

Sample Survey Design for Evaluation (The Design of Analytical Surveys)

© 2009 Joseph George Caldwell. All Rights Reserved.

Posted at Internet website <http://www.foundationwebsite.org> 20 March 2009, updated 15 April 2009, 10 June 2009, 26 June 2009, 14 July 2009, 13 January 2010, 18 January 2010, 11 February 2010.

Contents

1. The Design of Analytical Surveys	1
2. Design-Based Approach	3
3. Model-Dependent Approach	4
4. Model-Based Approach	6
5. Survey Design for Model-Based Applications	8
6. Construction of Control (Comparison) Groups	10
7. Sample Size Determination	24
8. Estimation Procedures	44
Appendix A. A Procedure for Designing Analytical Surveys	44
Selected References in Sample Survey Design	53

1. The Design of Analytical Surveys

Most sample surveys are conducted to make inferences about overall population characteristics, such as means (or proportions) or totals. The population under study is viewed as fixed and finite, and a probability sample is selected from that population. The statistical properties of the sample are defined by the sample design, the sample selection method and the population, not by any underlying properties of the process that created the population. These kinds of sample surveys are referred to as “descriptive” surveys.

In some instances, it is desired to make inferences about the process that generated the population, such as a test of the hypothesis about population characteristics (e.g., that two population subgroups (domains) could be considered to have been generated by the same probability distribution). In this case, the goal is to develop a mathematical model that is a reasonable description of the process that generated the population data, or that describes relationships among the variables that are observed on the population elements. The particular population at hand (the subject of the sample survey) is viewed as a sample from this process. Surveys conducted to assist the development of mathematical models are referred to as “analytical” surveys.

Note that it does not make sense, in a descriptive survey, to test the hypothesis that two finite-population subgroups have different means. The means of *any* two population subgroups of a finite population are virtually always different. Tests of hypothesis about differences make sense only in the context of an analytical survey and conceptually infinite populations. (For the same reason, use of the “finite population correction” is not applicable to analytical surveys.) Similarly, it does not make sense to test the hypothesis that the mean of the population or a subpopulation equals a specific value – for descriptive surveys, it is *estimates* (point estimates and interval

estimates) that are of interest, not tests of hypothesis. It is essential to decide on the conceptual framework for a survey (descriptive or analytical) prior to the design and analysis.

During the early period of the development of the theory of sample survey, up to about 1970, work in developing the theory of sample survey focused on the design of descriptive surveys. Standard textbooks on the subject, such as William G. Cochran's *Sampling Techniques* (Wiley, 3rd edition, 1977) and Leslie Kish's *Survey Sampling* (Wiley, 1965), describe methodology for designing and analyzing descriptive sample surveys. A popular elementary text on survey sampling is *Elementary Survey Sampling* 6th edition (Duxbury Press, 2005). The author's previous note, *Vista's Approach to Sample Survey Design* (1978) (<http://www.foundationwebsite.org/ApproachToSampleSurveyDesign.htm>) and *Sample Survey Design and Analysis: A Comprehensive Three-Day Course with Application to Monitoring and Evaluation (Day 3)* (1980) (<http://www.foundationwebsite.org/SampleSurvey3DayCourseDayOne.htm>) summarize methods for the design of analytical surveys. This article presents more detail on this topic.

The author specialized in the design of analytical surveys in his statistical consulting practice in the 1970s. At that time, there were no reference texts or articles on the subject. The methodology applied by the author to design analytical surveys was developed by drawing on his background in experimental design (in which he specialized in his PhD program as a student of Professor Raj Chandra Bose, the "father" of the mathematical theory of experimental design). During that time, he also promoted the use of experimental design to specify "run sets" for large-scale computer simulation programs. Since that time, a number of papers and books have been written on the topic of analytical survey design. These include "History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis" by J. N. K. Rao and D. R. Bellhouse (*Survey Methodology*, June 1990); *Practical Methods for Design and Analysis of Complex Surveys* 2nd edition by Risto Lehtonen and Erkki Pahkinen (Wiley, 2004); *Sampling* 2nd edition by Steven K. Thompson (Wiley, 2002); *Sampling: Design and Analysis* by Sharon L. Lohr (Duxbury Press, 1999); and *The Jackknife and Bootstrap* by Jun Shao and Dongsheng Tu (Springer, 1995).

The classification of surveys into two types – descriptive and analytical – was described in Cochran's *Sampling Techniques*. The distinction between descriptive and analytical surveys is a little "fuzzy." For example, a survey designed to describe the incomes of various population subgroups could be referred to as a descriptive survey, but if statistical tests of hypotheses about differences in income levels among the groups are to be made, the survey could be called an analytical survey. Rao and Bellhouse classify surveys and survey methodology into a slightly different and somewhat finer categorization: (1) design-based approach; (2) model-dependent approach; and (3) model-based approach or model-assisted approach. In the design-based approach, a probability sample is selected from the population under study (a fixed, finite population), and the nature of the sampling procedure suffices to define reasonable estimates of the population characteristics. How the population was generated is irrelevant – no probability (or other) model is defined to describe the generation of the population items. In the model-dependent approach, a purposive (non-probability) sample of observations is selected. Each observation is considered to be a realization (sample unit) from a specified probability distribution. The sample is usually assumed to be a sample of independent and identically distributed (iid) observations from the specified distribution, and (in any case) the nature of the joint probability distribution function of the sample determines what are good estimates for the quantities of interest, using standard sampling theory (e.g., as presented in *Introduction to the Theory of Statistics*, 3rd edition, by Alexander Mood, Franklin A. Graybill and Duane C. Boes (McGraw-Hill, 1950, 1963, 1974). In the model-based (or model-assisted) approach, a probability model is

specified for the population units, *and* a probability sample is selected from the finite population under study. The sample is analyzed in a way such that the estimators are reasonable *both* for estimation of characteristics of the finite population under study *and* for estimation of the parameters of the assumed model and tests of hypothesis.

It could be argued that the model-dependent approach has nothing to do with “survey sampling,” which typically involves analysis of a probability sample from a fixed and finite population, and should not be considered as a separate category of survey sampling (leaving it to standard sampling theory and experimental design). Including it, however, facilitates discussion of the other two categories.

Note that, from a theoretical viewpoint, in the design of an analytical survey it is not necessary that all population items be subject to sampling, or even that the probabilities of selection be known. (It is required, however, that the probability of selection not be related to the model error term.) These conditions apply, however, only if the analytical model is *correctly specified (identified, in the terminology of economics)*. A practical problem that arises is that this condition (of correct specification) can never be proved. It is always possible that the model specification differs, in unknown ways, for different segments of the population (e.g., male / female, urban / rural, or different agricultural regions). The only sure defense against this possibility is to use probability sampling, where the probability of selection of all population elements is known and nonzero (it is not practical for all the probabilities to be equal in the design of analytical surveys). These probabilities may be set, however, in ways that enhance the precision of the sample estimates of interest (and power of tests of hypotheses of interest).

2. Design-Based Approach

The design-based approach is the standard approach to design and analysis of descriptive sample surveys. All of the older books on sample survey consider only this approach. As mentioned, the objective is estimation of means or totals for the population or subpopulations of interest. The major survey-design techniques for achieving high precision are stratification, cluster sampling, multistage sampling and sampling with varying probabilities of selection (e.g., selection of primary sampling units with probabilities proportional to size). Stratification may be used either to increase the precision of estimates of the overall population mean or total or to assure specified levels of precision of estimates for subpopulations of interest. Cluster and multistage sampling may be used for administrative convenience, to improve efficiency, or because sample units at different stages of sampling are of intrinsic interest (e.g., a survey that must produce estimates for schools and for students). In cluster and multistage sampling, precision may be increased by setting the probabilities of selection of first-stage sample units proportional to unit size or a measure of size. Determination of number of strata and stratum boundaries may be done to improve precision of overall estimates or because certain strata are of particular interest.

The more information that is available about the population prior to conducting the survey, the better the job that can be done in survey design. In some instances it may be desirable to conduct a preliminary first-phase survey to collect data that may substantially improve the efficiency of the full-scale survey (double sampling, or two-phase sampling). In dealing with populations that are geographically distributed, it is usually the case that a simple random sample is not the best survey design, and large gains in precision or decreases in cost may be achieved through use of the survey-design methodologies mentioned.

A common problem in the design of descriptive sample surveys is that a number of estimates may be of interest, and the optimal design will vary for each of them. The survey designer's goal is to determine a survey design that produces an adequate and efficient return of precision for all important survey estimates.

With respect to estimation of means (or totals) and standard errors, closed-form formulas are available for all standard descriptive-survey designs. It is possible to use simulation methods to estimate sampling errors, but this is not necessary for standard descriptive-survey designs.

3. Model-Dependent Approach

In the model-dependent approach, the investigator has reason to believe that a particular probability distribution or stochastic model adequately describes the population, and taking this into account can improve the precision of estimates. The estimates may be estimates of population means or totals, but what is more likely are estimates of parameters of the probability distribution and estimates of differences (linear contrasts). In this approach, the population under study is viewed as having been generated by an underlying or hypothetical process. The population at hand is just one realization of a conceptually infinite set of alternative populations that might have been generated (in the "realization" of our world). In the model-dependent approach it is often the case that the investigator is interested in estimating relationships between variables, such as the relationship among several dependent variables observed on each sample unit, or on the relationship of a dependent variable to various independent (explanatory) variables.

In using the model-dependent approach, there is a basis for believing that the observations may be considered to be generated in accordance with an underlying statistical model, and it is the objective of the survey to identify (estimate) this model. This is a different conceptual framework for design-based surveys, where the objective is simply to describe the particular population at hand. This conceptual approach is the basis for statistical quality control, where the observations are produced by a manufacturing process. It is also the conceptual framework appropriate for evaluation research, where the outcomes of a program intervention are assumed to be produced by an underlying causal model. (For discussion of causal models, see Judea Pearl's *Causality: Models, Reasoning and Inference* (Cambridge University Press, 2000).)

The model-dependent approach may be applied to both descriptive and analytical surveys. A few examples will illustrate this. In the first example, let us assume that we wish to estimate the mean and variance of the population income distribution, and that it is known (or reasonable to assume) that income follows a log-normal distribution (i.e., the logarithm of income is normally distributed). In the usual descriptive-survey approach, a probability sample may be selected, using one or more of the sample-design procedures identified earlier. For simplicity, let us assume that a simple random sample is selected (with replacement).

The standard approach in descriptive survey analysis is to make no assumptions about the probability distribution that describes the population elements, and to base the sample estimates (means and standard errors) on the sample design and the particular population at hand. The estimate follows from the design; how the population elements came about and the properties of any probability distribution that may be considered to have generated them is irrelevant. In the case of simple random sampling from a finite population, the sample mean is the estimate of the population mean, and the sample variance is the estimate of the population variance. The

standard error of the estimated mean is the square root of the ratio of the sample variance to the sample size.

In the model-dependent approach, however, we take advantage of the fact (assumption) that the underlying distribution of income is log-normal. We assume that the population represents a sample of independent observations from the same (lognormal) distribution, i.e., that the simple random sample of observations represents a sample of independent and identically distributed observations from this distribution.

Statistical theory tells us in this case that the best (minimum-variance, unbiased) estimates of the parameters of the lognormal distribution are obtained by estimating the mean and variance of the logarithms of the observed values. The estimates of the mean and variance of income are then obtained from these by appropriate transformations (i.e., the mean income is $\exp(\text{mean} + \text{variance}/2)$ and the variance of income is $\exp(2 \text{ mean} + 2 \text{ variance}) - \exp(2 \text{ mean} + \text{variance})$, where “mean” and “variance” denote the mean and variance of the logarithm (which has the normal distribution).

The preceding is a very simple example of how information about an underlying probability distribution might be taken into account in determining population estimates. In most cases, the situation is more complex. What is usually the case is that the investigator wishes to estimate relationships among variables. In such cases, he is probably not interested in estimating overall characteristics (means or totals) of the population at hand. Interest focuses on the *process* (real or hypothetical) generating the observations, on relationships among variables, and on tests of hypothesis about the process that generated the particular finite population at hand. This underlying process may be described simply by a univariate probability distribution (e.g., the lognormal distribution in the example presented earlier), or a linear statistical model (e.g., multiple regression, experimental design), an econometric (structural equation) model, or a more complex model (e.g., a “latent variable” or path-analysis model).

In order to estimate the model parameters, the investigator typically engages in an iterative process of model specification, parameter estimation and model testing. If an estimated model does not pass the tests of model adequacy, the model is respecified and the process repeated until an adequate model is obtained. A key assumption in this approach is that the observed sample is an independent and identically distributed sample from the posited model, i.e., that the model is “correctly specified.” If this assumption holds true, good results will be obtained (for a sufficiently large sample). From a practical point of view, the problem is that the investigator usually does not know the correct model specification, and tries to determine it empirically from the data analysis. The difficulty that arises is that if the model is incorrectly specified, then the parameter estimates will be incorrect. Note that this may or may not affect the quality of certain estimates derived from the model (e.g., least-squares forecasts corresponding to a model may be unbiased, even though estimates of the model parameters are biased).

In the model-dependent approach, it is not necessary to select a probability sample from the population (i.e., a sample in which every item in the population is selected with a known, nonzero probability (or the same unknown probability)). The investigator may select any sample he chooses, as long as he may reasonably assert that the sample is an independent and identically distributed sample from the posited distribution (or, if the sample items are not independent, he must specify the nature of the dependence). Note that this precludes selecting a sample in a way that the likelihood of selection is related to the model error term (i.e., to the dependent variable). In the earlier example of the lognormal distribution of income, this is achieved by selecting a simple random sample from the population (i.e., a probability sample was in fact selected). In the

case where a linear regression model describes the relationship of a dependent variable to an independent (explanatory) variable, all that is required is that the model be correctly specified and that a reasonable amount of variation be present in the independent variable (and the collinearity among the explanatory variables be low) – any such sample from the population, whether a probability sample or not, will suffice.

Note that, as discussed above, the condition that the model be correctly specified is difficult or impossible to guarantee in practice. One of the usual conditions is that the model error terms be stochastically independent. This condition may be relaxed, as long as the nature of the dependence is taken into account. For geographically distributed populations, such as human populations, nearby (spatially proximate) population elements may be dependent (related). This fact should be taken into account when selecting the data sample (and analyzing the data). Observations taken over time may also be (temporally) correlated. The methods of time series analysis address means for modeling spatial and temporal correlations (e.g., time series analysis (Box-Jenkins models), geostatistics (kriging)).

The problem in designing the sample in the model-dependent approach is to have a good idea of what variables are important in the model, to assure a reasonable amount of variation in each of them, and to have low correlation among them. This is the same problem faced in experimental design. The optimal sample design depends on what model is assumed. For example, if a zero-intercept regression model is assumed ($y_i = b x_i + e_i$, where e_i are a sequence of iid random variables of mean zero and constant variance), then the best sample to select is the n items in the population having the largest values of x , where n denotes the sample size. If in fact this model is incorrect, and the correct model involves an intercept ($y_i = a + b x_i + e_i$) then the best sample is the one for which half the observations have the largest values of x and half have the smallest values of x . In this case, the previous sample (of the n items having the largest values of x) is a terrible sample design. If the model is not linear but curvilinear, then the second sample will be poor (we would want some observations in the middle of the range). As additional variables are added to the model, the difficulty of constructing a good sample design increases – this is the subject of the field of experimental design (e.g., fractional factorial designs; see, e.g., *Experimental Designs*, 2nd edition, by William G. Cochran and Gertrude M. Cox (Wiley, 1950, 1957)).

The important thing to realize with the model-dependent approach is that the optimal sample design, and good estimates, derive from the *model* assumed to generate the population units, not from the structure of the particular population (realization) at hand. In physical experiments, the experimenter generally has much control over the specification of combinations of experimental conditions, whereas in dealing with finite populations this is often not the case (e.g., it may be impossible to orthogonalize the variables).

4. Model-Based Approach

In the model-based (or model-assisted) approach, it is desired both to estimate overall population characteristics *and* to estimate parameters of the underlying probability model assumed to adequately describe the generation of that population (and to test hypotheses). In this case it is desired to select a probability sample from the population at hand, and to construct that sample such that it will produce good estimates of the population characteristics *and* the model parameters and tests of hypotheses. To accomplish the former, it is generally desired that the probabilities of selection be as uniform as possible. To accomplish the latter, it is desired that there be substantial variation in all dependent variables of interest, and that the correlation

between independent variables that are causally (logically) unrelated to the dependent variable be low. For a probability sample from a finite population to produce such a sample, the selection probabilities will usually vary considerably (often with some selection probabilities equal to one).

Books on sample survey design for descriptive surveys describe a variety of different types of estimates. Two estimation procedures that may seem related to the model-based approach are ratio estimates and regression estimates. While the model-based approach may involve the specification and estimation of ratio and regression models, the ratio and regression estimates of descriptive survey data analysis have little to do with ratio and regression models of analytical survey data analysis. In descriptive survey data analysis, ratio and regression estimates are used to develop improved estimates of population means and totals, with no regard to any underlying probability model that may be considered to generate the population units. The ratio and regression estimates are simply numerical procedures used to produce improved estimates by taking into account ancillary data that may be correlated with the variable of interest. There is no consideration of an underlying model and whether it is correctly specified. The objective is to obtain good (accurate: high precision and low bias) estimates of population means and totals, not of parameters or properties of hypothetical models (such as regression coefficients or treatment effects) or tests of hypotheses (e.g., about a double-difference measure of program impact). (In particular, as mentioned earlier, it is not the objective in a descriptive survey to test hypotheses about equality of distributions or means of subpopulations, because for finite populations those distributions (or means) are (virtually) always different – there is nothing to test.) Furthermore, in ratio and regression estimation in descriptive-survey applications, the theory is developed for a single independent variable. In model-based applications, there are typically many independent variables (e.g., in a survey intended to develop an econometric model).

Note that, as mentioned, the “finite population correction” (FPC, the reduction in the variance for simple random sampling without replacement, owing to the fact that the population is finite) is not applicable to the estimation of underlying models (i.e., to analytical surveys). The inferences are being made about the *process* generating the finite population at hand, not about this particular realization of the process.

