Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Cumulating/Combining population surveys

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canadä

by Leslie Kish

December 1999

Statistics Canada Statistique Canada

Canadä

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 129-138
Statistics Canada, Catalogue No. 12-001

Cumulating/Combining population surveys

Leslie Kish [1]

Abstract

Designs for and operations both of multipopulation surveys and of periodic surveys have become more common and important. The needed large resources, both financial and technical, have been organized only in recent decades, and the great values of both became recognized. For both types of designs the developments have concentrated on comparisons between surveys. Yet the coordination and harmonization needed for comparisons also makes the combinations of the survey statistics possible, desirable, and practiced. But the combinations of surveys have been achieved and presented largely without a theoretical/methodological framework, and often poorly. Here such a framework is attempted. Some closely related designs are also discussed: multidomain designs, rolling samples, combining experiments, and combining several distinct survey sites.

Key Words: Multipopulation design; Multidomain design; Periodic surveys; Rolling samples; Combining experiments.

1. Introduction: Multipopulation models

A paraphrase of the standard model in all books on survey sampling goes roughly thus: "The aim of survey samples is to produce an estimate of the total Y (or the mean $\bar{Y}$) for a variable Y_i in a population of N elements." Such statements are misleading because they fail to describe the actual purposes and practices of survey sampling. First, most surveys treat many variables, and second, survey results use diverse kinds of statistics; thus sample surveys are "multipurpose" on several dimensions (Kish 1988). But instead of discussing all the omissions of the standard model, I want in this paper to concentrate only on its insufficiency and inadequacy because of its restriction to a single, finite population. Among the several examples below of multi-population expansions that are possible with a new and different model, I begin with two important examples of survey samples that achieve a variety of treatments and results on different dimensions. First is the emergence of multinational designs since 1965, best illustrated by the World Fertility Surveys (section 3), which involve combinations across national spatial boundaries. Second are combinations of periodic samples, best illustrated by "rolling samples" (sections 4 and 5), which concern combinations across temporal dimensions.

The designs and operations for periodic surveys require large resources and new methods. Those for multinational surveys are even more demanding. Both of these types of complex surveys are rather late arrivals among sample surveys and both types are growing in numbers and in importance. Furthermore, both types have been designed and used mostly for comparisons: temporal and spatial comparisons, respectively. The concept of using them additionally for combinations and cumulations is new, and is often encountered initially with doubt and disbelief (sections 3, 4, 5). For both types, the variations between the populations are commonly affirmed as obstacles to combinations or cumulations, and thus are then used for restricting the sample estimates to single populations, because typically methods for combining them are unknown or unavailable. Or even when they are combined, only *ad hoc* methods are used, without justifying them. References to several papers indicate my concern for designs of multinational surveys and of rolling samples. In this paper the emphasis will be on combinations for multinational samples and on cumulations of periodic and rolling samples.

You may notice that I use the terms "cumulating" and "combining" interchangeably and perhaps confusedly. "Combining" seems to fit the multipopulation and multidomain situations better, whereas "cumulating" seems better for periodic and rolling samples. It would be better to have one word to cover both spatial and temporal combinations/cumulations, but neither seems to be exactly right. Also "combinations" serves uses other than joining populations – the usage I wish to emphasize here.

I am also not clear if it is better to consider the enlargement of the scope of samples from one population to several as a new model or as a paradigm shift. Discussions with a few philosophers here left me confused about this choice. And my fellow statisticians probably do not care whether we write the word model or paradigm. In any case, a new model instead of the standard model of sampling from a fixed frame of a stable, finite population is the radical proposal I am pursuing in this paper.

2. Multidomain designs

Statistics for national samples are commonly based on combinations of domains, and these are often quite diverse. But because these combinations are simple and familiar,

[1] Leslie Kish, ISR, University of Michigan, Ann Arbor, MI 48106.

they can also serve as heuristic examples for the less familiar combinations I want to discuss, such as multinational and multiperiodic statistics. The diversity of domains may be recognized within national sample designs; *e.g.*, provinces, which may number from 5 to 20 in most countries. In samples of smaller populations (cities, institutions, firms, *etc.*) similar partitions into major domains also are typical. But for smaller and more numerous domains (*e.g.*, the 3,000 counties of the USA) deliberate sample designs are not feasible for most samples of limited size. For these small domains, methods of "small area estimation" have been developed (Kish 1987, 2.3; Platek, Rao, Särndal and Singh 1987 pages 267-271). There are great practical differences in both design and estimation between large and small domains, and it is careless to use the adjective "subnational" to cover both. Furthermore, these distinctions between large and small domains exist not only for national designs, but also for samples of smaller populations also. It seems that the structured (nonrandom, grainy) natures of populations persist also on smaller scales. This conforms to the proposed new model of populations, and is supported with empirical analysis of multistage components (Kish 1961).

Although practical for provinces, deliberate designs are not feasible for most domains, whether few and large or many and small, of the kind we call "crossclasses," such as sex and age or occupation, social class, education, *etc.* These "crossclasses" are often important both for their relations (correlations) with the survey variables and for their great diversity. Thus samples of national (or other) populations are mosaics of domains that are diverse and often highly variable; and we must depend on the properties of large probability samples to yield reliable representations of them. In this sense we perceive that all population samples consist of combinations of subpopulations.

