

Vista's Approach to Sample Survey Design

© 1978, 1988, 2006, 2007, 2009 Joseph George Caldwell. All Rights Reserved.
Posted at Internet website <http://www.foundationwebsite.org>. Updated 20 March 2009 (two references added).

Contents

1. The Role of Sample Survey Design	1
2. General Procedures for Development of a Sample Survey Design	3
3. Development of OMB Clearance Request Package	5
4. Considerations in the Design of Analytical Surveys.....	6
5. Vista's Experience in Sample Survey Design	8
Selected References in Sample Survey Design	9

1. The Role of Sample Survey Design

Sample survey is concerned with the problem of making inferences about the characteristics of a collection of items -- a "population" -- based on the characteristics of a portion of the population -- a "sample." If the sample is selected from the population in such a way that the probabilities of selection are known, then the theory of statistics may be used to make these inferences, i.e., to estimate population characteristics and to test hypotheses about the population.

There are several reasons for employing proper statistical design techniques in a research study. First, the use of statistical sample survey design techniques not only assures sample estimates having satisfactory levels of precision, but it also enables *measurement* of what that level of precision is. The usual measure of precision is the standard error of the estimate. Second, the use of sound sample survey design procedures improves the ratio of cost to level of precision, by enabling higher precision for a given level of cost, or lower cost for a desired level of precision. The increase in precision or decrease in cost is effected by allocating the sampling effort in ways that take into account the costs of sampling, the variability of the target population, and special features of the target population (such as the occurrence of natural "clusters" of the population, e.g., cities, schools, and households). Third, the use of sound sample survey procedures assures high validity of the estimates, i.e., the estimates will have low bias. A final advantage in the use of statistical design concepts is in enhancing the usefulness of the survey results by assuring (prior to implementation of the survey) that the level of precision of the sample estimates will be sufficient to permit meaningful interpretation.

(Note: *Precision*, or *reliability* refers to the variability of an estimate -- i.e., the extent to which the estimate fluctuates around the average value obtained through repeated sampling. *Validity* refers to how close this average value (obtained in repeated sampling) is to the "true" (population) value of the parameter being estimated (and also to how well that parameter represents the concept that is desired to be measured). The difference between the average value and the population value is called the *bias*. Precision is usually measured by the standard error of the estimate. Validity is measured by the bias. *Accuracy* refers to a combined measure of precision and validity. Accuracy is usually measured by the *mean squared error*, equal to the square of the standard error plus the square of the bias.)

Construction of an efficient survey design is accomplished by taking into account the characteristics of the population being surveyed, the costs of sampling, and the precision requirements of the study. It is important to keep in mind the study requirements, in order to assure that the precision obtained for key variables is sufficient, but not unnecessarily high. Furthermore, it is desirable to obtain comparable levels of precision for estimates of variables that are of comparable importance.

The development of a sample design is accomplished by taking into account the identification of the basic estimates of interest together with prior information concerning the variability of the population with respect to the variables of interest and the intercorrelations between these variables. Because most projects have multiple inference objectives, the design is usually not "fine-tuned" to provide maximum precision for a single estimate. Rather, the design is generally structured to provide acceptable accuracy for a number of relationships of interest.

There are two basic types of sample surveys – descriptive sample surveys and analytical sample surveys. In a descriptive sample survey, the objective is to obtain estimates of basic descriptive characteristics (such as means or proportions) for the whole population or particular subpopulations of interest. In an analytical sample survey, the primary objective is to be able to estimate the relationship of a particular variable (a "dependent" variable) to other variables (the "independent," or "explanatory," variables): for example, the relationship of client income to program inputs, client characteristics, and regional demographic and economic characteristics.

The distinction between a descriptive and an analytical survey has a profound effect on the nature of the survey design. In a descriptive survey, attention usually centers on the estimation of the state of a finite population at some point in time. The estimation objectives of the survey are simple -- usually just estimation of means and proportions for the population and various subpopulations, with perhaps some estimation of basic relationships, usually through presentation of results by stratum or by crosstabulations. The sample design process usually proceeds without difficulty -- the basic requirement is that the sample size is sufficiently large for all subpopulations of interest.