In the model-based approach, the role of a regression model is conceptually quite different from the role of a regression estimator in a descriptive (design-based) survey. In a descriptive survey, the regression estimate is nothing more than a computational mechanism for producing high-precision estimates. In the model-based approach, the objective is to determine a statistical model that is a valid representation of a process considered to have generated the population at hand. The validity (adequacy) of the model is assessed by conducting various tests of model adequacy, such as by examining the model error terms (“residuals”), and revising (respecifying and re-estimating) the model, if indicated. The book by Sharon L. Lohr, *Sampling: Design and Analysis* (Duxbury Press, 1999) discusses these concepts at a general level. For more on the topic of model adequacy, see any of the many books on statistical model-building, including books on econometrics and regression analysis. Examples include the books cited earlier (on regression analysis and the general linear statistical model) and: *Mostly Harmless Econometrics: An Empiricist’s Companion* by Joshua D. Angrist and Jörn-Steffen Pischke (Princeton University Press, 2009); *Micro-Economics for Policy, Program, and Treatment Effects* by Myoung-Jae Lee (Oxford University Press, 2005); *Counterfactuals and Causal Inference: Methods and Principles for Social Research* by Stephen L. Morgan and Christopher Winship (Cambridge University Press, 2007); *Econometric Analysis of Cross Section and Panel Data* by Jeffrey M. Wooldridge (The MIT Press, 2002); *Matched Sampling for Causal Effects* by Donald B. Rubin (Cambridge University Press, 2006); and *Observational Studies* 2nd edition by Paul R. Rosenbaum (Springer, 2002, 1995). These books relate to econometric modeling for evaluation. A comprehensive review of

econometric literature is presented in “Recent Developments in the Econometrics of Program Evaluation” by Guido W. Imbens and Jeffrey M. Wooldridge (*Journal of Economic Literature* 2009, Vol. 47, No.1, pp. 5-86). References on the general subject of econometrics (structural equation modeling) include: *Econometrics* 2nd edition by J. Johnston (McGraw Hill, 1963, 1972); *Econometric Models, Techniques, and Applications* by Michael D. Intriligator (Prentice-Hall, 1978); *Principles of Econometrics* by Henri Theil (Wiley, 1971); *Introduction to the Theory and Practice of Econometrics* 2nd edition by George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chou Lee (Wiley, 1982, 1988); and *The Theory and Practice of Econometrics* 2nd edition by George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chou Lee (Wiley, 1980, 1985). (These are from my personal library – there are many newer references available.)

5. Survey Design for Model-Based Applications

As described above, the problem in designing an analytical survey is to construct a survey design that has substantial variation in the independent variables of interest, low (or zero) correlation among causally (logically, intrinsically) unrelated independent variables. Also, all units of the population must be subject to sampling, and the probabilities of selection should be as uniform as possible, subject to achievement of the previous condition. (The probabilities of selection must be nonzero if it is desired to obtain unbiased estimates of population means or totals. Keeping the selection probabilities uniform generally increases the precision of these estimates (unless there is a good reason for varying them).)

There are several cases that may be considered, depending on the objectives of the investigation and the control that the survey designer has over selection of the sample units. In some applications, the goal is simply to develop a set of tables or a multiple regression model that describes the relationship of one or more dependent variables to a set of independent (explanatory) variables. An example of this would be the collection of survey data to develop an econometric or socioeconomic model of a population. In other applications, such as program evaluation (evaluation research, impact evaluation), it is of primary interest to estimate a particular quantity, such as a “double-difference measure” of program impact (in statistical terms, the interaction effect of program treatment and time), but because it is often not possible in socioeconomic evaluations to employ randomization to eliminate the influence of non-treatment variables, it is also desired that the survey enable the estimation of the relationship of program effects to other concomitant variables (ancillary variates, “covariates,” uncontrolled variables). Finally, it may be possible to make use of the principles of experimental design to configure the design to reduce the bias or increase the precision of particular estimates, by means of techniques such as “blocking” or matching (to promote “local control”).

If it were not for the fact that we are sampling from a finite population (so that not all combinations of explanatory variables are possible), and the desire to control the sample selection probabilities, the survey design problem (for a model-based application) would simply be an exercise in experimental design. The investigator would specify combinations of independent-variable values that corresponded to good variation in all variables and low correlation among them, and select units having these variable combinations from the population. This could be accomplished, for example, by employing an experimental design (e.g., a fractional factorial experimental design or a balanced incomplete block design), the Goodman-Kish method of “controlled selection,” or Cochran’s method of stratification on the margins. While these methods (of stratification) work well for small numbers of independent variables (such as two), they do not “scale” to situations

involving large numbers of independent variables, as is common in the field of evaluation research. (The number of independent variables known in advance of the survey, and of interest in the data analysis, may be very large, including data from previous related surveys, from government statistical systems, or from geographic information systems. The number could easily be several hundred variables. The methods proposed by Kish and Cochran for multiple stratification are of no use in such a situation.) When the number of independent variables is large, the number of combinations of stratification values (stratum cells) becomes very large. Furthermore, a problem that arises in survey applications is that not every combination of variables exists (is possible or is represented in the population at hand), so that it may not be possible to accomplish orthogonality. Depending on how the selection is made, the probabilities of selection of the sample units may be poorly controlled, i.e., could be zero for some sample units and much more variable than desired or necessary for others.

Note that whereas in descriptive-survey design attention focuses on the *dependent* variables, in analytical-survey design attention focuses on the *independent* (explanatory) variables.

There is another very significant difference between descriptive surveys and analytical surveys. Descriptive surveys tend to deal mainly with *independent samples* and minimizing correlations among sample units (to obtain high precision for estimates of totals and means), whereas analytical surveys tend to deal with *correlated samples* and introducing correlations into sample units (to obtain high precision for estimates of differences). Correlations certainly occur in descriptive surveys, such as in the case of cluster or multistage sampling (from intracluster / intraunit correlation), but it is generally sought to minimize these correlations. They tend to decrease the precision of the estimates of interest (means and totals), and are introduced because they have substantial cost advantages (e.g., lowering of travel costs, administrative costs, or listing (sample-frame construction) costs). In analytical surveys, the introduction of correlations into the sample (in special ways) can increase the precision of estimates of interest (differences, regression coefficients). In such surveys, clusters (or first-stage sample units) are particularly useful in this regard, since information about them is often available prior to sampling, and may be used as a basis for matching or stratification (e.g., census enumeration areas, villages, districts).

The most important tool for increasing precision and reducing bias in analytical surveys is matching, and matching can be done only when some information (unrelated to the dependent variables) is available on the sample units. For ex-ante matching (before the sample is selected), some information is generally known about population aggregates (groups, clusters, areas), such as census enumeration areas or districts or regions, but it is typically not known about the ultimate sample units – obtaining data on them is the primary purpose of doing the survey. For this reason, ex-ante matching can usually be done only on clusters, or “higher-level” sample units. Matching (pruning, matching, culling) may be done on the ultimate sample unit after the survey data are collected (ex-post matching), but it can be done only on clusters prior to collecting the sample data. A descriptive survey uses clusters in sampling *despite* the intracluster correlation (because they enable cost savings which compensate for the associated precision loss); an analytical survey uses clusters *because of* it. Clusters enable matching, and are therefore the vehicle by which correlations are introduced into the sample.

Note that the intracluster correlation tends to decrease as the size of the cluster increases. For this reason, it is most desirable to match on the smallest units for which (pre-survey) data are available for matching. For this reason, determination of sample size, which will be discussed in detail later, focuses on the lowest-level unit for which pre-survey data (suitable for effecting matching) are available. Descriptive surveys seek clusters with low intracluster correlations;

analytical surveys seek clusters with high intracluster correlations (to increase the precision and decrease bias of comparisons, through matching).

During the 1970s, the author investigated alternative approaches to the design of analytical surveys. On the one hand, he investigated the use of formal optimization theory to determine sample allocations that minimized the variance of estimates of interest. That approach proved to be unfruitful. The approach that proved most useful was a technique of setting the selection probabilities to effect an “expected marginal stratification.” With this approach, a design is iteratively constructed that satisfies a large number of expected stratification constraints, subject to keeping the probabilities of selection nonzero for all population items, and as uniform as possible. There are two major ways in which this algorithm is implemented, depending on whether it is desired to configure the design to maximize the precision of a particular treatment comparison (as is typically the goal of a program evaluation, such as estimation of a double difference). The method involves specifying the selection probabilities such that the expected numbers of units in each stratum cell are as desired. The method is general and can be applied in any situation.

Appendix A describes the procedure in the case where it is desired to use the survey data to estimate a general linear model (e.g., analysis of variance, multiple linear regression), including the case in which it is desired to estimate a particular treatment comparison (e.g., a double-difference estimate of impact). It addresses the goal of the model-based approach of designing a survey that addresses both the estimation of model parameters and differences and tests of hypothesis (e.g., of a double-difference estimate of program impact) *and* estimation of overall population characteristics such as means and totals.

It is noted that, as in the case of design of descriptive surveys, the best design for a particular dependent variable will not be the best design for another dependent variable. For descriptive designs, this fact is accommodated by examining good designs for each of the important dependent variables, and selecting a design that is adequate for all of them. This is accomplished in the design of analytical surveys by including all independent variables for all models (i.e., associated with all dependent variables) in the algorithm simultaneously.

Note that the design should be matched to the analysis. This paper does not address analysis, except in very broad terms (e.g., referring to model development). This is a major topic, and will be addressed in a later paper. If the design is not carefully considered, the analysis may be much more difficult, and its quality degraded. If the analysis does not take the design fully into account, much of the value of the design may be lost. It was reported recently, for example, that a high proportion of articles in medical journals analyzed data as matched (unpaired) samples, when they should have analyzed the data as matched pairs (“A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003,” by Peter C. Austin, *Statistics in Medicine*, vol. 27: pp. 2037-2049, 2008.)

6. Construction of Control (Comparison) Groups

A useful experimental design in evaluation research is a “pretest / posttest / control group” design. In this design, the program treatment is applied to a random sample of population units, and a similarly sized set of population units are selected as “controls.” A “baseline” survey is conducted to measure outcomes of interest prior to the program intervention, and a “follow-up” survey is conducted at a later time to measure outcomes after the program intervention. To increase “local control,” a “panel” survey (longitudinal survey) is usually attempted, in which the same units are

measured before and after the program intervention. The measure of program impact is the difference, between the treatment and control units, of the difference in outcome before and after the program intervention. In statistical terminology, this effect is called the interaction of treatment and time. In evaluation research it is referred to as a “double difference” or a “difference-in-difference” estimate (or measure). Because randomization is used to determine which units receive the program treatment, the influence (effect) of all variables other than the treatment variable(s) are removed from the estimate.

A problem that arises in evaluation of social and economic programs is that it is often not feasible to randomize the allocation of the program treatment to population units. This problem arises for many reasons, including the fact that all members of the population may be eligible for a program, or the program may be offered on a voluntary basis and not all members of the population choose to apply. In this case, what is usually done is to select a sample of population units that are similar to the treated units in known variables that may have an effect on program outcome or may have affected the likelihood of selection into the program. The process of selection of the comparison-group items is usually done by selecting a sample of treatment units and then selecting a group of items (the comparison group) for which the empirical probability distribution is as similar to that of the treatment group as possible. To increase the precision of the estimate, it is generally desirable to match each (individual) treatment unit with a similar comparison unit, i.e., to use matched pairs as a means of promoting local control, not simply to ensure that the probability distributions of the treatment and comparison groups are similar (matched). Since the treatment is not allocated by randomized assignment, the design is referred to as a “quasi-experimental” design, and the term “comparison group” is usually used instead of “control group” (although this usage convention is not universal).

The process of selecting the comparison group when randomized assignment to the treatment group is not possible is called “matching.” (Note that the term “matching” may refer either to matching of the probability distributions or to matching of individual units (matched pairs).) There are a variety of different methods of matching. They are described in articles posted on Professor Gary King’s website (<http://gking.harvard.edu>), including Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart, “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference,” *Political Analysis*, Vol. 15 (2007), pp. 199-236, posted at <http://gking.harvard.edu/files/matchp.pdf> or <http://gking.harvard.edu/files/abs/matchp-abs.shtml> , and “MatchIt: Nonparametric Preprocessing for Parametric Causal Inference,” by Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart (July 9, 2007), posted at <http://gking.harvard.edu/matchit/docs/matchit.pdf> . The preferred method of matching, called “exact matching,” involves finding, for each treatment unit, a unit that matches it (exactly) on all known variables (or, more correctly, on categorical variables derived from the original variables, since it is unlikely to find units that match on a large number of interval-scale variables). One advantage of exact matching is that it represents an attempt to construct a comparison group for which the *joint* probability distribution of the matching variables is similar for the treatment and comparison groups. Another advantage, as noted earlier, is that it generates sets of “matched pairs,” from which a more precise estimate of difference-in-means can be obtained, than from simply comparing the means of two unrelated samples. With other types of matching, the matching is done so that the marginal probability distributions of the matching variables are similar for the treatment and comparison groups (this may be tested by a Kolmogorov-Smirnov test of the equality of two probability distributions). In these cases, the matching is usually done by constructing a measure of similarity, or closeness, between units, and selecting as a match the unit that has the most similar value of this score (i.e., is “closest” or “nearest” to each treatment unit).

A common matching method is the so-called “propensity-score” matching (PSM). With this method, a logistic regression model is constructed to describe the probability that a unit is included in the treatment group. This probability is called the “propensity score.” After the model is estimated, a propensity score may be estimated for each population unit. For each member of the treatment group, a non-treatment unit having the closest propensity score is selected as its matching unit (this is called “nearest-neighbor” matching). Propensity-score matching, or any other matching method based on a single composite (scalar, one-dimensional) measure of similarity, is not as good as “exact” matching (multidimensional matching in which the matched units are identical for each matching variable). The problem is that exact matching may not be possible, because it may not be possible to find, in a finite population, a unit that matches another on all match variables (even when they are categorical). Also, exact matching becomes difficult to implement when many match variables are involved (the so-called “curse of dimensionality”).

The goal of matching is that the treatment variable and the unit response be conditionally independent given the observed covariates (match variables). This is a *distributional property* – a property of the *process* that generates the treatment and control groups. The fact that individual treatment and comparison units match on the propensity score is neither necessary nor sufficient for this distributional property to hold. Just because the unit response and treatment variable are conditionally independent with respect to the propensity score does not imply that individual units having similar propensity scores are similar with respect to the component variables of the score. Even if the propensity score matching is done over the entire population using “nearest neighbor” matching, the “nearest neighbor” matches may not match well at all with respect to the component variables of the score. Two individual units may have very similar propensity (or identical) propensity scores, yet differ markedly with respect to the component (match) variables. For propensity score matching to work (to form similar treatment and control groups), the matching must hold (be done) for *the complete distribution* of the propensity score. Even then, all this method does is produce similar treatment and control *groups* – not *individually matched pairs* of treatment and control units. Propensity score matching can reduce the bias associated with a lack of randomization (by forming treatment and control groups that are similar overall), but it does nothing to improve the precision of estimates of interest. Considering the general inadequacy (shortcoming) of this matching method, the extent to which it is used is rather amazing. It is a good “check” on whether *groups* are well matched, but it is not at all sufficient for *individual matching* (forming matched pairs). If units do not have similar propensity scores, then they are not good matches, but if they have similar propensity scores, it cannot be concluded that they are good matches – they may be terrible matches.

This point deserves reiteration for emphasis. There are conditions under which matching on propensity scores can result in well-matched distributions (samples). See the paper “The central role of the propensity score in observational studies for causal effects,” by Paul R. Rosenbaum and Donald B. Rubin, *Biometrika* (1983), vol. 70, no. 1, pp. 41-55, for discussion of this. These methods involve *conditional distributions and expectations*. They produce well-matched distributions (matched samples – matched treatment and control groups), but these methods are not suitable for one-on-one matching (i.e., for forming matched pairs).

Propensity-score matching may be appropriate for distributional matching (for bias reduction), but it is not appropriate for the formation of matched pairs for precision or power enhancement. The reason for this is that two individual units may have similar propensity scores yet differ markedly with respect to the values of the variables comprising the propensity score. Properly applied, propensity-score matching can be used *to reduce bias* (by forming distributionally similar treatment and control groups), but it can produce terrible results if used to match individual units (to increase precision). An important feature of powerful evaluation designs is the introduction of correlations

between units in the treatment and control group. This is done by matching of individual units (i.e., by forming matched pairs) – not by matching the treatment and control groups overall. Forming matched individual pairs on the basis of propensity scores is, in general, a very poor approach – matched pairs must be formed by comparing the units on all match variables, not simply on a composite (scalar, one-dimensional) score. In view of the fact that individual matching to form matched pairs is a crucial ingredient of research design for evaluation, it is a wonder that PSM is ever used, except as a check on results. The PSM approach can reduce bias, but it is not an effective means for increasing precision. Good evaluation designs should involve matched individual treatment-control pairs (to increase precision and power), not just matched treatment-control groups.

Propensity-score matching should not be used for small sample sizes (i.e., small treatment and control groups). Since matching is usually done on first-stage sample units (for which data are available for them prior to the survey), the sample sizes involved in matching may in fact sometimes be quite small. The problem that arises is that propensity-score matching is a distributional property. It involves conditional expectations. The quality of the match is determined by the matching *process*, not by the individual matches that it produces. This situation is similar to assessing randomness – randomness is a property of the *process* that generates a sample, not of a particular sample itself. The situation is also similar to invoking the law of large numbers and the central limit theorem in sample survey – these theorems “work” only if the sample sizes are large. If the number of treatment and control units is small, then the small number of match-sets formed by matching propensity scores may be very poor matches indeed. For small matched groups, matching should be done using a matching procedure that takes into account the value of each and every important match variable (such as exact matching, or a multidimensional match score that considers “nearness” of units with respect to every match variable).

It is interesting to note that propensity-score matching is often misapplied. First, it is a poor method of forming matched individual pairs. Second, a recent survey of published articles reveals that the analysts did not even use the correct statistical methods for analyzing matched-pairs data. (See “A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003” by Peter C. Austin, *Statistics in Medicine* 2008, Vol. 27, pp 2037-2049.) In view of the poor results that may be obtained by using propensity scores as the basis for forming matched pairs, it is quite possible that these incorrect analyses resulted in the additional loss of little precision and power – it is likely that the opportunity for achieving high power and precision was thrown away when PSM was used in the design.

(In the cited article, Rosenbaum and Rubin prove that if treatment assignment is strongly ignorable given a set of covariates, x , then it is strongly ignorable given any balancing score based on x (of which the propensity score is one). (A treatment assignment is strongly ignorable given a set of covariates, x , if treatment assignment and response are conditionally independent given x .) This is a sufficient condition. It means that matching of treatments and controls on a propensity score based on strongly ignorable covariates will result in an unbiased estimate of the average treatment effect, if there are no hidden variables affecting the response. The practical difficulty associated with this result is that it refers to properties of *distributions* – it produces matched *samples*, not matched *pairs*. The theorems that Rosenbaum and Rubin present apply to *random samples*, where the matching is done by random sampling on x . The results are conditional on the value of the propensity score as a function of the *random variable* x . This is a crucial assumption. The matching may be done on the propensity score only for a *random sample* of x . To select one or more groups having a specified value or range of values of the propensity score, one must *randomly sample* the conditional distribution of x conditioned on this value. This process produces matched *samples*, not matched *pairs* – even if “nearest neighbor” matching is employed.)

Because propensity score matching is often misunderstood and misused, some further comments will be made about it. Propensity scoring was originally developed to assist the selection of comparison groups for clinical trials, to reduce the bias introduced when treatment and control groups are not formed by randomization and may hence differ with respect to variables that affect outcome. A comparison group was selected by finding, for each person having a medical condition, a non-ill person having the same propensity score (i.e., likelihood, or propensity, of having the disease). While this may be a reasonable approach for constructing comparison groups for clinical trials, it is not a reasonable for constructing comparison groups for evaluation research studies in general. (The problem is that it generates matched *samples*, not well-matched *individual pairs*. While it may not be feasible in medical studies to form matched pairs of individuals, it is often feasible in evaluation studies to form matched pairs of primary sample units.) The goal of propensity-score matching is to select a comparison group such that the (joint) probability distribution of the comparison group and the treatment group are the same, over all available match variables. As mentioned, this may be done in either of two ways: (1) by matching the probability distribution, or (2) by matching individual units (in which case the probability distribution also matches), in which case the matched pairs can be used to effect paired-comparison tests, which are more precise than comparisons between unrelated samples. Properly applied, propensity-score matching may be a useful procedure for reducing bias, but it is a poor method for increasing precision or power (via matched pairs). Just because two units have the same probability of being included in the treatment group does not at all mean that they are similar with respect to quantities that may affect program outcome.