Subclasses designate the representations in the entire sample of the domains that compose the whole population. Crossclasses are commonly the most common types of subclasses in survey analysis: partitions of the sample, for which deliberate selection designs are not feasible. For example, occupation and education classes, behavioral and attitudinal categories, and so on. These can be strong explanatory variables for survey analysis; yet we lack the data and resources not only for pre-stratification, but also even for post-stratification methods. From that extreme of lack of controls at one end, we can move to the other extreme of strong controls by separate samples, which can be designed for major provinces.

For example, different methods of sampling can be used in the different provinces. But more common are designs that use different sampling rates; for example, higher sampling rates for small, or for especially important provinces. Sometimes equal sample sizes $n_f = n / H$ are designed for all H provinces in order to obtain (approximately) equal precisions for all provinces, regardless of their sizes. This equal allocation results in sampling fractions n_h / N_h that are inversely proportional to province sizes. But for fixed total sample size n the consequences are higher variances for the entire sample, as well as for crossclasses; see section 8 (Kish 1988). We assume here, that the statistics $\bar{y}_h$ of the provinces (domains) get weighted with population weights $W_h = N_h / \sum N_h$ for the overall statistics $\bar{y}_w = \sum W_h \bar{y}_h$, as is commonly practiced for national statistics. This serves as a useful introduction to the multinational statistics coming next.

3. Multinational sample designs

National "representative" samples were started by Kiaer (1895) only in 1895 and, after much opposition, they became widespread only after 1945 (Kish 1995). Since then the efforts of the samplers were encouraged and supported by statistical agencies of the United Nations, especially the UN Statistical Office and the FAO. Their spread then led naturally to multinational comparisons of surveys; yet the deliberate design of multinational samples that could provide valid comparisons is recent, starting only around 1965 (Szalai 1972; World Fertility Surveys (WFS) 1984; Kish 1994). The new demands for survey designs for multinational comparisons create many new difficulties: in resources – financial, institutional, cultural; and also in methods. Those difficulties encountered with comparisons reappear also in similar form for multinational combinations, our main concern in this paper.

It is interesting to compare these difficulties with ones with which we are familiar in multidomain designs. From a theoretical perspective, combining the provinces of a country is similar to combining the nations of a continent. Indeed we should profit from those similarities by using metaphorical arguments from the familiar multidomain designs to the proposed multinational combinations. However, from a practical view we find great differences between the two efforts because of five fundamental practical obstacles that make multinational designs much more difficult to achieve, discussed below.

1. The centers of decisions reside in separate national offices, both for setting policy targets and for obtaining funds. Further, within any nation the agencies for policy setting and for resource allocation may be distinct and separate; *e.g.*, the Education Ministry may share participation in a school survey, but the Parliament or the Finance Ministry may fail to allocate funds.

2. The needed technical resources reside in and are staffed and developed by separate national offices. These separate offices may have very different levels and types of technical development, as well as distinct organizational structures and different social connections.

3. The survey variables can vary immensely across national boundaries, due to different cultures, religions, economic and educational levels, legal and social relations, *etc.* Achieving comparable results demands immense efforts – but the task is not impossible, as multinational surveys have shown.
4. The crossnational translation of concepts and of questionnaires, also of codes and analysis, are daunting challenges that need ingenuity, knowledge, and devoted effort.
5. Separate samples must be designed and operated to meet distinct national conditions, with local resources, sampling frames, and field operations. This subject needs volumes; more discussion and study than is possible here.

Multinational comparisons probably go back many years, based on diverse kinds of observations – by travel, wars, conquests, *etc.* But probability sample surveys of entire nations have become common over all continents only during the past half century. As the second phase of development, those national surveys soon led to multinational comparisons. The third phase of deliberate multinational designs dates only from 1965: the Time Use Surveys of 1965 (Szalai 1972); the World Fertility Surveys of 1972-82 (WFS 1984; Cleland and Scott 1987); the Demographic and Health Surveys since 1985 (DHS 1991); the Labour Force Surveys of the European Community (Verma 1992, 1999); see Kish (1994). Other multinational survey designs are also emerging, with the funding and technical resources increasingly meeting the growing effective demands. I am heartened and amazed at the emergence of the International Surveys of Psychiatric Epidemiology, a field that I had feared was beyond the reach of probability surveys in my lifetime! (Heeringa and Liu 1999).

Now, for the new fourth phase, I propose deliberate designs for combinations of multinational surveys. Multinational combinations of surveys are now being produced and published; *e.g.*, European unemployment rates or birth rates; African or Sub-Saharan birth rates or death rates; world growth rates; and many other rates, means, and totals. The data for each nation may be based on probability samples (phase two), or even designed for multinational comparisons (phase three). But the methods used for combining them seem to be completely *ad hoc*; and current usage for the relative weights for combining national statistics seem to be in order of A, B, C, D, E from most common to the least. I made no actual counts, nor an empirical study, but glaring examples appear weekly. Very often the methods and weights for combining the national samples are not even mentioned in the media, even in respectable and scientific journals. To the contrary of the above order, our preferences may be almost in the reverse order of E, F, D, C, B, A – and very much depending on the situation, sample sizes, *etc.*

Allow me, with due modesty, to propose that phase five should be the development of solid theory for choosing among those preferences, and also others. But the need for methods for combinations cannot wait for the future better theory; and it is usual in statistics (and in the sciences) for practice and methods to develop before, and thus both to precede and to stimulate theory. Meanwhile, the discussions below may lead to some improvements in methods, even if they are not quite “optimal.”