In an analytical survey, on the other hand, attention centers on the estimation of relationships between (dependent and independent) variables, for a conceptually infinite population. Attention centers on making inferences about the *process* generating or acting on the population, not about the population itself. The purpose of the survey is to develop a model of the process. Usually, a parametric statistical model such as the general linear model (regression model) is used. The analysis involves the iterative steps of model identification, estimation, tests of model adequacy, and respecification.

The major differences between an analytical survey and a descriptive survey are the following. In a descriptive survey, we want large sample sizes in subpopulations of interest. In an analytical survey, on the other hand, we want variation and balance in the explanatory variables, and orthogonality (low correlation) between variables that are not causally related. In addition, the finite population correction factor is irrelevant in an analytical survey. This fact can have a substantial impact on sample size determination.

A sample survey that has the estimation of relationships as its primary objective differs from the majority of sample surveys, whose objective is simply to compute descriptive statistics (such as percentages) about the populations and subpopulations under study. The relationships between the study variables are often complex, and there are numerous instances in which the independent variables may be correlated to a substantial degree. In some instances, the relationship of the

correlated independent variables to a dependent variable may be quite difficult to assess. In order to obtain a good estimate of these relationships, the sample design must be structured to "orthogonalize" these "confounded" independent variables, i.e., to reduce the correlation between them.

We shall now describe a general procedure for determining a sample design. This description is in two parts. First, we discuss general procedures that apply to both descriptive and analytical survey design. That discussion is followed by a discussion of additional considerations that apply to the development of an analytical survey design.

2. General Procedures for Development of a Sample Survey Design

A standardized procedure for developing a sample design is presented below. This general procedure may be followed in developing the sample design for most applications.

The Elements of Survey Design

1. Specify population of interest
2. Specify units of analysis and estimates of interest
3. Specify precision objectives of the survey; resource constraints; political constraints
4. Specify other variables of interest (explanatory variables, stratification variables)
5. Review population characteristics (distributional, cost)
6. Develop instrumentation (development, pretest, pilot test, reliability and validity analysis)
7. Develop sample design
8. Determine sample size and allocation
9. Specify sample selection procedure
10. Specify field procedures
11. Determine data processing procedures
12. Develop data analysis plan
13. Outline final report

The paragraphs that follow describe each of the steps listed above.

1. Specify Population of Interest The population about which it is desired to make inferences is called the *target population*. It is defined by four quantities -- content, units, extent, and time. (For

example, we may be interested in estimating the *income, of US citizens, residing overseas, in the past year.*)

In many surveys, there are several different *units of analysis* (e.g., hospitals and patients, or schools and students). The *survey population* is a subset of the target population, that recognizes practical constraints, such as inaccessibility or lack of data, cost, and political or legal constraints.

Prior to development of a good sample design, it is helpful to summarize what information is known about the survey population, such as means and variances of the population and subpopulations of interest; sampling cost information; and the cluster structure of the population.

2. Define Estimates of Interest The variables that are of interest must be identified (e.g., age, race, sex, earnings, status). The design effort should specify subpopulations that are of interest (e.g., women, the unemployed). Surrogate variables must be identified if desired variables cannot be obtained (e.g., earnings vs. income). The method of observation (record scan, personal interview, telephone interview, or mail questionnaire) must be identified, as must the measures of interest (e.g., mean earnings change in the past year).

3. Specify Objectives and Constraints The precision objectives of the survey may be specified in terms of standard errors or in terms of confidence intervals on key parameter estimates, for high-interest populations. Resource constraints, political and legal constraints, and data measurement constraints must be identified.

4. Specify Other Variables of Interest Variables which can be used to stratify (categorize) the population should be identified. Stratification can be used to improve precision, to reduce costs, to assure adequate precision for subpopulations of interest, and to enable measurement of precision. Also, explanatory variables that can be used in crosstabulations, ratio and regression estimates, and nonresponse analysis should be specified.

