Encyclopedia of Social Measurement
Kimberly Kempf-Leonard, Editor-in-Chief,
San Diego, CA: Academic Press, 2003

SAMPLE DESIGN

W. Penn Handwerker
University of Connecticut

GLOSSARY

autocorrelation correlation of case residuals, ordinarily among cases adjacent in time or space.

confidence interval (limits) interval identified by specific upper and lower bounds, or limits, that contains the population parameter a specific proportion of the time (e.g., 95%) for the same variable measured for all possible samples of a specific size drawn from a specific population at a specific time.

case a member of a population, the primary sampling unit on which one makes measurements.

enumeration (listing) unit spatially discrete sampling unit which contains sets of cases.

ethnography, ethnographic analysis identification and description of cultures based on analysis of behavioral and cognitive similarities among cases

parameter the value of a variable that characterizes a population.

population the set of cases to which one generalizes sample findings.

point estimate your single best guess about a parameter of interest.

sampling distribution the frequency distribution of all possible statistics calculated from measurements of a specific variable made on all possible samples of a specific size drawn from a specific population at a specific time, the average variability in which is summarized by a number called a standard error rather than a standard deviation.

sampling frame a list of all cases (members of the population) either individually or by enumeration unit.

statistic the value a variable calculated from sample data that constitutes a point estimate of a parameter.

Sample design refers to the means by which you select the primary units for data collection and analysis appropriate for a specific research question. These units may consist of states, cities, census enumeration districts, court records, cohorts, or individuals. Irrespective of the kind of unit, we always collect data at specific times and places about a specific set of cases (a *sample*) which comprises a selected subset of a larger set of cases, times, and places (a *population*). Answers to research questions thus take the form of inferences from samples to populations. A useful sample design warrants the conclusion that your inferences are both accurate and appropriately precise.

I. RESEARCH QUESTIONS CAN TARGET VARIABLES OR CULTURES

Decisions about sample design depend on your research question. Some research questions call for answers that come from the analysis of variables:

- What is the *percentage* of students at Humboldt State University in Fall 1987 who believe that we need to preserve wilderness areas?
- Has the degree to which Barbadian couples share child care and household responsibilities *changed* between 1950 and 1980?
- Does traumatic stress experienced in childhood *increase the risk* of depression in adulthood?

Other questions call for answers that come from the analysis of cases, or *ethnographic analysis*:

- Is there a pattern of behavior that constitutes a culture of drug-use, or a cultural model of what constitutes sexual behavior?

A. Sampling and Inference for Variables

We call the set of cases, times, and places from which we sample a *population* (or, universe). Populations are characterized by values we call *parameters*. Each question posed earlier identifies a specific parameter and the pertinent population. For example, the population in the first question consists of all students enrolled in Humboldt State University in the school year 1987-88; the parameter is the percentage who believe we need to preserve wilderness areas. Instead of the number of students enrolled in a university, the population might consist of the people who live in St. Johns, Antigua, or a cohort of women living in Brazil. Instead of a percentage, the parameter might be an *incidence* rate (the incidence of child abuse), an *average* (the average number of children born to women by age 50), or another univariate statistic. Answers to questions like these ordinarily come from synchronic observational studies that employ a cross-sectional design. As will become clear shortly, accurate and precise answers depend on samples of sufficient size that employ random selection criteria.

The population in the second question consists of all the couples living in Barbados between 1950 and 1990; the parameter is a measure of historical differences in the variable "the

degree to which Barbadian couples share child care and household responsibilities." Instead of the couples who live in a specific geographical region during a specific historical period, the population might consist of all countries in the world from 1960 to 1990, or court records for the last 12 months. A measure of historical differences remains the parameter of interest, but the variable that may or may not change may be infant mortality, or the proportion of drug-use charges made against people from ethnic minorities. Answers to questions like these come from diachronic studies that employ a retrospective or prospective panel or time series design. Accurate and precise answers depend on samples of sufficient size that employ random selection criteria.