Here then follow six possible alternative ways and weights for combining national statistics.

A. Do not combine: publish only separate national statistics. This is the most common treatment for several reasons. 1. The authors have not thought of the possibility or need for combination, or rejected them. 2. Perhaps they could not decide on the “best” method, and wanted to leave that to the reader, user, customer. This may be defended by “caveat emptor,” or “Bayesian” arguments. However, I reject them. The authors should do no worse in choosing than the average users – who in any case can reject the authors’ combination if the national statistics are also published. I believe that when the reader’s eye roves over the usual horizontal (or vertical) bars in graphs or over data in tables, it tends to yield a simple mean, hence this roving reduces Method A to Method C in effect. This tendency can perhaps be improved if the width of the bars is made proportional to population weights.

B. Even in the absence of combining populations, designs for multinational comparisons should be “harmonized” in survey measurement methods, to allow for proper comparisons (Kish 1994).

C. Use equal weights $(1/H)$ for every country. This method is also common and also avoids (like A) the difficult questions of how to choose population weights W_h, with h denoting country. Probably its use is seldom based on deep reflection, but is widespread mostly because it appears to be a “common sense” approach. Perhaps it would be justified with models, where the between-country variation is paramount, and the population sizes are not relevant. However, I have no faith in such models.

D. Weight with sample sizes n_h. Thus $\bar{y}_w = \sum n_h \bar{y}_h / \sum n_h$, which results automatically from simply cumulating sample cases from separate countries, or sites, or surveys. This is also done frequently, and can be justified when elements are drawn essentially from the “same population” or when per-element variance is the only (or prime) component of variation. It denotes “cumulating cases,” as distinguished from combining statistics (Kish 1987, 6.6). This approach can be extended to situations where there are serious differences of element variance due to “design effects”; and then “effective sample sizes” n_h/deff_h may be substituted for the n_h. The "effective sample size" may also be applied if the σ_h^2 differ between populations in order to use weights with precisions $n_h/\sigma_h^2 = 1/(\sigma_h^2/n_h)$. In most situations, however, the variations in sample sizes n_h depend on arbitrary, haphazard factors; and C may be a worse choice than using equal weights $1/H$ for all countries (surveys, sites).

E. Use population weights W_h. Thus $\bar{y}_w = \sum N_h \bar{y}_h / \sum N_h$ and $W_h = N_h / \sum N$. This method has the most commonly understood meaning when the N_h represents total numbers of persons in population h. However, sometimes the population content may be quite different. For example, for grain (or wine) production it may be total number of farmers, or wheat (or grape) farmers, if those numbers are available, or can be estimated; these populations may yield potentially interesting meanings either for comparisons or for combinations. The population extent also needs to be determined; for example, all persons, or only adults, or only women, or only married women; only urban or rural, or both? Also the timing (date) of the surveys needs standardization, *e.g.*, censuses are conducted in '0 (or '9, or '1) years. Often the population weights are not persons, but acres of land, or tons of steel, or barrels of oil, and so on.

F. Use post-stratification weights. Often in multipopulation situations we encounter the same problem as described later in section 7 for multiple sites. And we may consider the same hierarchy of alternative treatments, the last of which (F) is using "post-stratification" weights. We may well have comparable surveys from several diverse countries of a continent (or the world), but neither all the countries, nor a probability sample of them. (For example, the African or South American countries in the World Fertility Surveys or the Demographic Health Surveys.) One may think of constructing "pseudo-strata" from which the available countries would be posed as "representative selections." Some one stratum could have only a single, available, large country. Another stratum could have 2 (or 3) countries, but with only one available representative that would get the weight of all 2 (or 3) countries. This artificial "pseudo-stratification" procedure may be preferable to simply adding up the available countries into an artificial combination with $W_h(E)$ or with $1/H(C)$. The rationale for this preference is not very different from methods of adjustments for nonresponses.

Several questions and decisions remain concerning the choice among alternative weights. First: the choice should be made chiefly on substantive grounds. What must the combination represent mainly? My own preferences tend strongly toward D and E, and I deplore the prevalence of A and B that we encounter daily. However, I cannot support my preferences on technical grounds. Also I have faced grave problems with the extremes posed by the giants China and India, each more like a continent, and neither solutions E or C seem adequate. I advise defying the geographer's classification of Asia and leave both of them out of Asia, considering them as separate entities. For example, I have omitted all four countries greater than 200 millions in total population (including the USA and USSR) in my computations in 1970 (Kish 1976, Table 4; Kish 1987, 7.3D).

Second: Is the bias due to using incorrect weights important? This would be difficult to prove, as the bias is a function of correlations between the weights and specific survey variables. However, the proof should belong to the denial, as it does with the biases of nonresponses or of poor sampling methods. Ignorance of sources of bias does not imply their absence. I believe that using equal weights instead of population weights can often lead to important biases.

Third: When samples are (roughly) equal-sized, weighting up to population sizes can greatly increase variances. These increases in variances due to unequal weights can be measured quite well (see section 8). They should be balanced against probable biases in models for reducing mean-square errors. In small samples the large variances may dominate the MSE.