5. Review Population Characteristics Whatever information is available on distributional breakdowns (frequency distributions, crosstabulations) should be reviewed. This information can suggest what stratifications might be of value in the design stage. Information on within-stratum variances and intracluster correlation coefficients can also assist the design effort, as can information on sampling costs.

6. Develop Instrumentation The major steps in developing the survey instrumentation are:

- o Determination of the type of instrumentation
- o Development of the instruments
- o Pretest of the instruments
- o Pilot test of the instruments

In developing the survey instruments, a number of factors should be kept in mind. These factors are as follows:

- o Question content
- o Question order
- o Question wording
- o Question structure/format (open or closed, number of categories)
- o Questionnaire length

- o Questionnaire layout
- o Questionnaire instructions
- o Opportunities for interviewer comment on validity of response

7. Develop Sample Design The process of determining the specific sample design is a creative one. The variety of choice is essentially limitless, and the number of design objectives are usually multiple and conflicting. The recommended procedure is to synthesize several design alternatives which emphasize different design objectives, to characterize the alternatives in terms of precision, cost, and operational problems, and to achieve a consensus on the best overall design. The design alternatives include various combinations of stratification, clustering, multistage sampling, and double sampling.

8. Determine Sample Size and Allocation This step involves specification of items such as the number of first-stage and second-stage sample units, the number of sample units per stratum, and the proportion of the panel to be replaced in panel sampling.

9. Specify Sample Selection Procedures It must be decided how the sample units are to be selected -- with or without replacement, probability-proportional-to-size sampling, systematic sampling, Rao-Hartley-Cochran sampling, or controlled selection.

10. Specify Field Procedures "Field procedures" include the number and spacing of questionnaire waves, nature of the initial contact, recall procedures, field edit procedures, and transmittal to central headquarters.

11. Specify Data Processing Procedures "Data processing procedures" include procedures for logging, manual edit, coding, keying, machine edit, data base design, treatment of non-response and missing values, and data base documentation.

12. Develop Data Analysis Plan The data analysis plan should include consideration of a preliminary descriptive analysis of the data, nonresponse counts, and treatment of nonresponse. The directed analysis of the data includes computation of estimates of interest, crosstabs, and tests of hypotheses. In an analytical survey, the iterative process of model building must be specified.

13. Report Preparation The report of the analysis should include estimates, tables, charts, commentary, and interpretation. Also, the report may contain tables of generalized variances, a characterization of nonresponse, and discussion of possible sources of bias.

3. Development of OMB Clearance Request Package

For US survey efforts requiring collection of data from more than nine individuals or organizations, clearance must be obtained from the Office of Management and Budget (OMB). This clearance is required by the Federal Reports Act; the clearance request is effected by submission of Standard Form 83, "Clearance Request and Notice of Action." If a pretest is conducted involving more than nine respondents, it must also be approved by OMB. Vista staff are very familiar with the procedures required to OMB clearance. We have developed a procedure that has proved effective in obtaining OMB clearance without incurring additional delays. While OMB clearance was formerly accomplished in about six weeks, this process now takes up to four months, with additional time often required by the requesting agency's own review (although the survey

contractor customarily prepares the justification package, it is the agency that actually submits it to OMB). Vista staff have prepared clearance requests that have "passed" OMB review without a single provision or requirement! We are able to assure smooth accomplishment of the clearance process by adherence to a number of steps, which include:

