.3.3)

Comparisons of the variances produced by these three means are available [Cochran 1977, Ch. 7; Hansen, Hurwitz and Madow 1953, 11.2; Murthy 1967, Ch. 11; Kish 1965, 12.3B, 11.8]. Ratio means may be viewed as forcing the linear regression line to go through the origin, so that y = 0 when x = 0; whereas the computed regressions y = a + bx usually have $y = a$, positive more often than zero, for x = 0. The regression estimators are shown to have somewhat lower variances, but with assumptions of SRS and linearity. But the coefficients b_p may also be computed from multiple regressions $y = \Sigma b_p x_p$ from large complex samples, benefiting from several ancillary variables.

The above three estimators for means may also be used for estimating aggregates and the formulas are similar, but with estimates of totals in the place of means. The differences of efficiency among the three are complex and subtle, and ratio means are more often used, I believe, because of their simplicity, but often they all can be vastly superior to the simple unbiased expansion estimator y/f. Comparisons of the efficiencies of (Xy/x) and (y/f) may be stated most simply in terms of the relvariances C_y^2 and C_x^2 of the sample totals y and x:

$$\frac{\text{Var}(Xy/x)}{\text{Var}(Fy)} = \frac{C_y^2 + C_x^2 - 2R_{yx}C_yC_x}{C_y^2} = 1 + \{\frac{C_x^2}{C_y^2} - 2R_{yx}\frac{C_x}{C_y}\} . \quad (12.3.4)$$

Large gains occur when the bracketed quantity is large and negative, especially when $0.5 < C_x/C_y < 1.3$ and $R_{yx} > 0.7$ [Hansen, Hurwitz and Madow 1953, 4.19; Kish 1965, 6.5]. This occurs often in practice when the sample size x is highly variable and also a principal determinant of the total for y found by the sample. For example, two stage selection with constant factors $f = f_a \times f_b$ produces variable sample sizes n from unequal clusters; and ratio estimators Ny/n may be much more precise than y/f.

We next need to examine situations when the ancillary information X is subject to error; either to sampling error, as in two-phase sampling (12.4), or to possible biases, as in adjustments by weighting (12.5). Ratio adjustments are similar in form to ratio means, but they differ in fundamentals. In *poststratification* all three elements y_h, n_h and N_h come from the same population and the estimators of

$$\text{totals } \hat{Y}_w = \Sigma N_h \bar{y}_h \text{ or means } \hat{Y}_w = \Sigma W_h \bar{y}_h \quad (12.3.5)$$

serve to adjust these $\bar{y}_h = y_h/n_h$ for failure to select the sample cases n_h in proportion to the N_h: identification of the N_h population elements may have been unavailable, inconvenient, or ignored at the time of selection. Poststratification by N_h or $W_h = N_h/N$ of the sample means $\bar{y}_h = y_h/n_h$ in the strata will yield almost as low a variance for an SRS of size $n = \Sigma n_h$, as if it

were proportionately stratified. Furthermore these methods may be applied more broadly to other sampling methods, to other kinds of subclasses, and to other types of variables $\Sigma(y_c/x_c)X_c = \Sigma r_c X_c$. These resemble the adjustments by *reweighting* we discuss later (12.5).

The estimator $r = y/x = \Sigma\bar{y}_h/\Sigma\bar{y}_h$ and its expansions, Xr and $\bar{x}r$, are called *combined ratio* estimators, because they are *ratios of sums* of the $\bar{x}_h$ and $\bar{y}_h$. Each sum receives some stability from averaging over random variables. It is a basic, simple and relatively stable statistic that is used most frequently in survey practice. The *separate ratio* estimator, on the other hand, is seldom used in practice, but it appears in the literature because it can have much lower variance than $r = y/x$:

$$r_{sep} = \Sigma W_h r_h = \Sigma W_h y_h/x_h \,. \qquad (12.3.6)$$

This represents an *average of ratios* computed separately in strata and then combined into a weighted mean. It has several disadvantages: 1) Computing the ratios r_h for each stratum can be complicated for many statistics. 2) Reliable, unbiased weights W_h are seldom available for many strata for most statistics. 3) The separate ratios r_h can often be unstable (because of errors in the bases x_h) and the ratio bias of each stratum can add up to a considerable bias for the sum.

The bias of the ratio estimator is a vast subject in the literature of sampling, but it is seldom of great concern for the combined ratio estimators in practice. This technical bias occurs because the denominators x of $r = y/x$ are random variables. It is readily shown that the expected Bias of $(r) = E(r - R)$ can be stated in *relative* terms as:

$$\text{Bias ratio of } r = \text{Bias}(r)/\text{Ste}(r) = -R_{rx}C_x. \qquad (12.3.7)$$

The bias of r as a ratio of its standard error equals $-R_{rx}\,C_x$. Correlations between the ratio r and the denominator x probably exist often, but are probably small; $|R_{rx}|$ is probably much closer to 0 than to 1. Therefore, $|\text{Bias}(r)/\text{Ste}(r)| \leq C_x$, the coefficient of variation of x. It is not often feasible to measure R_{rx} but estimates of the approximation

$$\text{bias}(r)/r = [\text{var}(x)/x^2 - \text{cov}(y,x)/yx] \qquad (12.3.8)$$

have shown that the bias ratio is seldom important.

The coefficient of variation of x, cv(x) = ste(x)/x, serves a critical control on the validity of combined ratio means r = y/x. It is a useful and safe check on the bias of r; and also on their standard errors, ste(r), (13.1). Therefore, routine computation is recommended to check that cv(x) <0.2. This statistical advice fits well with common sense: the ratio y/x should not be used if x is unstable and highly variable, and this may also be viewed as a statistical refinement on mathematical taboos against dividing by zero [Cochran 1977, 6.8–6.12; Kish 1965, 6.6B].

12.4 TWO–PHASE SAMPLING, SCREENING CALIBRATION.

"It is sometimes convenient and economical to collect certain items of information on the whole of the units of a sample, and other items of information on only some of these units, these latter units being so chosen as to constitute a sub–sample of the units of the original sample. This may be termed two–phase sampling. Information collected at the second or sub–sampling phase may be collected at a later time, and in this event, information obtained on all the units of the first–phase sample may be utilized, if this appears advantageous, in the selection of the second–phase sample. Further phases may be added as required.

"It may be noted that in multi—phase sampling, the different phases of observation relate to sample units of the same type, while in multi—stage sampling, the sample units are of different types at different stages.

"An important application of multi—phase sampling is the use of the information obtained at the first—phase as supplementary information to provide more accurate estimates (by the method of regression or ratios), of the means, totals, etc., of variates obtained only in the second phase." [UN, 1950.]

This UN definition of two—phase sampling is generally accepted in survey sampling, where it is also called *double sampling*, though elsewhere those words may refer to sequential sampling. The method may also be extended to more phases in *multiphase* sampling, but discussions of two phases will suffice. The central concept involves selecting the second phase and basic sample of n elements not directly from the population of N elements, but from a larger first phase sample of n_L elements. Ancillary information, not available (and too costly to obtain) from the population of N elements, is obtained for the large sample of n_L elements, in order to improve the statistics based on the final sample of n elements.

In the two phases we have the total cost = $cn\ (1 + n_L c_L/cn)$ where c/c_L, the ratio of basic to ancillary information per element cost is a constraining factor on the utility of two phase selection, because large ratios (say c/c_L over 10 or 100) are needed to justify the use of the first phase. For example, where the correlation is 0.8, two—phase regression estimators can reduce the variance by 0.8 only when $c/c_L > 7$; and reduce it by 0.5 only when $c/c_L > 55$. In these comparisons the total cost is fixed, so that a two phase sample costing $(cn + c_L n_L)$ is compared with the cost of a one phase sample of cn_o, and $n_o/n = 1 + (n_L/n)(c_L/c)$. This reduction of the sample size prevents frequent use of two phase sampling [Kish 1965, 12.1–12.2; Cochran 1977, Ch. 12]. On the other hand, when c_L is very cheap, and c/c_L very large, the entire population of N may be used instead of the first phase n_L.

The basic theory of two phase sampling has been developed chiefly for SRS, and chiefly for application to proportionate sampling in the second phase, and to regression estimators. However, two phases of selection may be and has been applied more broadly, also to larger sampling units, and introduced into designs complicated with stratification and clustering. Some of these applications are listed below.

1) *Proportionate* stratified selection of the n elements based on information for the n_L elements. It is not likely that the gains of PRES would justify the cost of two—phase sampling.

2) *Disproportionate optimal allocation* in the second phase may be based on information from the large sample n_L in the first phase. This may be called a *screening* operation in some situations; for example, three different sampling fractions had been applied in the second phase to dwellings in three strata of dwellings, distinguished by socio—economic ratings, which were assigned in a first—wave screening operation [Kish 1965, 11.4].

3) *Calibration*, post—survey checks, quality checks are names given to *remeasurements* with better and more costly techniques for a subsample n selected from a larger sample n_L observed with less costly techniques. The chief aim here is better measurements, whereas the emphasis in two—phase sample is on improved selection procedures, but the two aims can be combined. Furthermore, here we usually consider the smaller subsample n as auxiliary to the larger basic sample n_L; whereas in two phase sampling the n_L is considered auxiliary to the first sample n. Crop—cutting measurements can be related to both and *crop yield* surveys can be related to areas under cultivation [Yates 1981, 7.14].

4) *Regression and ratio estimators* can be applied to two—phase sampling and to calibration measurements. Two phase regression estimators can be efficient only when the correlation of the two measurements and the cost ratio c/c_L are both very high, as noted earlier; this may occur with remeasurements of the same variable.

For these two phase estimators the variance component of the first phase must be added to the variance of the regression or ratio estimators (13.4) [Kish 1965, 11.8B, 12.2].

12.5 WEIGHTED ESTIMATES.

For self—weighting designs with EPSEM selections with the overall probability f for all elements, the weight can be $1/f = F$ for all sample elements to estimate totals $\hat{Y} = \Sigma y_j/f = Fy$ and 1/n to estimate means $\bar{y} = \Sigma y_j/n$. That simple uniform weight serves as a great convenience for self—weighting samples, though other weights *may* be introduced for nonresponses, or for adjustments if so desired. The estimators still remain simple if f/w is used instead of f, which is, thus decreased in the ratio w = 1/(coverage x responses). with w>1 to reflect less than perfect responses and coverage rates (15.3): this would be needed for expansions like $\hat{Y} = wy/f$, but not for the means $\bar{y}$. However, estimation becomes more complex when different weights are assigned to separate subclasses. *Several reasons and sources for weights should be distinguished because they have distinct effects on weighting.*

a) *Disproportionate sampling fractions* can be introduced deliberately, either to decrease variances or costs with "optimal" allocation among strata, or in order to produce larger samples for separate domains. These differences in the sampling rates (f_h) should be large (factors of 2 or 5 or even 100) and must be compensated by inverse weights in the sample estimates to avoid bad biases in the statistics (5.6).

b) *Inequalities in the selection frames* and procedures may create unequal selection probabilities, if not adjusted during selection. These may be serious if they affect more than a small portion of the sample; they can be corrected with weights inversely proportional to those selection probabilities, if these have been carefully obtained and maintained (4.2, 4.4).

c) *Differences in nonresponse* and noncoverage rates between parts of the sample can be balanced with unequal weights, inversely proportional to the response weights. Such variations between subclasses have different effects than a uniform nonresponse correction noted above; also they can have different effects on the overall mean, on subclass means, on their comparisons, and on analytical statistics (15.3). With small nonresponses, the differential corrections between the parts should also be small, and often not needed for small samples. They require knowledge about the sizes of nonresponse within defined parts; also assumptions based on past data and models are needed for those imperfect methods to compensate for inadequacies of field operations. Techniques for compensating for 1) *item nonresponse* may differ from techniques for 2) *total nonresponse*; also for 3) *noncoverage* and for 4) *deliberate exclusions* (15.3).

d) *Statistical adjustments* of sample data can be made in order to decrease variances, or to reduce biases, or for standardization to fit some model. Unequal weighting for different parts of the sample may be done with various techniques; and they all involve using auxiliary sources of data from outside the sample itself. 1) *Poststratification* corrects the sizes of strata with outside data to compensate for random variation, as for nonresponse, and especially for noncoverage. SRS selections of size n poststratified with the estimators $\hat{Y} = \Sigma N_h \bar{y}_h$ or $\Sigma (N_h/N)\bar{y}_h$ are practically as precise as a PRES selection of $n = \Sigma(N_h/N)n$ would be, as shown in sampling texts. In practice, however, adjustments for biases of nonresponse and especially noncoverage are the chief reasons for the widespread uses for poststratification and for ratio estimation [USCB, 1978, Ch. V]. 2) *Ratio estimators* Σ X*y*/*x* are more general forms for diverse variables X; the formula resembles poststratification, but differences have been noted before (12.3). 3) *Standardization* is often used by economists, demographers and others to adjust means and rates found in samples to some "standard population" that differs from the frame population. It is also used to remove "disturbing" base variables from subclass comparisons [Kish 1987, 4.5, 7.4]. 4) In *multiple classification* problems the cells of two or more dimensions

of classification can be adjusted (with least–squares?) to agree with unidimensional marginal totals [El–Badry and Stephan 1955]. These four examples illustrate possible adjustments of data; but these are not strictly sampling problems, because complete censuses can be similarly adjusted. 5) *Adjustments for biased selection* methods (quota samples, judgment samples) formally may resemble 1 or 2 above, but the common weakness of these procedures is that within adjustment classes the selections cannot be assumed to be "random" or unbiased. Those assumptions are often unstated, unjustified and misleading; "exchangeability" within classes is lacking (1.4, 1.8).

Procedures for weighting differ, because each has disadvantages, and their effects can differ between specific situations. 1) *Separate weights* w_j for each element, on the data tape and applied in all statistics, is the most practical and practiced procedure, increasingly feasible with modern computers. 2) *Weighted statistics* $\Sigma W_h \bar{y}_h$ with uniform self–weighting within classes may be preferred for a small number of strata and for only a few, simple statistics; for example, for two strata with selection rates $f_1 = 1$ for one and f_2 for the other. 3) *Random duplication* of sample elements may be used to prepare *self–weighting tapes*, which may be convenient for some situations and some statistics. With those replications the weights within classes are approximated. 4) *Elimination* of cases may be used instead of duplication, or combinations of duplication with elimination. The effects of these procedures are explored below (12.6).

Computing the weights w_j for all sample elements needs to be done only once. First compute F_j, the inverse of the selection probability through all stages: e.g., $1/F_j = f_j = f_{ha} \times f_{hab} \times f_{habc}$ for three stages of sampling units in the h–th stratum. Next, adjustments for nonresponse and noncoverage rates may be introduced, usually within reasonable subclasses c: $w'_{cj} = F_{cj}$ (response$_c$ x coverage$_c$)$^{-1}$; for example 0.95 response and 0.92 coverage yields $w'_{cj} = 1.144\ F_{cj}$. Next, further adjustments may also be introduced to produce

$w_{cj} = w'_{cj}$ x adjustments, either in the same or different subclasses [UNCB 1978, Ch. V]. Finally these weights may be standardized to any convenient proportion, so that $W_j = w_j/\Sigma_j w_j$.

12.6 EFFECTS OF WEIGHTING.

Weighting can have one or more of the following effects on statistics: 1) reduction of biases; 2) possible introduction of other biases; 3) reductions of the variance; 4) increases of the variance, 5) complications of computations and of statistical analyses.

Reducing potential biases, such as from nonresponse, and especially from noncoverage, are good reasons for weighting, but this can also introduce biases: e.g., when survey data are adjusted with census counts, there are also possibilities for introducing other biases (15.3). Reducing variances with poststratification and with ratio estimators are also discussed elsewhere (12.3).

Weighting introduces complications in computations and statistical analyses, despite beliefs that computers can now easily deal with them. "a) Machines seldom if ever make mistakes, but man—machine systems often do. In one situation many person—months were lost due to computing with inverted weights. A more painful example concerned false results that were analyzed, published and "explained" – and then retracted. b) Data tapes may be reanalyzed later by researchers without sure access either to good computers or to the reasons for the weights. The risks of mistakes increase with the separation in time and personnel from the collection and coding of data. c) Problems of weighting arise for complex statistics, such as multivariate analysis. Weighting appears "simple" only to minds fixed on simple aggregates $\hat{Y}$ or simple ratios $\hat{Y}/\hat{X}$. There exist both theoretical (inferential) and procedural problems for statistics like regression coefficients, or like "design effects," for which the value of n in σ^2/n creates problems. d)

Weighting involves the costs of accurate records to obtain, maintain, and properly use accurate weights. Cheap records may result in biases from improper weights" [Kish 1977].

Now we concentrate on increases in variances due to various techniques for weighting. The increases are viewed as due to weights given to randomly chosen elements, and as departures from equal, proportionate weights. This simple model is convenient and reasonably justified by experience. The contrasting problems of optimal allocations and conflicts with them are treated elsewhere (5.6, 9.5). The increases in variances due to "random" weights discussed below tend to be similar for means and totals for the entire sample, also for subclasses, for their comparisons, and for most statistics. Thus, for example, increases by 1.25 of the variance have effects similar to reducing n to $n/1.25 = 0.8n$, or by 20 percent of the sample.

Increases due to random weights may be stated most conveniently in terms of the frequency distribution of relative weights. *If proportions W_h (when $\Sigma W_h = 1$) of the population are given the relative weights k_h, variances are increased by the factor*

$$1 + L = (\Sigma\, W_h k_h)(\Sigma\, W_h/k_h). \qquad (12.6.1)$$

L denotes the "relative loss" over the minimal *1*, which obtains for uniform k_h. Note that the relative values of k_h cancel in the product. Furthermore, the inverse of the weights, proportional to selection probabilities, yields the same factor. Sometimes it is more convenient to deal with relative sample sizes, where the proportion n_h of the sample has weight k_h and $n_h k_h/n = W_h$. The relative increases in variances are

$$1 + L = \frac{n\Sigma\, n_h k_h^2}{(\Sigma\, n_h k_h)^2} = \frac{n\Sigma\, k_j^2}{(\Sigma\, k_j)^2}, \qquad (12.6.2)$$

where the last expression sums elements individually rather than in classes. We may also view the increase L due to weighting as the *relvariance* $s_k^2/\bar{k}^2$ of sample weights k_j, because $\Sigma\, k_j = n\bar{k}$ and

Table 12.6.1 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
$U_i \propto 1/k_i$	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, 1976.

[b]Dichotomous, $1 + L = 1 + U(1-U)(K-1)^2/K$. 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.

$$L = \frac{n\Sigma k_j^2}{\Sigma k_j^2} - 1 = \frac{(\frac{1}{n}\Sigma k_j^2 - \bar{k}^2)}{\bar{k}^2} = \frac{s_k^2}{\bar{k}^2}. \qquad (12.6.3)$$

Thus only small relative losses L are incurred for 1) small relative differences in weights k_j; or 2) small proportions n_h/n of the sample with very different weights, or 3) for samples mostly in the center of the range for the relative weights k_j. Table 12.6.1 illustrates these statements and can be useful.

Replication of cases may be used to replace element weights w_j in order to produce self–weighting sample data. The *relative* weights $k_h = (1 + W_h)$, where $1 \leq k_h \leq 2$ and $0 \leq W_h \leq 1$, for a set of n_h cases may be replaced by giving weights of 2 to a random selection of $W_h n_h$ cases and weights of 1 to the residual $(1-W_h)n_h$, so that the total weight will be $n_h[1(1-W_h) + 2W_h] = n_h(1 + W_h)$. If relative weights of $k_h = (k + W_h)$ are needed (where k is an integer), then $(1-W_h)$ should get weights of k and W_h get k + 1.