The population in the third question consists, potentially, of all people at all times and places; the parameter is a function that maps "traumatic stress experienced in childhood" onto "depression in adulthood." Answers to questions like these come from a variety of both synchronic and diachronic research designs applied to specific sets of people at specific times and places. Accurate and precise answers depend on samples of sufficient size that employ random selection criteria.

B. Sampling and Inference for Ethnographic Analysis

The population in the fourth question, like the third, consists potentially of all people at all times and places; the parameter, however, is a construct that summarizes behavioral and cognitive similarities among a set of people. This shift in the meaning of the parameter means that answers to *ethnographic* questions come from the analysis of *autocorrelation* among cases.

Cultures consist of recurrent patterns of behavior rationalized by shared domain-specific schemas. They exert effects because the sensory inputs they generate constitute an environment to which people must respond. The effects of cultures come from how people respond to ecological contingencies that influence the consequences of behavior. In short, people construct the cultures in which they participate, and make them evolve, through social interaction. The socially-constructed properties of cultures means that any one person who knows about a particular culture participates with other experts in its construction and evolution. Cultures thus inescapably embody spatial and temporal autocorrelation. What one cultural participant does or tells you will correspond closely to what another cultural participant does or tells you. The errors you make in predicting what one cultural participant will do or say will correspond closely to the errors you make predicting what any other cultural participant will do or say.

This means that a random sample of people does not constitute a random sample of culture. The culture of an individual consists of configurations of cognition, emotion, and behavior that intersect in multiple ways the culture of other individuals. Hence, random samples of individuals will yield a random sample of the intersecting configurations of cognition, emotion, and behavior (i.e., the cultures) in a population. But random samples (defined by case-independence) of cultural phenomena (which necessarily contain case-dependence) cannot exist: they constitute mutually exclusive alternatives. To identify and describe cultures, ethnographic analysis aims to accurately characterize spatial and temporal autocorrelation. Like answers to the third question, answers to ethnographic questions come from a variety of both synchronic and diachronic research designs applied to specific sets of

people at specific times and places. Accurate and precise answers depend on samples designed to actively search for cultural variation that comes from specific forms of variation in life experiences. Sample size depends on the degree of similarity among cases.

II. SELECTION CRITERIA PROVIDE THE INGREDIENTS FOR SAMPLE DESIGNS

You may select cases on the basis of one or more of six criteria,

(1) availability,
(2) to fulfill a specific quota,
(3) random (or known probability) selection,
(4) case characteristics,
(5) presence in specific enumeration units, or
(6) presence along transects or at specific map coordinates.

We call all samples that utilize random (or known probability) selection *probability samples*. If you do not employ random selection, you produce one of four different forms of *non-probability samples*.

A. Non-Probability Samples

For example, if you select a predetermined number or proportion of cases with specific case characteristics, or from specific enumeration units, transects, or sets of map coordinates, you produce a *Quota Sample*. If you select cases on the basis of case characteristics to acquire specific forms of information, you produce a *Purposive (Judgment) Sample*. If you select cases simply because they will participate in your study, you produce an *Availability (Convenience) Sample*. If cases become available because one case puts you in contact with another, or other cases, you produce a *Snowball Sample*.

B. Probability Samples

We distinguish probability samples from nonprobability samples because the former exhibit known sampling distributions that warrant parameter estimation with classical statistical tests (e.g., Chi-squared, t-test, F-ratio). By convention, we identify parameters with Greek letters like β (beta), α (alpha), ε (epsilon), ϱ (rho), and σ (sigma). Samples, by contrast, yield *statistics*. By convention, we identify statistics with Latin letters and words (like b, median, percentage, mean). Each statistic constitutes a *point estimate* of a parameter – your single best guess about the value of the parameter.