1. Careful development of the survey instruments, including consideration of item and instrument reliability and validity (with special attention to layout or forms design, and to question content, wording, and order); compatibility of answer categories with traditional government response categories; inclusion of only those data elements required to satisfy the survey objectives; and capitalization on experience with similar questions / instruments in earlier surveys.
2. Careful development of the survey sample design, with full consideration of objectives, sampling costs, and the nature of the target population in developing an efficient sample design.
3. Meticulous attention to the *Instructions for Requesting OMB Approval under the Federal Reports Act*, as revised. Figure 1 contains a list of the items required in the approval application.
4. Consideration of the terms of relevant legislation, such as the Privacy Act of 1974, the Freedom of Information Act, the Department of Commerce's "Directives for the Conduct of Federal Statistical Activities," and agency regulations in developing the survey instruments and procedures.
5. Close coordination with agency staff, OMB liaison staff, agency counsel, OMB clearance officials, and other government officials *before* submission of the clearance request form.
6. Consultation with appropriate interested parties. A key item in most surveys is item 5(c), "Consultation with State Government Officials." This consultation is required. Failure to consider the interests and concerns of interested organizations (advocacy groups, professional associations, interstate commissions) can cause serious delays – even total failure -- in large-scale survey efforts.

By following the above guidelines, we anticipate no problems in obtaining OMB approval of the survey instruments and design. In essence, our approach consists in capitalizing on previous experience in this area. A key resource in the preparation of the design-related portions of the qualification statement is the fact that one of the principals of Vista (Dr. Caldwell) is a Ph.D. statistician with over ten years' experience in sample survey design.

Survey Design Report

A final survey design report generally includes revised pretested instrument, a recommended sample design (including a description of the sample selection procedure and the probabilities of selection for the sample units), and an OMB clearance request package. The instruments and OMB package are included as appendices.

4. Considerations in the Design of Analytical Surveys

As mentioned earlier, the purpose of an analytical study is essentially to develop a model that will describe the relationship of one variable (the dependent variable) to a number of other

(independent or explanatory) variables. In order to be successful in this objective, we need to include in the model all variables that are expected to have a substantial impact on the dependent variable. The list of potential variables that could reasonably be included in the model may be long.

In order to be able to develop a model that adequately describes the relationship of a dependent variable to various independent variables, it is necessary that the independent variables exhibit two basic properties:

1. There is good "balance" -- that is, adequate variation -- in the independent variables (i.e., there are comparable numbers of "high" and "low" values of each independent variable); and
2. There is a high degree of orthogonality between independent variables that are not causally related.

The problem that arises in the development of an analytical survey design (i.e., a survey design to support development of an analytical model) is that there are usually several independent variables of interest, each represented by several levels, and the total combination of levels of variables (or "cells") is usually far larger than the allowed sample size. The standard sampling approach of multiple stratification hence cannot be applied (since the number of stratum "cells" will exceed the sample size).

An approach that has often been used in the past in this situation is to select a "judgment" sample of observations. This has a serious drawback, however, in that the theory of statistics can no longer be applied to make inferences from the sample data. Moreover, judgment sampling may easily be avoided, since there is a statistical procedure for sample selection that possesses all of the flexibility of judgment sampling, but has none of the associated drawbacks. This procedure is called *controlled selection*.

A final note concerning the sample selection concerns ways in which prior information on dependent variables or surrogate dependent variables (i.e., manpower levels) can be used in the sample selection process. In "descriptive" surveys, this kind of information is often used as a basis for stratification. In "analytical" surveys, this kind of information is usually not used -- the design generally centers on the *independent* variables, not on the *dependent* variables.

For analytical surveys, we generally utilize the statistical sampling procedure of controlled selection as the basis for the sample design. This procedure was developed for the situation in which multiple stratification results in a number of cells that exceeds the allowable sample size. The procedure is illustrated in Figure 4. In this example, there are a total of three variables of stratification. Two of the variables have three levels, and one of the variables has two levels. There are a total of $3 \times 3 \times 2 = 18$ cells in the multiple stratification, but three of the cells contain no population elements.

It is desired to select a sample of $n=10$ units from the 15 occupied cells of the table. The way that this is done is to specify two "sample patterns," either one of which is satisfactory from an analysis point of view (i.e., has adequate variation and balance in each variable, and adequate orthogonality between the variables), and all of which "cover" the entire cross-stratification grid table (i.e., each cell is contained in at least one pattern). Probabilities of selection are then specified for each pattern, and one of the patterns is selected. The selected pattern is the sample to be used. Since every sample pattern was specified such that it was "desirable" from an

analysis point of view, the sample pattern that is selected is guaranteed to be completely acceptable. Since every unit was contained in at least one pattern, every unit has a nonzero probability of selection, and probability theory can hence be used to support analysis of the data.