Fourth: It seems clear that the combination of population surveys into multipopulation statistics needs a good deal of research, both empirical and theoretical – and especially together.

4. Cumulating periodic surveys

Periodic surveys have been designed and used mainly for measuring periodic changes, and also for "current" estimates, exploiting the advantages of partial overlaps. But here we shall explore their design and use for cumulated estimates. Furthermore, I include periodic surveys here in order to emphasize their basic similarities to surveys combined over space, such as multidomain and multipopulation surveys. We cannot enter here into the philosophical issues involved in repeated studies of the "same" population, except to note that the "stability" of any population differs greatly for diverse variables (Kish 1987, chapter 6); and the stability for any one variable will also differ greatly, depending on the length of the periods, which may be weekly, monthly, or quarterly. These are common and useful man-made periods. But there exist only two global "natural" cycles of variations: the diurnal and annual cycles, based on the earth's rotation around its own tilted axis, and around the sun.

I must note four practical, rather than theoretical, differences between cumulating periodic surveys and combining multinational or multidomain surveys.

1. Periodic surveys are designed for the "same" population, which tends to retain some stability between periods. The "sameness" and "stability" are only

relative, and with many exceptions; *e.g.*, epidemics in health data or fluctuations in stock prices. They differ greatly between variables and decrease for longer periods.

2. These stabilities imply positive correlations between periods, encouraging designs with "overlapping" sampling units in order to reduce both unit costs and variances for estimates of change and of current values. These overlaps are not desirable for cumulations, so this conflict between the two designs must be resolved.
3. Because similar methods and designs are feasible and generally preferred, they are used over all the periods; on the contrary, harmonization of methods is difficult to achieve between national samples. I emphasize here cumulating periodic surveys, but these aspects also apply to comparisons.
4. Methods for periodic surveys for comparisons have been widely published, in contrast to the novelty both of multinational designs and of periodic cumulations.

There now exist several cumulated representative samples (CRS) of national populations: samples designed for cumulations over large populations. These remain restricted within selections of primary sampling units in order to reduce field costs, whereas "rolling samples" (section 5) are spread deliberately over all sampling units in the population. The Health Household Interview Surveys (HHIS) of the USA are separate weekly samples of about 1,000 households, cumulated yearly to 52,000 households (National Center for Health Statistics 1958, pages 15-18). These samples are selected by the US Census Bureau within their large sample of PSUs. The yearly samples of over 150,000 persons constitute a remarkable example of multipurpose surveys, representing even rare diseases. The Australian Population Monitors have quarterly nonoverlapping samples that are cumulated to yearly samples, and these are also confined into primary sampling units (Australian Bureau of Statistics (ABS 1993)). The new Labour Force Surveys of the United Kingdom publishes each month the cumulation of three separate, nonoverlapping monthly samples (Caplan, Haworth and Steele 1999). There are other examples as well, and the applications of cumulative representative samples (CRS) are increasing in scope and diversity, although until now they have lacked a common name and literature. Nevertheless, I propose to differentiate the CRS from rolling samples for practical reasons (section 5).

Two problems and methods associated with cumulated samples deserve brief mentions, but with references to more adequate treatments. Asymmetrical Cumulations refer to proposals and some actual practices of reporting large aggregates frequently, but reporting on small domains only after cumulating over longer intervals. For example, the HHIS above may report some national averages each week or monthly, but smaller regions, or specific diseases, only for annual aggregates (Kish 1990).

A serious conflict can arise if periodic samples are to be used (as they should) both for measuring periodic changes and current levels and for measuring cumulations over the periods. This double use has been proposed and practiced, although I do not yet know of any deliberate double designs. Most periodic surveys use partially overlapping samples with some kind of rotation design. One reason often given for these overlaps is the reductions in variances per sample element both for measuring changes between periods and for making current estimates. These reductions depend on positive correlations between the overlapping sampling units. Such reductions are well documented in sampling textbooks and articles since the original papers on this topic (Jessen 1942; Patterson 1954). But even greater reductions are possible in element costs, when the later interviews are much cheaper than the first contacts; for example, if the later contacts are by telephone. On the other hand, separate new samples will be much preferred for cumulations in order to avoid the positive correlations. One may imagine different compromises that may be efficient, when: (a) most of the positive correlations are not high; (b) reinterview costs are not much cheaper; and (c) reinterview response rates are discouraging.

However, consider also a new design that I call a Split Panel Design (SPD) that adds a panel *p* to a parallel series of nonoverlapping samples *a-b-c-d etc.*; with the combination then denoted as *pa-pb-pc-pd etc.* The panel replaces the overlaps of rotating designs and provides the useful correlations for measuring net (macro) changes. Further, it also serves to measure individual (gross) changes, which are lacking in the usual designs of overlapping sampling units, because of the mobility of persons and households. Including panels of individuals (persons, elements) would bring considerable advantages for SPD over all current overlapping samples, which usually use merely the same sampling units (Kish 1987, 6.5; Kish 1990).

Another considerable advantage of SPD is that these overlaps would be based on the correlations from all periods, rather than only for the arbitrarily chosen periods for the rotation designs. How arbitrary are these? Some decisions use one-month groups, some three months, others 12 months, *etc.*, *etc.* It is most unlikely that these disparate overlaps are actually "optimal" for those countries. It seems most likely that the "optimal" overlap cannot be predetermined for any single variable, and a single optimal period is even less likely for multipurpose designs.