A simple example will illustrate this point. Suppose, for example, that people are drafted into the Army based solely on height, and that particularly short and particularly tall people are rejected (so that fitting in uniforms is simplified). In this case, a very short person and a very tall person have about the same propensity score. Suppose further that we are interesting in measuring the ability of draft rejects to jump high, i.e., our “treatment” group is draft rejectees. For this measure of performance, short people will perform much less well than tall people. The treatment group (draft rejects) will include both short and tall people. If we select a matched sample based on propensity score, however, it is possible that we could match a short person with a tall person, because they have the same propensity score. This would be a terrible individual match and this matching procedure would produce a terrible comparison group: the matching process would have introduced a massive amount of variation between “matched” units, not decrease it. In this case, it would have been much better to match on height, not on propensity score. The outcome measure (ability to jump high) is highly related to height, and not related at all well to propensity score. In this example, the propensity score is perhaps the worst possible one-dimensional matching variable of any that might be reasonably considered. It amplifies differences between matched units, rather than reducing them – it would have been far better to do no matching at all. Matching on the basis of the propensity score may produce treatment and control groups that are similar with respect to observable variables (i.e., such that the unit response and treatment are conditionally independent of observed covariates), but this procedure can produce *absolutely terrible* matched pairs.

This simple example shows how important it is that a composite matching score be constructed such that the matched units (either individuals or groups) are similar with respect to variables related to the outcome measures of interest and to selection for treatment, not simply with respect to the *probability* of selection into the treatment group, i.e., on the propensity score (or on any other single (scalar, one-dimensional) attribute). If they are properly matched, then they will match on the propensity score, but simply because they match on the propensity score does not imply that they are matched on variables that are related to outcome or selection for treatment.

It is important to remember that the precision of difference estimates may be dramatically improved if the matching is done for individual units, rather than just for distributions (so that a matched-pairs comparison may be made). All that is required (to reduce bias) is for the distributions to match, but the improvement in precision of the difference estimate and the power of a test of differences can be tremendous if comparisons are made between individually-matched pairs instead of between distributionally-matched samples. When the matching is done on individual units, the samples are also matched (i.e., the samples are matched on groups defined by the match variables). It is very important to realize, however, that the use of matched pairs increases precision and power only if the matches are “good,” in the sense that there is a substantial correlation associated with the matched pairs. If the matching of pairs is poorly done, such as using “nearest neighbors” based on propensity scores, there may be little or no increase in precision and power. (It is worth emphasizing here that if the design involves matched pairs, then the analysis must recognize this design feature. The Austin article cited earlier reports that researchers often fail to do this.)

In summary, propensity-score matching should not be used as a basis for forming matched pairs in evaluation designs. Many evaluation studies involve estimation of a double-difference (or “difference-in-difference”) estimate of program impact. In order to increase the precision of double-difference estimates and the power of double-difference tests, it is necessary to introduce correlations between individual units of the treatment and control groups. In view of the fact that the formation of matched pairs (of treatment and control units) is the way in which correlations are introduced into evaluation designs, and in view of the fact that propensity-score matching is a terrible method of forming matched pairs, it is concluded that propensity score matching should never be used as the basis for constructing evaluation designs, except perhaps as a check on the distributional similarity of large treatment and control groups. To reiterate: Propensity-score matching can be used to reduce selection bias by forming comparison groups that are similar to treatment groups, but it is an inappropriate method for forming matched pairs (to increase precision and power). Once a set of variables is available for matching, it is ridiculous to use them solely to reduce bias, but not to increase precision and power. Propensity-score matching can do only the former and not the latter. A different method of matching must be used to form matched pairs (viz., one that takes into account the similarity of units on specific match variables). Since only one matching procedure is used (i.e., it is not the case that one method is used to form matched groups and a different one to form matched pairs), that method cannot be propensity-score matching. For this reason it is of little use in evaluation design (except as a check on the similarity of large treatment and control groups).

For a probability sample of units, each population item must have a known, nonzero probability of selection (or all selection probabilities must be equal, if they are unknown). If matching is done after selection of the treatment sample, then it is not possible to determine the selection probabilities. In some applications, the treatment sample will have been selected prior to the researcher’s involvement. In such cases, it is not possible to determine the ex-ante selection probabilities for the treatment units – they are taken (ex-post) to be one. Appendix A describes a matching procedure for which the selection (inclusion) probabilities can be determined.

Now that we have discussed some of the major aspects of sample design for analytical surveys, it is appropriate to summarize the four major types of analytical survey designs used in impact evaluation studies.

1. Experimental design. Pretest-posttest design with control group selected by randomization. Experimental design is the best approach to measuring causal effects, but

it is generally not feasible to allocate the treatment (program intervention) using randomization in socio-economic programs. If randomized assignment of the treatment is possible, then the influence of all other variables on outcome and the impact estimate is removed. Examples of experimental design are presented in William G. Cochran and Gertrude M. Cox's *Experimental Designs* (2nd edition, Wiley, 1957). (Other related texts include George E. P. Box and Norman Draper's *Evolutionary Operation* (Wiley, 1969) and Raymond H. Myers and Douglas C. Montgomery's *Response Surface Methodology* (Wiley, 1995). E. S. Pearson and H. O. Hartley's *Biometrika Tables for Statisticians* (Cambridge University Press, 2nd edition, 1958) contains tables of orthogonal polynomials.)

2. Quasi-experimental design. Pretest-posttest design with a comparison group selected by matching. If possible, to promote local control over time, the pretest and follow-up surveys are implemented as panel surveys on the same units (e.g., households, businesses). Ideally, exact matching is employed to construct the comparison group, in which each treatment unit is matched to a similar non-treatment unit. This individual-unit matching ensures that the joint probability distribution of the treatment units and the non-treatment units are similar, and also allows the use of a "matched-pairs" estimate of impact (via the double-difference estimate). (In this article, attention has focused on the pretest/posttest/comparison-group quasi-experimental design. There are many other kinds of quasi-experimental designs. See *Experimental and Quasi-Experimental Designs for Research* by Donald T. Campbell and Julian C. Stanley (McGraw Hill, 1963, 1966), or *Quasi-Experimentation: Design and Analysis Issues for Field Settings* by Thomas D. Cook and Donald T. Campbell (Houghton Mifflin, 1979) for examples.)
3. Analytical model. A general linear statistical model (e.g., a multiple regression model) is specified that describes the relationship of program outcome to a variety of explanatory variables related to program intervention, but there is not a well defined comparison group. This type of design is appropriate, for example, in the evaluation of road-improvement projects, where program impact can be viewed as a continuous function of travel time or travel cost (which can be measured directly, reported by survey respondents or estimated by a geographic information system). For example, a "path-analysis" (hidden variable) model may be used to describe the relationship of income to travel cost, and an engineering model may be used to describe the relationship of travel cost to road characteristics (which are affected by the program intervention). The ease with which an analytical model may be specified varies. For a road-improvement program, it may be generally agreed that the preceding model is a reasonable representation of reality. For a training program, it may be much more difficult to specify a reasonable model (so that use of a randomized experimental design is much preferred, and it is not necessary to worry about the relationship of impact to omitted variables). References on the general linear model include Norman Draper and Harry Smith's *Applied Regression Analysis* (Wiley, 1966); David W. Hosmer and Stanley Lemeshow's *Applied Logistic Regression* (Wiley, 1989); and C. Radhakrishna Rao's *Linear Statistical Inference and Its Applications* (Wiley, 1965 (the last book is theoretical).
4. Attribution of causality via open-ended questions. It may be that no experimental or quasi-experimental design is feasible, and it is not clear how to specify an analytical model describing the relationship of program outcome to program intervention (or no data relating to the explanatory variables of such a model are available prior to the survey). In this case, it may be that the best that may be done is to directly ask the respondents what they attribute observed changes (in impact variables) to. This is similar to the "focus-group" approach. (Note that including open-ended questions about the reasons underlying

observed changes is also useful in the two preceding design types, since once we depart from a true experimental design with randomized control groups, there is always a question about the cause underlying observed changes.) For this type of design, the analytical-survey design methodology described in this article is not helpful, since it is not clear what variables should be used for an analytical model (or data on them is not available). In this case, the survey design will be similar to a descriptive-survey design (e.g. a simple random sample, stratified sample, or multistage design). The design may be stratified into domains for which it is suspected that the model specification may differ, but that is probably all that is done with respect to tailoring the design to assist identification of an underlying analytical model. This approach is really too “weak” (vulnerable to threats to (internal) validity) to be considered for a formal “impact evaluation,” but it may be useful as part of a monitoring system that may suggest hypotheses to test in a future rigorous impact evaluation and collect data useful for its design.

The methodology described in this article (and, in particular, in Appendix A) assists the development of analytical survey designs for cases 1-3 above.

It is important to recognize that randomization (random selection of units and random assignment of treatment values to units) is not sufficient to guarantee good results, and elimination of all bias. The essential ingredient of a designed experiment (randomized trial) is that treatment assignment and response are independent. Randomized selection and randomized assignment of treatment values does not assure independence of response. A simple example serves to illustrate this. Suppose that we wish to evaluate a worker-development (training) program, which teaches basic job skills to workers, such as proper dressing, proper speech, résumé preparation, interview skills, dependability, and the like. The objective of the program is to increase employment and job retention. Let us suppose that the program is in fact very effective in increasing the likelihood that a person lands and keeps a job. Suppose however, that the number of jobs in the program area is fixed. If the graduates of the program get jobs and keep them, they are simply taking these jobs away from others. In this situation, even the use of a before / after / randomized-control-group experimental design would show that the program was very effective helping workers to get and keep and in increasing employment. The training program is effective only in redistributing existing jobs, not in creating new ones. The problem is that there is an interaction among the experimental units (individuals receiving training) – if one person gets a job, someone else has to lose one. This effect has been called the “Stable-Unit-Treatment-Value-Assumption” (or “SUTVA”). It is also called the “partial equilibrium assumption” or the “no-macro-effect” assumption. (For discussion, see Rubin, Donald B., “Bayesian Inference for Causal Effects: The Role of Randomization,” *Annals of Statistics*, vol. 6, no. 1, pp. 34-58 (1978); or Imbens, Guido W. and Jeffrey M. Wooldridge, “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, vol. 47, no. 1, pp 5-86, (2009); or Morgan, Stephen L. and Christopher Winship, *Counterfactuals and Causal Inference*, Cambridge University Press, 2007). Other examples of this effect are easy to cite. For example, a farmer-development program might teach farmers how to grow tomatoes, but if the demand for tomatoes is fixed, the increased production may serve simply to drive prices and farmer income down.

Some Additional Comments on Matching

This section discusses some of the problems and procedures associated with matching. See the referenced article by Daniel E. Ho et al. for detailed discussion of matching.

Matching may be done before the sample is selected (“ex ante”) or after the sample is selected (“ex post”). If done prior to sampling, it involves just the variables that are known prior to sampling.

This set of variables may be substantially smaller than the set available after the survey is completed, and may be available only for higher levels of aggregations (e.g., for census enumeration areas, or for districts). The article and computer program (MatchIt) of Ho are directly concerned with ex-post matching (although much of their exposition pertains to ex-ante matching as well). The present article is concerned primarily with matching as a survey design tool (i.e., ex ante). Since ex-post matching involves dropping of observations (from the sample), it is also referred to as “pruning,” or “trimming.”

When matching is done ex ante, it is generally attempted to have treatment and control groups of similar size. If individual-unit matching is done, the treatment and control groups will be of exactly the same size. When matching is done ex post, many comparison units (or even treatment units) may be dropped from the sample. For this reason, when doing ex post matching it is generally desired that the sample of comparison units be substantially larger than the sample of treatment units.

Matching is done for two reasons – to increase the precision of estimates, and to reduce bias of estimates. In the latter case, it is usually employed when randomized assignment of treatment to experimental units is not feasible. In either case, the primary goal of matching is to make the joint (multivariate) probability distributions of the treated units and the untreated units as similar as possible, with respect to all variables that may have an effect on outcome or may have influenced the selection of the treatment units. This reduces bias. The second goal of matching is to increase precision, e.g., by use of individual matching to enable “matched pairs” estimation.

The precision of an estimate is measured by the standard deviation (usually called the standard error when it is referring to an estimate) or its square, the variance. (The mean (or expectation) is the expected value of the (population or sample) elements. The variance is the expected value of the squares of the deviations of the elements from the mean.) Bias is the difference between the expected value of a parameter estimate and the true value of the parameter. Accuracy is a combination of precision and bias. The usual measure of accuracy (of a parameter estimate) is the mean squared error (MSE), or expected value of the squared deviations of the elements from the true value of the parameter. The mean squared error is the variance plus the square of the bias. To achieve the goal of high accuracy, it is important that both the variance and the bias be low. It may be the case that by sacrificing a little in precision, a substantial reduction in bias can be achieved, so that the accuracy is improved.

The precision of estimates of differences between groups is increased when comparisons are made between units that are similar to each other with respect to all variables except the treatment variable. Bias is reduced when the probability distributions of the treatment and non-treatment variables are similar. Of the objectives of matching (precision improvement or bias reduction), matching to reduce bias (e.g., a selection bias, such as when roads are selected for improvement, or individuals volunteer for a program) is the more problematic. It is not that the procedure is more complicated, but that its effectiveness is usually more limited. Furthermore, unlike the case of precision, which may be easily measured, it is generally not possible to estimate the bias. (We are referring here to selection bias, not bias associated with estimators, which may be reduced by procedures such as jackknife estimation.) Matching may be very effective or of limited help. For ex-ante matching, its effectiveness depends on what information is available prior to the survey. For ex-post matching, its effectiveness depends on how many observations may be dropped (pruned, trimmed) from the sample. In either case (ex post or ex ante), it usually helps and it may be quite effective. In the case of matching to increase precision (e.g., reinterview of the same households in a panel survey, or identification of matched pairs to which treatment will be

randomly assigned to one), things usually don't go very wrong (unless the investigator applies a terrible matching procedure, such as propensity-score matching in the example presented earlier).

In the propensity-score matching example presented earlier, it was shown how matching for variance reduction (precision improvement) could in fact make things terribly worse. A similar disaster can occur in matching for bias reduction. In the case of ex-ante matching, this occurs in the case in which matching is based on an unreliable pre-measure of the outcome variable, in which case a "regression effect" bias is introduced into the impact measure (using a pretest / posttest / comparison-group design). This phenomenon is discussed, for example, in Elmer L. Streuning and Marcia Guttentag's *Handbook of Evaluation Research*, vol. 1, pp. 183-224 (Sage Publications, 1975), and summarized (graphically) in my notes at <http://www.foundationwebsite.org/ApproachToEvaluation.htm> or <http://www.foundationwebsite.org/ApproachToSampleSurveyDesign.htm> . In the case of ex-post matching, bias may be introduced or increased if observations are dropped depending on the value of the dependent variables. For ex-post matching, the decision to drop observations may be made based only on values of the independent variables, not the dependent variables or any variables dependent on them.

When randomized assignment of treatment is done, the effects of all other (omitted) variables are eliminated (assuming that the unit responses are independent). When randomized assignment is not possible, all that can be done is to try to reduce the effect of uncontrolled variables on the outcome. This is attempted in a number of ways, including ex-ante matching, covariate adjustment, and ex-post matching (or "pruning"). The fundamental difficulty with these procedures as substitutes for randomization is that they are weak "second bests." One can never be sure how effective matching or covariate adjustment have been in reducing bias, and experience suggests that they are not very effective. Data may not be available on important omitted variables, so that no control is possible. There is, quite simply, no effective substitute for randomization (and even randomization does not solve all problems, such as a lack of independence).

When matching is done prior to the survey, it is desired to maintain knowledge of the probabilities of selection of the population units. The reason for this is that it is possible to construct the standard descriptive-survey estimates (e.g., of population means or totals) only if those probabilities are known and nonzero (or known to be equal). When matching (or pruning) of data is done ex post (i.e., on the sample), knowledge of the selection probabilities is lost. Once the probabilities of selection are unknown, the ability to estimate overall population characteristics, such as population (or subpopulation) means, totals, or an average treatment effect (ATE) from the sample data is lost. At this point attention centers on estimation of model parameters and differences, and estimates that are conditional on the sample (such as the average treatment effect on the treated (ATT)).

In an analytical survey, estimation of population means or totals based on parametric models (of an underlying probability distribution) is difficult or impossible, since these models usually describe relationships among variables, not overall characteristics of the population. Once knowledge of the selection probabilities is lost, other procedures, such as synthetic estimation (usually used in the field of demography) are more appropriate than the usual methods of sample survey, for estimation of overall population characteristics. The situation is analogous to that of time series analysis: Once the data have been filtered (differenced) to achieve stationarity, it is no longer possible to estimate the mean level of the process from the sample (filtered) data. The stationary model used for analysis no longer contains information about the mean level of the process. Similarly, a model of the relationships among variables will typically not contain any information

about the mean of the population (since without probability sampling it is not possible to make statistical inferences about the population mean from the sample mean).

Randomized assignment of treatment eliminates the possibility of bias from all uncontrolled sources (assuming that the responses of individual units are independent). Without randomization, bias may be introduced by any uncontrolled source, or “omitted variable.” As observed by Ho, a variable must be controlled for if it is causally prior to treatment, empirically related to treatment, and affects the dependent variable conditional on treatment. It is not generally realized, however, that the only effective control is inclusion of the variable as an independent variable in a controlled experiment. If it is desired to predict how a system will behave when a forced change is made in a variable, the prediction model must be derived from data in which forced changes are made in the variable. (Paul Holland and Donald Rubin coined the insightful aphorism, “No causation without manipulation” mentioned on p. 959 of “Statistics and Causal Inference” by Paul Holland, *Journal of the American Statistical Association*, Vol. 81, No. 396 (Dec. 1986), pp 945-960.) A model derived from passively observed data cannot reliably be used to predict what will happen when forced changes are made. This is the reason why econometric models are so notoriously poor when used for forecasting (prediction). They may “fit” past data very well, but that is not very helpful. As the securities sellers invariably comment, “Past performance is not a guarantee of future performance.”

In most surveys, much of the data on explanatory variables is not known until after the survey has been completed. Often, the data that are available before the survey are for aggregate administrative or geographic units, such as census enumeration areas, regions or localities, rather than for the ultimate (lowest-level) sample unit, such as a household, business, school or hospital. This means that whatever matching is done will be done for sample units at a relatively high level of aggregation, and for “surrogate” variables rather than the variables of primary interest. Under such conditions, matching may not be highly effective (unless the intraunit correlation coefficient is very high).

Evaluation literature often implies that matching is an effective alternative to randomization, and that it “removes the bias” associated with nonrandomized selection for treatment. Such statements are false. While matching *may* reduce bias, it may not, or it may not reduce it very much, or may not reduce it to an acceptable level. An important omitted variable may be unknown or overlooked, or it may be that no information is available about an important omitted variable. Whether matching is effective depends on many factors, and it is generally never known how effective it is (except in simulation studies, where the model is specified by the investigator). Matching and covariance estimation are poor substitutes for randomization. (As noted earlier, even randomization is not a “silver bullet,” e.g., if unit responses are not independent.)