We see then, that controlled selection possesses all of the advantages of judgment sampling (i.e., every pattern is a "judgment" sample, and exhibits the desired variation in the variables), but none of the disadvantages (selection biases, inability to apply statistical inference to a non-probability sample).

The design task is concerned with determining exactly which variables are to be included in the sample design, and the number of categories of each variable.

Specification of the Regression Model

An analytical survey is generally concerned with using the sample survey data to estimate the parameters of a linear regression model, such as a multiple regression model. The general form of a regression model is:

$$y = b_0 + b_1X_1 + b_2X_2 + \dots + b_mX_m + e$$

where

y = dependent variable

x = independent variable, or explanatory variable

b = regression coefficient (model parameter), to be estimated in the regression analysis

e = model error term, representing random variations in the dependent variable that are not explained by the rest of the model

In regression analysis, there are generally two approaches. One is to attempt to develop a single "combined" regression model that represents the relationship for all members of the population. The other is to break the population into a number of different categories, and to develop a different model for each category. This latter approach (the "separate regressions" approach) is used if the nature of the relationship is quite different for the different categories. If the total number of observations in this analysis is probably not very great, it is prudent to avoid the use of separate regressions, since the precision of the coefficient estimates would generally be unsatisfactorily low. Differences in the nature of the relationship among different categories is generally treated by including "interaction" terms in the model.

5. Vista's Experience in Sample Survey Design

Sample surveys that Vista staff have designed include the following:

- o The 1976 Survey of Institutionalized Persons

- o Survey of the Impact of National Health Insurance on Bureau of Community Health Service Users
- o Nationwide Hospital Cost Data Study
- o Professional Services Review Organization (PSRO) Data Base Development Study
- o Follow-up Study of Vocational Rehabilitation Clients
- o Elementary and Secondary School Civil Rights Survey
- o Needs Assessment of State Social Services Programs
- o Survey of Discrimination in US Sales and Rental Markets
- o Survey to Collect Data for the US Highway Capacity Manual
- o Survey of Coffee Production in Haiti

The Hospital Cost Data Study, the PSRO Data Base Development Study, the Housing Market Discrimination Study, and the US Highway Capacity Manual Study involved the development of analytical survey design; the others involved descriptive survey design.

In addition to the above one-time sample surveys, Vista staff developed the sampling manuals used by states to collect data for the Social Services Reporting Requirements, for Utilization Review of Medicaid, and for Office of Child Support Enforcement Reporting Requirements.

In addition to experience in the design of sample surveys, Vista has capabilities in technical training in the area of sample survey design and analysis. Dr. J. G. Caldwell developed the popular seminar, "Sample Survey Design and Analysis," which has been attended by members of government and industry. A copy of a flyer advertising this seminar is included at the end of this section.