5. Rolling samples and censuses

These should be considered as special types of the related cumulative representative samples (CRS); but rolling samples (RS) should be distinguished, because they are designed for different and specific functions. CRS have been confined to designs of PSUs. They are spatially restricted for cost reasons and for fitting the designs of labor force surveys, and other surveys associated with them. However, RS designs must aim at a much greater spread in order to facilitate maximal spatial range for cumulations over time. Rolling samples must be designed specifically to readily yield good estimates for all small spatial units, when the periodic samples are cumulated into annual or decennial larger samples or censuses.

First let us define a rolling census: it consists of a combined (joint) design of F separate (nonoverlapping) periodic samples, each a probability sample with fraction $f = 1/F$ of the entire population, and so designed that the cumulation of the F periods yields a detailed census of the whole population with $f' = F/F = 1$. Intermediate cumulations of $k < F$ periods should yield rolling samples with $f' = k/F$ and with details intermediate between 1 and F periods.

Imagine a weekly national sample, each designed with epsem selection rates of $f = 1/520$. The cumulations of 52 such weekly samples would yield an annual sample of $52/520 = 10$ percent. Then ten of these annual samples would yield a census of 520/520. I have proposed in several papers to have these rolling samples replace both kinds of the most important forms of official statistics that are either used or planned in many countries: the monthly surveys of population and labor force and the decennial censuses. Even more important, these surveys could also provide annual detailed data, perhaps with 10 percent samples, which are badly lacking, and needed in many countries (Kish 1990, 1997, 1998). Providing spatially detailed annual statistics for a variety of economic and social variables, not a mere population count of persons, would be the chief aim of rolling samples in many countries. These are needed even in countries that can provide fairly good estimates of population counts and a few simple statistics either from registers or with estimation methods. In countries without good frequent (monthly or quarterly) surveys of labor force and population, rolling samples could also serve them as efficient vehicles.

I must admit that the above basic ideas provide merely the skeleton for any actual national design for rolling samples. But such actual national samples have been recently designed – the largest and best of which is the American Community Survey (ACS) – now undergoing a 37-area pilot study by the US Census Bureau (Alexander 1999). This aims to provide monthly surveys of 250,000 households and detailed annual statistics based on 3,000,000 households, after year 2003; and also to provide quinquennial and decennial census samples later. The National Statistical Office of France is working on plans for a Census Continué (Isnard 1999). The Labour Force Surveys of the United Kingdom are now based on cumulated monthly surveys. Some other countries are examining different but generally similar possibilities.

It is also proper to add references to two early publications describing "rolling samples" of large sizes, although not national in scope (Mooney 1956; Kish *et al.* 1961). Others probably exist that I have not seen.

How to cumulate periodic surveys? This topic must receive serious technical consideration in the future, because so far they have been done only with *ad hoc* procedures. Perhaps for cumulating over a single year, epsem samples with the same sampling fraction f, and simple cumulation of cases may serve as a simple model: averaging over seasonal and random variations may outweigh secular trends. However, averaging annual statistics over 10 years may have to consider secular trends in population size.

Consider several alternative sets of weights W_i to be assigned to yearly means $\bar{y}_i$ for a decennial mean $\bar{y}_i = \sum W_i \bar{y}_i (i = 0, 1, 2, ..., 9)$ and $\sum W_i = 1$.

a) $\bar{y}_{ta} = \bar{y}_9$, with $W_9 = 1$ and the other nine $W_i = 0$, utilizing only the final year. This could be used for national and large domain estimates, and for highly fluctuating variables (unemployment, epidemics, stock prices), where the need for timeliness dominates sampling precision.

b) $\bar{y}_{tb} = \sum W_i \bar{y}_i$, with all ten $W_i = 0.1$. For variables without time trends, and for small domains, obtaining a stable average over time may be good strategy.

c) $\bar{y}_{tc} = \sum W_i \bar{y}_i$, with $W_0 \le W_1 \le W_2 \le ... \le W_9$, monotonically increasing (or nondecreasing) W_i. The curve of increase may be determined with a model or with empirical data. Thus $\bar{y}_{ta}$ and $\bar{y}_{tb}$ may be viewed as two extremes of $\bar{y}_{tc}$. They all seem better than the present practice of giving full weight $W_0 = 1$ to a decennial census that may be from 1 to 10 years old and obsolete.

Furthermore, with rolling censuses, the statistical office need not wait to publish only decennially. It can publish annually the results of the latest rolling samples, with several available alternatives from those above: either the latest year $\bar{y}_{ta}$; or $\bar{y}_{tc}$ an average that favors the latest years. Or "asymmetrical cumulations" favoring $\bar{y}_{tb}$ for smaller domains, but $\bar{y}_{ta}$ for larger domains and totals. It could conceivably publish both $\bar{y}_{ta}$ and $\bar{y}_{tb}$ and let the reader choose (perhaps publish electronically). Clearly technical research will be needed to search for "optimal" solutions to support the applications already appearing.

6. Combining experiments

A) This topic has been the subject of three early and good papers by Cochran and has also received attention from both Yates and Fisher at Rothamsted (Cochran 1937 and 1954; Yates and Cochran 1938). These dealt with experiments relating crop yields (predictands) to fertilizers (one or more predictors), conducted over different populations, fields, and years. They used ANOVA methods for statistical analyses and for combining the several independent experiments.