After a survey is completed, it may become much clearer how much a “matched” comparison group differs from the treatment group, with respect to independent variables. Ex-post matching may be implemented at this time (i.e., after the survey data are available) to increase the similarity of the probability distributions of independent variables for the treatment and nontreatment groups. The goal of ex-post matching is usually bias reduction (or reduction of model dependence for estimates of interest), although precision may also be increased. Usually, the limited amount of sample data places severe restrictions on what can be done. Although data are available on many more variables than before the survey, and at lower levels of aggregation, the number of units that are available for matching (or pruning) is very limited (i.e., limited to the sample, rather than to the population). In ex-ante matching, the entire population may be “tapped” to seek matching units for treatment units. In ex-post matching, the two groups of units (treatment units and comparison units) are both small. About all that can be done is to “prune” the groups (delete observations) so

that they have a common support (the same range of variation for each explanatory variable). As long as the model is correctly specified, any observations may be deleted from the sample, without introducing bias into the parameter estimates, as long as the criteria for doing so are a function only of the independent variables (including the treatment variable), not the dependent variables. Of course, dropping observations may decrease precision, but this may be viewed as an acceptable cost if it means that biases may be reduced. Note that pruning the data may have the dual effect of increasing precision (even though the sample size is reduced) *and* reducing bias. Pruning may increase precision if it reduces the correlations between the treatment variable and other independent variables. *To allow for pruning (ex post matching), it is advantageous for the sample size of the comparison group to be somewhat larger than the sample size of the treatment group.* Dropping observations to the point where the sizes of the treatment and comparison groups are comparable will have little effect on precision of a difference estimate. (Note that pruning of the sample is done only by dropping observations based on the values of the *independent* variables. Although this allows for dropping of observations based on the values of the treatment variables (since they are independent variables), this is generally not done. Note that dropping of variable is not appropriate for a descriptive survey – it is appropriate only for model-based (analytical) surveys.)

As mentioned, it is desired that the joint probability distribution of the explanatory variables be similar for the treatment group and the nontreatment group. If the explanatory variables are related, attainment of this goal is practically impossible. A much more realistic goal is to work with a set of stochastically independent explanatory variables, in which case all that is required is that the marginal probability distributions be similar (for the treatment and nontreatment groups). In the matching method presented in Appendix A, it is attempted to determine a set of uncorrelated variables, so that matching on the marginal distributions is reasonable.

There are many methods of matching, such as those implemented in Ho et al.'s MatchIt computer program. An advantage of exact matching of individual units is that this procedure helps assure that the *joint* probability distributions of the independent variables are similar for the treatment and nontreatment groups (with respect to the match variables). Most other matching methods simply aim to make the marginal probability distributions similar. Another advantage of exact matching is that when individual units are matched then “paired-comparison” tests of differences may be done (which are generally far more precise than tests of unrelated samples).

Matching prior to sampling is more difficult than matching (pruning) after sampling, since the goal is to do the matching in such a way as to maintain knowledge of the sample selection probabilities. As mentioned, when matching (pruning) is done on the sample, knowledge of the selection probabilities is lost. Once the goal of keeping track of the sample selection probabilities is abandoned (i.e., when matching / pruning the sample data), the ability to construct overall-population estimates such as means or totals from the sample data alone is essentially lost. From this point on, the estimates are conditional on the particular sample. Attention now centers on estimation of model parameters and differences (e.g., a double-difference estimate of program impact). The goal is to make the estimates as “model independent” as possible. Once the goal of keeping track of the selection probabilities is abandoned, the job of matching (or pruning) becomes much easier. The matching can be done on one variable at a time, iteratively, until all of the marginal probability distributions match. If matching is done on the sample, it is desirable for the comparison-group sample to be larger than the treatment-group sample, so that after trimming, the two samples are of comparable size.

If we are dealing with uncorrelated explanatory variables, the goal in matching is for the marginal distributions to match. The equivalence of two probability distributions may be tested, for example,

with a Kolmogorov-Smirnov test (or a chi-squared test). Some matching procedures involve matching particular distributional characteristics, such as the mean, the variance, or the support (the range over which observations occur). If matching is done on scalar (one-dimensional) characteristics (such as a mean, propensity score, Mahalanobis distance, or other distance measure), it is essential to compare the complete distributions of each independent variable (for the treatment group vs. the nontreatment group) after matching is done, to make sure that they are similar overall. If matching is done on one variable at a time, it is important to check that the matching of previously considered variables is still satisfactory (although if the variables are unrelated (uncorrelated), then this problem is reduced).

It should be recognized that while the quality of some estimated population characteristics, such as means or totals, may be sensitive to the form of the underlying parametric model (which describes the distribution of the dependent variables as a function of the independent variables), the quality of others, such as a difference (or double difference) in means, may not be. Increasing the likelihood that they are not is, in fact, the goal of matching – to reduce model dependence for estimates of interest to acceptable levels. (In general, however, estimates based on correctly specified models are better than those based on incorrectly specified models.) This is analogous to the situation in econometrics where a forecast derived from a model may be unbiased even though the model parameter estimates may be highly biased. The goal of matching (or pruning) is to reduce the *model dependence* of the estimates of primary interest (such as a double-difference estimate), so that the estimates of interest are not adversely affected by the choice of parametric model.

Note that if the parametric model (of the underlying distribution considered to have generated the observations) is correctly specified, the parameter estimates will be correct even if the distributions of the treatment and nontreatment units are not identical. Furthermore, if the (joint) probability distributions of the treatment and nontreatment units are identical, the estimated difference between treated and untreated units, adjusted to common values of the other variables, is consistent (converges to the correct value as the sample size increases). Unfortunately, it is in general not possible to prove whether a parametric model is correctly specified, and it is not possible to prove that the joint probability distributions of all omitted variables (i.e., any variables that may affect outcome or may have influence selection for treatment) are similar. Because of the presence of correlations among variables (collinearity), it is generally not possible to determine the “correct” model. For this reason, it is necessary to rely on matching, rather than model specification, as the primary method reducing the bias of survey estimates. After a matching procedure has been applied, the results can be viewed to assess the quality of the match (e.g., by comparing the marginal distributions). Assessment of the correctness of a model specification is substantially more difficult to achieve. If the model dependence of the estimates has been decreased to a low level, then this becomes less of a concern. The goal of matching (and pruning) is to reduce model dependence.

As mentioned earlier, matching may be done in two ways – by matching individual units, or by matching distributions (i.e., matching groups or samples). Which is done depends on the situation. When there is a choice, matching of individual units is preferable, for two reasons: it produces matched pairs (which generally leads to high precision for estimates of differences), and it leads to a match on the *joint* probability distribution of the match variables, not just the marginal distributions (so that interrelationships among the variables are matched). Matching of individual units may be done to increase precision (by promoting local control, e.g., through a “matched pairs” design), or to reduce bias, by obtaining a comparison group that is a substitute for a randomized control group (i.e., that is “statistically” (distributionally) similar to a treatment group). Individual matching accomplishes the dual objectives of increasing precision and reducing bias.

Matching of individual units applies to the first two types of evaluation designs identified earlier, viz., experimental designs and quasi-experimental designs.

In an evaluation context, it is often the case that the “treatment” units (units subjected to the program intervention) are not randomly selected. Moreover, it may be impossible to find any similar units that may be used as suitable candidates for matching. An example of this is a development project to improve roads – the roads to be improved are usually selected according to political or technical criteria, not by randomization. In this example, marginal distributions may be controlled to achieve variation in variables that are important in road selection or program outcome, and to achieve low correlation among them, but matching of individual units may not be a reasonable objective because the treatment areas or roads are unique (e.g., the road being improved is the country’s only superhighway, or the improvement program takes place in one region concurrent with several other development initiatives). In this example, there is no population similar to the treatment population, from which comparison items may be selected. In this situation, a reasonable alternative approach is to develop an analytical model in which treatment is represented by multiple levels, rather than by just two levels (treatment and non-treatment). For example the effect of a road improvement may be represented as a function of change in travel time caused by the road improvement. In this case, the methods of Appendix A could be applied to achieve desirable distributional characteristics for the sample (e.g., spread, balance, orthogonality), but no matching of individual units would be done.

Control of marginal distributions (to achieve spread, balance and orthogonality) is useful in all three of the “quantitative” evaluation designs identified earlier: experimental design, quasi-experimental designs, and analytical models. It is optional in experimental designs (e.g., the design may be intended simply to assess the overall impact of a program, with no desire to develop models of the relationship of impact to explanatory variables), highly desirable for quasi-experimental designs (to adjust for the effects of covariates), and essential for analytical models (where neither randomization nor matching are available to eliminate or reduce the effects of extraneous variables).

Note that if matching of individual treatment and non-treatment units is done, then (ideally) all other variables are made orthogonal to them (i.e., to the treatment indicator variable). The question may be asked why one would attempt to achieve orthogonality by working with marginal distributions, when matching of individual units is easier. There are several reasons. First, matching introduces orthogonality of the treatment variable with respect to the other design (match) variables, but it has no effect on the orthogonality among the other design variables. (It is desired that the correlations among explanatory variables used in a model be low (low “multicollinearity”), in order to increase the precision and decrease the correlation of model parameter estimates.) Another reason is that while matching leads to orthogonality with respect to treatment, it does nothing about the spread or balance of the design variables, or about other stratifications that may be desired. Those aspects are also important from the viewpoint of model development (determining the relationship of impact to explanatory variables). While the design tools of matching of individual units and control of marginal distributions are related, the objectives in using them are not identical, and there are situations in which individual matching is not feasible. Ideally, both techniques are used to construct an analytical sample design, but that is not always possible. Matching of individual units is a powerful technique for increasing the precision of difference estimates and for assuring orthogonality of the treatment variable with respect to the other design variables. If all that is to be done is to estimate the average treatment effect, then individual-unit matching of the treatment and non-treatment groups is all that would be required. Since most evaluations are concerned with estimation of the relationship of impact to design

variables, however, control of the spread, balance and orthogonality of the other design variables is virtually always of interest. Note that when both procedures are used (i.e., matching of individual units and control of spread, balance and orthogonality by marginal stratification), the match variables would typically be the same variables as used to control marginal stratification.

7. Sample Size Determination

Sample Size Determination for Descriptive Surveys

Statistical program packages such as Statistica, Stata, SPSS or SAS contain modules for determining sample sizes in survey applications, but they are applicable mainly to descriptive surveys and usually do not apply directly to the double-difference estimator (i.e., the pretest / posttest / comparison group quasi-experimental design). There are numerous free computer programs available on the Internet for calculating sample size. For example the SampleXS program provided by Brixton Health, posted at <http://www.brixtonhealth.com/SXSetup.exe>. Most of these programs, too, calculate sample sizes for simple descriptive surveys. They usually take into account design features such as stratification, clustering and matching only indirectly, through specification of the “design effect” (deff), which is the ratio of the variance of an estimate using a particular sample design to the variance using simple random sampling (with replacement). Furthermore, they often apply the “finite population correction” (FPC) to adjust the variance. As discussed earlier, the FPC is applicable only to descriptive surveys, not analytical surveys.

The Design Effect, “deff”

Some comments are in order about the role of the design effect (deff) in the sample-size formulas. The value of deff is determined by the nature of the sample design. In a descriptive survey, it is determined by the design features, such as stratification, multistage sampling, and selection with variable probabilities (e.g., selection of primary sampling units (PSUs) with probabilities proportional to size). For simple random sampling, the value of deff is 1.0. In general, the value of deff may be less than one or greater than one, depending on the design. If a design incorporates stratification in an effective way, the sample estimates could be much more precise than for simple random sampling, and the design effect would be less than one. If stratification is used to determine estimates of comparable precision for subpopulations, the estimate of the population mean could be much less precise than for a simple random sample, and the value of deff would be greater than one. For most socio-economic surveys, the design effect has a value greater than one, such as two or three or more. The principal reason for this is that most such surveys involve multistage sampling, and the PSUs are usually internally more homogeneous than the general population, and this decreases the precision of the survey estimates (of population means and totals). This effect is measured by the “intra-unit (or intracluster) correlation coefficient.” Although there are formulas that show the relationship of the variance of a survey estimate to the intra-unit correlation coefficient, its value is often not known (unless a similar survey has been done before), and so these formulas may not be very helpful. Stratification may increase or decrease the precision of population estimates, depending on how it is being used. For large surveys, the analysis of variances will often include estimation of the deff, and that can be used to assist the design of later surveys. Note that the deff is different for each estimate. If no information on the deff is available, and no data are available to estimate it, then judgment will have to be used to estimate its value. The survey designer must judge whether each aspect of the design would likely cause the precision of estimates of the population mean or total to be more precise or less precise than if a simple random sample had been used, and set the value of deff accordingly.

Some Comments on Determining Sample Size for Single-Stage Cluster Sampling and Two-Stage Sampling

While it is difficult to estimate the effect of stratification on the deff, some useful comments can be made about the effect of single-stage cluster sampling or multistage sampling on it. First we consider single-stage cluster sampling (in which all of the elements of a cluster are included in the sample). In cluster sampling there are two population means of interest – the mean of the cluster (or unit) totals, and the mean per element. Let us denote the mean per element by \bar{Y} . The variance among elements is

$$S^2 = \frac{\sum_{i,j} (y_{ij} - \bar{Y})^2}{NM - 1}$$

where N denotes the total number of clusters in the population and M denotes the number of elements per cluster. (See W. G. Cochran, *Sampling Techniques* 3rd edition (Wiley, 1977) for the formulas presented here.) The variance of the sample mean per element,

$$\bar{y} = \frac{\sum^n y_i}{nM}$$

where n denotes the sample size, is

$$V(\bar{y}) = \frac{1-f}{n} \frac{NM-1}{M^2(N-1)} S^2 [1 + (M-1)\rho]$$

where $f = n/N$ and ρ denotes the intraclass correlation coefficient, defined as

$$\rho = \frac{E(y_{ij} - \bar{Y})(y_{jk} - \bar{Y})}{E(y_{ij} - \bar{Y})^2} = \frac{2\sum_i \sum_{j < k} (y_{ij} - \bar{Y})(y_{jk} - \bar{Y})}{(M-1)(NM-1)S^2}$$

The formula for a sample of nM elements drawn using simple random sampling is the expression on the right-hand-side of the preceding formula preceding the brackets. Hence the bracketed expression, $[1 + (M-1)\rho]$, indicates how much the variance differs for cluster sampling from the variance for a simple random sample of the same size (nM). This is Kish's deff for cluster sampling.

If it is assumed that we are sampling clusters from a conceptually infinite population of clusters, then this formula reduces to

$$V(\bar{y}) = \frac{1}{n} \frac{M-1}{M^2} S^2 [1 + (M-1)\rho]$$

The formula for ρ is a little complicated:

$$\rho = \frac{(N-1)M^2S_1^2 - (NM-1)S^2}{(NM-1)(M-1)S^2}$$

where S_1^2 denotes the variance between unit (cluster) means (note that Cochran uses a slightly different formula, involving the variance, S_b^2 between the cluster totals on a single-unit (element) basis). For N large (or assumed infinite), this simplifies to

$$\rho \approx \frac{MS_1^2 - S^2}{(M - 1)S^2}$$

This gives the following approximation for S_1^2 in terms of ρ :

$$S_1^2 \approx S^2 \frac{\rho(M - 1) - 1}{M}$$

We also have, for N large, the following approximation for S_2^2 in terms of ρ :

$$S_2^2 \approx S^2(1 - \rho)$$

where S_2^2 denotes the within-unit (within-cluster) variance.

Note that ρ can be negative only if M is small. For M large, ρ is approximately equal to S_1^2/S^2 , which is positive.

In most applications, the values of S_1^2 and S_2^2 are not known, but reasonable assumptions can be made about the value of ρ . If the clusters are relatively internally homogeneous, the ρ is large, e.g., .5 to 1. If units within clusters vary about as much as the general population, then ρ is small, e.g., 0 to .3. The value of ρ is used to estimate the value of $deff$, which is entered into the formula (program) for estimating sample size.

For two-stage sampling, where a sample of m elements is randomly selected from each cluster, the formula for the variance of the sample mean per element is

$$V(\bar{y}) = \frac{1 - f_1}{n} S_1^2 + \frac{1 - f_2}{mn} S_2^2$$

where $f_1 = n/N$ and $f_2 = m/M$ denote the first- and second-stage sampling fractions, and where S_1^2 denotes the variance among primary unit means and S_2^2 denotes the variance among subunits within primary units:

$$S_1^2 = \frac{\sum_{i=1}^N (\bar{Y}_i - \bar{\bar{Y}})^2}{N - 1}$$

and

$$S_2^2 = \frac{\sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{Y}_i)^2}{N(M - 1)}$$

If we assume that N is large, this simplifies to

$$V(\bar{y}) \approx \frac{1}{n} S_1^2 + \frac{1 - f_2}{mn} S_2^2$$

If M is large, this simplifies further to

$$V(\bar{y}) \approx \frac{1}{n} S_1^2 + \frac{1}{mn} S_2^2$$

In terms of the intracluster correlation coefficient, the variance of the sample mean per element is (for N and M large)

$$V(\bar{y}) \approx \frac{1}{nm} S^2 [1 + (m - 1)\rho]$$

(This expression is obtained by using the approximations (for N and M large) $S_1^2 \approx (1 - \rho)S^2$ and $S_2^2 \approx \rho S^2$, where S^2 was defined earlier.) Since the expression outside of the brackets is the formula for the variance of a simple random sample of size nm, the expression in the brackets shows the change in the variance for two-stage sampling (from simple random sampling), and is hence the (approximate) value of the design effect, deff. In the case of single-stage cluster sampling, the deff, $(1 + (M-1)\rho)$, was defined in terms of M, which is known. Here, the deff, $(1 + (m-1)\rho)$, is defined in terms of the element sample size, m. In order to know the value of deff, m must be specified. We will now show how to do this.

As with single-stage cluster sampling, in most applications the variances S_1^2 and S_2^2 are not known, and the value of the intra-unit correlation coefficient is used to determine sample size. For two-stage sampling, however, there are two sample sizes to be determined: the number, n, of first-stage sample units (primary sampling units, PSUs) and the number, m, of elements to select per first-stage unit. We shall consider the case in which a constant number of elements (m) is selected per first-stage unit (this would be appropriate, for example, if the first-stage units were of constant size, or were selected with probabilities proportional to size).

The standard approach to determining the within-unit sample size, m, is to specify a sampling cost function and determine the value of m that minimizes the variance of the estimated mean per element subject to specified cost or that minimizes the cost subject to specified variance. If the cost function is

$$C = c_1 n + c_2 nm$$

then the optimal value of m is

$$m_{opt} = \frac{S_2}{\sqrt{S_1^2 - S_2^2/M}} \sqrt{c_1/c_2}$$

Since the values of the variances S_1^2 and S_2^2 are usually not known, this formula is of limited value. If we use the approximations given earlier for S_1^2 and S_2^2 in terms of ρ (and S^2), this formula can be approximated (for ρ not equal to zero) by:

$$m_{opt} \approx \sqrt{\frac{c_1(1 - \rho)}{c_2\rho}}$$

For many applications, this optimum value is rather "flat," i.e., it is not highly sensitive to small variations in ρ . For many applications, such as sampling households from villages, the number of households is in the range 15-30. Note that the value of m_{opt} is set independently of the value of

the first-stage unit sample size, n . Once m has been determined (as m_{opt}), the value of n is then determined by solving the cost equation or the variance equation, depending on whether the cost or the variance has been specified. (Note that if $m_{opt} > M$ or $S_1^2 < S_2^2/M$, set $m = M$ and use single-stage cluster sampling.)