Duplications can be used to replace weighting for nonresponses because of the convenience of self–weighting "decks" of cases, and they are especially useful for *imputation for item nonresponses*. First, it would be difficult to assign different weights to the several items (variables) of the same case, depending on nonresponse rates for the items. Second, available valid responses on many items permit close matching of pairs of cases, so that instead of sets of cases ($n_h > 1$), single individuals ($n_h = 1$) on case–by–case bases can be used for imputations for the missed items.

Duplications of the fraction W of cases in set increases variances by the factor $1 + L=[1 + W(1 - W)/(1 + W)^2]$. The maximal factor is $1 + L = 1.125$ and it comes for $W = 1/3$. These increases due to duplication are additional to increases due to weighting, which are noted above.

Three extensions of duplication should also be noted.

1) *Imputation with case–by–case matching* may allow for less bias in duplicating for nonresponse. It may be particularly useful when 'good data are available for better matching, as there are for item nonresponses. This technique is frequently used in "hot–deck procedures."

2) Instead of duplicates for the portion W, it is possible to use *multiple replication* by selecting the random portion W for (k + 1) replications, whereas the residual portion (1 − W) receives only k replications. "Thus the unit variance increases by $W(1 - W)/(k + W)^2$, when the portion W receives (k + 1) replications and the portion (1 − W) receives k replications" [Kish 1965, 11.7B]. The maximal increase with k = 3 is reduced to 1/48, whereas with k = 1 for duplications it is 1/8. Thus with multiple replications it is possible to almost eliminate the increases in variances, additional to those for weighting.

3) *Elimination of cases* can also be used to produce self–weighting decks. For example, suppose that in a national EPSEM selection with f, a small domain (province, city, etc.) receives a much larger rate kf; it may be better to designate a subsample with kf/k = f, for this domain of the national sample, setting aside the residual (k − 1)f.

4) *Combination of duplication and elimination* may also be used for differences of subclasses. Elimination may be viewed as the case k = 0 in the formulas above, where the increase $W(1 - W)/(0 + W)^2 = (1 - W)/W$ may be viewed as e/(1 − e), where e = (1 − W) is the portion to be eliminated. When e is small this is not much greater than the increase $W(1 - W)/(1 + W)^2$ when W is small. Thus eliminating a *small* portion e is (surprisingly?) not much less efficient than duplicating it. This knowledge may be used for producing self–weighting decks for small differences in response rates for subclasses.

CHAPTER 13. COMPUTING VARIANCES FOR COMPLEX SAMPLES

13.1 VARIANCES FOR RATIO MEANS

Combined ratio means may be expressed in several useful and instructive forms:

$$r = \frac{y}{x} = \frac{\Sigma y_j}{\Sigma x_j} = \frac{\Sigma w_j y_j'}{\Sigma w_j x_j'} = \frac{\Sigma y_h}{\Sigma x_h} = \frac{\Sigma\Sigma y_{h\alpha}}{\Sigma\Sigma x_{h\alpha}} = \frac{\Sigma(y_{ha}+y_{hb})}{\Sigma(x_{ha}+x_{hb})} \qquad (13.1.1)$$

The mean r represents the ratio of two random variables y and x, each the sum of the n element values y_j and x_j. These may be weighted values $y_j = w_j y_j'$ and $x_j = w_j x_j'$. In surveys the x_j usually represent count variables, so that $x=n$ or $x=\Sigma w_j$. These may represent either the entire sample or only subclasses. Furthermore y often represents a dichotomy, a subset of the count n, and then $r=p$, a proportion (12.3). The first four expressions are used for computing r and the last three for computing variances. The stratum totals y_h and x_h, represent H independent sets of selections ($h=1,2...H$). Within each stratum the $y_{h\alpha}$ and $x_{h\alpha}$ denote independent primary selections ($\alpha=1,2...a_h$). The numbers of primary selections a_h may vary between strata: a_h may be 2 in one stratum, but 3 or 4 or more in another. For computing variances within the strata $a_h \geq 2$ are needed. The last term shows each stratum with exactly two primary selections, a and b. Such designs of *paired selections* from strata are often preferred, because they permit a) most stratification for fixed number of primary selections $a=2H$, and b) simple variance computations, as we shall see.

Variances for the ratio of two random variables $r=y/x$ can be computed from

$$\begin{aligned} \text{var}(r) &= x^{-2}[\text{var}(y)+r^2\text{var}(x)-2r\text{cov}(y,x)] \\ &= x^{-2}[\Sigma dy_h^2+r^2\Sigma dx_h^2-2r\Sigma dy_h dx_h] \\ &= x^{-2}\Sigma dz_h^2 . \end{aligned} \qquad (13.1.2)$$

The first line expresses the variance of the function r=y/x as a function of variances for simpler terms, y and x. This "propagation of variances" refers to an asymptotic method of approximate variances for functions of random variables in large samples; it is also called a "Taylor series approximation," or the "delta method," and it has been applied to other complex, multivariate functions (13.2). The series are intractable mathematically, but empirical results have shown that the approximations are good for large and even moderate sized samples. The convenient, useful, perhaps necessary control is to check that the coefficient of variation of the denominator x, for cv(x) = ste(x)/x <0.2. This caution is needed to guard against a large bias ratio for the estimator r; it is not too restrictive, but neglecting it can be dangerous. For example, suppose the variability of individual cluster sizes x_{hm} is $C_x = S_x/\bar{x} = 1$; and CV(x) = Ste(x)/x for 25 paired clusters will be $CV(x) = C_x/\sqrt{25} = 1/5 = 0.2$, just on the borderline. But if C_x can be reduced to C_x=0.1, with appropriate controls within strata and with PPS selections; then the CV(x)=0.1/5=0.02. This can become a problem especially for subclasses [Sukhatme and Sukhatme 1970; Kish 1965, 14.2].

When the denominator x is a fixed constant n the last two terms vanish and $var(y/n) = var(y)/n^2$. When the two variables x and y are independent the covariance disappears; these covariances arise in surveys because the y and x are based on the same sampling units. The relative size of this covariance term, $cov(y,x) = \rho_{yx}\sqrt{[var(y)var(x)]}$, depends on the correlation ρ_{yx} between the two variables; when these relations are high the covariances may drastically reduce the var(r). For example, the x_{hm} often measure highly variable sizes of sample clusters, and if the ratios y_{hm}/x_{hm} in clusters are rather evenly distributed then the correlations may be high. On the other hand, for rare items, for example, the y_{hm} may occur largely at random, with only low correlations with the x_{hm}.

The second line shows that the two variances and the covariance represent summations over the H strata of squares of the dy_h and dx_h terms, which denote variations within strata of the $y_{h\alpha}$ and $x_{h\alpha}$ terms. These terms, often called "*ultimate clusters*," are sums for the primary selections of the pair of variables: $y_{h\alpha}=\Sigma_j y_{h\alpha j}$ and $x_{h\alpha}=\Sigma_j x_{h\alpha j}$. The identifications of strata and of primary selections ($h\alpha$) are *necessary* for variance computations. For lack of these identifications on data tapes, many probability samples fall short of "measurability." They are also *sufficient*, because with those primary values ("ultimate clusters") the computations may ignore lower stages (secondary, tertiary) of selections. The computing units are the deviations dy_h and dx_h, defined as follows for both the general case with a_h selections and also for the special case with paired selections ($a_h=2$, $\alpha=a,b$). First compute the stratum sums $y_h=\Sigma y_{h\alpha}$ and $x_h=\Sigma x_{h\alpha}$ and then:

$$dy_h^2 = (a_h \Sigma y_{h\alpha}^2 - y_h^2)/(a_h-1) \text{ or } (y_{h\alpha}-y_{hb})^2$$

$$dx_h^2 = (a_h \Sigma x_{h\alpha}^2 - x_h^2)/(a_h-1) \text{ or } (x_{h\alpha}-x_{hb})^2$$

$$dy_h dx_h = (a_h \Sigma y_{h\alpha}x_{h\alpha} - y_h x_h)/(a_h-1) \text{ or } (y_{h\alpha}-y_{hb})(x_{h\alpha}-x_{hb}) . \qquad (13.1.3)$$

Alternatively, you can first compute $z_{h\alpha}=y_{h\alpha} - rx_{h\alpha}$ and $z_h=\Sigma z_{h\alpha}$ and then

$$dz_h^2 = (a_h \Sigma z_{h\alpha}^2 - z_h^2)/(a_h - 1) \text{ or } (z_{h\alpha} - z_{hb})^2 .$$

For paired selections (a and b) the computations are convenient, especially because the stratum sums need not be computed; convenient, but not necessary (as some believe). They are often used in variance computations, and the data can come from various selection designs (6.4): a) Two selections, perhaps with replacements, from each stratum; or b) single selections from pairs of half strata; c) collapsing of pairs of strata with single selections from each; d) systematic sampling of primary selections, with H/2 simulated half strata.

For systematic selections it is also possible to use all the (H−1) overlapping pairs (e.g. 1–2, 2–3, 3–4,...) instead of only the H/2 nonoverlapping pairs (1–2, 3–4,...). In that case the summations of (13.1.2) are over (H−1) terms; the paired computations can be used and $dz_h^2 = (z_1 - z_2)^2 + (z_2 - z_3)^2$, $+ (z_3 - z_4)^2 + \ldots$; and the factor $a/2(a - 1)$ multiplies the summation:

$$\text{var}(r) = x^{-2}[a/2(a - 1)]\Sigma\, dz_h^2 . \qquad (13.1.4)$$

The terms $z_{h\alpha} = y_{h\alpha} - rx_{h\alpha}$ express the *deviations from the expected values* $rx_{h\alpha}$, *which the actual values* $y_{h\alpha}$ would have if the mean y content of the primary selection were equal to the average r value; if the $y_{h\alpha}/x_{h\alpha} = \bar{y}_{h\alpha}$ were equal to r, then the $z_{h\alpha}$ would vanish. The magnitudes of $z_{h\alpha}$ measure the departures from the expectations of uniformity of means of sample clusters within strata. The $z_{h\alpha}$ terms make the computations easier to do by hand and to check because $\Sigma\Sigma\, z_{h\alpha} = \Sigma\, z_h = 0$. However, on computers this convenience is not important, and to have the three separate components of the variance may be revealing and useful. Particularly var(x) is needed for computing $cv^2(x) = \text{var}(x)/x^2$, which we need for the tests that $cv(x) < 0.2$.

The *finite population correction* (fpc) or $(1 - f)$ can be multiplied to each term, if it is appropriate and not negligible, for EPSEM with f the overall probability of selecting all elements. Similar procedures hold for $(1 - f_h)$ for EPSEM within the strata, but f_h different for the strata. If the f or f_h vary within strata and are not negligible, there may be difficulties with the single "ultimate cluster" computations.

Unstratified samples of a cluster are seldom actually selected, but they may be used as approximations or models. The formulas above can be used, of course, with a single stratum. However, a simpler formula may also be used, because without strata the sums y and x cancel out from (13.1.3) because $\Sigma y_\alpha - r\Sigma x_\alpha = y - rx = 0 = y^2 + r^2x^2 - 2ryx$, and we are left therefore with only:

$$\text{var}(r) = x^{-2}(\Sigma y_\alpha^2 + r^2 \Sigma x_\alpha^2 - 2r\Sigma y_\alpha x_\alpha) = x^{-2}\Sigma z_a^2 . \qquad (13.1.5)$$

Computing programs are increasingly better and more available, among them CLUSTERS from the International Statistical Institute.

13.2 SIMPLE VARIANCE PROCEDURES

Paired selections are denoted in Table 13.2.1 for 10 strata (col 1) with $\alpha = 1$ and 2(col 2). For the single ratio mean r only the 10 pairs of values for y and x (cols 3 and 4) are needed. The auxiliary columns for Dy and Dx (cols 7 and 8), and for z and Dz (cols 11 and 13) are helpful for three different procedures for the same results.

Another similar set of variables y' and x' is also given (cols 5 and 6), with similar auxiliary variables (cols 9, 10, 12, and 14). With the two sets of variables we can compute all the three terms of var(r − r') = var(r) + var(r') − 2 cov(r,r). These data represent a difference of two crossclass means from the same survey, but they could equally represent differences of means from two periodic surveys from the same primary selections. These and other differences are computed frequently in the analysis of survey data, where the effects of covariances need to be computed.

Data for the ratio means are found readily at the bottoms of the columns (3,4,5,6):

$$r - r' = \frac{y}{x} - \frac{y'}{x'} = \frac{149}{255} - \frac{77}{156} = 0.5843 - 0.4936 = 0.0907 = 9.07\text{x}10^{-2}.$$

The y variables denote proportions p−p' and these occur most frequently in survey data. They can be also expressed as percentages: 58.43−49.36 = 9.07 percent difference. We can also write 10^{-2} for percentages and for their standard errors and 10^{-4} for the variances. This notation helps to reduce the number of zeros we must write and the mistakes they tend to incur. Furthermore, for this size sample and this precision it may be enough and more reasonable to write the difference as 58−49 = 9 percent.

To compute the var(r) = 5.66 x 10^{-4} and ste (r) = 2.38 x 10^{-2}, from columns 3 and 4 we can use one of three procedures.

A. Compute the Dy_h and Dx_h (cols 7.8) and then

$$\text{var}(r) = x^{-2}[\Sigma Dy_h^2 + r^2 \Sigma Dx_h^2 - 2r\Sigma Dy_h Dx_h]$$

$$= (255)^2[217 + 0.584^2(475) - 0.584(586)]$$

$$= 5.66 \times 10^{-4} = (2.38 \times 10^{-2})^2 .$$

Check: From col 7, $\Sigma\, Dy_h = +7 = \Sigma y_h - \Sigma y_{h2}$, from the sums of first and second selections in col 3. Similarly for $\Sigma Dx_h = +15$ (cols 8,4). This procedure uses columns 7.8 only.

Table 13.2.1 For Computing Variances of Two Ratio Means and Their Differences [Kish 1965, 6.5]

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
h	x	y	x	y'	x'	Dy	Dx	Dy'	Dx'	z	z'	Dz	Dz'
1	1	11	19	5	12	2	3	−1	3	−0.102	−0.923	0.257	−2.481
1	2	9	16	6	9					−0.349	1.558		
2	1	8	10	1	1	2	0	−6	−12	2.157	0.506	2.000	−0.077
2	2	6	10	7	13					0.157	0.583		
3	1	6	13	2	10	−9	−7	−7	0	−1.506	−2.936	−4.910	−7.000
3	2	15	20	9	10					3.314	4.064		
4	1	13	23	7	12	8	15	3	6	−0.439	1.077	−0.765	0.039
4	2	5	8	4	6					0.326	1.038		
5	1	9	13	3	5	5	7	2	−1	1.406	0.532	0.910	2.494
5	2	4	6	1	6					0.496	−1.962		
6	1	8	10	6	13	−3	−3	4	9	−1.843	−0.417	−1.247	−0.443
6	2	7	13	2	4					−0.596	0.026		
7	1	5	7	3	6	−2	−3	0	2	0.910	0.038	−0.247	−0.988
7	2	7	10	3	4					1.157	1.026		
8	1	4	8	0	1	−1	−4	−4	−9	−0.674	−0.494	1.338	0.442
8	2	5	12	4	10					−2.012	−0.936		
9	1	9	12	2	13	0	−3	1	12	1.988	−4.417	1.752	−4.923
9	2	9	15	1	1					0.236	0.506		
10	1	9	20	10	18	5	10	9	16	−2.686	1.115	−0.843	1.102
10	2	4	10	1	2					−1.343	0.013		
Σ_{m_1} or $\Sigma_{m'}$		78	135	39	91	+22	+35	+19	+48	−0.881	−5.919	+6.247	+4.077
Σ_{m_2} or $\Sigma_{m''}$		71	120	38	65	−15	−20	−18	−22	+0.884	+5.916	−8.012	−15.912
Σ_m		149	255	77	156	+7	+15	+1	+26	+0.003	−0.003	−1.765	−11 835

B. Compute Dh_h and Dx_h (cols 7,8), then $Dz_h = Dy_h - rDx_h$ (col 13), then

$$\text{var}(r) = x^{-2}\Sigma Dz_h^2$$

$$= 255^{-2}(36.7724) = 5.66\text{x}10^{-4} = (2.38\text{x}10^{-2})^2.$$

Check: From col 13, $\Sigma Dz_h = -1.765 = \Sigma Dy_h - r\Sigma Dx_h$, from cols 7,8. This procedure uses columns 7,8.13.

C. Compute $z_{h\alpha} = y_{h\alpha} - rx_{h\alpha}$ (col 11), then $Dz_h = z_{h1} - z_{h2}$ (col 13), then

$$\text{var}(r) = x^{-2}\Sigma Dz_h^2$$

$$= 255^{-2}(36.7724) = 5.66\text{x}10^{-4}$$

Check: From col 11, $\Sigma\, z_h = 0.003 = 0$, except for rounding errors. This procedure uses columns 11,13.

The variance of r', denoted var(r'), is similar to var(r) and uses the other 5 columns (5,6,9.10,14). Its value is $\text{var}(r') = 36.25 \times 10^{-4} = (6.02 \times 10^{-2})^2$. Thus, if we use $t_p = 2.23$ for 0.95 confidence intervals with 10 degrees of freedom, we state $58.43 \pm 2.23(2.38)$ and $49.36 \pm 2.23(6.02)$ for r and r' for intervals in percentages.

For $\text{var}(r - r') = \text{var}(r) + \text{var}(r') - 2\text{cov}(r,r')$ we need the covariance also. This can be computed readily from the Dz_h and Dz'_h (cols 13,14), but care must be taken to distinguish the values with unlike (+ and −) signs, marked (*) in the column; then:

$$\text{var}(r - r') = x^{-2}\Sigma\, Dz_h^2 + x'^{-2}\,\Sigma\, Dz'^2_h - 2(xx')^{-1}\Sigma\, Dz_h Dz'_h$$

$$= \frac{36.7724}{255^2} + \frac{88.2080}{156^2} - \frac{55.3484}{255 \times 156}$$

$$= (5.66 + 36.25 - 13.91) \times 10^{-4} = 28.00 \times 10^{-4} = (5.29\text{x}10^{-2})^2.$$

The coefficients of variation of the denominators serve as useful checks on the stability of ratio estimators, needed for their approximations, and $\text{cv}(x) < 0.2$ has been used for upper limits. Here we have

$$\frac{\sqrt{var(x)}}{x} = \frac{\sqrt{475}}{255} = 0.086 \text{ and } \frac{\sqrt{var(x')}}{x'} = \frac{\sqrt{756}}{156} = 0.176$$

These meet the limit, but they are rather large and we can see reasons for their instability. The data are based on 10 differences only, and we may note that only 1 or 2 of these account for most of the $\Sigma\, Dx_h^2$ (cols 8,10) and the $\Sigma\, Dz_h^2$ (cols 13,14). This illustrates the value of having printouts of sample details even in this age of computers (14.3).