B) Fisher's test for combined probabilities, from 2×2 Chi-square tests of the "same" null hypothesis is even older. It can use entirely different populations, and even diverse variables, for testing the "same" null hypothesis. This well-known test can be found in most statistics textbooks.

C) Methods of meta analysis are newer, and increasingly used. They combine experimental results from different samples and populations for the same predictand (outcome) variable from one or more predictors (inputs). (Glass 1976, Hodges and Olkin 1985.)

Methods for combining sample surveys are just emerging, much later than methods for combining experiments. The two fields, however, have many similar aims, which should be noticed, in order to see useful relations between the two distinct topics. Perhaps these relations can be best perceived by looking at the differences between the aims and the problems that have been the subjects of the two methods. There seem to be three main differences between the two methods, as they have been applied.

1. Combining surveys (CS) needs a great deal of advance preparation, planning, and coordination. This is true of multinational surveys for both the comparisons, which have been already achieved, and for their combinations, which are new. For national multidomain surveys the coordination comes naturally, but for multinational surveys the coordination of the separate national designs is difficult, but necessary (Kish 1994). On the contrary, a great virtue of combined experiments (CX) is that they can be performed on the reports of experiments already performed, as the name meta analysis signifies. That analysis is based on the relations of the predictand/predictor pair of experimental variables. The Fisher test needs only the probabilities P_i achieved by the tests of significance.
2. The second difference between the two methods is related to the first. The CX are based on experiments, whereas CS concentrates on surveys. Thus CX emphasizes experimental control through randomization of variables over subjects. However, CS are based on probability sampling with randomized selections of subjects – not variables – from defined populations. Usually these two kinds of randomizations are difficult to achieve in any research study and one must be sacrificed (Kish 1987, 1.1). The population base of CS is specified, whereas those for CX usually are not and cannot be.
3. Third, CS involves a full statistical analysis, and even a full survey method, designed for similarity and comparability in order to facilitate the joint analysis. On the contrary, the methods of CX can use the very end of the statistical analyses, often even from published statistics. The extreme of this kind of abstraction is shown by the combined Fisher test, based only on the terminal P_i values of the separate statistical tests.

Because of the large, consistent and interrelated differences between Combined Experiments and Combined Surveys, it may be best to keep the two methods separate. Some may propose that the gap between the two subjects is only an historical accident and that the gap can be closed sometimes. But I believe that it is more useful to maintain the separation of the two methods, even if sometimes a compromise may be usefully adopted.

That still leaves open the question whether the three methods of combined experiments (A, B, and C above) should be called "Combined Experiments," as Cochran, Yates and Fisher called them since the 1930s or if it is better to distinguish them all as "Meta-Analysis," now a widely known and accepted joint designation. Happily we need not decide here, but perhaps meta-analysis is the best, provided we also recognize the earlier successes.

7. Combining separate sites

Suppose that similar data have been collected in several sites of a combined population, but not in all of the sites, nor in a probability selection of them. The sites may be cities, provinces, or districts of one country. Or they may be institutions, such as schools, or hospitals, or factories. Or the sites may even be entire countries of a continent. I have seen a variety of such situations when the sites are either chosen arbitrarily, or are simply "volunteers." Often the sample sizes per site are similar, though the population sizes of the sites vary greatly. Here follows a list of possible alternative treatments of the data.

A. Separate survey estimates $\bar{y}_i$ may be presented only. Usually this is all that is done, especially if the data have not been coordinated, or "harmonized." Any comparisons and any combinations of the separate statistics are left to the readers, to use their own methods or resources.

B. Comparisons between the separate sites require harmonization (of variables, measurements, timing, populations) to render the differences $(\bar{y}_i - \bar{y}_j)$ meaningful.

C. Simple cumulations $\bar{y}_t = \sum y_i / \sum n_i$ of all sample cases amount to assuming that the populations N_i of the sites can be considered parts of the same population of $\sum N_i$ elements. Note that the sample means $\bar{y}_i$ are weighted by the sample sizes n_i. Often these are nearly equal and then C approaches D.

D. Equal combination $\sum \bar{y}_i / k$ of k sites weight each of the sites equally, disregarding both the sample sizes n_i and the population sizes N_i.

E. Weighted combinations $\bar{y}_w = \sum W_i \bar{y}_i / \sum W_i$ weight the sites with some measure of their relative importance.

Population sizes N_i seem reasonable, but others may be used. However, we may object to the combination of an arbitrarily selected set of sites.

F. Post-stratification weights $W_i \propto \sum_j N_{ij}$ can save attempts to overcome the above objections by constructing pseudostrata $\sum_j N_{ij}$, composed of "similar" sites, from which the unit N_i may be considered a valid selection. Thus the total sample then is considered a sample from the larger population of total size $\sum_i \sum_j N_{ij}$. Such model building resembles the attempts to reduce nonresponse bias with nonresponse classes.

Three sets of decisions must be made, and this order is chronological in activity, but not necessarily in planning. a) The allocation of sample sizes, especially whether equal sizes for the sites, or proportional to relative population sizes (W_i). b) Whether the samples should be combined, or to merely accept alternative a). c) What weighting to use among alternatives b) to f).