Now that the value of m is specified, the value of deff, $(1 + (m-1)\rho)$, can be calculated and used in the sample-size formula (program).

(The relationship of ρ to S_1^2 and S_2^2 is complicated because of the finite populations. If both N (the unit population size) and M (the cluster size) are large, the formulas become simple, and it is easier to see what is going on. In this case, a simple model of the situation is that the element response is

$$x = x_1 + x_2$$

where x_1 denotes the mean of the unit to which the element belongs and x_2 denotes a deviation from the mean. Both x_1 and x_2 are random variables, independent of each other. The value of x_1 is independent from unit to unit, and all values of x_2 are independent. The mean of the x_1 is the population mean, μ , and the mean of the x_2 is zero within each unit. Denote the variance of x_1 as σ_1^2 and the variance of x_2 (the same for every unit) as σ_2^2 . Then, by independence of x_1 and x_2 ,

$$\text{var } x = \text{var } x_1 + \text{var } x_2$$

where “var” denotes “variance,” or

$$\sigma^2 = \text{var } x = \sigma_1^2 + \sigma_2^2$$

By the definition of ρ , which is

$$\rho = E(x - \mu)(y - \mu) / (\sigma_1 \sigma_2)$$

where E denotes expectation and x and y are two elements in the same unit, it is easy to show that

$$\rho = \sigma_1^2 / \sigma^2$$

and hence $\sigma_1^2 = \rho\sigma^2$ and $\sigma_2^2 = (1 - \rho)\sigma^2$. The formula for the variance of the sample mean, x_{bar} , is

$$\text{var}(x_{bar}) = \sigma_1^2 / n + \sigma_2^2 / nm.$$

Substituting the values for σ_1^2 and σ_2^2 in terms of ρ , we obtain

$$\text{var}(x_{bar}) = (\sigma^2 / nm) (1 + (m - 1) \rho),$$

which is the approximate expression obtained earlier for two-stage sampling for N and M large.)

Sample-Size Determination for Analytical Surveys

Note that the use of techniques recommended for analytical survey designs – panel sampling and matched comparison groups – will cause the design effect for overall-population estimates to be substantially increased. This is not a great concern, for the objective in analytical surveys is to

estimate differences and relationships, not overall population characteristics (means and totals). The formulas used to estimate sample sizes for descriptive surveys *should not* be used to estimate sample sizes for analytical surveys – they do not explicitly reflect the features that are present in analytical survey designs (panel sampling, matching of comparison groups), nor do they account for the fact that the estimates of interest are differences (or double differences), not overall population characteristics. (Theoretically, it may be possible to use the descriptive-survey formulas for determining sample sizes for analytical surveys, but in practice the designs are so complex that it is not possible to estimate a reasonable value for the *deff*. (A typical analytical survey design for an impact evaluation will include not only stratification, multistage sampling and selection of units with variable probabilities, but also panel sampling and construction of comparison groups by matching of individual units.) What is needed is a sample-size estimation procedure that takes into account the special features of an analytical survey design, such as panel sampling and construction of comparison groups by matching of individual units).

There are two main ways of determining sample size in surveys: (1) to determine the sample size required to provide a specified level of precision (e.g., measured by the size of the standard error of an estimate or the size of a confidence interval) for an estimate of a particular quantity (such as a double-difference estimate); and (2) to determine the sample size required to provide a specified power for a specified test of hypothesis (such as the probability of detecting a double-difference of a specified size). The former method is generally used in determining sample sizes for descriptive surveys, and the latter method is generally used for determine sample sizes for analytical surveys. The latter method of determining sample size is usually referred to as “*statistical power analysis*.” Depending on their structure, they are intended to specify either the number of lowest-level (ultimate) elements, or the number of highest-level sample units (first-stage (primary) sample units). The Brixton program mentioned above determines sample size based on specification of a level of precision; the Statistica program (for example) determines sample size based on specification of power. To use these programs effectively, it is necessary to know something about the structure of the population, such as unit and element variances, intra-cluster correlation coefficients, and spatial and temporal correlations. In addition to depending on population characteristics, the sample-size formulas depend crucially on the nature of the sample design. Computer programs for determining sample sizes vary substantially in how much information is specified about the sample design. In some cases, features of the design are explicitly specified, and in others, the effect of design features is implicitly reflected in the value of the design-effect parameter, *deff*.

The determination of sample size by specifying precision of an estimator involves specification of two factors: the level of precision desired (e.g., as specified by the magnitude of the standard error of an estimate or by the width of a confidence interval of a prescribed confidence coefficient (such as 95 percent)) and the variability of the population. The determination of sample size by specifying the power of a test of hypothesis involves specification of four factors: the magnitude of the effect size to be detected, the significance level of the test (probability, or size, of the Type I error), the level of power (probability of rejecting the null hypothesis, i.e., of detecting the effect), and the variability of the population.

The relationship between the significance level and the power is very important, and it is often overlooked. Here follows a quote from Lehmann’s *Testing Statistical Hypotheses*, 2nd edition (Wiley, 1986, pp. 69-70): The choice of a level of significance α will usually be somewhat arbitrary, since in most situations there is no precise limit to the probability of an error of the first kind that can be tolerated. Standard values, such as .01 or .05, were originally chosen to effect a reduction in the tables needed for carrying out various tests. By habit, and because of the convenience of standardization in providing a common frame of reference, these values gradually became

entrenched as the conventional levels to use. This is unfortunate, since the choice of significance level should also take into consideration the power that the test will achieve against the alternatives of interest. There is little point in carrying out an experiment which has only a small chance of detecting the effect being sought when it exists. Surveys by Cohen (1962) and Freiman et al. (1978) suggest that this is in fact the case for many studies. Ideally, the sample size should then be increased to permit adequate values for both significance level and power. If that is not feasible, one may wish to use higher values of α than the customary ones. The opposite possibility, that one would like to decrease α , arises when the latter is so close to 1 that α can be lowered appreciably without a significant loss of power (cf. Problem 50). Rules for changing α in relation to the attainable power are discussed by Lehmann (1958), Arrow (1960), and Sanathanan (1974), and from a Bayesian point of view by Savage (1962, pp. 64-66). See also Rosenthal and Rubin (1985)."

The reason why there are so many "false alarms" of studies that claim to show the efficacy of some medicine, later to be discredited, is more likely to be the result of the setting of the significance level much too low, e.g., .05, than of a faulty research design. At the .05 level, there is a one-in-twenty chance that the study will show a "significant" effect, simply by chance, when there is none. It would appear that it would be much more cost-effective for socio-economic studies to use much higher levels of α , such as .001. The increased risk that this carries that some positive effects may be unnoticed is not very troubling, since these "missed" effects will be small. On the other hand, program managers do not want to see results that suggest that their programs are not effective, even if the effect is very small. So they will press for large values of α , such as .05, in program evaluation studies. (This will mean that "significant" results will be concluded, even when they are not true, about five percent of the time.) This accrues the additional advantage to the program manager of allowing smaller sample sizes (since if α is higher, the power is also higher for the same sample size; or, if α is higher, the sample size required to achieve a specified level of power is lower).

Descriptive surveys are concerned with estimation, and analytical surveys are concerned with hypothesis testing. Here follows a quotation from *Introduction to the Theory of Statistics* by Mood, Graybill and Boes (McGraw-Hill, 1963, 3rd edition 1974): "The power function will play the same role in hypothesis testing that mean-squared error played in estimation. It will usually be our standard in assessing the goodness of a test or in comparing two competing tests. An ideal power function, of course, is a function that is 0 for those θ corresponding to the null hypothesis and is unity for those θ corresponding to the alternative hypothesis. The idea is that you do not want to reject H_0 if H_0 is true and you do want to reject H_0 when H_0 is false." [The parameter θ specifies a probability distribution; H_0 denotes the null hypothesis. Mean-squared error is variance plus the square of the bias.]

As mentioned, the principal method of determining sample sizes for analytical surveys is statistical power analysis. The emphasis is on power since analytical surveys are involved with tests of hypothesis – descriptive surveys are concerned mainly with precision of estimates, not with the power of tests of hypothesis. Consideration of power is a fundamental aspect of the branch of statistics known as "testing statistical hypotheses." This is a very old branch of statistics. The fundamental theorem of hypothesis testing is the "Neyman-Pearson Lemma," which states necessary and sufficient conditions for a most powerful test of an hypothesis. This theorem was proved in the 1920s. (The major reference book on testing statistical hypotheses is *Testing Statistical Hypotheses* by E. L. Lehmann (Wiley, 1959, 2nd edition 1986). This is a "companion" to Lehmann's book on point estimation, *Theory of Point Estimation* (Wiley, 1983).) There are two parameters that may be specified for a test: the probability of making a Type I error, or rejecting a null hypothesis when it is true (this probability is called the "size" or "significance level" of the test,

and is usually denoted by α); and the probability of making a Type II error, or accepting a null hypothesis when it is false (this probability is usually denoted by β). The power of the test is the probability of rejecting the null hypothesis, or $1 - \beta$. In general, power analysis should address both the values of α and β , but it is customary to do power calculations for “standard” values of α , such as .0005, .01 or .05.

Note that the probability of rejecting a null hypothesis depends on which alternative hypothesis is true. In many applications, the null hypothesis is that a parameter (such as a mean or a double-difference) has a particular value, and the alternative is that the parameter differs from this value by a specified amount, D . In this case, the power of the test may be considered as a function of the value of D . This function is called a “power function,” or “power curve.” The power curve is the probability of rejecting the null hypothesis as a function of D . The one-complement of the power function, or the probability of accepting the null hypothesis as a function of D , is called the “operating characteristic” curve (or “OC” curve).

All basic-statistics books include discussions of the power of statistical tests of hypothesis. Consideration of power is a central focus of the field of statistical quality control (through plots of its complement, the probability of accepting the null hypothesis, via the operating characteristic curve). See, for example, *Quality Control and Industrial Statistics* by Acheson J. Duncan (Irwin, 1952, revised edition 1959, 5th edition 1986). It is a curious fact, however, that consideration of power is largely absent from older books on sample survey. For example, the classic text, *Sampling Techniques* 3rd edition by W. G. Cochran (Wiley, 1977) does not even include the word “power” in the index. Nor does *Elementary Survey Sampling* 2nd edition by Scheaffer, Mendenhall and Ott (Duxbury Press, 1979). Nor do any of the other older major texts on sampling. There is, of course, a reason for this: these texts deal with descriptive surveys, and descriptive surveys are concerned with estimation, not with tests of hypothesis. What is quite amazing, however, is that even recent sampling texts, such as Lohr’s, Thompson’s, and Lehtonen/Pahkinen’s, do not address the topic of statistical power analysis.

In recognition of the fact that books on sample survey did not consider power, Jacob Cohen wrote the book, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, 1969). This book is of little relevance to the field of sample survey, however, since it deals exclusively with simple random sampling.

Recently, funded by a grant from the William T. Grant Foundation, researchers at the University of Michigan conducted a project to develop computer software to conduct statistical power analysis. A report describing their work is *Optimal Design for Longitudinal and Multilevel Research: Documentation for the “Optimal Design” Software*, by Jessaca Spybrook, Stephen W. Raudenbush, Richard Congdon and Andrés Martínez, University of Michigan, July 22, 2009. The report and the software are posted at <http://sitemaker.umich.edu/group-based/home> or http://sitemaker.umich.edu/group-based/optimal_design_software. This software conducts statistical power analysis for a variety of sample designs, for randomized trials. Since randomization is used to assign treatment level to experimental units (the authors assume two treatment levels – treatment and control), the treatment and control groups are “statistically equivalent” (i.e., have the same joint distribution for all variables other than treatment), and a comparison between the treatment and control groups may be made with a single-difference estimate, rather than the double-difference estimate between these two groups at two different points in time. The Optimal Design software produces a wide range of output.

Another program for determining sample size for survey designs in evaluation (i.e., for analytical surveys) is available at the author’s website,

<http://www.foundationwebsite.org/JGCSampleSizeProgram.mdb> (this is a Microsoft Access program). The program determines the sample size of primary sampling units. It calculates sample sizes for a number of cases involving differences in means of population subgroups, including the “double difference” estimator. This program considers three different survey designs – random sampling of primary sampling units for estimation of a population mean; random sampling of two groups, for estimation of a difference in group means (a “single difference”); and random sampling of four groups, for estimation of a double-difference in group means. This last case corresponds to the “pretest-posttest-with-comparison-group” design that is often used (either as an experimental design (randomized comparison group) or quasi-experimental design) in evaluation research. Note that in addition to specifying parameters for the design (such as means, variances and correlations), the user may specify a value for a design effect (deff). The deff is intended to address all design features that are not already addressed by the specified design structure and parameters. For example, in the case of the four-group design, the various correlation coefficients requested would take into account correlations introduced by matching and by panel sampling, and the deff would take into account all other design effects (e.g., caused by stratification or multistage sampling).

It is emphasized that the formula for calculating sample size depends both on the estimate of interest and on the sample design. The most common design in evaluation studies is a pretest-posttest-with-comparison-group design. The referenced sample-size program calculates sample sizes for this design (and simpler designs).

An interesting issue that arises with evaluation designs is the following. Suppose that the objective is to estimate the change in income caused by a certain program, and that the treatment groups and control groups are determined by randomization (i.e., randomized assignment of treatment level (treatment or control) to units). In this case, the treatment and control groups are equivalent with respect to all variables except treatment level. Hence, they are equivalent at the beginning of the evaluation, and the impact of the program intervention may be estimated simply by calculating the difference in income means between the treatment and control groups at the end of the study. The interesting thing, however, is that if pretest and posttest data are available, most analysts would still use the double difference in means (difference, before and after the study, between the difference in means of the treatment and control groups) to estimate program impact. It is important to understand why this is so.

Ordinarily, with independent simple random sampling of four groups, it would be advantageous to use a single difference instead of a double difference because (as will be discussed later) the variance of a double difference is four times as large as a single difference based on the same number of observations. Because of the introduction of correlations between the treatment and control groups (introduced by use of matched pairs) and the before and after groups (introduced by use of panel sampling), this factor of four could be reduced substantially, in many cases to the point where the double-difference estimate may even be more precise than the single-difference estimate. This, however, is not the reason for using the double-difference estimator. There are several reasons for doing so.

First, when matching is done, it is usually done on clusters (aggregates, higher-level sample units), such as villages, districts, census enumeration areas, or other administrative units. The reason for this is that data required for ex ante matching are known (prior to the survey) usually only for aggregates, not for the ultimate sample unit of interest (such as a household). The typical design, then, is a multi-stage design, in which the number of primary sampling units is usually not very large. It could be as large as 100 or more, or it could be as few as five or ten. For small sample sizes, however, the two fundamental theorems so frequently invoked in statistical analysis – the

law of large numbers (which says that for large samples the sample mean is close to the population mean) and the central limit theorem (which says that for large samples the sample mean is approximately normally distributed) – do not apply. For a small sample of PSUs, it is quite possible for the treatment group and the control group to be rather different. For example, the means with respect to one or more variables could be somewhat different. In other words, the particular samples selected for the treatment and control groups might not be highly “orthogonal” (conditionally independent) with respect to some variables (observed or unobserved). (The theory of propensity-score matching, for example, relates to asymptotic (large-sample) properties of the matching *process* – it does not apply well for small sample sizes.) In this case, although the sample estimates are unbiased in repeated sampling, the particular sample may provide poor results, simply because it is small (the “luck of the draw”), not because the randomization process was flawed. A way to improve the precision of the estimator is to use a double difference estimator instead of a single difference estimator. The reason the double difference estimator performs better is because it is based on matched pairs with matching not only between both treatment and controls, but also between pretest and posttest units (via panel sampling). The double difference estimator is a more complex “model-based” estimate than the single difference estimator, and less sensitive to vagaries of sample selection. (In fact, for a correctly specified model, we do not need a probability sample at all – just a sample with good spread, balance and orthogonality on all variables of interest.)

The second reason why a double difference estimator would be used even for a design involving randomized selection of treatment and control groups is nonresponse. If some PSUs are “lost” from the sample (e.g., refusal to participate in the survey, lack of access due to rain or civil disturbances), it is usually desirable to substitute replacements for them, to maintain the sample size (in a matched-pairs design, both units of a matched pair should be replaced, if practical, when either one of the pair is replaced). While this may keep the precision level high, it may also introduce selection bias (if the outcome is in some way different for the nonresponders). In this case, we do not have the pure experimental design that we had planned. The double-difference model is less sensitive to selection bias than the single-difference model, for the same reason as discussed above.

In summary, for a pretest-posttest-with-randomized-control-group design, a double-difference estimator is used, even though a single-difference estimator is correct. The single-difference estimator would be used only if baseline (pretest, time 1) data were not available. (A double difference estimate of impact would *always* be used for a quasi-experimental pretest-posttest-with-comparison-group design, since there is no guarantee that the treatment and control groups are equivalent – they can be matched only on observables.)

As is clear from the preceding discussion, the determination of sample size for analytical surveys proceeds quite differently for analytical surveys than for descriptive surveys. Sample size determination for descriptive surveys generally focuses on specification of the level of precision for an estimator (such as a population mean or total), whereas for analytical surveys it focuses on specification of the power for tests of hypotheses (such as whether two populations could be considered to have the same probability distribution (or mean)).

There is another very significant difference. In order to calculate sample size for analytical surveys, it is essential to know something about the correlation between observations in different groups involved in estimates of interest. For example, the double-difference estimate of program impact for a pretest / posttest / comparison-group design involves a double difference (linear contrast) of four groups – the treatment and comparison groups before the program intervention (pretest) and these two groups after the program intervention. In a descriptive survey, the

observations of four different population subgroups (e.g., four different strata) would typically be uncorrelated (i.e., be selected independently of each other) – this would typically maximize the precision of estimates of means and totals. In an analytical survey, the observations involved in an estimate of interest (such as the double difference) would almost certainly be correlated, by intention (i.e., by design). The reason for introducing correlations among the four groups is to increase precision (via a “matched pairs” sample that includes matching of individual treatment units with individual control units, and individual time-1 units with individual time-2 units). In the example just mentioned, the survey designer would prefer to use a panel survey involving reinterview of the same sample units before and after the program intervention. This would dramatically increase the precision of the double-difference estimate, over the case in which the units of the second round of the survey were independently selected. Also, the survey designer would prefer to promote local control between the treatment and comparison groups by matching individual units (or “blocking”), rather than simply matching the probability distributions (i.e., by using unrelated (independent) samples). This would have a large impact on the precision of the estimate, depending on how effective the matching was in reducing the variation between paired treatment and comparison units.