The design effects are easily computed when the means are proportions. The Σy_j^2 need not be computed for S_y^2, since we use $S_y^2/n = r(1 - r)/(n - 1) = y(n - y)/n^2(n - 1)$ for the SRS variances in $deft^2 = var(r)/var_{srs}(r)$.

$$var_{srs}(r) = \frac{149 \times 106}{255^2 \times 254} = 9.53 \times 10^{-4} \qquad deft^2(r) = \frac{5.66}{9.53} = 0.59 = 0.77^2$$

$$var_{srs}(r') = \frac{77 \times 79}{156^2 \times 155} = 16.02 \times 10^{-4} \qquad deft^2(r') = \frac{36.25}{16.02} = 2.26 = 1.50^2$$

$$var_{srs}(r - r') = (9.53 + 16.02) \times 10^{-4} = 25.55 \times 10^{-4}$$

$$deft^2(r - r') = \frac{28.00}{25.55} = 1.10 = 1.05^2$$

Note: a) These $deft^2$ values are unstable, with only 10 differences (degrees of freedom) for the denominators (14.2). The four–fold ratio in $deft^2(r)/deft^2(r')$ is unreasonable, and the $deft^2(r) < 1$ is most probably merely sampling variability. b) The $var(r - r')$ is reduced by covariance, but the SRS models assume independence. c) The deft values (0.77, 1.50, 1.05) are damped by $\sqrt{\ }$ and they compare directly the effects on the standard errors. d) Values of deft are most useful for checking for gross errors (14.1).

13.3 COEFFICIENTS OF VARIATION, VAR $(R_1 - R_2)$ AND $\overline{V}AR\ (R_1/R_2)$

Coefficients of variation cv(r) and relvariances $cv^2(r)$ are often useful for designing sample sizes and for comparing sample results (9.1). They also permit more concise formulas for ratio means and functions based on them, as we do below. They denote *relative* measures of standard errors and of variances, respectively:

$$C_x = S_x/\bar{x} \text{ and } C_x^2 = S_x^2/\bar{x}\,. \tag{13.3.1}$$

The variations are expressed relative to their means $\bar{x}$. The words "coefficients of variation" and "relvariances" are commonly used to refer (confusedly) both to elements, as above, and to sample means, as below:

$$CV(\bar{x}) = Ste(\bar{x})/\bar{x} = DC_x/\sqrt{x} \text{ and } CV(\bar{x})^2 = Var(\bar{x})/\bar{x}^2 = D^2C_x{}^2/x^2 \tag{13.3.2}$$

Here $D^2 = Deft^2$, the design effects that modify the variances of designs, and $Deft^2 = 1$ for SRS (5.4, 6.6). The standardization removes the units of measurements from C_x, because they affect S_x and $\bar{x}$ similarly. It also provides another kind of convenience with invariance to the constant factors n, N, and f in $\bar{x} = x/n$ and $\hat{X} = N\bar{x}$:

$$CV^2(\bar{x}) = \frac{Var(\bar{x})}{\bar{x}^2} = CV^2(\hat{X}) = \frac{Var\ (\hat{X})}{\hat{X}^2} = CV^2(x) = \frac{Var\ (x)}{E(x)^2}\,. \tag{13.3.3}$$

Instead of the population values above, we usually deal with computed estimates, denoted by $cv^2(x) = var\ (x)/x^2$ for example, where $x = \Sigma x_j$, the sample total. These estimates serve as checks, with $cv(x) < 0.2$, for proper uses of the ratio means y/x.

The various expressions of CV^2 are useful and sensible only to the degree that the denominators, chiefly $\bar{x}^2$ and $\bar{y}^2$ and their multiples, are safely positive. This is true for combined ratio means of positive quantities like areas of holdings, household possessions and other physical variables. But CV^2 will not be useful for differences and changes, which often vary around zero.

We may begin with the basic variances of the ratio means $r = y/x$:

$$\frac{\text{var}(r)}{r^2} = \frac{\text{var}(y)}{y^2} + \frac{\text{var}(x)}{x^2} - \frac{2\text{cov}(y,x)}{yx} \quad \text{or}$$

$$\text{cv}^2(r) = \text{cv}^2(y) + \text{cv}^2(x) - 2\text{cv}(y,x)\,. \qquad (13.3.4)$$

Sometimes the products yx of two random variables are used, for example crop areas x yields:

$$\frac{\text{var}\,(yx)}{(yx)^2} = \frac{\text{var}(y)}{y^2} + \frac{\text{var}(x)}{x^2} + \frac{2\text{cov}(y,x)}{yx} \quad \text{or}$$

$$\text{cv}^2(yx) = \text{cv}^2(y) + \text{cv}^2(x) + 2\text{cv}(yx)\,. \qquad (13.3.5)$$

Note the similarity of these expressions in relvariance terms to the variances in $\text{var}(y \pm x) = \text{var}(y) + \text{var}(x) \pm 2\text{cov}(y,x)$. We must be reminded that whereas the variances denote complete expressions, the relvariances refer to approximations from Taylor expansions (13.1), which depend on large samples, also on denominators that are safely positive and relatively stable, though random.

Computing units dz_h^2 have been developed and illustrated for variances of ratio means and for their differences (13.2):

$$\text{var}(r) = \Sigma dz_h^2/x^2 \text{ and } \text{var}(r_1 - r_2) = \Sigma(dz_h/x)_1^2 + \Sigma(dz_h/x)_2^2 - 2\Sigma(dz_h/x)_1(dz_h/x)_2 \qquad (13.3.6)$$

Differences of ratio means are frequently used, and sometimes other linear combinations also. We can apply these relvariance forms to some of the most useful functions [Kish 1965, 12.11]. All these functions utilize similar basic computing forms, which can be combined into larger forms.

$$\text{var}\,(r_1 - r_2) = \Sigma\,(dz_{h1}/x_1 - dz_{h2}/x_2)^2\,.$$

$$\text{var}\,(\Sigma\, r_g) = \Sigma_h[\,\Sigma_g dz_g/x_g]^2\,.$$

$$\text{var}\,(\Sigma\, W_g r_g) = \Sigma_h[\Sigma_g W_h dz_{hg}/x_g]^2\,. \qquad (13.3.7)$$

The last refers to a weighted index, which combines several ratio means from survey samples. Changes and differences of the index become extensions of this concept.

In addition to their linear combinations, the ratios of ratio means, "double ratios," are also used [Yates 1981, 10.5; Kish 1965, 12.11B; Deming 1960, 390–396]. For example, instead of comparing two subclass means with their difference $(r_1 - r_2)$, their ratio $R = r_1/r_2$ may be often used; e.g., the holdings, or crop yields, or fertilizer uses for two subclasses may be compared as ratios. Again we must assume that the r_2 are positive, large, and stable enough to be used in the denominators. The preceding methods can be used to show that:

$$\text{var}\,(R) = r_2^{-2}[\text{var}(r_1) + R^2\text{var}(r_2) - 2R\text{cov}(r_1,r_2)] \quad \text{or}$$

$$\text{cv}^2(R) = \text{cv}^2(r_1) + \text{cv}^2(r_2) - 2\text{cv}(r_1,r_2)]\,. \qquad (13.3.8)$$

13.4 VARIANCES FOR COMPLEX STATISTICS

Variances for complex functions can be well approximated with complex variance functions of simple sums by using Taylor approximations, similar to those for var(y/x) (13.1). "Propagation of variances refers to an asymptotic method of approximate variances for functions of random variables in large samples. Let the joint distribution of the statistics $\bar{y}_1\bar{y}_2\ldots,\bar{y}_k$ tend to the k−th variate normal form with mean values $\bar{y}_1$, $\bar{y}_2\ldots$, $\bar{y}_k$ and dispersion matrix V_{ij} (all finite). This means that the variables $(\bar{y}_1 - \bar{y}_1)$, $(\bar{y}_2 - \bar{y}_2)\ldots$, $(\bar{y}_k - \bar{y}_k)$ are in the limit distributed as a k−variate normal distribution with zero mean values and dispersion matrix V_{ij}. If $g(\bar{y}_1,\bar{y}_2\ldots,\bar{y}_k)$ is a continuous function with continuous first partial derivatives (not all simultaneously zero), then the variable

$$u = g(\bar{y}_1,\bar{y}_2,\ldots,\bar{y}_k) - g(\bar{y}_1,\bar{y}_2,\ldots,\bar{y}_k)$$

is distributed normally in the limit with zero mean and variance

$$\sum_i \sum_j V_{ij} \frac{\delta g}{\delta \bar{y}_i} \frac{\delta g}{\delta \bar{y}_j}. \qquad (13.4.1)$$

Here $V_{ij} = \text{Cov}\,(\bar{y}_i,\bar{y}_j)$ and $V_{ii} = \text{Var}(\bar{y}_i)$, terms in the k x k covariance matrix of the k variates. $\delta g/\delta \bar{y}_i$ represent partial derivatives of the function. Thus the variance of functions of variates can be expressed approximately and asymptotically in terms of the variances and covariances of the variates. Illustrations are given. Applications to complex functions that contain many covariance terms should be accompanied with investigation about the quality of the approximation" [Kish 1965, 14.2]. The examples in Table 13.4.1 may be useful.

Variances for medians, quartiles, deciles and other *quantiles* are frequently used in economic and social research. "*Median* designates a value Y_M of a variable such that

$Y_i < Y_M$ for half of the population elements. It is the most frequently desired quantile or percentile – which are general terms for position measures. Other familiar position measures are deciles and quartiles. The median is the second quartile and the fifth decile. Although this discussion centers on the median, it is applicable to other quantiles as well.

"We may want to estimate, for example, the yearly income which is not attained by half of the families. Interest in the median is most common for highly skewed distributions, when the median diverges considerably from the mean. In such distributions the variance may be less for the median than for the mean, because the latter is greatly influenced by large values on the far tail of the distribution.

Table 13.4.1 Examples of the Taylor (delta, linearization) method

g(y,x)	$\frac{\delta g}{\delta y}$	$\frac{\delta g}{\delta x}$	$(2)^2$	$(3)^2$	2(2)(3)	Assume E(y) = Y and E(x) = X Computable Variances
(1)	(2)	(3)	(4)	(5)	(6)	
$r = y/x$	$1/x$	r/x	$1/x^2$	r^2/x^2	$-2r/x^2$	$x^{-2}[\text{Var } y + R^2\text{Var } x - 2\text{ Cov}(y,x)]$
yx	x	y	x^2	y^2	$2yx$	$X^2\text{ Var } y + Y^2\text{ Var } x - 2\text{ YXCov}(y,x)$
ky^2	$2ky$	–	$4k^2y^2$			$4k^2\,Y^2\text{ Var } y$
			$k^2/4y$			$k^2\text{ Var } y/4Y$
$k\sqrt{y}$	$k/2\sqrt{y}$					
k/y	$-k/y^2$		k^2/y^4			$k^2\text{ Var } y/Y^4$

"The variance of a median can be computed conveniently and approximately by an indirect method, based on computing variances for proportions for complex samples. It consists of several steps, which are justified elsewhere" [Kish 1965, 12.9; Hanson, Hurwitz, Madow 1953, 10.17–18]. First, obtain a table and graph of the cumulated frequency distribution of sample cases and compute the sample median value y_m. Second, compute the standard error of a proportion p^*_m near the median values. Third, find the desired lower and upper limits p_l and p_u. Fourth, on the cumulated sample frequency graph find the y_l and y_u values that correspond to p_l and p_u.

Trichotomies and matched dichotomies are often used in surveys and they have similar variances, although their sources as variables appear very distinct. 1) Scale values 0, 1, 2, (or 1, 2, 3) may be assigned to *ranked variables* (low, medium, high), and the proportions p_0, p_1, and p_2 are measured, but the variances for mean scores are the same as for $(p_2 - p_0)$. 2) Differences between *two categories of a multinominal* may be measured by $d = (p_2 - p_0)$, disregarding all other categories. 3) *Before—after* measures on the same individuals may measure the net difference $d = (p_{10} - p_{01})$, disregarding the unchanged cases p_{00} and p_{11}, and we may conveniently consider them as $d = (p_2 - p_0)$ as above. 4) *Comparisons of dichotomies* for

two variables for the same individuals have characteristics similar to 3), because $d = (p_{10} - p_{01})$ disregards the p_{11} and p_{00} cases; e.g., comparisons of preference for a specified fertilizer versus a specified seed. 5) *Matched pairs of individuals* compared on a dichotomy can be analyzed similarly to the overlapping analysis of 3).

These very diverse cases have in common the correlations for individuals in the difference $(p_2 - p_0)$. The var $(y/x - y'/x')$ presented in (13.1) and (13.2) for complex samples include the correlations even when the x and x' cover the same individuals. It is only in the SRS formula used for $deft^2$ that the binomial formula $pq/(n - 1)$ must be changed to the $var_{srs}\ (p_2 - p_0) = [(p_2 + p_0) - (p_2 - p_0)^2]/(n - 1)$ [Kish 1965, 12.10].

13.5 REPEATED REPLICATIONS, RESAMPLING: JRR, BRR, BOOTSTRAP

These three names denote three distinct but related methods for computing sampling errors for complex statistics, alternatives to the Taylor approximations presented above. Each method relies on *replicates* comprising the sample of basic units; in complex surveys these are the primary selections (or ultimate clusters), e.g. 20 replicates in the sample of Table 13.2.1; in SRS the n elements are the replicates. Those replicates are combined into larger subsets to form *replications*; these replications comprise subsamples of the entire sample and they are selected in different ways for each method. Statistics (means, regressions, etc.) are computed from these replications, similar to the statistics computed from the entire sample for which the variances are wanted. For example, in Table 13.2.1, random selection of either a or b from each of H strata would yield two half-samples as two replications, and their differences would estimate the variation of the entire sample. But with only one degree of freedom it would be a very poor, unstable estimate. Therefore, *each of the methods has developed procedures for repeated replications* or "resampling" methods for obtaining more stability (degrees of freedom) for the variance estimates.

Jack—knife methods were the earliest of these procedures but they referred to selecting from SRS samples of size n, "pseudo—samples" of size (n − 1), selected repeatedly n times. This was adopted to complex samples with "*jack—knife repeated replications (JRR)* [Kish and Frankel 1974]. For 2H paired replicates from H strata, a random replicate *a* from each stratum is replaced by its paired replicate *b*. Thus the replication contains the entire sample of a+b replicates in H − 1 strata plus two *b* replicates and zero *a* replicates in one stratum. These jack—knife estimates $g(J_h)$ are repeated for all h strata and their variance $var_{JA}\{g(J)\}$ estimates the variance of the statistic g(S) based on the entire sample. Even better and simpler is the estimate $var_{JB}\{g(S)\} = \Sigma[g(J_h) - g(S)]^2$. And even better is the variance that *also* uses the complement replications CJ_h, which now discards the other replicates *b* and doubles the *a* replicates instead. Then $var_{HC}\{g(S)\} = \Sigma\,[g(J_h) - g\,(CJ_h)]^2/2$. These jack—knife computations are simpler and easier to learn than other methods of replication.

For example, consider the 20 replicates in 10 strata in Table 13.2.1, and compute $var_{JC}\{r\}$ for r=0.5843. We may compute $y+y_{1a}-y_{1b}=149+11-9 = 151$ and $x+x_{1a} - x_{1b} = 255+19 - 16=258$ and $J_1 = 151/258 = 0.5853$. Then $(J_1 - r) = 0.5853 - 0.5843 = 0.0010$. The ten values of $(J_h - r)10^2$ are {0.10,0.79,−1.98,−0.28,0.35,−0.49,−0.10,0.53,0.70,−0.32} and from these we may compute $var_{JB}\{r\} = \Sigma(J_h - r)^2 = 5.88x10^{-4}$. This compares well with $var(r) = 5.66x10^{-4}$ computed in (13.2). Furthermore we may also compute its complement $CJ_1 = (149-11+9)/(255-9+16) = 147/252 = 0.5833$. This $(J_1 - CJ_1) = 0.0020$ and the 10 values of $(J_h - CJ_h)x10^2$ are {0.20,1.57,−3.86,−0.60,0.72,−0.97,−0.20,1.04,1.38,−0.67}. From these we can compute $var_{JC}\{r\} = \Sigma(J_h - CJ_h)^2 = 5.67x10^{-4}$, very close to the $5.66x10^{-4}$ we first computed.

Half—sample replication leads to another method of repeated replication. It also can be illustrated Table 13.2.1 by constructing two half samples from the a and b selections respectively from each of the ten strata. Thus $\Sigma y_{ha}/\Sigma x_{ha}$

= 78/135 = 0.5778 and $\Sigma y_{hb}/\Sigma x_{hb}$ = 71/120 = 0.5917. Then d_i = (0.5778 − 0.5917)^2/4 = 0.4830 x 10^{-4} is one estimate of error, an example of how unstable one degree of freedom (or a few d. of f.) can be and another estimate from that table was 41.00 x 10^{-4} (14.2). There are 2^{H-1} ways of choosing the H pairs for the half samples, and their average would yield all the stability the sample can provide, but that effort would be too large, e.g. 2^9 = 512. However, *balanced repeated replications (BRR)* provide a method for balancing so as to obtain all the available stability from H strata from only a few more than H patterns.

Bootstrap is an attractive name for a newer technique of *resampling*, which has much in common with repeated replications, I believe. However, its development seems to be more theoretical and rather distinct from the first two. But no simple procedure seems to be available for ready applications to stratified, clustered survey samples.

JRR, BRR and Bootstrap each rely on distinct methods for repeated replications or repeated resampling from the entire sample base. JRR and BRR have been and can be used as alternatives to TAYLOR approximations and a relative appraisal seems appropriate [condensed from Kish and Frankel 1974].

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

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

3. "The BRR method was consistently the best when judged by the criterion we believe most significant: the closeness to expected P_t of the actual proportions of t(s) values. The BRR performed consistently better than

JRR, and JRR performed better than TAYLOR. The BRR's better performance is particularly noticeable for simple and partial correlation coefficients, where JRR and TAYLOR are less satisfactory.

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

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

Since 1974, programs have been written that make TAYLOR methods much more available, especially the SAS programs (Shah 1979) and SUPERCARP (Fuller 1975). In conclusion, there exist three methods that are increasingly available for either hand or machine computations. The CLUSTERS program from the ISI in the Hague is another. They are all satisfactory and much better in most situations than the evasions derided below (14.4). Choice among them may depend chiefly on convenience.

CHAPTER 14. GENERALIZED SAMPLING ERRORS

14.1 DESIGN EFFECTS: DEFT^2 AND ROH

According to the simple, clear view of "measurable" probability samples, the estimate $\bar{y}$ and its standard error ste($\bar{y}$) = $\sqrt{\text{var}(\bar{y})}$ can be both computed from the (large) sample itself, and statistical inferences can then be based on intervals like $\bar{y} \pm$ tp ste ($\bar{y}$), . However, inferences are more complex in most survey situations, aside from the problems of nonsampling errors and biases. The term "sampling errors" is often used in order to cover those broader needs of survey sampling: 1) *Mean—square errors* = MSE($\bar{y}$) = Var ($\bar{y}$) + Bias^2 ($\bar{y}$) is meant to include other sources of errors, in addition to sampling variances (Ch. 15). 2) *Coefficients of variation*, cv($\bar{y}$) = ste ($\bar{y}$)/$\bar{y}$ and the *relvariances* cv^2($\bar{y}$) = var ($\bar{y}$)/($\bar{y}$)^2 are useful for positive ($\bar{y}>1$) variables as *relative* measures of variability by removing units of measurement and especially for skewed variables (e.g., income, size of holdings). 3) The *cv(x) = ste (x)/x* of the denominators of ratio means r = y/x are useful as checks on their stability. 4) *Design effects*, deft^2 = var($\bar{y}$)/var$_0$($\bar{y}$), where the var$_0$($\bar{y}$) are the simple random variances s_y^2/n, are useful in several ways; they provide *relative* measures of variation by removing the effects of basic elemental variability s_y^2 and of sample size n. 5) The *ratios of homogeneity*, *roh* = (deft^2 − 1)/($\bar{b}$ − 1) are more portable than deft^2 by removing the effects of sample cluster sizes. 6) *Components of the variance* would provide theoretically sounder paths of importance, with separate components for each stage of selection and stratification. However, such computations are rare, because they would be too complicated and the residual results too unstable. 7))Averaging) of sampling errors (values of deft and roh) is frequently used for greater stability to facilitate generalizing and inferring; also for simplicity and brevity, and especially for greater stability (14.4). 8) *Tables of sampling errors* are often used for presenting averaged sampling errors. In research reports they may be the only feasible procedures to inform readers of the approximate values of sampling variability.