Statistics constitute point estimates of parameters because samples of populations cannot perfectly replicate the properties of the populations from which they come. Every sample yields different findings, and every statistic contains three sources of error (construct, measurement, and sampling). Construct error comes from trying to measure a construct that imperfectly fits the culture or cultures found in the population studied. Measurement error comes from imperfections in the means by which we assign a value to an observation from a set of possible outcomes. To the extent to which we can rule out significant construct and measurement errors, the difference between a specific statistic and the population parameter

constitutes *sampling error* in that specific sample. Measurements of the same variable made on a large number of samples of the same size drawn from the same population exhibits a characteristic *sampling distribution* of errors around the parameter. Some statistics underestimate the parameter. Others overestimate the parameter.

Sampling errors may reflect chance or bias. Sampling errors that come from chance exhibit characteristic distributions. Many such sampling distributions (the family of t-distributions and the normal distribution) are symmetrical and are summarized by a mean of 0 and a standard deviation of 1. We call the average amount of error in a sampling distribution the standard error rather than standard deviation, to distinguish sampling distributions from the frequency distributions of the variables we study in social science research.

Although some statistics underestimate the parameter and others overestimate it, when cases are selected independently and have the same probability of inclusion in any one sample, sampling errors come solely from chance. When this condition applies, the sampling distribution of all possible statistics reveals that most statistics come very close to the parameter, and the average amount of sampling error is 0. With statistics that exhibit a normal sampling distribution, for example, 68% of all sample statistics fall within +/- 1.00 standard errors of the parameter; 95% of all sample statistics fall within +/- 1.96 standard errors of the parameter.

Small samples contain large amounts of sampling error because randomly selected extreme values exert great effects. Large samples contain small amounts of sampling error and thus estimate parameters very *precisely.* Sample precision is measured by the size of *confidence intervals. Accurate* samples yield confidence intervals that contain the parameter a given proportion (ordinarily 95%) of the time. Statistical test findings apply to samples of all sizes because they incorporate into their results the degree of sampling error contained in samples of different sizes. Confidence intervals for small samples are wider than confidence intervals for large samples, but statistics from both large and small samples estimate parameters equally accurately.

This generalization holds only for statistics that come from samples that are reasonably *unbiased. Unbiased* samples are those in which all members of the population have an equal chance of selection. The only way to reliably obtain a reasonably unbiased sample is to employ the random selection criterion.

1. Simple Random Samples

Simple Random Samples (SRS) constitute the reference standard against which all other samples are judged. The procedure for selecting a random sample takes two steps. First, make a list of all members of the population. Second, randomly select a specific number of cases from the total list. Random selection may rely on tables of pseudo-random numbers or the algorithms that generate uniform pseudo-random number distributions in statistical analysis software like SYSTAT. You may sample with or without replacing cases selected for the sample back into the population. Sampling without replacement produces unequal probabilities of case selection, but these are inconsequential except with very small populations. More importantly, even Simple Random Samples overestimate the true

standard error by the factor: $\sqrt{N/(N-n)}$. Application of the finite population multiplier, $\sqrt{(N-n)/N}$, will produce correct standard errors. The importance of this correction grows as the ratio of sample size (n) to population size (N) grows.

2. Random Systematic Samples

Random Systematic Samples (RSS) constitute a variation on SRS samples in which you substitute random selection of a starting point for random selection of all cases. For example, to select an RSS sample of 20% of a population, randomly select a number between 1 and 5, make your first case the one with the randomly selected number, and select every 5th case thereafter. To select an RSS sample of 5% of a population, randomly select a number between 1 and 20, make your first case the one with the randomly selected number, and select every 20th case thereafter.

Periodicity in a list of population members introduces significant bias into RSS samples. In the absence of periodicity, and with a known population size, to determine a sampling interval (*k*), divide the size of the population (N) by a desired sample size (n). RSS samples produce unbiased samples when *k* is an integer. The bias introduced when *k* is not an integer is inconsequential with large populations. However, if you know the size of the population, the following procedure always yields unbiased estimates:

(1) random select a number (*j*) between 1 and N;
(2) express the ratio (*j/k*) as an integer and a remainder (*m*);
(3) when m equals 0, select the case numbered (*k*) as your first sample element; when *m* does not equal 0, select the case numbered (*m*) as your first sample element.