To take the effect of panel sampling and matching of treatment and comparison units into account, it is necessary to specify the correlation between panel units and the correlation between treatment and comparison units. A simple example will illustrate the concept involved. Suppose, for example, that we wish to use a sample to estimate the overall population mean, and also to estimate the difference between two population subgroups. If the samples for the two subgroups are selected independently (e.g., strata in a descriptive survey), then the formulas for the variance of the estimated mean and difference are

$$\text{Var}(\text{mean}) = (1/4) (v_1/n_1 + v_2/n_2)$$

$$\text{Var}(\text{difference}) = v_1/n_1 + v_2/n_2$$

where v_1 and v_2 denote the within-stratum variances of units and n_1 and n_2 denote the stratum sample sizes. If $v_1 = v_2 = v$ (i.e., the strata are no more internally homogeneous than the general population, so that stratification is of no help in reducing the variance) and $n_1 = n_2 = n/2$ (a “balanced” design), then these estimates become

$$\text{Var}(\text{mean}) = v/n$$

$$\text{Var}(\text{difference}) = 4v/n,$$

which are the usual formulas for the variance of the mean and difference in the case of simple random sampling. These formulas illustrate that for descriptive surveys (independent sampling in different strata) *four times* the sample size is required to produce the same level of precision (as measured by the variance (or its square root, the standard deviation) for an estimated difference as for an estimated mean – a sample of size $2n$ is required for each of the two groups (strata) comprising the difference, instead of a single sample of size n overall.

(This may seem a little counterintuitive. If a sample size of n is required to produce a certain level of precision for an estimated population mean, then that same sample size is required to produce that level of precision for a subpopulation mean (assuming that the variance of the subpopulation is the same as the variance of the general population). Hence the sample size required to produce comparable-precision estimates of two subpopulation means, each equal in size to half the population, would be $2n$ (i.e., n for each subpopulation). The variance of the *difference* in

these two means, however, is twice the variance of each individual mean. Hence, to obtain the same level of precision for the estimated *difference*, the sample size must be doubled again, to $2n$ for each group.)

If steps are taken to pair similar items, such as in panel sampling (reinterview of the same unit at a later time) or matching of individual treatment and comparison units, then the sample items represent “paired comparisons,” and the variances of the estimated mean and difference change. If ρ denotes the correlation coefficient between the units of matched pairs, then

$$\text{Var}(\text{mean}) = (1/4) \{v_1/n_1 + v_2/n_2 + 2 \rho \text{sqrt}[(v_1/n_1)(v_2/n_2)]\}$$

$$\text{Var}(\text{difference}) = v_1/n_1 + v_2/n_2 - 2 \rho \text{sqrt}[(v_1/n_1)(v_2/n_2)].$$

What we see is that the presence of correlation between paired units increases the variance of the mean and decreases the variance of the difference. In analytical surveys, we are generally interested in estimating differences (or regression coefficients, which are similar), and so the presence of correlations between units in different groups involved in the comparison can reduce the variance of the estimate very much. If $v_1 = v_2 = v$ and $n_1 = n_2 = n/2$, these formulas become

$$\text{Var}(\text{mean}) = (1 + \rho) v/n$$

$$\text{Var}(\text{difference}) = (1 - \rho) 4v/n.$$

For example, if $\rho = .6$, then the variance of the mean is $1.6 v/n$ and the variance of the difference is also $1.6 v/n$. Because of the introduced correlation, the standard deviation of the difference has been reduced by a factor of $\text{sqrt}(1.6/4) = .63$ and the standard deviation of the mean has been increased by the factor $\text{sqrt}(1.6/1) = 1.26$.

In designing an analytical survey, it is differences rather than means that are of primary interest, and the survey designer will construct the design so that the correlations between units in groups to be compared are high. For estimating a double difference, the two correlations of primary interest are the correlation between corresponding units of the panel survey and the correlation between matched treatment and comparison units. (The presence of these correlations will introduce correlations among other groups, e.g., between the pretest treatment group and the posttest comparison group. These correlations affect the variance of the estimates, but they are not of direct interest, and are not controlled. This will be discussed in greater detail later.)

Keep in mind that while the introduction of correlations (via matching and panel sampling) will improve the precision of double-difference estimates, it will reduce the precision of estimates of population means and totals. In many survey applications, it is desired that the survey be capable of producing *both* kinds of estimates, and so the design will be a compromise between (or combination of) a descriptive survey and an analytical survey (or between the design-based approach and the model-dependent approach, i.e., it will be a model-based approach). A practical example is a sample survey designed to produce data for both program monitoring and evaluation (“M&E”). For program monitoring, primary interest focuses on estimation of means and totals for the population and for various population subgroups. For impact evaluation, attention focuses on estimation of differences (comparisons, linear contrasts, regression coefficients). From the viewpoint of efficiency, it may be desirable to address both concerns simultaneously, i.e., not to design a monitoring system based on one survey design and then have to develop a separate design for impact evaluation. In a monitoring and evaluation application, the monitoring system is an example of a descriptive survey, and the impact evaluation is an example of an analytical

survey, but both surveys involve the same population over the same time period, and so it is desirable from an efficiency viewpoint to address both objectives simultaneously.

The sample-size program posted at the author's website addresses the problem of determining sample size to estimate the double-difference of a pretest / posttest / comparison-group design, but it does not address more complex designs, such as an application involving multiple comparison groups. Such cases would involve forming matched "groups" rather than matched pairs, to improve the precision of comparisons among several groups, not just two. The matching methodology presented in Appendix A addresses the issue of constructing matched groups, as well as matched pairs.

Note that the "deff" referred to in the sample-size program is the design effect from design aspects *other than* those reflected in the correlations between the treatment and control groups, and between the pretest and posttest groups. The deff to be specified should reflect the design effect from other design features, such as stratification, clustering, multistage sampling and selection with variable probabilities, not the specified correlations.

Additional remarks on sample size determination and sample design for analytical surveys

This section will close with some additional discussion of considerations in determination of sample size, in the case in which it is desired to construct a model that emphasizes comparison of treatment and nontreatment groups, but also includes a number of other explanatory variables. This kind of model is appropriate, for example, when it is not possible to use a true experimental design with randomized selection of the controls (nontreatment units), and a quasi-experimental design is being used as an alternative. Because of the lack of randomization, data are also collected on a number of other variables (in addition to the treatment variable) that may have had an influence on selection for treatment or may have an effect on outcomes of interest. (The discussion in this section is a little technical, and may be skipped with little loss in the conceptual issues discussed in this article.) (It is noted that data may be collected on other (nontreatment) variables even if randomization is employed, to estimate the relationship of treatment effects on these variables.)

We assume that the model is a general linear statistical model, such as a multiple regression model or an "analysis of covariance" model. Let us consider the case in which there is a single explanatory variable in the model, viz., the treatment variable. Let us assume that the design is "balanced," so that half the observations are treatment units and half are nontreatment units. Let us denote the sample size as n .

It was shown earlier that if simple random sampling is used and sampling is done independently in different groups, the number of observations required to produce a given level of precision for estimation of a difference in means is four times that required to produce the same level of precision for estimation of the overall mean. Let us assume that we have a sample of n independent observations sampled from a distribution having mean μ and variance σ^2 . Let us denote the dependent variable of interest as y . Then the variance of the sample mean, \bar{y} , is

$$\text{var}(\bar{y}) = \sigma^2/n$$

and the variance of the estimated difference, $\bar{y}_1 - \bar{y}_2$, is (because of independence)

$$\text{var}(\bar{y}_1 - \bar{y}_2) = \text{var}(\bar{y}_1) + \text{var}(\bar{y}_2) = \sigma^2/(n/2) + \sigma^2/(n/2) = 4 \sigma^2/n .$$

Similarly, the number of observations required to produce a given level of precision for estimation of a double difference in means is *sixteen times* that required to produce the same level of precision for estimation of the overall mean, in the case of simple random sampling and independent groups:

$$\begin{aligned} \text{var}(\bar{y}_1 - \bar{y}_2 + \bar{y}_3 - \bar{y}_4) &= \text{var}(\bar{y}_1) + \text{var}(\bar{y}_2) + \text{var}(\bar{y}_3) + \text{var}(\bar{y}_4) \\ &= \sigma^2/(n/4) + \sigma^2/(n/4) + \sigma^2/(n/4) + \sigma^2/(n/4) = 16 \sigma^2/n . \end{aligned}$$

(It should be noted that a double difference is not a linear contrast. A linear contrast is a linear combination of the observations such that the sum of the coefficients is zero and the sum of the positive coefficients is one. For all linear contrasts having coefficients of equal magnitude (for each observation), the variance of the linear contrast is equal to $4 \sigma^2/n$. For a double difference, the coefficient of each observation is plus or minus $1/(n/4)$, and the sum of the positive coefficients is $(n/2)(1/(n/4)) = 2$, not one. A double difference is hence twice the value of such a linear contrast, and so its variance is four times as large, or $16 \sigma^2/n$ instead of $4 \sigma^2/n$.)

Hence we see that if we are comparing independent subgroups, the sample sizes required to estimate differences and double differences become very large. The way that the sample size is reduced to reasonable levels is by avoiding the use of independent groups, by introducing correlations between members in different groups. For estimation of a double difference from a pretest / posttest / control-group design this may be done by reinterviewing the same unit in the posttest (second round of a panel survey) and matching individual control units with treatment units. To keep the example simple, let us see what effect this has in the case of estimating a single difference (the more complicated case of a double difference is considered in the sample size estimation program mentioned earlier, and in an example presented later in this article).

In this case, the formula for the variance of the estimated difference is

$$\begin{aligned} \text{var}(\bar{y}_1 - \bar{y}_2) &= \text{var}(\bar{y}_1) + \text{var}(\bar{y}_2) - 2 \text{cov}(\bar{y}_1, \bar{y}_2) = \text{var}(\bar{y}_1) + \text{var}(\bar{y}_2) - 2 \rho \text{sqrt}(\text{var}(\bar{y}_1) \text{var}(\bar{y}_2)) \\ &= \sigma^2/(n/2) + \sigma^2/(n/2) - 2 \rho \text{sqrt}(\sigma^2/(n/2) + \sigma^2/(n/2)) = 4 (1-\rho) \sigma^2/n , \end{aligned}$$

where cov denotes the covariance of \bar{y}_1 and \bar{y}_2 and ρ denotes the correlation of \bar{y}_1 and \bar{y}_2 . That is, the variance is reduced by the factor $(1-\rho)$. By doing things such as reinterviewing the same sample unit in the posttest and matching of individual comparison units with treatment units, the correlation, ρ , may be quite high, such as .5 or even .9. That is, the variance of the estimated difference may be substantially reduced.

In a laboratory or industrial experiment, there may be many treatment variables of interest, and it is important to use a design in which they are orthogonal, such as a factorial or fractional factorial design, so that the estimates of the treatment effects will be uncorrelated (unconfounded) and readily interpreted. In many socioeconomic experiments, there is but a single treatment effect, viz., the effect of program intervention (e.g., increased earnings or employment from participation in a farmer training program, or decreased travel time or cost from a roads improvement program). In a laboratory experiment, there are usually several treatment variables and no covariates, whereas in an evaluation of socioeconomic programs there is often a single treatment variable and lots of covariates. Whichever is the case, it should be realized that differences in many variables may be represented in and estimated from the same data set. The greater the degree of orthogonality among the explanatory variables (i.e., the lower the correlation), the lower the correlation among the estimated effects, and the easier it is to interpret the results.

It is noted that in a well designed experiment it is possible to estimate the effects of a number of treatment variables simultaneously. All that is required is that the number of observations be substantially larger than the total number of effects (linear contrasts) to be estimated (main effects, first-order interactions, second-order interactions, etc.). In many social or economic evaluations, there is only a single treatment variable, viz., participation in the program, but if there are numerous program variations, evaluation of all of them can often be done simultaneously, in a single sample (experiment), without the need for separate samples for each treatment variable. If a descriptive-survey approach were adopted in this situation, a separate sample (or stratum) might be used for each treatment combination of interest, and a large sample size would be required. This "one-variable-at-a-time" approach would be an inefficient approach, if it is possible to construct a sample design that addresses multiple treatment combinations at the same time.

Estimation of a regression coefficient in a multiple regression model is analogous to estimation of a difference in group means in an analysis of variance (experimental design) model. This is easy to illustrate in the case of an experiment in which there is a single explanatory variable and there are just two values of the variable (e.g., treatment and control). We showed above the formula for the variance of the difference in means between the two groups. For the regression equation, the model is

$$y_i = \alpha + \beta x_i + e_i$$

where α denotes the intercept, β denotes the slope, and e is an error term (independent of each other and the x 's, with mean zero and common variance σ^2). The variances and covariance of the estimated parameters (a and b) are:

$$\text{var}(a) = \sigma^2 \sum x_i^2 / (n \sum (x_i - \bar{x})^2)$$

$$\text{var}(b) = \sigma^2 / \sum (x_i - \bar{x})^2$$

$$\text{cov}(a,b) = -\sigma^2 \bar{x} / \sum (x_i - \bar{x})^2.$$

where \sum denotes summation.

Let us define the values of the x 's so that their mean is zero and their range is one (so that the slope coefficient denotes the mean change in y per unit change in x), that is, $x_i = 1/2$ for treatment units and $x_i = -1/2$ for nontreatment units (controls). In this case, the three preceding quantities become

$$\text{var}(a) = \sigma^2 / n$$

$$\text{var}(b) = 4 \sigma^2 / n$$

$$\text{cov}(a,b) = 0.$$

In this simple regression model, the intercept is simply the overall mean, and the slope is the average difference between the treatment and control groups, and we see that the variances of these two estimates are exactly what we obtained earlier.

In a laboratory experiment, we could have many other treatment variables orthogonal to (uncorrelated with) the first treatment group, and the results would be the same for all of them. As

mentioned, in socioeconomic experiments (such as program evaluations), there is usually a single treatment variable, but there may be many additional covariates. The covariates will in general not be orthogonal to the treatment variable, even if we apply the survey design algorithm of Appendix A to reduce the correlation among the explanatory variables.

Let us now examine the variance of certain estimates of interest. The estimate corresponding to a specified value of x is obtained simply by substituting the value of x in the estimated regression equation, $y = a + bx$. Its variance is given by

$$\text{var}(y | x) = \text{var}(a + b x) = \text{var}(a) + x^2 \text{var}(b) - 2 \text{cov}(a, bx) = \text{var}(a) + x^2 \text{var}(b)$$

since the covariance is zero. (See Alexander Mood, Franklin A. Graybill and Duane C. Boes *Introduction to the Theory of Statistics* (3rd edition, McGraw-Hill, 1974) for the formulas for the variances and covariances of the regression-model estimates.) To estimate the overall mean, substitute $x = 0$, obtaining $y = a$ and

$$\text{var}(y | x=0) = \text{var}(a) = \sigma^2/n.$$

To estimate the value for the treatment, substitute $x = 1/2$, obtaining $y = a + 1/2 b$ and

$$\text{var}(y | x = 1/2) = \text{var}(a) + 1/4 \text{var}(b) = \sigma^2/n + 1/4 4 \sigma^2/n = \sigma^2/(n/2).$$

Note that this is exactly the same as the variance of the estimate of the mean of a sample of size $n/2$, which is what the set of treatment observations is. In other words, as had to be the case, the use of a regression model to estimate the treatment effect produces exactly the same result as estimating the effect from an analysis-of-variance model, i.e., as the mean difference between the treatment and nontreatment units (difference in means of the treatment and nontreatment groups).

This simple example illustrates well the fact that the variance of an estimator from a regression model depends very much on the value of the explanatory variable, and that estimates for extreme values of x , at the limit of the observation set, such as for the treatment group (treatment value $x = 1/2$), will be much lower than the variance for observations near the middle of the observation set, such as the overall mean (treatment value 0). The regression model does not “add” any information to the estimation process, over that reflected in the simple difference-in-means model.

If we introduce correlations into the design, e.g., by matching of individual units in the treatment and nontreatment groups, the formulas for the variances change. The variance of a will increase by the factor $(1 + \rho)$ and the variance of b will decrease by the factor $(1 - \rho)$, where ρ denotes the correlation between a unit in the treatment group and its corresponding (matched unit) in the comparison group. Note that the estimate of the mean for the treatment group (or the nontreatment group) will still be the same (since the $+p$ will cancel the $-p$ in the formula). That is, the introduction of correlations between the treatment and nontreatment groups (whether we match the group or match individual items) does not reduce the precision of the estimated group mean.

For a treatment variable having only two values ($1/2$ and $-1/2$), both values are extreme, and the variance of the group means will be relatively large (compared to that for a central value of x). For variables that have many values over a range, this may not be the case. For example, suppose that we wish to estimate a model for four different regions. There are two approaches to doing this. On the one hand, we may select independent samples and estimate separate models for each one. On the other hand, we may posit an overall model for the entire population and reflect

differences among the regions by interaction terms in this overall model. This may be a much more efficient use of the data, since data from all four groups is being used to estimate the various model parameters (in a single, combined model). It should be recognized, however, that the process of matching across groups will generally increase the precision of estimates of model parameters and differences at the expense of decreasing the precision of the estimate of the overall mean. In general, the precision of the estimates of the group means will be unaffected by matching across groups. (The precision of estimates of group means will be adversely affected, of course, by any matching done within the groups, such as between units in a treatment and control group).

The preceding examples have illustrated some of the considerations that should be taken into account in determining sample size and sample design for analytical surveys. There are many factors to be taken into account, and the process is not simple or easy. These examples illustrate that the variances of estimates can be influenced substantially by the survey design. It is important to focus on what the estimation goals for the survey are, and to take into account all available information to achieve high levels of precision and power for the sampling effort expended.

Some Examples

This section will present a number of examples to illustrate the use of the sample-size determination program referred to earlier. The examples include cases for descriptive surveys as well as analytical surveys. In the examples that follow we shall assume that the sample sizes are sufficiently large that the estimates are approximately normally distributed.

Example 1. Determine sample size by specifying a confidence interval for a population mean.

Suppose that a survey is conducted to estimate the mean or a total for a population. This is an example of a descriptive survey. Suppose first that the survey design is a simple random sample. In this example we shall assume that the population size is very large (the program does not make this assumption – if the population size is not large, the formulas are a little more complicated). The standard approach to sample-size estimation is to specify an “error bound,” E , for the estimated mean, which is half the width of a 95-percent confidence interval (an interval that includes the true value of the mean ninety-five percent of the time, in repeated sampling), and to determine the sample size, n , that will produce this error bound. The formula for the error bound is $z_{1-\alpha/2}$ times the standard deviation (standard error) of the estimated mean, or σ / \sqrt{n} , where n denotes the sample size, σ denotes the standard deviation of the population units, $1 - \alpha$ denotes the confidence coefficient (.95 in this case), $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution, and n denotes the sample size. For $1 - \alpha = .95$, $z_{1-\alpha/2} = z_{.975} = 1.96$, and we have:

$$E = z_{1-\alpha/2} \sigma / \sqrt{n} = 1.96 \sigma / \sqrt{n}.$$

Solving for n , we obtain

$$n = (1.96 \sigma / E)^2 .$$

For example, if $\sigma = 100$ and $E = 10$, then $n = 384$.

Note that in order to determine the sample size, it is necessary to specify the value of the standard deviation, σ . This may be known approximately from previous surveys. If it is not known, the usual approach is to determine the sample size required to provide a specified level of precision

for a proportion. This is “easier to do” since the standard deviation (standard error) of an estimated proportion depends on the true value of the proportion. If p denotes the value of the proportion, then the standard deviation of the underlying 0-1 (binomial) random variable is $\sqrt{p(1-p)}$, so that the formula for the sample size becomes

$$n = [(1.96 \sqrt{p(1-p)})/E]^2 .$$

This expression has its maximum value for $p = .5$:

$$n = (.98 / E)^2 .$$

For example, if $E = .03$, then $n = 1067$. This is about the usual sample size for television opinion polls, which are usually reported to “have a sampling error of three percentage points.”