Several reasons for averaging, pooling, generalizing sampling errors need to be mentioned, because the standard statistical literature tends to neglect this topic. 1) *Too many statistics* are presented in most survey reports for separate computations of standard errors for all of them. With modern computers it may be feasible to compute ste ($\bar{y}$) for the *overall means* of all important survey variables. But this is less feasible for all the *subclass means* often presented in survey reports, and impossible for all *comparisons for subclasses*. Furthermore, analysis and presentation of those results to the readers of the reports must be condensed some way, often in tables. 2) *Difficulties* of computing valid estimates of sampling variation from complex samples for complicated analytical statistics may lead to conjectures from other statistics of the same survey and from models based on other experiences (14.2 and Table 3.4.1 cell C3). 3) *Unstable estimates of standard errors*, because of low "degrees of freedom," often result from too few primary selections for samples even when the number of elements n are large; also for design domains of most samples: "pooled" average deft values may be preferable (14.4). 4) *Designs for future samples* depend heavily on values of $deft^2$ and roh computed from "similar" surveys. "Borrowing" the wrong values will decrease the efficiencies for the borrowing designs; but they can be computed later unbiased estimates from their own results. 5) But "*borrowing*" *sampling errors from other surveys*, instead of computing them, can result in biased estimates. For example, $deft^2$ values are often borrowed, but these can be biased, because they depend on cluster sizes; and roh values are more often "portable."

External needs, like the last two, should be distinguished from *internal needs* for design effects. $Deft^2$ is good for conjectures internal to the same sample base, design, and size. But for conjectures to crossclasses of different sizes, and even more to other samples with different weights and selection rates, roh values are more portable.

Clustering, stratification and weighting are three features of sample selection with important effects on sampling errors. Variances are not suitable for pooling, because the units of measurement must be removed, and $deft^2$ and roh are more generally suitable than coefficients of variations cv^2. Note that both $deft^2$ and roh are designed to yield rough, simple, single statistics for complex designs. Each may cover several parameters, such as unequal cluster sizes (B_i and b_i in the population and sample), and diverse selection methods and stratification in several stages. However, neglecting to separate the variance components is usually less important than the differences between variables and between subclasses.

$Deft^2(\bar{y}) = Var(\bar{y})/Var_o^2(\bar{y})$ is in most situations relatively simple to understand, also to compute as $var(\bar{y})/var_o(\bar{y})$. The $var_o(\bar{y})$ denotes SRS variance; and for sample means ($\bar{y}$) of self–weighting (EPSEM) selections $var(\bar{y}_o) = s_y^2/n$; and $s_y^2/n = p(1 - p)/(n - 1)$ where $\bar{y} = p$ for proportions. We need not discuss here whether the factors $(1 - f)$ and $(n - 1)/n$ should or should not be used, because these are of relatively small magnitude and discussions can become technical. Formulas and computing programs often provide usable values of SRS variances for statistics other than means, such as for various coefficients of linear regressions (Table 14.3.1). It is useful to *compute and present deft($\bar{y}$) values for the overall means only for all survey variables*. They can yield useful *conjectures of $deft^2$ values for other statistics* as well (14.3).

In addition to its direct use for $\bar{y}$ itself, and its indirect use for conjectures, values of $deft^2$ are useful in several more ways. They serve practitioners as *checks against gross mistakes in computations* for variances; usually in cluster samples we should expect $1 < deft^2 < U_s$. Values just below 1 may be due to sampling fluctuations, but very low values should indicate mistakes. The upper limit U_s should be guessed by the sampler; excesses may

indicate mistakes in computations. Excessive values may indicate that the design was poor and should be changed in the future; but for that purpose roh values may be more useful than $deft^2$.

Subclass means usually require flexible interpretation of the values of $deft^2$, and they can be as important as the overall means. *For design subclasses* (e.g. provinces), *the $deft^2$ values are useful directly*. They can differ, depending on both population distributions and on sample designs, and in complex ways specific to situations. For example, urban areas may have different selection designs imposed on different population distributions than rural provinces. The overall $deft^2$ should be close to the weighted average of $deft^2$ values for all the design subclasses comprising the total.

Crossclasses on the other hand behave differently and the value $deft^2 = [1 + roh(\bar{b}_c - 1)]$ provides a preferred approach to these situations. This *ratio of homogeneity* is computed as $roh = (deft^2 - 1)/(\bar{b}_c - 1)$, where $\bar{b}_c = n_c/a$ is the mean sample cluster size for the subclasses of sample size n_c spread into a primary selections. Because the sizes of n_c and $\bar{b}_c = n_c/a$ can vary greatly for different crossclasses, the $deft^2$ will also differ greatly even for the same survey variable. But computing and presenting $deft_c^2$ for all crossclasses is often not feasible. On the other hand, often roh does not vary greatly, and values of $roh_t = (deft^2_t - 1)/(\bar{b}_t - 1)$ based on the entire sample are relatively stable, and "portable" and they can be used for crossclasses. From many empirical computations we find that $roh_c = 1.,2\ roh_t$ are good rough averages for imputing roh_c from roh_t.

Roh is a useful extension of rho $= \rho$, the correlation of intraclass correlation, defined as $Deft^2 = [1 + \rho\ (B - 1)]$ strictly only for equal clusters, all of size B. But the extension to unequal cluster sizes b_i averaging $\bar{b} = n/a$ has withstood many tests. It is useful for crossclasses, and it also is more "portable" than $deft^2$ to other sample designs for the same population; also to other, "similar" populations, but that with more care.

Weighting, however, has different effects, which were ignored in discussing roh above. When $deft^2$ and var $(\bar{y})$ include the weighting effects, they are useful indicators of the overall effects of design. However, the increases $(1 + L)$ due to "random weighting" remain relatively constant, and do not decrease for crossclasses as clustering effects do (12.6). Thus for small crossclasses the effects of $(1 + L)$ may be considerably greater than the effects of roh due to clustering which decrease with size. Therefore, the following roundabout computation is suggested for computing var $(\bar{y}_s)$ for crossclasses (s) or for another sample from var $(\bar{y})$ of an overall weighted mean (note the intermediate $deftu^2$): var$(\bar{y})$ $\rightarrow$ $deft^2(\bar{y})$ $\rightarrow$ $deftu^2 = deft^2(\bar{y})/(1 + L)$ $\rightarrow$ roh $= (deftu^2 - 1)/(\bar{b} - 1)$ $\rightarrow$ $deft^2_s = 1 + roh\ (\bar{b}_s - 1)$ $\rightarrow$ $deft^2 = deftu^2_s\ (1 + L)$ $\rightarrow$ var$/(\bar{y}_s)$.

The values of roh and $(1 + L)$ may be relatively stable for crossclasses within the same sample. But for inferences to other designs, to other weightings and to other populations, those values may need to be changed.

14.2 APPROXIMATIONS, CONJECTURES, MODELS

We may summarize here the status of sampling errors for different kinds of statistics, based on several earlier discussions: in Table 3.4.1, in Section 5.4 on stratified element sampling, in Section 6.6 on clustered sampling, in Chapter 13, and in Section 14.1

1. *For means and totals based on the entire sample*, several satisfactory methods for estimating sampling errors are available when samples are large enough (14.4). Computing them for all survey variables or for most variables is recommended, because their sampling errors (variances, $deft^2$ and roh) can differ greatly.

2. *For means and totals of subclasses* the same methods as for the entire sample are available. However, the practical feasibilities may differ because usually there are too many estimates for all survey variables to compute variances for the many types and categories of subclasses analyzed.

Furthermore, the subclass sample sizes may become so small that the variance estimates may become unstable, especially for design subclasses (14.4). For design domains the $deft^2$ tend to resemble on the average the $deft^2$ for the entire sample. On the other hand, for crossclasses values of the $deft^2$ tend to approach 1 with decreasing sample bases, as the roh values are more portable because they tend to remain roughly constant.

3. *For differences between subclass means* the methods used for means can be and have been extended. For crossclass means the values of $deft^2$ approach 1, often rapidly. Because the numbers of possible comparisons are often so large, the possibilities of computing and presenting sampling errors for all of them vanish and strategies of computing, averaging and presenting them must be devised (14.3).

Table 14.2.1 Deft = $\sqrt{\text{Deff}}$ for Standard Errors of Five Types of Statistics from Three Complex Samples [Kish and Frankel 1974]

Sample Set	A	B	C
Means	1.11	1.80	1.44
Simple correlation coefficients	1.10	1.26	1.36
Regression coefficients	1.02	1.30	1.11
Partial correlation coefficients	1.04	1.40	1.36
Multiple correlation coefficients	NA	1.46	1.89

4. *For linear combinations of means* generally, as for their differences specifically, the techniques used for means can be further extended (13.4).

5. *For complex analytical statistics* based on clustered samples (cell C3 in Table 3.4.1) the situation is more difficult but deserves a brief discussion. For some statistics, e.g., linear regressions, both "Taylor" and repeated replications have been used to compute sampling errors (13.5, 14.4). First, the results have shown considerable design effects and ignoring them would lead to serious over confidence in sample results; Table 14.2.1 shows values of deft > 1 for diverse coefficients for regressions from three distinct samples. Second, deft values within data sets are related and they tend to be somewhat less on the average than for means; thus deft values for means may serve as convenient

upper limits for deft values for coefficients. These conjectures and approximations carry some risks, but they can be considerable improvements over alternatives [Kish, 1987, 7.1]. For example, using the standard errors based on SRS (or IID) assumptions, which are often available on "canned" computing programs, can lead to underestimates of actual errors.

14.3 STRATEGIES FOR SAMPLING ERRORS

If surveys would produce only a few statistics, $\bar{y}_g$, computing and presenting ste $(\bar{y}_g)$ for all of them would present few difficulties. However, most surveys concern many statistics and for many variables. Furthermore, beyond variances and standard errors other expressions of sampling errors are also needed often. A brief review of useful rules follows.

1. *Identification codes for strata (h) and for primary selections* (α) within strata must be available for all elements (j) in the sample in order to permit computation of variances. These should be supplied early at selection time and should be available on the data tapes.

2. *Any weights* w_j used for the elements in estimation should be available and used also for variance computations.

3. *Compute var($\bar{y}_g$) for many survey variables (g)*, based on the entire sample. These overall variances are important survey results, and they also serve as bases for conjectures for other statistics, such as subclasses. Variances are relatively easy to compute with modern programs simultaneously for many overall means. The variances will differ greatly between variables, and even the derived sampling error functions can vary considerably. The values of ste $(\bar{y}_g) = \sqrt{\text{var}(\bar{y}_g)}$ can be computed jointly, as well as the functions below.

4. *Sampling error functions deft*$^2(\bar{y}_g)$ and deft $(\bar{y}_g)$ should also be computed; also roh $=$ (deft^2 $-$ 1)/($\bar{b}$ $-$ 1) where appropriate; also cv$^2(\bar{y}_g) =$ var$(\bar{y}_g)/\bar{y}_g{}^2$ where appropriate, and cv$(\bar{y}_g)$. Computations of the values for

$deft^2(\bar{y}) = var(\bar{y})/(s_y^2/n)$ are simple for proportions $\bar{y} = p$, when they are $p(1 - p)/(n - 1)$; otherwise $s_y^2 = (\Sigma y_j^2 - n\bar{y}^2)/(n - 1)$ must be computed. The synthetic roh has reasonable interpretation when $\bar{b} = n/a$ refers to the mean primary cluster size of n elements selected with EPSEM in a cluster. However, for considerable departures from EPSEM the corrections (1 + L) for weights should be used (14.1 and 12.6).

5. *Computing and checking* $cv(x) < 0.2$ is a useful caution against unstable ratio means, $r = y/x$, where x is the variable denominator; often $x = n$, the sample size; but this may be Σw_j, when weighted.

6. *For crossclass means roh* $= (deft^2 - 1)/(\bar{b} - 1)$ is more useful than $deft^2$. A tabular display of computed values for the overall means of ste, deft and roh for all variables is useful (Table 14.3.1 and 14.3.2). Columns may be added for values of $deft^2 = 1 + roh\ (n_c/a - 1)$ for several (3 or 4) selected values of crossclass sizes n_c.

7. *Tables of ste (p) and ste* $(p_a - p_b)$ can be presented for several values of p and of sample sizes n_a, and n_b [Kish 1965, 14.1]. See also [Gonzalez et al 1975; USCB 1978, Ch. VIII].

14.4 STABLE SAMPLING ERRORS

Unbiased estimates for variances get a great deal of attention in sampling literature, and we must avoid bad biases for all aspects of sampling errors; for example, SRS variances for complex surveys can cause harmful underestimates of true variability. However, another common problem that is difficult to treat also needs attention: the lack of stability, the high variability, of estimates of sampling errors, which is often due to too few primary selections (ultimate clusters) on which the computations of variance are based. Below are four types of common problems among others.

Table 14.3.1 Design Factors (deft) for the Total Sample – by Country and Variable

	Nepal	Mexico	Thailand	Indonesia	Colombia	Peru	Bangla-desh	Fiji	Sri Lanka	Guyana	Jamaica	Costa Rica
NUPTIALITY												
01. %, currently married	1·14	1·68	1·06	1·35	1·26	1·02	1·05	1·07	1·06	1·12	1·12	1·11
02. %, exposed to child-bearing	2·11	1·13	1·59	1·36	1·07	1·25	1·06	1·36	1·15	1·19	1·07	1·04
03. %, with marriage dissolved	—	1·41	0·98	1·44	1·47	1·09	1·13	1·21	1·06	1·31	1·34	1·21
04. %, remarried	—	1·12	1·49	1·51	1·24	1·12	1·39	1·23	1·19	1·09	1·01	1·25
05. number of marriages	—	1·36	1·30	1·79	1·69	1·10	1·41	1·47	1·15	1·13	1·17	1·48
06. age at marriage	2·49	1·76	1·28	1·54	1·35	1·05	1·48	1·65	1·27	1·16	1·06	1·02
07. time spent within marriage	—	1·25	—	1·14	1·09	1·04	1·04	0·97	1·15	1·13	1·09	0·99
FERTILITY												
08. %, pregnant	1·39	1·06	1·02	1·31	0·86	1·26	1·10	1·30	1·17	1·04	1·05	0·96
09. children ever-born	2·08	1·78	1·52	1·34	1·15	1·07	1·05	0·98	1·21	1·04	1·07	1·12
10. living children	1·96	1·78	1·29	1·43	1·11	1·12	1·00	0·94	1·18	1·07	1·08	1·12
11. births in first 5 years	3·02	1·44	0·98	1·38	0·88	0·87	1·24	1·05	1·29	1·00	1·13	1·10
12. births during past 5 years	2·74	1·58	—	1·54	1·31	1·04	1·03	1·00	1·16	1·01	0·99	1·04
13. first birth interval	1·38	1·22	1·16	1·44	1·25	1·14	1·20	—	1·26	1·06	1·06	1·02
14. last closed birth interval	1·41	1·32	1·43	1·43	1·04	1·02	1·17	0·89	1·07	0·97	0·99	1·01
15. open birth interval	2·05	1·61	1·37	1·56	1·14	0·92	1·20	—	1·11	1·01	1·03	1·01
16. months breast fed child	2·08	2·05	1·98	1·46	1·52	1·50	1·12	—	1·23	1·09	1·08	1·08
17. 16, excluding dead children	1·26	1·45	1·75	1·28	1·50	1·09	1·16	—	1·21	1·04	1·21	1·00
18. %, of children who died	1·70	2·08	1·30	1·34	1·63	1·81	1·08	1·11	1·20	0·94	1·07	1·09
FERTILITY PREFERENCES												
19. %, last pregnancies unwanted	—	—	—	1·44	1·28	1·34	1·10	1·63	1·15	1·28	1·00	1·01
20. %, wanting no more children	2·39	1·36	1·29	1·71	1·12	1·18	1·28	1·62	1·14	0·93	1·03	0·91
21. %, expressing boy-preference	2·13	1·06	1·15	1·34	1·06	0·97	1·23	0·94	1·04	1·07	1·01	1·03
22. %, exceeded desired family size	2·37	1·49	1·15	1·31	1·46	1·21	1·17	1·29	1·30	1·18	1·18	1·00
23. additional children wanted	3·16	1·96	1·62	1·62	1·09	1·42	1·23	1·52	1·09	1·03	1·10	0·93
24. desired family size	3·76	1·74	1·63	1·86	1·18	1·32	1·26	1·22	1·24	1·31	1·15	1·12
CONTRACEPTIVE KNOWLEDGE												
25. %, knowing pill	3·25	3·20	2·74	2·32	2·84	1·74	1·70	—	1·31	1·57	1·35	0·97
26. %, knowing IUD	2·95	—	—	2·58	2·30	.	1·71	—	1·39	1·37	1·23	—
27. %, knowing condom	2·44	2·50	1·94	2·60	2·38	1·74	1·42	1·66	1·47	1·61	1·25	1·17
28. %, knowing modern method	4·19	2·74	2·77	2·24	2·46	2·10	1·80	—	1·25	1·51	1·18	0·85
CONTRACEPTIVE USE												
29. %, ever used pill	—	1·90	2·05	1·95	1·82	1·26	1·30	1·32	1·25	1·37	1·14	1·06
30. %, ever used IUD	—	—	—	1·90	1·21	.	1·05	—	1·19	1·02	1·02	—
31. %, ever used condom	—	—	—	1·56	1·13	—	1·52	—	1·27	1·25	1·10	—
32. %, ever used any method	2·23	2·19	2·28	1·79	2·35	1·84	1·56	1·42	1·34	1·50	1·14	1·10
33. %, ever used modern method	—	2·08	2·34	1·82	1·93	1·33	1·54	1·44	1·42	1·27	1·09	1·00
34. %, used in open interval	—	1·96	1·78	1·70	2·05	1·46	1·38	—	1·34	1·20	1·06	1·10
35. %, used in closed interval	—	—	—	1·30	1·69	1·47	0·99	1·26	1·31	1·37	1·10	1·05
36. %, currently using any method	—	1·96	2·08	1·67	1·98	1·40	1·36	1·40	1·30	1·18	1·04	1·15
37. %, currently using modern method	—	1·83	2·35	1·72	1·49	1·06	1·22	1·29	1·33	1·19	1·00	1·05
38. 37, for women wanting no more children	—	1·50	2·19	1·52	1·34	1·06	1·20	1·02	1·26	1·07	1·00	1·03
SAMPLE SIZE (ever-married women)	5940	6255	3820	9136	3302	5640	6513	4928	6810	3616	2765	3935
No. of effective PSUs	40	182	70	376	405	410	240	100	606	196	410	288
Coefficient of variation of cluster size, CV_{bs}	0·062	0·031	0·062	0·021	0·035	0·025	0·025	0·043	0·022	0·036	0·041	0·026

Table 14.3.2 South Korea Fertility Survey [1973]
Sampling Errors for 39 Variables