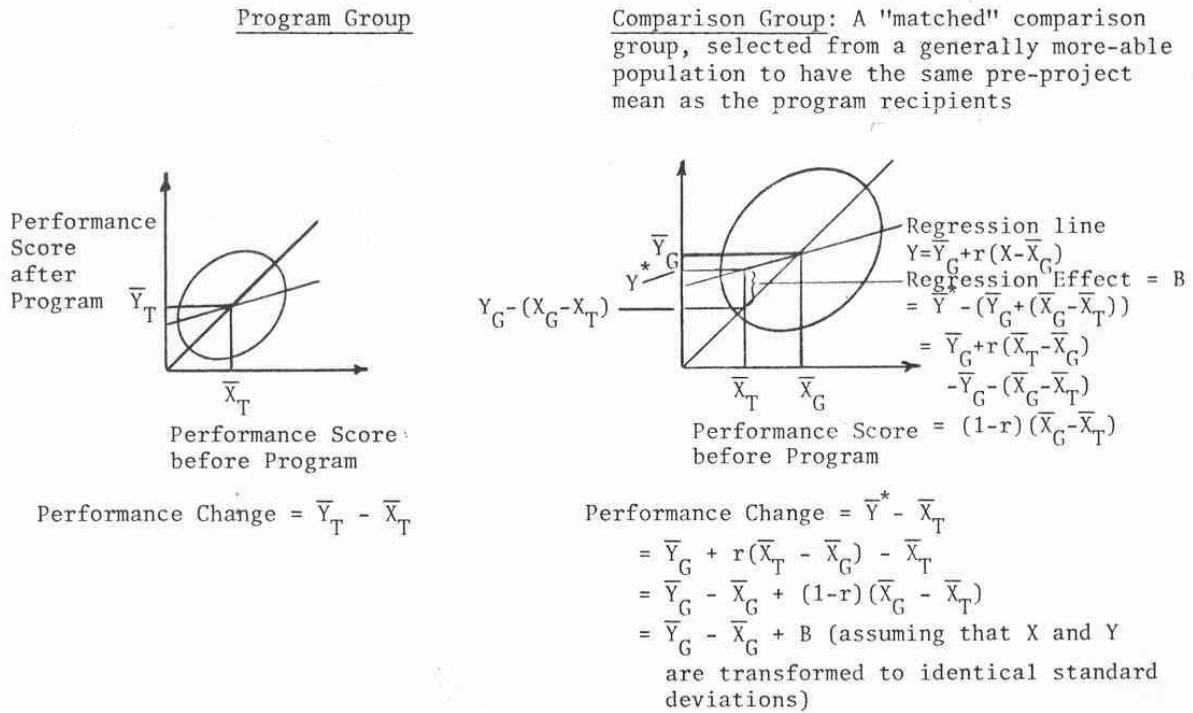
Selected References in Sample Survey Design

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. 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> .

5. Caldwell, Joseph George, *Vista's Approach to Evaluation*,
<http://www.foundationwebsite.org/ApproachToEvaluation.htm> or
<http://www.foundationwebsite.org/ApproachToEvaluation.pdf> .

6. Caldwell, Joseph George, *Sample Survey Design for Evaluation*,
<http://www.foundationwebsite.org/SampleSurveyDesignForEvaluation.htm> or
<http://www.foundationwebsite.org/SampleSurveyDesignForEvaluation.pdf>

Figure 1. Problems and Pitfalls in Evaluation Research:
How NOT to Select a Comparison Group (the Regression
Effect)



Suppose that the program has a real effect, E:

$$E = (\bar{Y}_T - \bar{X}_T) - (\bar{Y}_G - \bar{X}_G) = (\bar{Y}_T - \bar{Y}_G) - (\bar{X}_T - \bar{X}_G)$$

Then, using a "matched" comparison group, the observed program effect is:

$$E_B = \text{Change for program group minus change for comparison group}$$


$$= (\bar{Y}_T - \bar{X}_T) - (\bar{Y}_G + r(\bar{X}_T - \bar{X}_G) - \bar{X}_T)$$

$$= \bar{Y}_T - \bar{Y}_G - r(\bar{X}_T - \bar{X}_G) = E - B.$$

In general, whatever positive impact the program has will be biased low by the size of the regression effect, B. In particular, if the program is totally ineffective, $E_B = -B$, i.e., it will appear to have a negative impact, solely because of the regression effect. In other words, selection of "matched" comparison groups from generally more-able populations introduces negative biases into observed program impact measures.

FIGURE 4. SIMPLIFIED EXAMPLE OF CONTROLLED SELECTION

		1			2		
		1	2	3	1	2	3
Variable 1 Variable 2 Variable 3	1	/	o	x	x	o	x
	2	o	x	/	o	x	/
	3	o	o	x	o	o	x

 NO POPULATION

PATTERN X (MORE DESIRABLE): $P = .6$

PATTERN O (LESS DESIRABLE): $P = .4$

$P = 1.0$

ALL CELLS COVERED, THEREFORE ALL POPULATION UNITS HAVE A NONZERO PROBABILITY OF SELECTION