The above alternatives resemble those in section 3 and multinational combinations may be viewed as special cases of multi-site combinations, but a very special case, for the reasons given there. Furthermore, the alternatives listed above deal not with academic or idle speculation, but with many practical, actual problems. I have advised and argued on problems of every kind, and felt the need for and lack of dependable references on combinations and cumulations, whether technical and published or oral and authoritative. Some examples I have encountered:

a) The World Fertility Surveys had national sample sizes without much (any) relation to population sizes. Should they be combined and how? I thought yes and with $N_i(E)$.

b) Samples of several hundred households were selected in each of 12 large cities of the USA (which had "racial riots" in 1968). Should they be combined and how? I thought yes and with $N_i(E)$.

c) In each of 13 counties of the USA samples of a few hundred 4-year-old children were selected for a study of preprimary learning situations. They were combined with method F.

d) In 11 of China's 30 provinces probability samples averaging 1,000 4-year-old children were selected for studies of preprimary learning situations. They were combined with method F.

e) In 5 of Nigeria's 30 states small urban and rural samples were selected for studies of preprimary learning situations of 4 year olds. After examining the 5×2 small samples the sample cases were merged with Method C into urban and rural samples.

f) Coordinated survey designs and university resources are being planned for 5 to 8 large cities of China. The designs are planned both for comparisons and for combination, with either Method E or F.

8. Errors, losses, compromises

The Mean Square Error of a weighted combination of means may be written as

$$\mathrm{MSE}\left(\sum W_i \bar{y}_i\right)$$

$$= \mathrm{Bias}^2\left(\sum W_i \bar{y}_i\right) + \mathrm{Var}\left(\sum W_i \bar{y}_i\right)$$

$$= \left\{\sum W_i [E(\bar{y}_i) - \bar{Y}_i]\right\}^2 + \left(\sum W_i^2 D_i^2 S_i^2 / n_i\right).$$

This holds for distinct countries (i) and distinct domains like provinces. But for some domains there may also exist covariances (S_{ij}), positive or negative. The relative weights are W_i, and S_i^2 and n_i are element variances and sizes, with design effects D_i^2 to compensate for the effects of complex designs. On any study all these values can differ greatly between variables. Note that the bias of the combined mean is the weighted average of the individual biases. For periodic samples these may be fairly constant. For multipopulation and multidomain samples this emphasizes the need for reducing biases for the larger units, with large W_i. The variances of means decrease in proportion to the number of units being averaged, and thus they decrease in importance relative to the biases.

The situation is different for comparisons, where

$$\mathrm{MSE}(\bar{x} - \bar{y})$$

$$= \mathrm{Bias}^2(\bar{x} - \bar{y}) + \mathrm{Var}(\bar{x} - \bar{y})$$

$$= [E(\bar{x} - \bar{y}) - (\bar{X} - \bar{Y})]^2 + \mathrm{Var}(\bar{x}) = \mathrm{Var}(\bar{y})$$

$$= [\{E(\bar{x}) - \bar{X}\} - \{E(\bar{y}) - \bar{X}\}]^2 + D_x^2 S_x^2 / n_x + D_y^2 S_y^2 / n_y.$$

Note that the biases of differences tend to vanish if the biases are similar, even when not small. The variance is the sum of two variances (and a small n_x or n_y can increase it), hence may dominate the bias term. When there are overlaps (in periodic surveys) the covariance term $-2\,\mathrm{Cov}(\bar{x}, \bar{y}) = -2\, D_{xy} S_{xy} n_c / n_x n_y$ tends to decrease the variance.

I have emphasized in some detail elsewhere the need for the utmost "harmonization," for the coordination of survey methods: in variables, measurements, and in populations. On the other hand, there is great freedom to choose different sampling methods for the different populations, provided they are all based on good probability samples (Kish 1994).

In multipopulation combinations, frequent and serious conflicts arise, because the relative sizes W_i of the populations (of countries or of provinces) often vary greatly;

ranges of 1 to 50 or more are common. But the sample sizes may be (roughly) equal for all H populations. Then weights k_i may be introduced to adjust the combinations to the W_i. These inequalities of sampling rates increase the variances of combinations by a relative factor $1 + L = 1 + C_k^2$; where L denotes relative loss (increase in variances) and C_k^2 the coefficient of variation among the weights k_i. Both are zero when all k_i are the same, *i.e.*, for proportional allocation of the n_i to the W_i. But then the average variance of the populations and their comparisons suffer even greater losses than the sum. This conflict can be resolved with compromises, especially an "optimal" compromise with $n_i \propto \sqrt{(W_i^2 + 1/H^2)}$ (Kish 1976, Kish 1994).

A good numerical example comes from the 10 provinces of Canada, whose total population (in 1991) of 27, 211,000 with an epsem selection of $f = 1{:}2721$ would yield roughly these 10 values of n_i in row 1 for a total of $n = 10{,}000$. You see that the largest province of 3,706 cases is about 75 times greater than the smallest with 49. This range seems common for provinces within most countries. Also for multipopulation cases; *e.g.*, in the European Union, Germany is about 200 times the size of Luxembourg. The proportions are $W_i = n_i / 10{,}000$; and a proportional sample would yield an optimal value for $\sum W_i \bar{y}_i$ of $\sum W_i^2 \bar{y}_i^2 / n_i$, hence a relative loss function $1 + L = 1$, with loss $L = 0$. For simplicity and to concentrate on weights, we can assume that element variances $D_i^2 S_i^2$ and costs c_i are similar, or can be averaged out. However, these proportional n_i values would result for average provincial means $\sum \bar{y}_i / 10$ or for average comparisons of provincial values $(\bar{y}_i - \bar{y}_j)$ of $1 + L = H \sum (1/W_i) = 3.9785$ for a relative loss of 2.9785, a 300% increase in average variances. These losses come mostly from the 6 small provinces (Derivations in Kish 1976).