		1	2	3	4	5	6
Variable* Number	Variable Description	Mean	Std. Error	Deft	Roh	Mean Subcl. Roh	Ratio 5/4
513	Sales, Clerk, Prof., Husband Occ.	0.40	.030	2.674	.146	.145	1.00
512	School 10 + Yrs., Husband	0.41	.025	2.203	.091	.070	0.77
509	Urban Background	0.27	.022	2.204	.091	.122	1.34
530	Rural Background	0.63	.024	2.155	.086	.110	1.28
510	School H. S. +, Wife	0.18	.018	2.119	.082	.089	1.08
511	Wife Currently Working	0.09	.013	1.939	.066	.107	1.63
232	Family Planning Worker Contact	0.25	.019	1.898	.062	.131	2.11
315	Can Plan No. of Children	0.86	.015	1.900	.061	.057	0.94
428	Abortion Costs < 3,000	0.64	.021	1.728	.059	.043	0.72
516	Rich Living Status	0.29	.018	1.758	.049	.085	1.74
601	Age at Marriage < 21 Years	0.52	.020	1.734	.047	.066	1.40
333	Want Another Son, Given Qnly One	0.37	.019	1.711	.045	.095	2.05
321	Ideal Number of Children	3.18	.037	1.665	.042	.052	1.23
538	No Work Experience	0.70	.017	1.639	.040	.078	1.96
225	Ever Used Birth Control	0.55	.018	1.595	.036	.035	0.96
231	Visited Health Center	0.14	.012	1.563	.034	.047	1.38
207	No. of Abortions (1963–73)	0.61	.049	1.546	.033	.046	1.40
320	Ideal Number of Sons	2.33	0.55	1.527	.032	.019	0.59
226	No. of Children at First Contraception	3.10	.058	1.256	.026	.097	3.73
602	Marriage Duration (Yrs.)	11.47	.252	1.411	.024	.009	0.38
214	Mass Media Tells of Contraception	0.40	.016	1.400	.023	.032	1.41
105	Number of Live Births	3.39	.063	1.381	.021	.012	0.58
104	Number of Living Children	3.14	.054	1.315	.017	.010	0.62
319	No. of Children Desired	3.73	.042	1.296	.016	.019	1.17
224	Ever Used Pill	0.21	.011	1.204	.011	.020	1.78
422	Wife Should Do Contraception	0.69	.013	1.204	.011	.016	1.50
237	Age at First Contraception	29.37	.162	1.099	.009	.011	1.18
429	Believe Abortion OK	0.79	.011	1.164	.008	.018	2.41
108	No. of Pregnancies (1963–73)	2.85	.052	1.132	.007	–.004	–0.57
334	Want a Son, Given No Sons	0.69	.012	1.125	.006	–.007	–1.10
227	Using Contraception Now	0.14	.009	1.094	.005	.011	2.13
635	Age at Marriage < 25	0.92	.007	1.063	.003	–.001	–0.50
236	Marriage – First Contraception (Yrs.)	8.75	.179	1.027	.002		
103	Pregnant in 1973	0.31	.012	1.044	.002	.010	5.07
106	No. of Miscarriages (1963–73)	0.16	.012	1.041	.002	.003	1.58
340	Husband Decides Fertility	0.32	.011	1.026	.001	–.004	–3.81
139	No. of Living Sons	1.62	.028	1.022	.001	–.014	
223	Ever Used Loop	0.18	.007	0.796	–.009	–.015	1.71
318	Want ≤ 2 Children	0.19	.006	0.721	–.011	.002	–0.21
	Mean over 39 Variables			1.471	.0327	.0444	1.15
	Ratio of Means Col. 5/Col. 4					1.358	

* The first digit of the Variable Number denotes: 1) Fertility Experience, 2) Contraceptive Practice, 3) Birth Preferences and Desires, 4) Attitudes, 5) Socio-economic Background, 6) Demographic Variables.

We are concerned chiefly with sample surveys, but these problems are similar to experimental designs, and may be even more common and severe there, and should be so recognized; also in observational studies, or controlled observations. First, let me clarify the common and important problem by describing its several specific manifestations. Then we can discuss several possible approaches.

1. *Few sites.* Sometimes entire national (or provincial) surveys must be confined to 10 districts or 10 schools. Furthermore, statistics for provinces, states and other geographic subclasses are often based on few PSU's (districts, counties) even for large national samples.

2. *National samples of 20 to 50 PSU's.* Even for national samples of this size the computations of variances are based typically on 10 to 25 pairs of selections, or "degrees of freedom." These also yield unstable estimates of sampling errors, especially for variances of comparisons, such as var($\bar{x} - \bar{y}$) for two subclasses or two periods.

3. *Interpenetrating samples.* These designs have been advocated with 4 replications for national samples, or ten replications of complex samples have often been designed and used in order to facilitate computations of the ste ($\bar{y}$).

4. *Repeated replications (RR).* These methods for computing standard errors are often called balanced repeated replications (BRR), "half—sample" or "pseudo"—replications, or jack—knife repeated replications (JRR). Sometimes they are based on 16 pairs of combined replications, or 16 "degrees of freedom" (13.5).

What can we do about these problems? A single sampling unit is useless, but from two PSU's one can compute unbiased estimates of variances. Some are willing to stop there, especially since there is no other number beyond two PSU's (and 1 d.f), that can be denoted as a clear boundary for "measurability." Also 2H PSU's in "paired selections" from H strata serve as

bases for many designs of survey samples. With H "large" or "not too small" these are useful, although it is naive to believe (as many do) that such paired selections are either necessary or sufficient for good survey design.

Here are some alternative ways "to pool" sampling errors, "to borrow strength" where the replicates (or the d.f.) are too few. These methods imply "modelling" implicitly in different ways and to different extents. However, they may be better than "unbiased" but unstable variances; the sample errors of the computed ste $(\bar{y})$ are approximately $\sqrt{1/2d}$, where d = degrees of freedom. Thus for 4 interpenetrating samples C.V. [ste $(\bar{y})$] = $\sqrt{1/6}$ = 0.4, and for replications, $\sqrt{1/18}$ = 0.24, both much too large for many practical purposes.

a) *Pooling over periodic surveys* of the same design and variables would probably be the first choice when they are available; and those are used fruitfully for labor force surveys in the USA, Canada, Sweden, etc.

b) *Pooling over design classes of the same survey* may be the choice for major domains (regions, provinces) of one—time national samples. This may be readily done by computing errors for the entire (national) sample and increasing it in proportion to the reduced sizes of design subclasses. It will often require separation of strata (e.g., urban and rural) if these have different designs and different stratum sizes between the subclasses. The technique essentially assumes equal "design effects" ($deft^2$) across subclasses and the entire sample.

c) *Pooling over crossclasses* and the entire sample has been practiced frequently. Because the $deft^2$ vary greatly, rough and modified equality is assumed for values of $roh_c = (deft^2 - 1)/(\bar{b}_c - 1)$.

d) *Pooling over all variables* of a survey has been practiced. However such pooling of $deft^2$ is crude because those values vary a great deal. For example, with b = 100, and rho values ranging from 0 to 0.1, the $deft^2$ varies from 1 to 11.

e) *Pooling across similar variables and similar designs* poses difficult questions about the sets denoted as "similar." However, these are frequent sources of "borrowing" values of either deft^2 or roh. The World Fertility Surveys are sources of many examples, and are probably better than most (Table 14.3.1).

f) *Second stages of selection* (secondary sampling units, SSU's) can usually yield many more replicates; usually there are several blocks, segments, E.D.'s, etc., per PSU (district, county, etc.). These methods will yield biased underestimates of variances, because they disregard the clustering of the SSU's within the PSU's. Perhaps with modelling some components may be added to correct for those biases.

g) *Collapsing strata of PSU's* has been used to increase the d.f. of computations. The 2H PSU's in pairs from H strata have only H d.f.'s, but in triplets will have (2H)(2/3) = 1.3H, in quadruples (2H)(3/4) = 1.5H, and 2H − 1 without any strata. It has also been noted that systematic computations with 2H − 1 pairs will yield the precision of 1.3 H pairs. Collapsing strata yields biased overestimates by disregarding the stratification of PSU's (13).

h) *Modelling* of the sampling errors may go further than in the methods above, each of which also depends to somewhat limited extent on models for "borrowing strength." Most common are probably computations based on simple random sampling (SRS), but these often yield very biased underestimates of variances for clustered samples, except when the clusters are very small, including statistics for small crossclasses.

CHAPTER 15. BIASES AND NONSAMPLING ERRORS

15.1 BIASES AND VARIABLE ERRORS

This vast and critical subject poses extraordinary difficulties and contradictions. On the one hand, the problems belong mainly to survey design rather than to sample design (1.1). Systematic biases by definition are not strictly caused by sampling, because they can be expected to be similar even in complete censuses conducted under "similar essential conditions." The sampling statistician cannot be solely responsible for control of nonsampling errors and biases, which concern chiefly measurements. On the other hand, statistic theories and techniques are needed for measuring errors of all kinds and statisticians cannot neglect that vital task. Furthermore, sampling errors should be closely related to nonsampling biases when sample surveys are designed. Nonsampling errors are best investigated, measured and controlled in cooperation between sampling statisticians, survey technicians, and subject matter specialists.

In agricultural surveys nonsampling errors can be particularly difficult, diverse, and large. The problems of measurement are often formidable, and they can differ vastly for various crops, for factors of production and consumption; also for different provinces and different countries. (1.2 – 1.4) The subject is impossible to cover in a brief chapter, but also impossible to omit entirely from this book. What can be said here that has not been said better in the vast literature on the errors of agricultural censuses and surveys? Instead of a feeble attempt at comprehensive coverage, it may be better to emphasize a few topics that are too often neglected, although they are important.

The *variable errors of sampling and the nonsampling biases* cause more problems than their counterparts: that is, the sampling biases and variable nonsampling errors, which also exist. In agricultural surveys the effects of "interviewer variance" may be large, if the interviewers are few and not well standardized. These four classes result when from two dichotomies we

distinguish sampling from nonsampling errors, and biases from variable errors, although neither distinction is easy or entirely clearcut. The errors of sampling concern chiefly the selection of sampling units, but measurement operations fall largely outside its domains. The errors of statistical estimation and analysis, however, often become joint concerns for both statisticians and subject–matter specialists (12.1).

A widely accepted model in sampling theory combines variable errors and biases into the "total error," or root–mean–square error RMSE= $\sqrt{\text{MSE}}$, and this mean–square–error is

$$E[\bar{y}_c - \bar{y}]^2 = E[\bar{y}_c - E(\bar{y}_c)]^2 + [E(\bar{y}_c) - \bar{y}]^2 = VE^2 + \text{Bias}^2. \quad (15.1.1)$$

The expectation is taken over the distribution of all possible values of the estimator $\bar{y}_c$ determined by the sample design. The mean square deviations of all possible sample results from the target value $\bar{y}$ are analyzed into two components: the mean square deviations VE^2 of the variable errors around the average value $E(\bar{y}_c)$ of the sample design; plus the Bias^2 denoting the deviation of that average $E(\bar{y}_c)$ from the target value $\bar{y}$. There exist philosophical questions and doubt whether this target or population value $\bar{y}$ should be separated from a true value $\bar{Y}_{true}$ by $[E(\bar{y}_c) - \bar{y}_{true}]^2 = [E(\bar{y}_c) - \bar{y}]^2 + [\bar{Y} - \bar{Y}_{true}]^2$. Nevertheless the separation of biases from variable errors in Total Error = $\sqrt{(VE^2 + \text{Bias}^2)}$ is of primary and practical importance.

Between biases and variable errors we can make several broad, general, and useful distinctions. First, biases can be considered as a set of *constants, determined by the essential survey conditions*, although their values remain largely unknown. Biases have the same effect B_g on any sample estimate $\bar{y}_c$ regardless of sample size, also on their expected value $E(\bar{y}_c)$. Biases represent the difference $[E(\bar{y}_c) - \bar{y}]$ between the expected sample values and the population target value $\bar{y}$. Variable errors express the difference $\bar{y}_c - E(\bar{y}_c)$ between the estimate and its expected value; they would fluctuate, would be smaller or larger, plus or minus, if different samples were selected with the

very same design. Specific values of the differences $\bar{y}_c - E(\bar{y}_c)$ are unknown, but their "average" (long—run) value is measured by $\text{Ste}(\bar{y}) = \sqrt{E[\bar{y} - E(\bar{y})]^2}$, which is estimated by ste($\bar{y}$) from the sample data.

Second, *the total bias is the algebraic sum* ΣB_g of biases from all different sources g. Some may be positive distortions, others negative, thus often partially cancelling each other. On the other hand variable errors take the positive form $\Sigma S_v^2/m_v$, where S_v^2 is the unit variance and m_v the number of units selected for the sample for the component v. For example, S_a^2/m_a could represent the clusters and S_b^2/m_b the elements of a two—stage selection; and S_r^2/m_r the "interviewer variance" from m_r interviewers.

Third, biases can be reduced only by doing something *better*: by improving the quality of some operation, some "essential survey condition"; e.g. reducing nonresponses, better interviewing, etc. But the reduction of variable errors depends *chiefly* on selecting *more* of something: by increasing the number of units m_r of some kind, either sampling units, or observations, or observers. Sometimes the unit variance S_v^2 can be reduced also, as by stratification of clusters, or by better training of enumerators.

Fourth, variable errors can be estimated with designs and computations based on data from internal replications of units within the sample itself. These estimates require proper designs for replication of units — whether sampling units, or observations, or observers. On the other hand, *estimating biases depends on methods external to the survey* itself, with two alternatives, both of which assume better standards or "benchmarks," rather than correct, exact values. *Quality checks* with better methods can measure individual biases, their variations as well as their averages, and perhaps separate the diverse sources of biases. *Comparisons with external sources* can estimate only a *net average bias* that may seem the effects of several sources.

Fifth, biases and variable errors have different effects on various statistics: they can differ greatly in their absolute and relative values. Specifically, the ratio of Bias/VE is much greater on the overall means than on subclass means, and even less on the comparison of means where variable errors usually dominate (Fig. 15.2.1).

15.2 EFFECTS OF BIASES

The separation of biases from variable errors represents useful theoretical stances (models), which depend on specific situations. For example, for national samples, the sampling units like counties and districts (and smaller and more numerous units like ED's and segments) are assigned for random selections. But regions and provinces are treated as domains and omitting some would be considered as bias. Even more illustrative are the biases of individual enumerators, whose effects are better treated as "enumerator variance" to be reduced by taking larger numbers m_e to reduce their effects S_e^2/m_e. However, the effect of using a type of interviewer (male or female, level of training and education, etc.) is assigned to "essential survey conditions."

Treating the separate biases B_g and their sum ΣB_g as constants is also a stance that needs to be examined. If they were well known and accepted their values would be subtracted. Sometimes, however, enough is known or suspected about them that some adjustments are made and accepted. The Bayesian statisticians say that unimodal distributions about the B_g are more reasonable views than fixing them as constants. However, practical applications of this principle seem difficult and rare.

Clearly the importance of the magnitudes of biases B_g must be weighed in *relative* terms rather than absolute, and four frames of reference are in occasional use. 1) The *magnitude* of the statistic is relevant; for example, a bias of B tons may be neglected for large crops like rice, wheat or corn, but could be important for small crops like some spices or such. Those examples

were for totals, but the magnitudes of means and rates also provide scales of reference for judging the importance of biases. The ratio of Bias($\bar{y}$)/$\bar{y}$ or Bias($\hat{y}$)/Y resembles the coefficients of variation CV ($\bar{y}$) = Ste($\bar{y}$)/$\bar{y}$ that has been mentioned often. 2) *Policy implications* include both the magnitude of the gain or loss represented by the statistics, and the feasibility of corrective actions based on them. 3) The *structural basic variability* (seasonal, yearly, etc.) of the statistics also provides background. 4) The magnitude of *variable errors*, chiefly sampling errors, is perhaps the best standard for measuring biases. That is the aim of the "bias ratio": Bias($\bar{y}$)/Ste($\bar{y}$), or B/σ, or Bias/VE. This ratio of the two main types of errors also has the advantage that the denominator of variable (sampling) errors can and should be designed to take into account the first three kinds of magnitudes noted above.

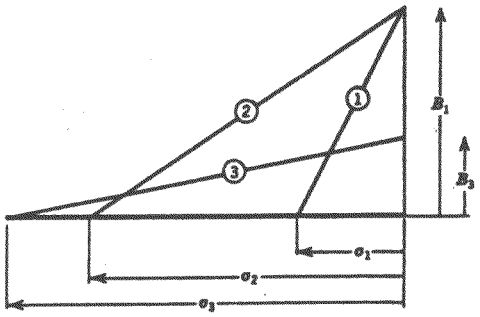


Figure 15.2.1 Variable errors (σ) and biases (B) in root mean square errors (RMSE).

The bases represent sampling errors and other variable errors (σ). For example, σ_1 may be the ste($\bar{y}$) for the mean $\bar{y}$ of the entire sample and σ_2 may be a larger ste($\bar{y}_c$) for a subclass mean, and σ_3 may be the ste($\bar{y}_c - \bar{y}_b$) for the difference between two subclass means.

The heights represent biases (B) and the hypotenuse denotes the RSME = $\sqrt{(\sigma^2 + B^2)}$. (1) For the entire sample that bias B_1 may be large compared with the variable error σ_1, thus taking larger samples would not decrease the $RMSE_1$ by much. (2) However, with the same bias B_1, but with a smaller sample in the subclass, the ratio changes and the σ_2 dominates the $RMSE_2$; and this is not much larger than for (1) despite a much smaller sample. (3) Furthermore, for the difference of means, the net bias B_3 may be much smaller; so that even with a larger σ_3, the $RMSE_3$ for the difference is but little greater than $RMSE_2$. This drastic change in the bias ratio B/σ tends to appear not only for differences between subclasses within the same sample, but also for differences between repeated surveys. [Kish 1987, 2.4.1]

The "total error" = RMSE = $\sqrt{(VE^2+Bias^2)} = VE\sqrt{(1 + Bias^2/VE^2)}$; the bias ratio is seen as the relative increase in the RMSE that is due to biases. In survey sampling the RMSE (the root—mean—square—error) or MSE is widely accepted as principal criterion of accuracy. Further justification rests in the relative invariance of error probabilities of confidence intervals for RMSE's instead of STE of the same size [Hansen, Hurwitz, Madow 1953, 2.2; Cochran 1977, 1.9; Kish 1965, 13.8].

The effects of biases can differ greatly for various statistics even from the same survey. There are, of course, very great differences in biases among different survey variables, but also for various statistics based on any survey variable. We can only illustrate this vast, complex subject by the very different effects of nonresponses on several major types of statistics; consider the simple model of $\bar{y}=Y/N=W_r\bar{y}_r + W_n\bar{y}_n$, with W_n and $\bar{y}_n$ representing the proportion and mean for nonresponses [Kish 1965, 13.4B]. 1) *Simple expansion totals* $\hat{Y}_r=y_r/f$ would have a relative bias $RB= (\hat{Y}-Y)/Y=-Y_n/Y=-W_n\bar{y}_n/\bar{y}$, proportional to the "y content" of the nonresponse stratum. For example, omitting from the frame small firms or small farms incurs in this simple expansion a bias to the degree that stratum is *not* "empty of y content." 2) *The mean* $\bar{y}$, and ratio expansions based on the mean, however, are subject to a different $RB=W_n(\bar{y}_r-\bar{y}_n)/\bar{y}$. The bias depends on the product of the nonresponse portion with only the *difference* of the two means. 3) *For subclass means* the RB will be similar to the overall means except for interaction terms between differentials for nonresponse and for subclasses. 4) *Subclass comparisons* show dependence on "interaction" terms: and the $RB = \{[W_n(\bar{y}_r-\bar{y}_n]_a-[W_n(\bar{y}_r-\bar{Y}_n)]_b\}/(\bar{Y}_a-\bar{y}_b)$. Comparisons of biased results (e.g., two subclasses of the same survey, or similar means from two periodic surveys) often benefit from (partially)"cancelling" biases. "Additivity" in components often tends to reduce bias terms of multivariate analyses. But

contradictory examples should caution against automatic assumptions of totally cancelling biases. 5) *Multivariate statistics* may also be subject to biases of response and nonresponse in ways too complex to explore here.