If the survey design is different from a simple random sample, a “design effect” factor, “deff,” is introduced into the formula. The design effect indicates by how much the variance of an estimate is increased for a particular sample design, over the variance for simple random sampling. In this case the formula for sample size is:

$$n = \text{deff} (1.96 \sigma / E)^2 .$$

If the sample is split randomly into two groups and the difference in means of the two groups is calculated, its variance (as discussed earlier) is four times that of the variance of the overall mean. This means that the sample size required to produce a specified level of precision (indicated by E) for an estimated difference is *four times* the sample size required to produce that level of precision for the estimate of the overall mean. Similarly, if the sample is randomly split into four groups and the double difference of means is calculated, its variance is *sixteen times* that of the variance of the overall mean. It is clear that if estimation of differences and double differences is based on independent samples, the sample sizes required to produce a specified level of precision for estimates of differences and double differences are much larger than those required to produce the same level of precision for the overall mean. Because of this, it is not reasonable to use a sample of independent observations as a basis for estimating means and double differences. Instead, as was discussed earlier (and will be illustrated in the examples that follow), samples intended for estimation of differences and double differences should not be comprised of independent observations, but should introduce correlations in such a way that the variances of the estimates of interest are reduced.

Example 2. Determine sample size by specifying the power of a test of hypothesis about the value of a population mean.

Suppose that it is desired to test whether the mean of a population is larger than a specified value. This example could refer either to a descriptive survey or an analytical survey. In the former case, we wish to test whether the mean of the finite population is larger than the specified value, and in the latter case we wish to test whether the mean of an infinite population that is considered to have generated this population is larger than the specified value. In this example, and in the program, it is assumed that the population size is very large. (For the analytical survey, this assumption would always hold.)

The power of a test about a distribution parameter is the probability of rejecting the hypothesis, as a function of the parameter, in this case the mean. Let α denote the probability of making a Type I error (rejecting the null hypothesis when it is true) and β denote the probability of making a Type II

error (accepting the null hypothesis when it is false). The power is $1 - \beta$. Let m denote the value against we wish to test (i.e., the “specified value” that the mean exceeds), and let D denote a positive number.

The power of the test, as a function of D , the amount by which the true population mean exceeds the specified value, m , is given by

$$\text{Prob}([\text{samplemean} - m] / \text{sqrt}(\text{deff} [\sigma^2 / n]) > z_\alpha \mid \text{popmean} = m + D) = 1 - \beta,$$

where “samplemean” denotes the sample mean and “popmean” denotes the population mean.

Solving for n , we obtain the following formula for the sample size (the value of m is irrelevant):

$$n = [\text{deff} (z_\alpha + z_\beta)^2 (\sigma^2)] / D^2 .$$

The preceding formula gives the same results as the determination of sample size based on specification of the size of a confidence interval, if (1) N (in the confidence-interval approach) is very large; (2) α for this (one-sided) approach is set equal to $\alpha/2$ for that (two-sided) approach (e.g., .025 here, .05 there); (3) β is set equal to .5 (i.e., $z_\beta = 0$); and D is set equal to E . In typical situations, in which it is desired to detect a small difference (D), this approach may yield sample sizes several times as large as the confidence-interval approach. For detecting large differences, this approach generally produces smaller sample sizes. (Example 1: Using the confidence-interval approach with confidence coefficient = .95 ($\alpha = .05$, $z_{1-\alpha/2} = 1.96$), $\sigma = .5$, $\text{deff} = 1$, $E = .05$ and $N = 1,000,000$ yields $n = 384$. Using the power approach with $\alpha = .025$ ($z_{1-\alpha} = 1.9600$), $\beta = .1$ ($z_{1-\beta} = 1.2816$), $\sigma = .5$, $\text{deff} = 1$, and $D = .05$ yields $n = 4,204$ (i.e., 11 times as large). Ex. 2: Same as Ex. 1, but $\alpha = .1$ ($z_{1-\alpha} = 1.286$) yields $n = 2,628$ (6.84 times as large).)

Example 3. Determine sample size by specifying the power of a test for the difference in population means.

Suppose that it is of interest to compare the mean incomes of two different groups, such as males and females, or workers belonging to two different ethnic groups, or a population of workers at two different times. This is an example of an analytical survey. The sample size will be determined by calculating the sample size required to produce a specified power for a test of the hypothesis that the mean for group 1 is greater than the mean for group 2 (i.e., a “one-sided” test).

Were this a simple descriptive survey, we would simply specify the level of precision desired for the estimate of the mean for each of the two groups, and determine the sample size for each group, independently, using the formula given in Example 1. Since it is desired to conduct a test of the hypothesis about the difference in group means, however, that approach is not appropriate. Instead, the sample size should be set at a level that assures a specified level of power for the desired test.

The formula from which the sample size is derived is

$$\text{Prob}([\text{samplemean1} - \text{samplemean2}] / \text{sqrt}(\text{deff} [\sigma_1^2 / n_1 + \sigma_2^2 / n_2 - 2 \rho \times \sigma_1 / \text{sqrt}(n_1) \times \sigma_2 / \text{sqrt}(n_2)])) > z_\alpha \mid \text{popmean1} - \text{popmean2} = D) = 1 - \beta,$$

where D denotes the true difference of the means of the two groups.

The user specifies the ratio n_2/n_1 (e.g., for $n_1 = n_2$, set the ratio = 1.0). If the two groups have the same variability and sampling costs then it is most efficient to have equal sample sizes for the two groups.

The formula for the sample size of the first group is

$$n_1 = [\text{deff} (z_\alpha + z_\beta)^2 (\sigma_1^2 + \sigma_2^2 / \text{ratio} - 2 \rho \sigma_1 \sigma_2 / \text{sqrt}(\text{ratio}))] / D^2 .$$

It is clear from this formula that the value of n is highly dependent on the value of ρ , the correlation between units in the two groups. In a descriptive survey, the two group samples would typically be selected independently. If it were desired to estimate the overall mean for both groups, they would surely be. But since the objective here is to test for a difference in means, it is advantageous to have a high degree of correlation between members of the two groups. This could be done, for example, by matching each member of group 1 with a similar member of group 1, based on available characteristics (other than group membership). If the comparison is between a population at two different points in time, the second-round sample could use the same members as were selected for the first round. In this case, the value of ρ could be quite high, and the sample size for comparing means would be substantially smaller than it would be for independently selected groups.

Example 4. Determine sample size by specifying the power of a test for a double difference in population means.

In rigorous impact evaluation studies, a preferred research design is the pretest / posttest / randomized-control-group design. If randomized allocation to the treatment group is not feasible, a reasonable second-choice is the pretest / posttest / matched-comparison-group design. In this case, the formula from which the sample size is determined is:

$$\text{Prob}([\text{samplemean1} - \text{samplemean2} - \text{samplemean3} + \text{samplemean4}] / \text{sqrt}(\text{deff}[\sigma_1^2 / n_1 + \sigma_2^2 / n_2 + \sigma_3^2 / n_3 + \sigma_4^2 / n_4 - 2 \rho_{12} \sigma_1 \sigma_2 / \text{sqrt}(n_1 n_2) - 2 \rho_{13} \sigma_1 \sigma_3 / \text{sqrt}(n_1 n_3) + 2 \rho_{14} \sigma_1 \sigma_4 / \text{sqrt}(n_1 n_4) + 2 \rho_{23} \sigma_2 \sigma_3 / \text{sqrt}(n_2 n_3) - 2 \rho_{24} \sigma_2 \sigma_4 / \text{sqrt}(n_2 n_4) - 2 \rho_{34} \sigma_3 \sigma_4 / \text{sqrt}(n_3 n_4)]) > z_\alpha | \text{popmean1} - \text{popmean2} - \text{popmean3} + \text{popmean4} = D) = 1 - \beta.$$

The user specifies the ratios n_2/n_1 , n_3/n_1 , and n_4/n_1 (referred to as ratio2, ratio3 and ratio4 in the formula below; set ratios equal to 1.0 for equal-sized samples).

The formula for the sample size of the first group is

$$n_1 = [\text{deff} (z_\alpha + z_\beta)^2 \times (\sigma_1^2 + \sigma_2^2 / \text{ratio1} + \sigma_3^2 / \text{ratio3} + \sigma_4^2 / \text{ratio4} - 2 \rho_{12} \sigma_1 \sigma_2 / \text{sqrt}(\text{ratio2}) - 2 \rho_{13} \sigma_1 \sigma_3 / \text{sqrt}(\text{ratio3}) + 2 \rho_{14} \sigma_1 \sigma_4 / \text{sqrt}(\text{ratio4}) + 2 \rho_{23} \sigma_2 \sigma_3 / \text{sqrt}(\text{ratio2 ratio3}) - 2 \rho_{24} \sigma_2 \sigma_4 / \text{sqrt}(\text{ratio2 ratio4}) - 2 \rho_{34} \sigma_3 \sigma_4 / \text{sqrt}(\text{ratio3 ratio4}))] / D^2.$$

Once again, it is clear that if positive correlations can be introduced between the members of different groups, the sample size may be reduced substantially. Let us assume that groups 1 and 3 are the “time-1” groups and 2 and 4 are the “time-2” groups, and that groups 1 and 3 are the “treatment” groups and groups 2 and 4 are the “comparison” or “control” groups. The most important correlations in this formula are ρ_{12} and ρ_{34} (the correlations between the time 1 and time 2 groups); and ρ_{13} (the correlation between the treatment and comparison groups at time 1). These are the correlations over which the survey designer has control. The correlations ρ_{14} , ρ_{23} and ρ_{24} are “artifactual” – they follow from the other ones, and will typically be smaller. It is expected that ρ_{12} and ρ_{34} (associated with interviewing the same units in both waves of a panel

survey) could be quite high, but ρ_{13} (associated with matching treatment and comparison units) would not be so high.

In order to use the preceding formula, it is necessary to set reasonable values for the variances, the correlations, and for deff . It is reasonable to expect that some data may be available that would suggest reasonable values for the variances. Similarly, it may be possible to estimate the value of deff (the change in the variance from design features such as stratification, clustering, multistage sampling, and selection with variable probabilities) from previous similar surveys. It is less likely that useful data will be available to assist the setting of reasonable values for the correlations. If a panel survey is done using the same households for the second round, then the values of ρ_{12} and ρ_{34} will typically be high (unless there is a lot of migration). How large the value of ρ_{13} is will depend on the survey designer's subjective opinion about how effective the ex-ante matching is.

8. Estimation Procedures

The sample design for an evaluation research study is likely to be a complex survey design. In most cases, it is unlikely that closed-form analytical formulas will be available to determine the standard errors of the sample estimates. For this reason, it will be necessary to employ simulation methods to calculate the standard errors (which are required to construct confidence intervals for parameters and make tests of hypotheses), even if closed-form formulas are available for the estimate itself. These simulation methods are referred to as "Monte Carlo" methods or "resampling" methods, and include techniques such as the "jackknife," the "bootstrap" and "balanced repeated replication. Texts on these methods include *The Jackknife and Bootstrap* by Jun Shao and Dongsheng Tu (Springer, 1995); *Introduction to Variance Estimation* by Kirk M. Wolter (Springer, 1985); and *An Introduction to the Bootstrap* by B. Efron and R. J. Tibshirani (Chapman and Hall, 1993). Resampling methods are included in popular statistical computer program packages, such as Stata.

This paper is concerned with design, not with analysis. For a review of literature related to analysis (and design as well), see the Imbens / Wooldridge paper referred to previously. For discussion of estimation for evaluation models ("causal models"), see Paul W. Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association*, vol. 81, no. 396, December 1986. The basic model used for evaluation is the general linear statistical model that has been used in experimental design since the early twentieth century, but when applied to evaluation it is usually referred to as the "Rubin causal model" or the "potential outcomes" model. For general information on statistical analysis of data from analytical surveys, refer to reference books on the general linear statistical model and econometrics, and to the more recent sampling texts cited earlier (e.g., Sharon L. Lohr, *Sampling: Design and Analysis* (Duxbury Press, 1999)). Some discussion of statistical program packages (e.g., Stata, SPSS, SAS, SUDAAN, PC Carp, WesVarPC) is presented in Lohr's book (page 364).

Appendix A. A Procedure for Designing Analytical Surveys

The general objective in designing an analytical survey is to have substantial variation in explanatory variables of interest and zero or low correlation among those that are inherently unrelated (i.e., not causally related). In addition, design features should ensure that estimates of

interest have high precision for the survey effort expended. The principles of designing analytical surveys are essentially those of experimental design, which are discussed in Cochran and Cox's *Experimental Designs*. The major principles of experimental design are randomization, replication, local control, symmetry and balance. "Randomization" involves random selection of items from the population of interest and randomized assignment of treatment to experimental units.

"Replication" and "local control" refer to having a sufficiently large sample size, and to the inclusion of similar units in the sample (by repeated sampling from similar stratum cells, or matching, or "blocking"). "Symmetry" refers to design characteristics such as drawing treatment and control units from similar populations, and having a high degree of orthogonality (low correlation) among explanatory variables (so that the estimates are not "confounded" (intermingled)). "Balance" refers to characteristics such as having a broad range of variation in explanatory variables, and comparable sample sizes for treatment and control groups. For an analytical design, it is important that the design be structured to provide good variation in the explanatory variables, and a high degree of orthogonality (low correlation) among those that are not closely related (from the viewpoint of their relationship to the dependent variable). In the physical sciences (laboratory experimentation), it is usually possible to achieve these features easily, since the experimenter can usually arbitrarily set the values of experimental variables (e.g., treatment time, temperature and concentration). In a survey context, there are usually constraints on how the variables may be set – they are determined by their occurrence in the finite population at hand. The main techniques for accomplishing the specification of variable levels are stratification (including matching) and setting of the probabilities of selection.

The concept of having substantial variation in an explanatory variable is not difficult to understand, but the concept of having orthogonality or "low correlation" between variables deserves some explanation. An example will be presented to illustrate this concept, before proceeding to describe the analytical survey-design methodology.

Suppose that we have just three explanatory (independent, experimental) variables, x_1 , x_2 and x_3 , and that we are able to select a sample of eight experimental units. Suppose further that we are interested simply in estimating the "linear" effect of each of these variables, i.e., the difference in effect (on a dependent variable, y) between a high value of the variable and a low value of the variable. In this case, a good design to use is a "factorial" design, in which all combinations of each variable are represented in the sample units. This may be illustrated in the following table listing the eight different types of observations according to their values of the three explanatory variables, where a -1 is used to designate the low value of a variable and a +1 is used to designate the high value.

x_1	x_2	x_3
-1	-1	-1
-1	-1	1
-1	1	-1
-1	1	1
1	-1	-1
1	-1	1
1	1	-1
1	1	1

The above design is a "good" one for two reasons: (1) there is good "balance" in every variable, i.e., the variable occurs the same number of times at the low value and the high value; and (2) there is good "symmetry," i.e., every value of any variable occurs with every possible combination of the other variables. Because of these features, it is possible to obtain a good estimate the

effect of each independent variable (x) on the dependent variable (y). Because of the good balance, the precision of these three estimates will be high. Because of the good symmetry, these estimates will not be correlated, i.e., each estimate will be independent of the other two. The correlation among the independent variables (as measured by the correlation coefficient, which is proportional to the vector inner product of the variables) is zero – the variables (vectors of observations) are said to be orthogonal. (The concept of orthogonality arises in many contexts in addition to experimental design, including error-correcting coding theory (to correct transmission errors in noisy communication channels), spread-spectrum coding (to increase bandwidth in communication channels, to reduce noise) and data encryption (to enhance security).)

Consider the following very poor experimental design:

X ₁	X ₂	X ₃
-1	-1	-1
-1	-1	-1
-1	-1	-1
-1	-1	-1
1	1	1
1	1	1
1	1	1
1	1	1

In this design, the values of the explanatory variables are perfectly correlated. The data analysis would not be able to isolate the effect of the three variables separately, because they vary “hand in hand.”

With the first design presented above, it is possible to estimate the average effect, or “main” effect, of each independent variable on the dependent variable, i.e., the mean (average amount) of change in the dependent variable per unit change in the independent variable. It turns out that with this design it is also possible to estimate “interaction” effects among variables – the difference in the mean change in the dependent variable per unit change in a particular independent variable for different values of another independent variable, or combination of values of other variables. (If it is desired to do this, however, we would need to “replicate” the experiment by observing two or more observations for each combination of the experimental variables.) This may be seen by forming column-wise products of the independent variables:

X ₁	X ₂	X ₃	X ₁ X ₂	X ₂ X ₃	X ₁ X ₃	X ₁ X ₂ X ₃
-1	-1	-1	1	1	1	-1
-1	-1	1	1	-1	-1	1
-1	1	-1	-1	-1	1	1
-1	1	1	-1	1	-1	-1
1	-1	-1	-1	1	-1	1
1	-1	1	-1	-1	1	-1
1	1	-1	1	-1	-1	-1
1	1	1	1	1	1	1

What we see is that each column has good “balance,” i.e., has the same number of high and low values. Moreover, the inner product of each column with every other column is zero, i.e., the correlation among the variables is zero. This means that we may estimate each effect (main or interaction effect) independently, and that the estimates will not be correlated.

In the discussion that follows, we shall present a procedure for ensuring that the correlation among non-causally-related variables is low in an analytical survey design. Motivated by the preceding example, this will be accomplished by ensuring that there is good variation in products of variables. (The example presented above is a greatly simplified example of an experimental design. In general, we are interested in more than two values of each variable.)

It is reiterated that the methodology presented below is intended to construct designs in which a large number of explanatory variables is involved. If only a few independent variables are involved, then the standard methodology of survey design (such as stratification or controlled selection) and experimental design (such as factorial designs, matching and “blocking”) may be employed.

As noted earlier, some applications may focus on estimation of a single or major item of interest, such as a “double-difference” estimate of program impact. This would be the case, for example, if it is possible to implement a pretest-posttest-control-group experimental design or quasi-experimental design, using the double-difference estimate (interaction effect of treatment and time) as the impact measure. On the other hand, randomized assignment of treatment may not be possible, and the goal may be to design a survey to assess the relationship of a number of explanatory variables on program outcome. In this case, the goal of the survey is to develop a general linear model (tables, multiple regression models). In either case, it may be desirable to increase the precision of estimates of interest (such as the treatment effect) by techniques such as matching or blocking. This is readily accomplished in the methodology that follows, but the procedure is a little more complicated if matching is used to enhance precision (e.g., by constructing a comparison group). The presentation that follows will present the methodology in general, and then discuss modifications to accommodate matching.