Row 1	3,706	2,534	1,206	935	401	363	331	266	209	49
Row 2	2,437	1,730	995	869	684	676	669	657	648	636

Thus, some people (in Canada and in other countries too) ask for equal size samples, $n_i = 1{,}000$, so that each province can provide the same precision. Then the means $\sum \bar{y}_i / 1{,}000$ will all have variances $\sum (1/1{,}000)$ and relative efficiency of $1 + L = 1$, with loss $L = 0$. However, the national mean will have a variance of $\sum W_i^2 / 1{,}000$, with a relative loss of $1 + C_k^2 = H \sum W_i^2 = 2.3003$, or a 130% increase in variance. We must also remember that all crossclasses, such as those by age, education, occupation, *etc*., will also tend to suffer similar losses.

However, some remarkably good compromises can be had, and the best is a least-square solution with the $n_i \propto \sqrt{(W_i^2 + H^{-2})}$. These give the $1 + L$ values of $1 + L = 1.2424$ and $1 + L = 1.2630$, for $\sum W_i \bar{y}_i$ and $\sum \bar{y}_i / H$, respectively, only a 25% loss for each! The n_i values in Row 2 show a "floor" between 600 and 700 for the n_i for the 6 small provinces, and a roughly proportionate increase (but below 10,000 W_i) for the largest 4 provinces. This optimal allocation has in fact been used for some of the surveys of Statistics Canada (Tambay and Catlin 1995). It is interesting that the mathematical solution also makes good common sense (Kish 1976, 7.6, Kish 1987, 7.3, Kish 1988). However, the mere common senses solutions of allocations proportional to $\sqrt{W_i}$ are less efficient than the optimal allocation.

References

Alexander, C.H. (1999). A rolling sample survey for yearly and decennial uses. *Proceedings of the 52nd Session of the International Statistical Institute*, Helsinki.

Australian Bureau of Statistics (1993). *The Australian Population Monitor*. Canberra: ABS.

Caplan, D., Haworth, M. and Steel, D. (1999). UK labour market statistics: Combining continuous survey data into monthly reports. *Proceedings of the 52nd Session of the International Statistical Institute*, Helsinki.

Cleland, J., and Scott, C. (1987). *The World Fertility Survey*. Oxford: The Oxford University Press.

Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, Series B, 4, 102-18.

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-29.

Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 5, 3-8.

Hedges, L.V., and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Heeringa, S.G., and Liu, J. (1999). Complex sample design effects and inference for mental health survey data. *International Journal of Methods in Psychiatric Research* 7, 56-65.

Isnard, M. (1999). *Alternatives to Traditional Census Taking*: *The French Experience*. Paris: INSEE.

Kiaer, A.N. (1895). *The Representative Method of Statistical Surveys*. English translation 1976, Oslo: Statistik Centralbyro.

Kish, L. (1961). A measurement of homogeneity in areal units. *Bulletin of the International Statistical Institute*. 33rd session, 4, 201-209.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society,* Series A, 139, 80-95.

Kish, L. (1987). *Statistical Research Design*. New York: John Wiley & Sons, Inc.

Kish, L. (1988). Multipurpose sample designs. *Survey Methodology*, 14, 19-32.

Kish, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 1, 63-71.

Kish, L. (1994). Multipopulation survey designs. *International Statistical Review*, 62, 167-186.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

Kish, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

Kish, L., Lovejoy, W. and Rackow, P. (1961). A multi-state probability sample for traffic surveys. *Proceedings of the Social Statistics Session*, American Statistical Association, 227-230.

Mooney, H.W. (1956). *Methodology in Two California Health Surveys, San Jose (1952) and Statewide (1954-55).* U.S. Public Health Monograph No. 70.

National Center for Health Statistics (1958). Statistical designs of the Health Household Interview Surveys. *Public Health Series,* 584-A2.

Platek, R., Rao, J.N.K., Särndal, C.-E. and Singh, M.P. (Eds) (1989). *Small Area Statistics*. New York: John Wiley & Sons, Inc.

Szalai, A. (1972). *The Use of Time*. The Hague: Mouton.

Tambay, J.-L., and Catlin, G. (1995). Sample design of the National Population Health Survey, *Health Reports*, Catalogue No. 82-003, 7, 29-38.

Verma, V. (1992). Household surveys in Europe: Some issues in comparative methodologies. In *Seminar*: *International Comparisons of Survey Methodologies*. Athens.

Verma, V. (1999). Combining national surveys for the European Union. *Proceedings of the 52nd Session of the International Statistical Institute*, Helsinki.

World Fertility Surveys (1984). *Major Findings and Implications.* The Hague: International Statistical Institute.

Yates, F., and Cochran, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science*, 28, 556-80.