Table 15.2.1 — Sources of Principal Types of Biases

Sampling Biases
- Frame biases
- "Consistent" Sampling Biases

Constant Statistical Analysis Biases

Nonsampling Biases
- Nonobservation biases
  - Exclusions
  - Noncoverage, missing units
  - Total nonresponses: refusals, not—at—homes, incapacity
  - Item nonresponse, unascertained, missing data
- Observational biases
  - Field data collection
  - Office data processing

The listing of the main sources of biases (Table 15.2.1) needs only brief comments. *Sampling biases* should be small in well—designed samples and they are of two contrasting types. The sources of *frame biases* have been described (Ch. 4) and their control may be chief contribution of the sampling statistician. Frame biases can be very bad but usually they can be controlled, (relatively well, if not perfectly) except for noncoverage, which is treated separately (15.3). On the other hand, "consistent" sampling biases should seldom be important in samples of reasonable magnitude and design. In statistical theory "consistent" refers to estimators whose bias disappears in large samples; for example the bias of ratio means (12.2).

But *constant* biases of statistical analysis differ from the sampling bias above. For example, using medians to estimate means (or vice versa) could result in bad "biases" (differences) for many of the skewed frequency

distributions often found in surveys. Such differences exist in the population and are not reduced by sample size. They belong to the joint domains of statistical and substantive analysis (12.1).

Nonsampling biases may be divided into errors of nonobservation (15.3) and of observation (measurement), which should be divided into two types. Those that arise during field operations (interviewing, enumeration, counting, measuring) are more difficult to control and to measure; and duplicate observations on elements are not feasible usually. On the other hand, processing, coding, tabulating and computing errors are easier to control, and to measure with replicate observations.

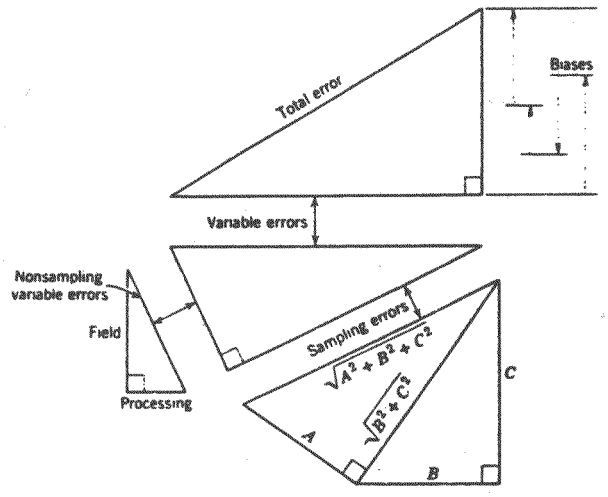


Figure 15.2.2 Classification of Sources of Survey Errors [Kish 1965, 13.2]
Sampling errors are shown arbitrarily with three components. Variable errors, sampling and nonsampling, combine with their summed squares. The total bias is the algebraic sum of all biases, sampling and nonsampling.

15.3 NONCOVERAGE AND NONRESPONSES

Biases of nonobservation result because of failure to obtain data from parts of the population, and because they don't occur with EPSEM, unlike the (N − n) nonselected elements after randomly selecting n elements. The many different sources of missing data may be combined into four main types usefully, because these four require different treatments.

1. *Exclusions, deliberate and explicit*, of portions of the population, which may have different justifications (besides saving effort), can be illustrated with a few examples. a) Some provinces or islands may contain few if any of the kind of holdings (or other elements) sought for the survey population. b) Larger cities may be excluded from some agricultural surveys. c) Areas above x thousand meters of altitude may have little or none of the crops being covered. Other examples are found in other situations and the portions of the areas and populations being excluded should be clearly identified. In some situations estimates may be computed for the effects of exclusions on statistics and perhaps even adjustments introduced, particularly for totals. Exclusions relate closely to *noneligibility* in defining the population coverage (e.g. by age, sex, residence, etc.), but it may be better to maintain the distinction. Furthermore, deliberately and explicitly identified exclusions differ from noncoverage.

2. *Noncoverage* denotes failure to cover in the actual, operational sampling frame, in contradiction of the population definition, some of the elements, or some clusters of elements, some sampling units. They are units missing from the sampling frame (4.2), and omissions due to faulty execution of survey procedures. They differ from nonresponses, because their location, their numbers, even their very existence, are usually unknown. To estimate them, measures *external* to the survey would be necessary; a) either some quality check, such as post—enumeration surveys for censuses, but these are expensive; b) or statistical checks against outside sources, such as demographic checks made on population census data. The population undercounts of even good decennial censuses can illustrate the problems, which are often even more

severe in surveys. Besides underestimates of aggregates, the means and other statistics are also subject to biases, because noncoverage occurs not at random, but in different proportions in various subclasses, which in turn are correlated with survey variables. To the degree that differential nonresponses may be estimated for meaningful subclasses, adjustments of results may be useful [USCB Ch V; Kish 1965, 13.3]. Finally, the net noncoverage is the result of gross *undercoverage minus gross overcoverage*; this latter may often be kept low, but it also can occur, and it has been investigated in crop—cutting measurements.

3. *Nonresponses* usually denote "total" nonresponses, as distinguished from "item nonresponses," which may require different treatments. Nonresponses refer to various sources of and reasons for failure to obtain observations (measurements, responses) on some elements designated for the sample. Thus they differ from noncoverage, because their numbers can be counted, and the *response rates computed, if and only if accurate accounts are kept* for all (eligible) elements designated for the sample. These procedures need some care and effort, and are necessary for estimating response rates, perhaps their possible effects, and also possible adjustments. Reporting the extent of nonresponses has become the accepted responsibility of better surveys. All these aims can be better served by sorting nonresponses into several major classes, as below.

4. *Item nonresponses* refer to *unascertained* items (variables) from cases (elements), where many (most) survey items have been obtained. Reasons for those missed items are varied: refusals or incapacity of respondents; error by enumerators; unusable, invalid, incorrect answers; lost or erased items. *Imputing* (editing, assigning) answers for item nonresponses seems more reasonable than for total nonresponses for two reasons. a) Because more data (variables) are available, imputing missing items can be done more accurately (as in multivariate regressions with good predictors). b) Multivariate statistics could have large proportions if missing cases, when these can result from any

single missing item; the product $r_1 r_{12} r_{123} \ldots$ can become small even if all response rates are large. Imputations aim at reducing biases from missing items without greatly increasing variances with unequal weights, and several methods are available [Kalton 1983]. Imputations may also be applied fruitfully to total nonresponses, especially in periodic surveys when responses on other waves make imputations both more feasible and more advisable. Also sometimes good auxiliary data may be obtained for total nonresponses from other data sources, or from neighbors, or from brief screening enumerations.

Classes of nonresponses are named here for interviews at households and holdings on location, but the terms can be translated to telephone or mail surveys and to other methods of data collection. All categories refer to eligible respondents; ineligibles (closed farms, vacant dwellings, stores, garages, etc.) should not be included in the counts of nonresponses. a) *Not—at—homes* (NAH) refer either to entire holdings or to households or to specific respondents, and these can vary greatly with type (holder, employed, housewife, any responsible adult, etc.), also with numbers of callbacks. There are operational differences also between temporary absences of the respondents and empty farms or houses. The NAH may be for an hour, day, week, season; NAH denotes temporary unavailability and deferral rather than denial of response. b) *Refusals*, on the other hand, denote denials of interviews and are less temporary and changeable, more fixed and unobtainable. c) *Inability* or *incapacity* may refer to physical or mental illnesses that interfere with responses for the entire survey period; or to illiteracy or language barriers on some surveys. d) *Not founds* can occur on mail surveys, *not attempted* in case of inaccessibility (due to costs, distances, dangers). e) *Lost schedules* denotes information lost or destroyed after field collection, when repeat efforts are not feasible [Kish 1965, 13.4].

15.4 CONTROLS FOR NONRESPONSE

Methods for reducing nonresponses must be based on knowledge of the sources of nonresponses and of differences among response rates. For example, city dwellers are more likely to be both NAH's and refusals than rural and farm people. Definition of *respondents* seems important: housewives (especially with children) and "any responsible adult" are more easily found and interviewed than specified, employed adults. Some *questions* (how many children?) are easier to obtain than others (what was your income?). Identification of the agency or *institution conducting the study* may matter also. These factors should be in the background while considering suggested methods for reducing the effects of nonresponses. This brief list cannot do justice to the vast volume of available material on this subject [Madow, Olkin, Rubin 1983; Kish 1965, 13.5; Zarkovich 1963, Ch. 7].

1. *Better procedures* is mentioned merely as a reminder, because specifics are impossible in this brief space, and because it is difficult to invent a good method (feasible and not too expensive) when so many have been tried already. But it is painful to see newcomers repeating mistakes that have been often exposed; for example, to see surveys without callbacks, or mail surveys without repeated mailings, when these have been shown to be so effective. But it is also harmful to use methods proven useful in some situations transposed to others, where they are quite inappropriate.

2. *Call—backs* (or repeated mailings, and other repeated efforts) are the most widely effective methods for improving response rates. With c calls it is possible to reduce not—at—home rates of q to q^c to the degree that the q remain relatively constant. This is often close enough for actual rates up to 3, 4 or even 5 calls; and thus not—at—homes of q can be brought below 10 percent. We must overcome a common mistake about the costs of callbacks. It is true that making c calls on a set of n addresses is more expensive than making single calls on each. But it follows not that achieving n interviews with

single calls would be much cheaper than with c calls. The actual increase is often slight! [Kish 1965, Table 13.5.II] Refusals present different and more complex problems.

3. *Subsampling* of nonresponses is only helpful when callbacks use much more expensive methods than first or earlier calls. This negative advice follows from ordinary callbacks not being much more expensive, being often productive, and also from increased variances from weighting for reduced sampling rates.

4. *Substitutions* for nonresponses seldom provide effective remedies because: a) substitutes are more like responses than nonresponse; b) they tend to reduce field efforts for better responses. For these reasons substitutions may be better justified a) when entire large (primary) units are substituted in the office, rather than in the field; b) if complete disclosures about substitutions are made,in order both to inform the readers and to deter the survey team from too many.

5. *Adjustments* for differential response rates may be made, although these may be less successful than for item nonresponses for the reasons stated above.

CHAPTER 16. SURVEYS ACROSS TIME

16.1 REPRESENTING TIME

Timing of surveys can be especially important in agricultural surveys, where so many activities of production and consumption are tied to seasonal and temporal production. Careful timing may be especially critical for less developed agricultural populations, where good records may be lacking. Representing time spans involves choosing reference periods for surveys.

To avoid confusion we need to distinguish three kinds of periods concerning any survey: a *collection period* during which data are collected; *reference periods* defining the data, which may differ greatly for diverse crops and statistics; and *reporting periods* which can consist of one or more reference periods. For example, for the U.S. Census the collection in period is a few weeks in April, but the reference and reporting periods are April 1 for current data, but the whole preceding calendar year for agricultural and economic data, etc. In multiround and cumulated surveys the reporting periods are pooled from reference periods. Reference periods may be as short as a single day or even a minute (in time studies) but they are cumulated for reporting; or a week (for employment) or month, or as long as a year (for income).

1. *Unique or special periods* may be accepted from natural forces; e.g. seasons for harvesting crops, also for lambs' births, for monsoons, etc. Dates fixed by laws, rules and customs – like Christmas, New Year, fiscal year, month's end, Sundays (or Fridays or Saturdays) – seem arbitrary, provincial and temporary, but they are fixed for the population, hence beyond the designs of researchers.

2. *"Typical" (representative) periods* are commonly used; perhaps too commonly in confusion either with uniquely fixed times (1) or with proper sampling of time (4). One good example is April 1 for the reference dates of decennial censuses of the USA, which is now traditional. Another is the choice of the third week to represent each month in the Current Population Surveys

[USCB 1978]. There are many examples of choosing "typical" (representative) periods by judgment in preference to sampling the time dimension; these resemble the "typical" areas, which had been commonly used also for spatial representation until the recent spread of probability area sampling.

3. *Complete and separate coverage* of all reference periods over the reference interval is a temporal analogue of a complete census over all administrative areas. These yield data for all periods, for changes between them, and also averages over them, e.g. the yearly survey over all 52 weeks of the Health Interview Survey [NCHS, 1958]; different examples arise from time series for some financial data. We can distinguish continuous from discontinuous periods over the entire intervals; the Current Population Surveys [USCB 1978] cover all 12 months over the year, but only one "typical" week to represent each month. Continuous collection of data is seldom feasible; but reference periods can be, as in multi—round surveys, and from these the aggregates and means for entire intervals can be computed. However these raise naturally the possibility of sampling instead of completely covering all reference periods.

4. *Sampling of separate periods over a time interval* can be an alternative to either confining the sample to one or a few "typical" (representative) periods (2) or complete coverage of the entire interval (3). Models of temporal variation can be made similar to spatial variation: as a target population varies in space, so we can consider time as another dimension of variation. Populations vary from year to year and week to week, as they vary among regions and among counties. Probability sampling spread over the population area serves as the accepted strategy to cover and counter spatial variation. But temporal variation can be even more important, especially for cyclical variations, e.g. seasonal, weekly, or even diurnal. Vast temporal fluctuations also occur in epidemics, economic fluctuations, social and political attitudes, and rapid and widespread changes have become common. To cover and to counter these

changes either complete coverage or sampling is needed. However, for many characteristics that have temporal stability but much spatial variation, spatial coverage may be more crucial.

5. *A temporal x spatial matrix* for averages (marginals) for both dimensions can be designed for periodic samples. A good example again is the Health Interview Surveys [NCHS, 1958] that yield weekly national averages, yearly statistics for small domains, and monthly and quarterly data for larger domains. The samples are too small to yield both spatial and temporal details simultaneously; but each period can be designed to sample the entire population area; furthermore the periodic samples can be so controlled that they cumulate to subtotals (regions) and totals (national) that are balanced (stratified).

16.2 CONCEPTS AND DESCRIPTIONS

Decennial censuses of populations have been used for decades or centuries, but periodic sample surveys are newer and increasing rapidly in numbers and scope. It is important to clarify basic concepts and terms to reduce the remaining confusion.

Repeated surveys denote "similar observations on the same population," whereas *periodic* surveys refer to surveys repeated at specified regular *periods* over a longer *interval* of time. The "same population" needs identification because populations change over time both in extent and in content: e.g., cities and countries change boundaries; for complex units (families, organizations) changes can be frequent, because their constituents (persons, adults) are born, die, and migrate. "Similar observations" must also be defined, operationalized and collected.

Overlapping designs refer to covering the same sampling units in repeated periods. The overlapping units may be defined as the elements of analysis (individuals, persons), or they may be larger units, such as area segments. Units such as holdings, households, composed of distinct elements, present problems of frequent and complex changes. Designs may require either

complete or partial overlapping; the latter permits gradual changes of the sampling units. In *nonoverlapping* designs the units are changed deliberately for each period.

Panel surveys refer to overlapping studies with *repeated observations on the same elements*, on the same persons or households or holdings. Panels face problems of learning, fatigue and losses from mortality and mobility; of moving and high locating costs; and of identification for complex units, like families; but they are needed for detecting the dynamics of *gross* (micro) *changes* of individuals. (though these get confounded with errors of measurement.) On the other hand, for measuring *net* (macro) *changes* of averages it may be easier and clearer to overlap simpler and larger units of sampling (such as area segments) and still retain much of the gains in the variances from correlations. (Some studies have done both: retain segments for clear net changes, but also follow moving individuals for gross changes.) The gains from correlations are also retained proportionately in partial overlaps. Net changes may be measured also with partial and nonoverlapping samples, though with higher variances. Panels have also been called longitudinal surveys and "strictly longitudinal studies."

A third use for overlapping and panel studies is for obtaining *incidence of new events* between two (or more) dates (periods), in contrast to measuring *prevalence* of all events at one time. These are called *multi—round* surveys by some and *prospective* studies by others; they stand in contrast to *retrospective* studies that depend on memories or records for past data. Such prospective designs should be panels for measuring individual changes, but they can be nonoverlapping studies for net changes. The collection of data on new events is sometimes aided with records (diaries, budgets) kept by respondents, or by others, or by machines.

Multi—round and prospective studies are usually designed to be *periodic continuous* studies, in order to cover the entire time interval of the study. Continuous registers can sometimes be used for this and retrospective studies

attempt this by relying on memory. However, other repeated and periodic surveys may give only disjoint "snapshots" of the time span; e.g. decennial censuses often yield data for census years separated by ten year gaps.

16.3 PURPOSES AND DESIGNS

In Table 16.3.1 we note five purposes and six designs, with pairings which call attention to designs that best serve each of the four purposes, with reduced variances. Most periodic studies have several purposes and thus we should face – not necessarily and completely solve – the difficult problems of multipurpose designs. Actually, current levels (A) and net (macro) changes (C) can be served with any of the six listed designs – but with some increase in the variances or in costs. But individual (gross, micro) changes (D) need panels, and cumulations (B) need some changes. The chief variation shown for these designs concerns the amount (and kind) of overlaps between periods. The rotation scheme of complete overlaps shows, with aaa–aaa, that the periods have all common parts; the nonoverlap with aaa–bbb shows none; and the partial overlap abc–cde–efg shows c and e as 1/3 overlaps between succeeding periods only.

Table 16.3.1 – Purposes and Designs for Periodic Samples

Purposes	Designs	Rotation Scheme
A. Current levels	A. Partial overlaps $0<P<1$	abc–cde–efg
B. Cumulations	B. Nonoverlaps $P=0$	aaa–bbb–ccc
C. Net changes (means)	C. Complete overlaps $P=1$	aaa–aaa–aaa
D. Gross changes (individual)	D. Panels	same elements
E. Multipurpose time series	E. Combinations, SPD	
	F. Master Frames	

This section concentrates on the effects of varying the proportions of overlaps P in diverse designs for different purposes; in complete overlaps $P=1$, in nonoverlaps $P=0$, and in partial overlaps $0<P<1$. Much of the discussion assumes for simplicity that the periodic samples are of the same size, or of the same sampling fraction; but changes in sizes, fractions and designs are possible, and even desirable in some cases, as noted below.

Current levels is one name for the most common type of estimates for single "points" in time, whether the point of reference period is a single day or even minute, or a week, month, or even a year; but "static" estimates and "cross—section" are other commonly used terms. Variances of current estimates are the same for complete overlaps $P = 1$ and for nonoverlaps $P = 0$; they can be expressed briefly for means as $Deft^2S^2/n$, where $Deft^2$ is the effect of the sample design on either the element variance S^2 or on the sample size n.

That simple formula also holds for *simple* means from partial overlaps ($0<P<1$). But statistics based on them can utilize the overlap P for a reduction of the variance with a complex mean: with help of the correlation R^2 between surveys within the sample overlap P, the portion $(1 - P)=Q$ of the *preceding* sample is combined with the current mean to improve it. The variances are reduced by the factor $[1 - Q)R^2]/[1 - Q^2)R^2]$. This is a clever technical contribution, much explored by sampling theory, though actual gains unfortunately tend to be modest in most practical situations [Cochran 1977, 12.11–12].