A General Methodology for Analytical Survey Design

1. Identify dependent variables of interest (“impact” variable, such as income, employment, environmental impact).
2. For each dependent variable, hypothesize a statistical model of interest, such as a multiple regression model that describes program impact as a function of explanatory variables, or a pretest-posttest-control group experimental design intended to produce a double-difference estimate of program impact. It is not necessary to specify a functional form for the model – what is required is to identify known variables that might reasonably be included in such models (i.e., that bear a relationship to a dependent variable, or may be related to the probability of selection of treatment units). If randomized assignment of treatment to experimental units is allowed, this model may include a single variable – the treatment variable. In program evaluation of socioeconomic programs, however, randomized assignment of treatment is often not feasible, and the impact evaluation models often contain many explanatory variables (either as covariates in a pretest-posttest-comparison-group design or as explanatory variables in a multiple regression model). Identify all variables known in advance of the survey that may reasonably be included in these models. Even though many of the variables of interest will not be known until the survey is completed, many variables will be known in advance of the survey, from existing data sources (geographic information systems, prior surveys, government data bases).
3. Construct a database containing all known values of the independent variables, for the primary sampling units (PSUs, such as census enumeration areas or localities). Categorize (classify, stratify) all explanatory variables, which we shall refer to as X’s. For

continuous variables, use a small number of categories (classes, strata) such as two or three. Define the stratum (category) boundaries by quantiles. For nominal-value (unordered) categorical variables, use natural categories. For ordinal-scale categorical variables, use a small number of categories (less than ten, typically two or three). Code all variable categories (e.g., 1, 2, ..., n_c , where n_c denotes the number of categories).

4. If necessary, reduce the number of explanatory variables to on the order of 20. There may be a large number of explanatory variables. There could be ten or twenty, but there could be many more, such as the set of all variables from a Census questionnaire, a national household survey, a business survey, or a geographic information system. As a practical limit, the number of variables must be somewhat less than 255, since that is the maximum number of fields allowed in a Microsoft Access database (the most widely used program for doing database manipulations). Reserving a number of fields for codes and working variables, 200 is a practical upper limit on the number of X 's. To reduce the number of explanatory variables, calculate the Cramer (nonparametric) coefficient of correlation among the X 's (the categorical variables), and combine or delete those that are highly correlated, causally related, and for which the relationship to impact (or selection for treatment) is similar or low. This may reduce the number of X 's somewhat or greatly, e.g., from 200 to 20. There is no need for the number to be much larger than 20. There is no need to use complex methods such as factor analysis or principal-components analysis to reduce the number of variables – examining correlations is quite sufficient.
5. Identify all variables for which an interaction effect is reasonable to anticipate (i.e., for which the variable effect (on a dependent variable) differs according to the value of another variable). For all hypothesized interactions (combinations of variables comprising an interaction), form a new variable which is the product of the component variables. For example, suppose that Y denotes income, X_1 denotes age, and X_2 denotes gender, and it is hypothesized that both X_1 and X_2 affect income but the magnitude of the age effect is different depending on gender. Then, in addition to the two main effects X_1 and X_2 there is an X_1X_2 interaction effect. Define a new variable $X_{1,2} = X_1 X_2$ (i.e., for each unit, the new variable is the ordinary arithmetic product of X_1 and X_2). Add these new variables to the set of independent variables (X 's). For ordinal-scale or interval-scale variables, it is generally easier to form the product variables prior to categorization. For nonordinal variables, the product variable is nonordinal, and its number of values may be as large as the product of the numbers of values of its components.
6. A categorization (stratification) has been defined for every independent variable (original independent variables plus the product variables). Specify the desired sample size for every category (stratum, "cell"). Let X_i denote the i -th independent variable and x_{ij} its value for the j -th category (i.e., for the i -th X , the values are $x_{i1}, x_{i2}, \dots, x_{in_{ci}}$, where n_{ci} denotes the number of categories for the i -th X). Let n denote the desired total PSU sample size, e.g., $n=100$. For each variable, X_i , specify the number, $n(X_i = x_{ij})$ of sample units desired for each value, x_{ij} , of the variable. These values are set to achieve a high level of variability for each variable. Note that unlike stratification in a descriptive survey, at this point the sample size may be zero for some cells; a modification will be made at the end to ensure that all category sample sizes are nonzero. Forcing a high level of variation in the original variables ensures that we will be able to estimate the relationship of impact to each variable with high precision. Forcing a high level of variation in the product variables ensures low correlation among explanatory variables (i.e., a high degree of orthogonality among them). For example, if it is desired to have 20 treatment units in each of five income categories (X_1), then $n(X_1 = j) = 20$ for all j . If there are two gender categories (X_2),

then $n(X_2 = j) = 50$ for all j . Let k denote an arbitrary category, or “cell” (i.e., a category of items (PSUs) having the same value for a given categorical explanatory variable).

7. Let n_k denote the desired number of sample units falling in this category (may be zero for some categories) and let p_k denote the probability of selection for units in the k -th category. The expected number of sample items falling in the k -th category is $E_k(p) = (\text{number of population items in category } k) \times (\text{probability of selection } (p_k))$. Now, apply a suitable numerical algorithm to determine a set of selection probabilities (p_k 's) that causes the expected number of sample items falling in each cell to be close to the desired number. It is not necessary (or even possible, in practical applications) to have the expected number equal to the desired number, since the requirement is simply to achieve a reasonable level of variation in each variable. A method that has proven both simple and effective is the following. First, sort the cells in order of their sampling fractions (ratio of the desired number of sample units in the cell to the number of population units in the cell). Starting with the cell having the largest sampling fraction, assign that probability of selection to all units falling in the cell. Since each unit may fall in other cells (of other variables of stratification), this probability of selection will then automatically be assigned to units in other cells. Move to the cell having the next-highest sampling fraction. Recalculate the sampling fraction by subtracting the already-assigned probabilities (of units processed in the previous step) from the original sampling fraction, and set the selection probability for all other units (in the cell) equal to this adjusted sampling fraction. Repeat this process for all cells having non-zero sampling fractions (i.e., containing population and sample units). (At the end of this process, if the categorization has positive desired sample sizes for all categories, then all population units will have a nonzero probability of selection. If the categorization does not have nonzero desired sample sizes for all categories, then some of the items would have been assigned a zero probability of selection. At the end of the process, set the minimum probability of selection for all population units equal to a low nonzero value.)
8. The result of this (or other suitable) process is a set of selection probabilities, p_i – one for each unit (i) of the population. The value of p_i is the same for all units within a particular category. A sample of PSUs is then selected using these probabilities. Since the PSUs may vary in size, and it is often desired to select an equal-sized sample of lower-level units from each PSU, the PSUs would normally be selected with probabilities proportional to a measure of size (PPMS), e.g., by using the Rao-Hartley-Cochran (RHC) selection procedure. A problem that arises is that the selection probabilities are determined by the above optimization procedure, and they will not be proportional to size (PPS). To address this problem, include PSU size as one of the “explanatory variables,” and impose a constraint on the sample size for various size categories of PSU (such that the sample will be approximately PPS or PPMS).
9. Select a probability sample from the population of PSUs, using the constructed selection probabilities. To do this, select a (uniformly distributed) random number for each unit and include the unit in the sample of this random number is less than or equal to the selection probability for the unit. Tally the number of sample items in each category and compare to the desired sample sizes for each category.
10. Upon completion of this process, the total expected number of sample items falling in a cell may be larger than desired (for two reasons: (1) because the sampling fractions for some categories may be larger than desired, in order to assure a sufficient sample size for some other categories, and (2) because the selection probability was raised from zero to a small

nonzero value for all categories having zero selection probabilities), and the total expected sample size may exceed the total desired sample size. Apply the appropriate multiplicative factor to all of the selection probabilities that are less than unity to match the desired total sample size.

The success of the method depends on how successful we are in adjusting the selection probabilities to meet the design constraints (on total sample size and category allocation). This problem is a constrained optimization problem, and it is rather difficult to solve since the objective function is not “cell-separable” (decomposable) – the selection of a particular unit affects the allocation in many cells (and many constraints), not just one. There are numerous interrelated and conflicting constraints, and there may be thousands of selection probabilities to work with (i.e., one for each population unit). While the method is not a formal optimization method, and does not produce a solution that is optimal with respect to a specific objective function, it has been observed to work quite well (i.e., it has very fast computer execution times and produces useful results).

Note that the probabilities of selection for an analytical survey design are usually quite different from the probabilities of selection for a descriptive design. That is, a design that is oriented toward estimation of means and totals for the population and major subpopulations of interest is usually quite different from one oriented to estimation of many relationships in an analytical model. It is usually the case, in fact, that the selection probabilities for an analytical survey design will be quite inappropriate for a descriptive design. In order to satisfy all of the stratification constraints, severe “distortions” will be introduced into the selection probabilities, compared to those for simple random sampling or ordinary stratification for descriptive-survey estimates. As noted earlier, when constructing an analytical model it is theoretically not necessary to keep track of the selection probabilities, or that they be comparable in magnitude for subgroups of comparable interest. In the data analysis, two kinds of estimators will be examined – unbiased estimates, that take into account the selection probabilities but may have low precision, and biased estimates that ignore the selection probabilities (e.g., are conditional on the particular sample, not on the population) but have high precision. Since the design is complex, closed-form formulas will not be available for estimation of sampling variances (and confidence intervals). Resampling methods will be used to estimate sampling variances (both relative to the sample and to the population).

Modification to Allow for Matching

If matching of units is to be done, Steps 8-10 of the process are modified. Matching may not be required at all, as in the development of an analytical model in which treatment is represented by a level, rather than by a dichotomous “treatment” variable, or in a situation in which there is no acceptable population from which to select comparison units. If it is desired to determine the treatment units by randomization, as in an experimental design, then matching is optional. It is not required – the control group could be selected randomly without matching – but it is often desirable to use matching as part of the randomized selection to increase the precision of difference estimates (by forming matched pairs). If the treatment units are to be randomly selected using matching, then match all population items (PSUs) to a nearest neighbor (using all explanatory variables as the matching variables), form strata each of which consists of two matched items, and randomly select one of the matched pairs as the treatment unit and the other as the control.

It would appear that there are two methods for selecting a matched sample: (1) sort the entire population (using matching) into sets of matched pairs (or triplets, or however many comparison groups are desired), randomly assign one of each match-set to treatment, and select a random sample of match-sets; and (2) specify the treatment and control populations (which may or may not overlap), match each treatment unit with the best-matching control unit (or units), and select a

random sample of match sets. In fact, both methods are equivalent to case (2), if in the first case we simply apply case (2), defining the treatment population and the control population each to be the entire population.

In either case, the sample is selected by randomly selecting matched pairs. The sample selection process proceeds in the usual fashion, but whenever a unit is selected, its match is also included in the sample. For the analysis, it is necessary to know the ultimate probability of selection of each sample unit, corresponding to this process. Since the sample draws are independent, the actual probability of selection resulting from this procedure is easily calculated (by the formula $\text{prob}(\text{unit } i \text{ is included in sample}) = p_1 + p_2 - p_1p_2$ where $p_1 = \text{prob}(\text{unit } i \text{ is selected in the draw})$ and $p_2 = \text{prob}(\text{unit } i \text{ is added as a matching item when its nearest neighbor is selected}) = \text{prob}(\text{unit } i \text{'s nearest neighbor is selected in the draw})$). Note that we are selecting the sample by making use of the original probabilities of selection of individual units (before adding the matched units to the sample), not the probabilities of inclusion after adding the matching units (as given by the preceding formula). The ultimate probabilities of inclusion of the sample units (as given by the formula) are used for the analysis, not for the sample selection. (For applications involving match sets of more than two units, the formula for calculating the probability of inclusion is similar to the one just presented for the case of matched pairs (e.g., $p(\text{inclusion}) = p_1 + p_2 + p_3 - p_1p_2 - p_1p_3 - p_2p_3 + p_1p_2p_3$ for the case of matched sets of size three.)

A number of algorithms are available for matching, and could be used in this application. A simple general-purpose method for matching items that have been coded into a small number of categories (either ordinal or nonordinal or both) is the following. Units are matched by calculating a distance measure (or function) from each population unit of the treatment population to every yet-unmatched unit of the control population (recall that these two populations may be distinct, overlap, or be identical), and selecting the closest one. The matching process starts from the top of a list of treatment population units in random order. Once the nearest match to a unit has been identified, both items are removed from the list. The process ends when half of the units in the population have been processed. This procedure, sometimes referred to as “greedy matching,” assures that half of the units will be matched to “nearest neighbors.” The other half (the units selected as nearest neighbors) may not be matched to *their* nearest neighbor, but to a “nearby” neighbor (i.e., greedy nearest-neighbor matching is not a symmetric relation).

The distance between two units is determined by calculating a distance “component” for each design variable and forming a weighted sum of these distance components, where the weights reflect the importance of the various matching variables relative to the dependent variable. The distance component is defined differently for ordinal and non-ordinal variables. For non-ordinal variables, the distance component is defined as equal to one if the sample units being compared have the same value of the variable, and zero otherwise. For ordinal variables, the distance component is defined as the difference between the values of the variable for the two units divided by the range of the variable. For both types of variables, the maximum value of the distance component is one and the minimum value is zero.

Note that when stratification is involved, it may be the case that two nearest neighbors do not fall in the same stratum cell (i.e., they match on some, but not all, of the variables). There is hence an issue of deciding to which stratum cell (category) a matched pair belongs. This apparent issue is obviated by selecting the sample in the way described above (i.e., by selecting individual units (not pairs) with specified probabilities of selection, and then adding the nearest-neighbor match for each selected item to the sample). The matching unit may or may not be in the same stratum, but it probably will be, since the same variables are typically used for matching as for stratification.

Note that with this procedure, the probability of selection is known for each unit of the sample. This is not the case for ex post matching (i.e., matching after the treatment sample has been selected).

Upon completion of this process, the total expected sample size will usually exceed the total desired sample size, and the total expected number of sample items falling in a stratum cell may not be as desired, and. (The former condition happens also in the case of no matching, in order to assure that every population unit is subject to a nonzero probability of selection.) This happens because whenever a PSU is selected into the sample, its matching item(s) are also included in the sample, and it may belong to a different stratum cell (if it does not match its match on all criteria). Apply the appropriate multiplicative factor to all of the selection probabilities to match the desired total sample size. (One adjustment is required in non-matching applications, and about ten iterations are required in matching applications (since the allocation is not a continuous function of the selection probabilities, units are added to the sample in pairs, and the matching unit of a selected unit may already be in the sample (especially in cells having small populations)).)

Whether the actual sampled stratum allocations match the desired stratum allocations exactly is not of great concern. The objective is to use stratification to achieve a reasonable level of balance, spread and orthogonality in the explanatory variables of a model. The allocation will never be “perfect,” because we are selecting the sample from a finite population, and some desired combinations of variable values may occur infrequently or not at all. As long as the balance, spread and orthogonality are reasonably good, the precision of the model estimates will be satisfactory, and the precision of the estimates of interest (e.g., the average double-difference estimate of impact, a general regression estimate of impact, or the relationship of impact to explanatory variables) based on this model will be satisfactory. (Note that the estimated impact is not very sensitive to errors in model specification or estimation (e.g., a maximum likelihood estimate of a parameter is invariant with respect to reparameterization) – it is omitted variables that are of greater concern than the functional form of observables.)

As mentioned in the main text, matching on individual units leads to orthogonality of the treatment variable (usually of two values, treated vs. control) with respect to the match variables, but it does nothing to control the spread and balance of the observations over the design variables, nor does it affect the orthogonality among other design variables. In general, design of an analytical survey *always* involves control for spread, balance and orthogonality (through marginal stratification), and it may or may not include matching of individual units. Note that control of orthogonality through marginal stratification is a form of matching, but it matches groups, not individual units. When there is a role for matching of individual units, it should always be done in preference to matching by marginal stratification (because it increases precision of estimates of differences and because it matches the joint distribution of the match variables, not just the marginal distribution). But matching on individual units has the distinct limitation that it introduces orthogonality only of the design variables with respect to the treatment variable, and has no effect on the orthogonality of other variables (i.e., between explanatory variables that may be included in a model).

Whenever it is attempted, by whatever means, to increase the degree of orthogonality between or among variables, the result is that the distributions are “matched” – the distribution of one variable is the same, independent of the value of the other. For this reason, when marginal stratification is used to promote orthogonality (e.g., by achieving uniform spread in a product variable), it is in fact a “matching” procedure (matching of samples, not of individual units). On the other hand, the use of marginal stratification may have nothing to do with promoting orthogonality, e.g., if used only to promote spread and balance.

Selected References in Sample Survey Design

1. Cochran, W. G., *Sampling Techniques*, 3rd edition, Wiley, 1977
2. Kish, L., *Survey Sampling*, Wiley, 1965
3. Des Raj, *The Design of Sample Surveys*, McGraw Hill, 1972
4. Cochran, William G. and Gertrude M. Cox, *Experimental Designs*, 2nd edition, Wiley, 1950, 1957
5. Campbell, Donald T. and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, 1966. Reprinted from Handbook of Research on Teaching, N. L. Gage (editor), Rand McNally, 1963.
6. Cook, Thomas D. and Donald T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* Houghton Mifflin, 1979
7. Rao, J. N. K. and D. R. Bellhouse, "History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis," *Survey Methodology*, June 1990
8. Imbens, Guido W. and Jeffrey M. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, vol. 47, no. 1, pp 5-86, (2009)
9. Holland, Paul W. "Statistics and Causal Inference," *Journal of the American Statistical Association*, vol. 81, no. 396, December 1986
10. Rubin, Donald B., "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, vol. 6, no. 1, pp. 34-58 (1978)
11. Risto Lehtonen and Eriikki Pahkinen, *Practical Methods for Design and Analysis of Complex Surveys*, 2nd edition, Wiley, 2004
12. Thompson, Steven K., *Sampling*, 2nd edition, Wiley, 2002
13. Shao, Jun and Dongsheng Tu, *The Jackknife and Bootstrap*, Springer, 1995
14. Efron, B. and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, 1993
15. Wolter, Kirk M., *Introduction to Variance Estimation*, Springer, 1985
16. Mood, Alexander M., Franklin A. Graybill and Duane C. Boes, *Introduction to the Theory of Statistics*, 3rd edition, McGraw-Hill, 1950, 1963, 1974
17. Cohen, Jacob, *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, 1969
18. Scheaffer, Richard L., William Mendenhall and Lyman Ott, *Elementary Survey Sampling*, 2nd edition, Duxbury Press, 1979 (6th edition, 2005)

19. Kusek, Jody Zall and Ray C. Rist, *Ten Steps to a Results-Based Monitoring and Evaluation System*, The World Bank, 2004 (In French: *Vers une culture du résultat: Dix étapes pour mettre in place un système de suivi et d'évaluation axé sur les résultants*, Banque Mondiale, Nouveau Horizons, Éditions Saint-Martin, 2004)
20. Iarossi, Giuseppe, *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*, The World Bank, 2006
21. Caldwell, Joseph George, *Approach to Sample Survey Design*, 1978, 2007, <http://www.foundationwebsite.org/ApproachToSampleSurveyDesign.htm>
22. Caldwell, Joseph George, *Approach to Evaluation Design*, 1978, 2007, <http://www.foundationwebsite.org/ApproachToEvaluation.htm>
23. Caldwell, Joseph George, *Sample Survey Design and Analysis: A Comprehensive Three-Day Course with Application to Monitoring and Evaluation*. Course developed and presented in 1979 and later years. Course Notes posted at Internet website <http://www.foundationwebsite.org/SampleSurvey3DayCourseDayOne.pdf> , <http://www.foundationwebsite.org/SampleSurvey3DayCourseDayTwo.pdf> and <http://www.foundationwebsite.org/SampleSurvey3DayCourseDayThree.pdf> .