Cumulations refer to the purpose and practice of accumulating, pooling and aggregating sample cases of individuals. Means based on several periodic samples covering a longer interval is the purpose we treat here, but the implications are similar for other statistics, such as regressions and other analytical statistics. The aims of cumulations are threefold. First, they obtain greater precision, with lower variances from larger sample bases, especially important for smaller domains. Second, from the larger sample

bases of cumulations we expect also greater spatial spreads of the design, so they can better cover small domains. Third, they can cover temporal variations – seasonal, cyclical, irregular – over longer intervals that include several periods.

Samples with no overlaps, $P = 0$, are best for cumulations. They are simpler and also yield lowest variances: $S_j^2/2$ for two periods and S_j^2/J for J periods, where the S_j^2 are variances for single periods assumed to include $DEFT^2$ and factors like $(1 - f)$. For overlapping samples, however, positive correlations R between periods increase those variances for means, totals, etc. Thus cumulations can be had even with partially overlapping samples; good compromises can be obtained, for example with $P = 1/3$, which is optimal for current levels and also good for net changes. However, optimal allocations of $P = 0$ for cumulations remain in conflict with optimal $P = 1$ for measuring changes.

Net changes refer to the differences $d = (\bar{x}_1 - \bar{x}_2)$ of means between two periods,; whereas *gross changes* deal with the total changes of individuals, some of which remain hidden (because they cancel) in the net change of means. Measuring net changes are common and important aims of surveys and studies, and they are also related to other uses of the data. Perhaps the most common forms are differences in dichotomies, denoted by proportions, such as $d = (p_1 - p_2)$, and in similar rates and ratios. We can also use the form $d = (\bar{x} - \bar{y})$, which denote aspects of design where, happily, statistics can yield great gains. The variance of $(\bar{x} - \bar{y})$ can be greatly reduced when the pair of variables have high positive correlations R in overlapping samples, and we now turn to several aspects of great *flexibility* that may be explored in statistical designs for differences.

1. The variances of mean differences are reduced by factors $(1 - R)$ in complete overlaps, which is the extreme (with $P = 1$) of the factors $(1 - PR)$, which may be obtained from partial overlaps for minimizing $\text{var}(\bar{x} - \bar{y})$. But

partial overlaps are used in practice: a) for reasons of feasibility, to reduce burdens, fatigue and biases of respondents, and b) to reduce variances of other statistics in multipurpose designs.

It is simple to think of the variances as $(2S^2/n)$ for differences between pairs of samples of size n without overlaps; $(2S^2/n)(1 - R)$ with complete overlaps; and $(2S^2/n)\ (1 - PR)$ with partial overlaps P. The S^2/n assumes simple random sampling. but for complex samples design effects $Deft^2$ should be included, and reductions obtained from overlaps in complex samples may be even greater than indicated by the factors $(1 - PR)$.

2. One may obtain almost the full reductions of complete overlaps even from partial overlaps by using improved estimators for the differences. In those estimators the overlap portion P gets larger weights by factors $1/(1 - R)$ than the nonoverlap portion $1 - P = Q$, because the overlapping elements contribute that much less to the variance. This improved estimator of the difference is

$$\hat{D}(\bar{y} - \bar{x}) = [P(\bar{y} - \bar{x})_p + Q(1 - R)(\bar{y} - \bar{x})_q]/(1 - QR). \qquad (16.3.1)$$

Its variance may be expressed, for two srs samples of size n_i as:

$$VAR[\hat{D}(\bar{y} - \bar{x})] = \frac{(1 - R)}{(1 - QR)} \frac{2S^2}{n} \qquad (16.3.2)$$

These effects are shown in Table 16.3.2 with $a = (1 - PR)$ for the simple difference and $b = (1 - R)/(1 - QR)$ for the weighted difference. This factor approaches $(1 - R)$ for high values of R (where most important) and for highef values of P, say $P \geq 2/3$, as seen in the last two rows of Table 16.3.2. High values of R are common for stable characteristics that can be well measured, but not for volatile or poorly measured characteristics or attitudes. Negative values of R must be rare, but that side of the table with negative values, can be used instead to see what happens to sums of two means $(\bar{x} + \bar{y})$

when the factors are (1 + PR). We also add again (as in note 1 above) that factors $Deft^2$ in complex samples may enhance considerably the gains from overlaps, because $Deft^2$ are less for the differences.

Table 16.3.2 Effects on Variance of Differences of R_{xy} for Several Proportions of Overlap (P) [Kish 1965, 12.4]

P		Negative Values of R_{xy}					0	Positive Values of R_{xy}						
		−1.0	−0.8	−0.6	−0.4	−0.2		0.2	0.4	0.6	0.8	0.9	0.95	1.0
1/3	*a*	1.33	1.27	1.20	1.13	1.07	1.00	0.93	0.87	0.80	0.73	0.70	0.68	0.67
	b	1.20	1.17	1.14	1.11	1.06	1.00	0.92	0.82	0.67	0.43	0.25	0.14	0
1/2	*a*	1.50	1.40	1.30	1.20	1.10	1.00	0.90	0.80	0.70	0.60	0.55	0.52	0.50
	b	1.33	1.29	1.23	1.17	1.09	1.00	0.89	0.75	0.57	0.33	0.18	0.10	0
2/3	*a*	1.67	1.53	1.40	1.27	1.13	1.00	0.87	0.73	0.60	0.47	0.40	0.37	0.33
	b	1.50	1.42	1.33	1.24	1.12	1.00	0.86	0.69	0.50	0.27	0.14	0.07	0
1.0		2.00	1.80	1.60	1.40	1.20	1.00	0.80	0.60	0.40	0.20	0.10	0.05	0

These effects are $a = (1 - PR)$ for the simple difference (12.4.8'), and $b = (1 - R)/(1 - QR)$ for the weighted difference. Two equal, unrestricted saples are assumed.

3. Great flexibility can be used in choices of sampling units for the overlaps. Using elements as sampling units is needed for gross changes from panels and they yield generally the highest values of R, hence the lowest variances. But they also have great problems, and therefore larger units, clusters of elements, must be used instead for the overlapping units, in many situations.

Compact area segments, containing several dwellings and their occupants, have been widely used for overlapping samples [USCB 1978; Kish 1965, 9.5, 10.4, 12.5C]. Identification of dwellings and persons with the segments are feasible if well done. Each period's sample retains its character as a probability sample of the population, despite the moves of households, families and individuals; despite births, deaths and migration, the stability of area segments retains representativeness of its inhabitants. It is true that, due to those changes and moves of the elements, the correlations R between periods

is proportionately reduced; but the reduction affects only the changing portion. Hence overlaps based on segments retain most of the correlations R for measuring net changes.

4. Greater flexibility may be used in the second and later waves of interviewing and generally in the data collection in the field. The first wave must bear the initial costs of selection, contact, cooperation, and some basic, core information that later waves may reduce or omit. Therefore, in later waves the costs per case (element, interview) can be made lower (little or much) than on the first wave. They may be done sometimes by different methods, perhaps by telephone or mail instead of personal interviews. This helps to explain the popularity of the large overlap portions P for periodic surveys, larger perhaps than are indicated by variances per case (n). Thus in the Current Population Surveys overlaps of P = 7/8 are used, with the last 7 of 8 waves conducted mostly by telephone interviews [USCB 1978]. In some situations responses may also be better in later waves, but that is a complex and difficult subject.

Panels denote samples in which the same *elements* are measured on two or more occasions for the purpose of obtaining the *individual* changes $d_i = (x_{i2} - x_{i1})$. From a good sample of the d_i we can estimate the distribution of individual changes for the N elements in the population. Furthermore, from the mean of these *internal* changes of individuals we can also estimate the net, mean, external change: $\Sigma (x_{i2} - x_{i1})/n = \Sigma x_{i2}/n - \Sigma x_{i1}/n = (\bar{x}_2 - \bar{x}_1)$. However, from the net change of means one cannot estimate (directly) the gross change of individuals. This duality of changes has various names: individual/mean, gross/net, internal/external, micro/macro.

Only panels can reveal the gross changes behind a net change; for example, a +2 percent net change of behavior may hide x+2 percent positive and x negative changes, where x may be small or large, unknown. Strong models could substitute for panel samples in theory; but in reality these exist only for some special individual variables: age, parity (births) for women; some

incurable, chronic diseases and infirmities; some acquired and permanent immunities; years of education, etc. Sometimes changes can be traced reliably from memory or from records. But often models and memory are both lacking or unreliable, and only panels can yield the data needed for individual, micro changes. These are needed not only for their frequency, but also for the dynamics of relationships and causation.

On the other hand, panels may be too difficult and not feasible for diverse reasons (mortality, mobility, refusals). Often, however, neither advantages nor disadvantages seem absolute; rather they should all be weighed against each other. Here we need to clarify also differences between panels and complete overlaps. Panels define special cases of complete overlaps when the sampling units are the elements themselves. But sampling units such as area segments used for overlaps differ from panels because of mobility and mortality in the population. Overlapping samples based on stable area segments can be preferable for good current estimates and net changes; they have been so used in many surveys, e.g. the CPS [USCB, 1978]. Area segments are more stable in rural portions, less in cities, and even less on their suburban fringes. Such stability (in degree and in time span) also describes their value for measuring changes [Kish 1965, 9.5,12.5C].

With their unique advantages, panels can reveal results undiscovered by other methods, but they are not common because of their difficult problems. Even less common are complete overlaps ($P = 1$), because they would have most of the problems without the completeness of panels. Since area segments have fair stability of people in short periods (about 82 percent of households over a year in the USA 1985), the variance of mean (net) change $(\bar{x}_2 - \bar{x}_1)$ is reduced by the factor $(1 - R')$ where R' is little less than the R from panels. Other benefits (lower costs) and some, if any, disadvantages (i.e., refusals) are also proportionately inherited [Kish 1987, 6.4; Duncan and Kalton 1986].

Multipurpose and Combined Periodic Designs. Most periodic surveys can, should, and do serve several purposes. Current estimates and net changes can be readily satisfied jointly using any proportion of overlaps. *Partial overlaps* can be designed for both current levels and for net changes. High overlaps are better for net changes, but low overlaps also (e.g. P = 1/3) can be made to yield low variances with improved estimators, and they are better for current levels (6.2A2 and Table 6.2.2). But high overlaps are often used because of the lower costs of later waves. *Split panel designs (SPD)* [Kish 1987, 6.5] would incorporate two separate designs that have conflicting properties, advantages and faults. A portion, say P = 1/4 or 1/3, would be for a panel for individual changes; it would also provide overlaps and thus reduce variances for mean changes and for current levels, with *correlations (R) with all periods.* The other portion (1 − P) would provide nonoverlapping samples to permit cumulation; hence they should have increased spread for cumulating periods. The two sample designs could be quite distinct to suit efficiently the needs of each. But the measurements would need to be similar to permit the combination of the two sets of results into single series of statistics.

16.4 PANEL STUDIES

We described (16.1) five alternative ways of representing time in reference periods. Now we distinguish *four major alternative ways for the collection of data over time.*

1) *Retrospective data* refer to methods based on the memory of respondents to report data over lengthy periods; from a year to a lifetime, for example (because all responses even for short periods are retrospective in a trivial sense). 2) *Registers, records or direct observations may* sometimes be preferred, but are often too difficult. 3) *Longitudinal studies, follow—ups, multi—round surveys* are terms used for repeated observations of

populations over longer intervals. 4) *Panel studies* involve repeated (periodic) observations (interviewing) of the same elements (subjects, persons, families) [Kish 1987, 6.1C, 6.4].

Registers and records are too specialized a subject and of too many possible forms for a brief treatment here. It would be best to concentrate on a twofold comparison of panel studies. On the one hand, a panel study of k periods can be compared to k distinct samples, and this comparison is more basic. Thus a sample of n elements (households or holders) may be observed in k periods for a panel, compared to a total of kn elements in k nonoverlapping periodic samples of n each. The costs per interview are somewhat cheaper for a panel [Freedman, Thornton, Camburn 1980; Duncan 1984]. On the other hand, there are interesting comparisons, chiefly in epidemiology, of *retrospective* studies versus *prospective* studies, each confined to one sample. In retrospective studies memory and records are used to retrieve the needed longitudinal information, and the costs per element are much higher for the panels than for retrospective studies based on single visits to the sample individuals.

A. Panels versus Distinct Samples

1. *Initial self–selections.* Any sample of humans probably involves some form and some amount of volunteering, hence self–selection, hence potential bias in representation. However it has been noted often that the rate of refusals is increased considerably when respondents are asked for cooperation in a long–continued panel after the first call.

2. *Attrition* continues after the first call, but at a much reduced rate. This attrition has two forms: refusals due to "panel fatigue", and nonresponses due to disappearances that cannot be traced. We distinguish these from losses due to temporary nonresponse, or mortality, or changes, or mobility, all treated separately below. The refusal at the first call may be, let us say, as high as 20 percent, but the attrition after that may be as low as 1 or 2 percent on each

call. Nevertheless, these small losses can also accumulate to a sizeable total after many calls. But these effects are extremely variable; and fatigue and refusals are much less in rural and in less developed areas (see 14 below).

3. *Temporary nonresponse*, either not—at—home or refusal, may be considerably higher than attrition; say 3–6 percent versus 1–2 percent, depending greatly on timing and kind of procedures. Hence they must be included in later calls, and their data interpolated with retrospection and with imputation.

4. *Mobility* must be treated distinctly from inevitable attrition. First, mobility may be much greater than attrition, depending on the population, and on the time interval covered; hence losses could accumulate to prohibitive levels. Second, they can be much reduced, with enough care, effort, ingenuity; and the literature conveys much good advice, specific to situations, but translatable to others. Mobility has entirely different effects on panels than on overlapping sampling units, such as area segments, which are self—correcting and reflect (in expectation) the changing population. Longitudinal studies of restricted sites with permeable borders, such as a single area, will reflect great mobility.

5. *Changes* of the elements can be considerable for complex units, like families, holdings, organizations, institutions, firms. Dealing with them in a panel requires much skill, knowledge and experience. Such changes generally reflect similar changes in the entire population, as does mortality.

6. *Mortality* affects the entire population, and its treatment would be different and simpler for a study defined by and confined to the initial sample and population. Within that definition the panels suffer no worse defects than changing samples, either from mortality, or from other forms of outmigration, or from changing elements. That is why we separate the panel effects of changes mortality, and births (5,6 and 7), from attrition and other specific defects of panels (1,2,3,4).

7. *Births and immigration* should be introduced into longitudinal studies defined by ever changing, living and complete populations, with births into as well as deaths from it. They must include some method for introducing births and migration, in contrast to studies confined to the initial population (6 above); also in contrast with non—panel methods defined by stable units, such as area segments, even in overlapping samples.

8. *Retest reactivity*, panel bias, panel contamination, sensitizing or learning are all names given to the fear that the experience of past interviews (observations, enumerations), and the anticipation of future ones, may change the behavior and attitudes, opinions of the individuals in the sample. Any effects would depend on many factors involving the nature and timing of observations, of the study variables, and of the population. See 14 below for beneficial effects.

9. *Reinterview laxity* has been raised as a possible source of bias: that both the respondent and the interviewer may be subject to inertia, and to similar answers they must (unintentionally) recollect from past interviews. Interviewers may also become generally somewhat less careful on return visits. The "rotating group bias" of the Current Population Surveys is the best known example, though a rather confusing problem [USCB 1978]. But see 14 below for advantages of familiarity.

10. *Checks and controls* are desirable to guard against possible biases from the use of panels. These can take so many forms and are so dependent on specific situations that listing them here seems futile. Checks can generally be of two kinds: comparisons in available background variables, like age and sex, and in the study variables that are more critical but also more difficult to validate. But see 13.

After those ten possible defects we come to four possible and considerable benefits of panels; three of these advantages (12,13,14) are also shared in good portion by overlapping sampling units.

11. Only *Panels* yield data on individual changes.

12. *Lower costs* per interview than for changing, nonoverlapping samples are common, in spite of the widespread hostility to panels. First, only the first wave bears the sampling costs, both in the office selection and designation and in the field work of identification and gaining access and cooperation. Second, the basic background data concerning individuals ("face sheet data") are mostly borne or more costly in the first wave. Third, acquaintance with the element (holding, holder, household) facilitates contact; for example, the timing of calls (interviews). Fourth, (and this is most variable), later calls can cost *much less*, if done by telephone, mail, or some other cheaper procedure on all or on most of the sample, when this seems not feasible for the first wave. (This in my view is the real, though largely neglected, reason for the large overlaps in some current labor force surveys.)

13. *Errors removed or reduced* can be a considerable advantage of panels, if procedures are introduced to check for differences and for consistency.

14. *Familiarity* with the sampling units and with the individuals can often have positive results, in contrast to the negative and feared effects of attrition (2), reactivity (8) and laxity (9). For demographic surveys in developing countries it has been noted that: "The survey staff will master their duties better and learn to know the sample areas and even the population. For their part the respondents, meeting interviewers they already know, become more relaxed and willing to answer questions. It has been reported in several surveys which have lasted three or more years that initial suspicion and reserve have with repeat visits given way to trust and the interviewers have been received with pleasure as old friends [Cantrelle 1974; Nepal 1976; Iran 1978]" [Kannisto 1983; UN 1984].

B. Prospective Panels versus Retrospective Studies

This contrast differs greatly from the contrast of a panel with a similar total number (kn) of visits. Here instead, the use, value and cost of a panel of several waves is contrasted with a one wave study, which depends on retrospective recollection of data over time. This contrast is best developed in the literature of epidemiology and public health as "retrospective" versus "prospective" studies of diseases and risks. The two kinds of contrasts give extremely different views. This is especially true of costs, because panels seem less costly per interview than a similar number of new waves, but prospective panels are much more costly per individual than a one time retrospective study. We now turn to a listing of the problems of prospective panels that lead often to using retrospective studies (1,2,3), followed by the doubts inherent in their use (4,5,6,7,8).

1. *Higher costs.* Panels of several waves are bound to incur considerably higher costs than a retrospective study of single observations (interviewers) on a similar number of individuals.

2. *Rare events.* Prospective studies of panels with several (many) waves can become especially expensive when one proportion of "susceptibles," or of "diseased" and especially of "susceptibles with disease" is small.

3. *Delayed results.* This may often be the principal reason for using retrospective studies, instead of waiting for years or decades, which the full unfolding of a prospective panel would require. It may be possible sometimes to do a retrospective study soon and then also begin a prospective long—range project to allay eventually the doubts from the former (6,7,8).

4. *Panel fatigue, bias, mortality.* Mortality and other selection biases are likely to be greater in retrospective than in prospective studies (7).

5. *Lack of randomization.* Random assignments of treatments seldom seems feasible, and often probability selection of individuals is also too expensive. However, it would be unreasonable to hold these imperfections against prospective panels, since these factors are likely to face greater hazards and doubts in retrospective studies.

6. *Biases of memory, recall, retrieval.* These cover the principal objections to retrospective studies and much is written about them.

7. *Mortality biases.* These refer to biases in the population arising from possible differential mortality (and other losses from attrition) between the two contrasts of comparisons. These biases may have greater effects in retrospective studies, because they may be traced in prospective panels.

8. *Selection biases.* In addition to biases in the population, retrospective studies may also suffer more from selection biases.

CHAPTER 17 CENSUSES AND SAMPLES

17.1 COMPARING SAMPLES WITH CENSUSES

The costs and the advantages and disadvantages of using a sample compared to a complete census to cover a national population is our main concern here. The contrasts between samples and complete censuses is more striking and decisive for large,. national populations than for a province or district. The comparisons may also be different for a population of households than for an inclusive agricultural count of many products with different seasonal variations. If the total cost is fixed, a different problem would be faced in comparing a modest sized national sample versus a complete census of one province, or a few provinces. Also the problems would be limited for limited populations; for example, for mailed questionnaires to members of some association.

This discussion must focus on censuses of population and housing, which are sources and frames for sampling dwellings and households. These are often used for agricultural and food surveys. Agricultural censuses, of course, have also been used for samples of holdings, though lists and locations of holdings may become obsolescent faster than frames for dwellings [FAO 1977, 1978a].

However, to cover the holdings or households of a large and widespread national population requires great efforts and the contrasts between censuses and samples are striking. "Complete censuses are nevertheless relatively expensive and slow, and even with today's modern, efficient procedures it can take four years to get most of the census data into the hands of users. These are the basic reasons for not taking censuses more often, or with greater depth and richness of data.

Table 17.1.1
Eight Criteria for Comparing Three Sources of Data
[from Kish 1979]

Criteria	Samples	Census	Adm. Registers
Rich, Complex, Diverse, Flexible	***		
Accurate, Relevant, Pertinent	*		?
Inexpensive	*		***
Timely, Opportune, Seasonal	**		*
Precise (large and complete)		*	*
Detailed		**	*
Inclusive (coverage), Credible, P.R.		*	?
Population Content	**	*	

"Therefore the primary objective of a census is typically to obtain a detailed and complete picture of the number (size) and basic structural and related characteristics of the population, and to provide as much detail as possible for small domains and especially for local areas...

"By contrast, inquiries confined to samples of the population can, by virtue of their smaller sizes, be designed to obtain a wide variety of complete data for studies of interrelationships and changes. Such data are not gathered in complete censuses: attempts to do so would result in very high costs and, even more important, in low quality. Furthermore, sample surveys can be tailored flexibly to fit a variety of needs with appropriate methods of collection. Choice of timing, of respondents, and of methods can be suited to the needs of data collection. The content of the study population can be better controlled and directed toward the specific survey aims; such flexibility may be prohibited by the public aspects of the census. Sample surveys are much cheaper, and they can be made much more timely. They can be repeated more often to provide information on rapidly changing or fluctuating variables" [Kish 1987, 5.2].

Table 17.1.1 presents subjective judgments on eight criteria about relative values for three major sources of data, which may be competing alternatives. Though this list is imperfect and incomplete, its use may help to avoid choices based on only one or two criteria. The relative and complementary advantages of samples and censuses are shown (*) to depend on different criteria: samples seem better in five criteria and complete censuses on three others. The importance of the criteria, as well as the relative advantages of the two sources, will depend on and vary greatly with actual situations, countries and times.

The quality of registers is even more variable and extreme differences may exist in their accuracy and inclusiveness for diverse variables and in different situations. Utility records (electricity, water, gas, telephones), tax records, birth and death records may be complete and accurate or bad; and even good records are often out of date for the current location of holders and households. They are inexpensive (***) because other needs pay their costs. But they seldom contain the data needed for agricultural and food surveys and they are included here only for completeness.

17.2 COMBINATIONS OF CENSUSES WITH SAMPLES

Samples and censuses should be viewed not only as competing methods, but also as methods that may be combined to produce better and cheaper data than either method can produce alone. This prospect is strengthened by the complementary advantages noted on several criteria for the two sources in Table 17.1.1 "Therefore the primary objective of a census is typically to obtain a detailed and complete picture of the number (size) and basic structural and related characteristics of the population, and to provide as much detail as possible for small domains and especially for local areas. For example, the population census provides information on the size, age–sex composition, geographic distribution, and basic demographic and socioeconomic characteristics of the population; similarly, agricultural censuses are designed

to provide basic data on structural and related characteristics of agriculture, including numbers of holdings by size, location, type and land—use (Khamis and Alonzo 1975). By contrast, inquiries confined to samples of the population can, by virtue of their smaller sizes, be designed to obtain a wide variety of data for the study of changes and interrelationships" [Kish and Verma 1986].

There are several ways to combine samples with censuses to improve data collection and statistics. And here it would be good to consider the relatively large "sample censuses" as censuses, although they are based on samples rather than on complete censuses. By combining reduced costs and richer data with greater area detail *large sample censuses (e.g., 1 or 10 percent) themselves should be considered as combinations of samples with censuses.* This may be said of the 1970 and 1980 rounds of sample censuses of agriculture [Khamis and Alonzo 1975].

"The census forms the basis for subsequent surveys in a number of ways: by providing the sampling frame; by providing auxiliary information for improved estimation, especially estimation of population totals through regression and ratio estimates; and by mobilizing resources for the development of infrastructural facilities for conducting subsequent sample surveys. Samples attached to the census can also serve as the basis for a programe of continuing surveys" [Kish and Verma 1975]. See (4) in Table 17.2.1.

A. Sampling frame. "Good samples need and are based on census data, especially in countries where alternative sources such as population registers are not available. The population census is the chief source of the sampling frame not only for household surveys covering a variety of demographic, social, and economic topics, including surveys of households and agriculture, but often also for establishment surveys, especially in sectors of small, informal businesses."

The "enumeration areas" or "districts" (EA's or ED's) of censuses serve several functions: to partition the population's total area into small areas with clear, identifiable, and stable boundaries based on maps and descriptions; to

facilitate complete and unique coverage of the units in the population; to create equitable and feasible workloads; to facilitate organization and control of census operations; to provide flexible area units for administrative area statistics at several levels. They also provide a basis for scientific and efficient sample selections for later surveys. They can do this for sampling households and holdings, which are stable enough for area sampling. On the other hand, census addresses and personal identifications are both changeable and confidential, and thus not fit to serve as sampling frames.

EA's can serve as frames best when they are relatively small and uniform in population, also with clear and identifiable boundaries that are available for sample surveys, together with information on population sizes and on basic characteristics. These are needed for measures of size in PPS selection and for stratification, also perhaps for estimation.

B. Estimation. Ratio and regression estimators, and others, can be used to improve sample statistics in combination with census data (12.3). Totals and aggregates can be especially improved with census counts, if the sample statistics like ratio means are based on small samples. The estimation problem in especially acute for small domains, particularly small area domains (14.5).

C. Census bases for continuing surveys. "The census is a major operation which can provide great impetus to the development of national statistical organizations. In addition, large—scale surveys attached to the census can provide a convenient and efficient basis for launching continuing survey programs. Later surveys can be smaller in scale but more varied and complex in content, or they can be specially designed to monitor changes, as for example in multi—round demographic surveys. The larger baseline survey can provide a master sample for more efficient and convenient subsampling for and estimation from the subsequent, smaller surveys. We hope that greater attention will be paid to these possibilities in future rounds of population censuses" [Kish and Verma 1986].

17.3 SAMPLES WITHIN CENSUSES

Five kinds of connections of samples with censuses are listed in Table 17.3.1. Of these set number 4, censuses as auxiliary data for samples, was discussed above (17.2). Joint uses of several sources for postcensal estimates for small domains (5) is discussed later (17.5). Here we note three ways for using samples to improve censuses: to supplement, to evaluate and to better utilize the data of censuses (1, 2, 3). These too brief notes clearly need details and deeper discussions from elsewhere [UN 1971, 1982], and a list is available in [Kish and Verma 1986].

Table 17.3.1
Samples Connected with Censuses [Kish 1979]

1. Sample enumerations to supplement complete censuses:
   (a) Obtain richer, more diverse, detailed, deeper data
   (b) Reduce costs of collection and of tabulation
   (c) Obtain more accurate data, perhaps with special enumerators
   (d) Reduce aggregate social burden on respondents

2. Samples added to complete censuses to evaluate and to improve them:
   (a) Evaluation studies of content (Post Evaluation Studies)
   (b) Coverage checks: dual coverage
   (c) Pilot studies of questions and techniques before the census
   (d) Quality control of individual enumerators, coders, processors

3. Samples from census records, microfilms, tapes:
   (a) Early (advanced, preliminary) tabulation and releases
   (b) Complex, multivariate analyses of relations
   (c) Public use tapes for further, deeper analyses,
   (without identification of respondents)

4. Census as auxiliary data for samples:
   (a) Data for selections: measures of size, stratifiers, maps of enumeration areas; seldom addresses or names
   (b) Data for improved estimation with ratios, regressions
   (c) Samples added to censuses to serve as bases for continuing surveys

5. Joint uses of several sources:
   (a) Current estimates for local areas and small domains
   (b) Rolling (rotating) monthly samples of 1/120 (weekly 1/520)

Perhaps all of these methods have been used somewhere, sometime, but their use is still somewhat sporadic and arbitrary, and this listing may be useful to remind those who are planning a census, whether on a 100 percent or on a large sample basis.

Sample enumerations to supplement censuses. On complete censuses each question is expensive, because it is multiplied by the sizes of the populations (holdings, households, persons, etc.). Hence complete censuses should be kept brief and simple, but more diverse data may be obtained with samples of the entire census. These samples (1:100 or 1:10) can be much smaller than the complete census but still much larger than most sample surveys. The sampling units may be elements (holdings, households) or entire EA's. The timing may coincide with the census or it may be done separately. If done separately, perhaps special teams of enumerators may be hired and trained. These sample supplements not only save costs but they can also obtain richer, deeper data and with higher quality.

Samples added to censuses to evaluate and to improve them. Samples for evaluating and improving the entire census are discussed later (17.4). Quality control, evaluation, and correction of specific individual enumerators need different procedures, because they need individual attention and specific treatments which must be suited to actual field conditions, and to procedures of supervision. The quality control of editors and coders in the office is another matter that is better treated elsewhere. Both of these controls will differ greatly between organizations and situations.

Samples from census schedules. "Whereas in classes 1 and 2 we discussed sampling of the data collected in the field, in class 3 we are concerned with sampling from the already collected census data. There are three distinct purposes for such samples, and their timing differs greatly; hence they need different methods of selection.

"Where early tabulations and releases are wanted, it is convenient to base them on selections of entire EAs (or even administrative districts) in accord with the system of returns from the field collection. The selections should be predesignated and speeded along. They should represent good and valid samples, not merely the first arrivals, which are bound to be biased portions of the population.

"Continuing advances in both statistical and computing methods have made it both desirable and possible to conduct more complex analyses of census data, and demands increase for deeper multivariate analyses of relations. For some of these it is convenient to select samples from the entire census to reduce computations, though this need for sampling may be reduced with faster machines and better programs. The analyses can vary in nature, scope, and timing. They are usually done from tapes in the statistical offices to preserve the confidentiality of the data.

"Public—use tapes are also prepared from census tapes for the use of researchers. Data that could identify individuals are removed from the tapes, and random selections help greatly to prevent identification. Samples of households are preferred for these uses; spreading the sample reduces the level of sampling errors, and it also facilitates the estimation of those errors by avoiding clustering. Households are easier to select than persons, and they provide samples of persons, families, and households. The clustering of individuals in households matters little in analyses, which seldom group multiple members of the same households into the same cells. Such public—use tapes are gaining in use and several countries are preparing them. The spreading availability of computers and related skills is chiefly responsible for this growth. Furthermore, public—use tapes are also being prepared from schedules of old censuses for historical analyses. It is also true (and sad) that the releases of "current" census data may need several years, making their analysis somewhat "historical" for rapidly changing variables" [Kish 1987, 5.3D].

17.4 CHECKS, EVALUATIONS, ADJUSTMENTS, PES

Evaluation surveys can have multiple purposes: not only to check the quality of the census, but also to evaluate, and furthermore even to measure the sizes of errors and its components on various, perhaps all, questions. Those checks and evaluations should serve to improve future censuses. The boldest use of measures of errors would be to use them to adjust the census returns, but that is still not done often.

The sizes of these samples, 1:100 or 2:1,000 or 1:10,000 of complete censuses are usually much smaller than the supplements. Yet the sample sizes (n's) needed for usable results are still too large for attachment to most sample surveys.

The evaluations try to measure two broad kinds of errors: errors of *content* (observation, response, nonresponse) and errors of *coverage* (omission and duplication). The emphasis in evaluation surveys tends to be on systematic biases, though variable errors may also be measured (15.1). Several options must be examined, some of them related, in designing evaluation surveys.

Timing. Post—enumeration surveys (PES) are best known and most common for evaluations. However, evaluations conducted simultaneously with the main census are possible.

Enumerator teams. Either the regular census enumerators, or separate teams may be specially hired and trained. Sometimes either the supervisors or the best enumerators may be used for the PES.

Overlap of the evaluation with the census. The evaluation may be an additional task for the selected sample of respondents, or it may replace the ordinary census enumeration in the sample areas.

Independence of evaluation observations. If the evaluation is additional to the census enumeration, the PES enumerators, if they not the Census takers, may be kept ignorant of the Census answers; or they may have them in order to check and improve on them.

Combining evaluation of coverage and of content. These may be combined when entire EA's are selected for the evaluation surveys.

Design and size of the evaluation surveys. If these must serve for quantitative measures of the biases, and especially for adjustments, they must be based on probability samples. If funds are lacking and only informal checks of quality are wanted, then restricted areas may be satisfactory. In these cases, perhaps particular emphasis may be put on the presumed most difficult, critical areas of greatest possible biases.

"*Pilot studies* are required to test the adequacy of census questionnaires, instructions, training programs, enumeration procedures, field organization, etc. They serve as practical training for the nuclear staff and supervisors, and provide information on operational aspects (costs, time) of enumeration. For pilots, it is usually difficult to insist on good samples of the entire country: the common practice is to choose areas which are convenient but also expected to yield a good test of questions and techniques in diverse circumstances" [Kish and Verma 1986].

17.5 POSTCENSAL ESTIMATES FOR SMALL DOMAINS

Census data are usually obsolete, data from registers inadequate, and sample data lacking in detail, especially for local areas. Since the strengths and weaknesses of the three sources are complementary, it seems reasonable to try to combine the strengths of the three sources to obtain estimates for small domains, especially for local areas; estimates that are current, pertinent, and accurate. To the general needs of researchers have been added the needs of social planners, of administrators, and of policy makers for valid, current data for small domains and local areas. *Local area estimation* has become a

fast—developing field, being pushed by increasing demands, and simultaneously pulled along by new developments in computing technology and new statistical techniques. These problems of "postcensal estimates" are treated currently as technical problems for estimates of the total population in small local administrative areas, with a new, large, but specialized list of publications; a few references can be the key to the longer list [Purcell and Kish 1979, 1980; Heeringa 1982; Platek, Rao, Sarndal, Singh 1987].

REFERENCES

Brewer KRW and Hanif M [1983], Sampling with Unequal Probabilities, New York: Springer-Verlag.

Cochran WG [1977], Sampling Techniques, New York: John wiley and Sons, 3rd ed.

Dalenius T [1957], Sampling in Sweden, Stockholm: Almquist and Wicksell, Ch. 9.

Darroch JN [1958, 1959], The multiple recapture census, I and

Duncan GJ and Kalton G [1986], Issues of design and analysis of surveys acrosss time, Int. Statistical Rev., 54, II, Biometrika, 45, 343–59 and 46, 336–51.

El-Khorazaty et al [1977], Estimating the total number of events with data from multiple record systems, Int. Statistical Rev, 45, 129–57.

FAO [1977], Report on the 1970 World Census of Agriculture,Rome: FAO.

FAO [1978a], Taking Agricultural Censuses, Rome: FAO, Economic and Social Development Paper No. 1.

FAO [1978b], Collecting Statistics on Agricultural Population and Employment, Rome: FAO, Econ. and Soc. Devt. Paper No. 7.

FAO [1982], Estimation of Crop Areas and Yields, Rome: FAO, Econ. and Soc. Devt. Paper No. 22.

FAO [1986], Food and Agricultural Statistics in the Context of a National Information System, Rome: FAO.

Gonzalez ME et al [1975], Standards for discussion and presentation of errors in survey and sample data, JASA, 70 (part II), 5–23.

Groves RM et al [1988], Telephone Survey Methodology, New York: Wiley and Sons.

Groves RM and Kahn RL [1979], Surveys by Telephone, New York: Academic Press.

Groves RM and Lepkowski JL [1985], Dual frame, mixed mode survey designs, Journal of Official Statistics, 1, 263–86.

Hansen MH, Hurwitz WN and Madow WG [1953], Sample Survey Methods and Theory, Vol. I, New York: John Wiley and Sons.

Hartley HO [1962], Multiple frame surveys, Proceedings of the Social Statistics Section, Am. Stat. Assn., 203–206.

Hess I [1985], Sampling for Social Research Surveys, Ann Arbor, MI: Inst. for Social Research, x + 294.

Hess I, Riedel DC, and Fitzpatrick [1975], Probability Sampling of Hospitals and Patients, Ann Arbor, MI: Health Administration Press.

Hiridoglou MA and Srinath KP [1981], Some estimators of a population total containing large units, JASA, 690–95.

Kalton G and Anderson D [1986], Sampling rare populations, JRSS(A), 149, 65–82.

Kalton G [1983], Compensating for Missing Survey Data, Ann Arbor: Institute for Social Research.

Kalton G [1979], Ultimate cluster sampling, JRSS(A), 142, 210–22.

Khamis SH and Alonzo DC [1975], Changes in methods, scope and concepts in the 1980 World Census of Agriculture, Bulletin of the Int. Statistical Inst, 45(2), 54–82.

Kish L [1977], Robustness in survey sampling, Bull. of Int. Stat. Inst., 47,3, 515–20.

Kish L and Frankel MR [1974], Inference from complex samples, JRSSS(B), 36, 1–37.

Kish L [1965], Survey Sampling, New York: John Wiley and Sons.

Kish L [1961], A measurement of homogeneity in areal units, Bull. Int. Stat. Inst., 3rd ed, Vol. 4, 201–209.

Kish L and Scott A [1971], Retaining units after changing strata and probabilities, JASA. 66, 461–70.

Marks ES, Seltzer W and Krotki KK [1974], Population Growth Estimation, New York: The Population Council.

Madow WG, Olkin I and Rubin DB [1983], Incomplete data in Sample Surveys, 3 Vols, New York: Academic Press.

Murthy MN [1967], Sampling Theory and Methods, Calcutta: Statistical Publishing Society.

Platek R, Rao JNK, Sarndal CE, and Singh MP [1987], Small Area Statistics, New York: John Wiley and Sons.

Purcell NJ and Kish L [1980], Postcensal estimates for local areas (or domains), Inst. Stat. Rev., 48, 3–18.

Rossi PH, Wright JD and Anderson AB [1983], Handbook of Survey Research, New York: Academic Press. Ch 2 Frankel MR, "Sampling Theory."

Sanchez-Crespo JL [1977], Selection with graduate variable probabilities with replacement, Bull. Int. Stat. Inst., 47,4, 458–61.

Sirken MG [1970], Household surveys with multiplicity, JASA, 65, 257–66.

Sukhatme PV and Sukhatme BV [1970], Sampling Theory with Applications, 2nd ed., Rome: FAO, and Ames: Iowa State U. Press.

U.S. Census Bureau [1978], The Current Population Survey: Design and Methodology, Technical Paper 40.

United Nations Statistical Office [1980], National Household Survey Capability Programme, UN ST/ESA/STAT. 92/Rev 2.

United Nations Statistical Office [1950], The Preparation of Sample Survey Reports, UN Series.

Verma V, Scott C and O'Muircheartaigh C [1980], Sample designs and sampling errors for the WFS, JRSS(A), 143, 431–473.

Wright T and Tsao HJ [1983], A frame on frames: an annotated bibliography, in T. Wright, Statistical Methods and the Improvement of Quality, Orlando, FL: The Academic Press.

Yates F [1981], Sampling Methods for Censuses and Surveys, London: Griffin & Co., 4th ed.

Zarkovich SS [1965], Sampling Methods and Censuses, Rome; FAO.

Zarkovich SS [1963], Quality of Statistical Data, Rome; FAO.