3. Stratified, Cluster, Transect, and Case-Control Samples

All other probability samples incorporate SRS or RSS samples into the selection process. *Stratified Samples*, for example, consist of a series of simple random or random systematic samples of population sectors identified by case characteristics (e.g., age, class, gender, ethnicity) or combinations of characteristics (e.g., old and young women; old and young men). *Disproportionally stratified samples* employ a quota criterion to oversample population sectors that might otherwise be insufficiently represented in the final sample. *Cluster Samples* consist of samples in which cases are selected from simple random or random systematic samples of enumeration units that contain sets of cases, like households, hospitals, city blocks, buildings, files, file drawers, or census enumeration districts. *Probability Proportional to Size (PPS) Samples* are cluster samples in which the number of cases selected from specific enumeration units matches a quota proportional to the size of unit relative to the entire population. *Transect Samples* consist of samples in which cases or enumeration units are selected from simple random or random systematic samples of units that lie along randomly drawn transects or randomly selected map coordinates. *Case Control Samples* consist of a set of purposefully (judgmentally) identified cases, a small set of which may be selected randomly, plus a set of randomly selected controls. This sampling procedure originated in epidemiology, where cases are characterized by specific health conditions not experienced by controls. But the procedure is readily generalizable by defining cases and controls by

reference to a binary variable that distinguishes cases with a specific experience from controls without that experience.

III. HOW TO CREATE PRACTICAL, USEFUL SAMPLING DESIGNS

Practical sampling designs balance feasibility, cost, and power. What constitutes a useful sampling design varies with the properties of the parameter to be inferred and the data collection context.

A. Inferences About Variables

When your research question calls for an inference about a variable's parameter, differences between parameters, or the parametric relationship between variables, accurate and precise answers depend on samples of sufficient size that employ random selection criteria. But the primary kinds of such samples (SRS, RSS, stratified, cluster, transect, and case-control) vary dramatically in their feasibility, cost, and power for the issue at hand. For example, SRS samples cannot be drawn in the absence of a complete list of primary sampling units. Such lists commonly do not exist; it may not be possible or cost-efficient to create one. If cases with a specific experience can be distinguished from controls without that experience, a case-control sample may be selected by Simple Random Sampling. Even when SRS samples can be drawn, however, it may not be cost-efficient to search out and contact independently selected primary sampling units.

Primary sampling units almost always may be identified either within a spatially-bounded region or by enumeration units, however. When the population occupies a spatially-bounded region, samples based on transects or sets of map coordinates become efficient choices. When a comprehensive list of enumeration units can be assembled efficiently, cluster samples of one kind or another becomes both feasible and relatively cheap in time and resources. However, both cluster samples and SRS samples exhibit large standard errors, compared to stratified samples. Stratification thus makes it possible to achieve the same power with smaller sample sizes.

Power refers to your ability to precisely identify a parameter, or to detect either differences in a parameter over time or space, or the parametric influence of one variable on another, if the effect is real. Sample design determines the population to which you can validly generalize. But you will waste your time if you don't put in the effort to select a sample large enough to estimate parameters with the requisite precision.

Power varies with the risk of a Type I or α (alpha) error that you're willing to accept, sample size, and the size of the effect that you want to be able to detect. The probability of making a Type II or β (beta) error – of *not* detecting a real relationship between variables – is 1-Power. For a fixed sample and effect size, when you lower α you simultaneously raise β. When you want to rigorously avoid concluding, for example, that traumatic stress in childhood influences the risk of depression in adulthood when, in fact, it doesn't, you might set α at .01. But when you do so, you reduce your chances of seeing a relationship that's real, but weak.

Figure 1 illustrates the interdependencies between sample size, power, and effect size, when you set α at .01 (assuming the standard errors of Simple Random Samples). Figure 2 illustrates the interdependencies between sample size, β, and effect size, when you set α at .01 (assuming the standard errors of Simple Random Samples). As sample size goes up, your ability to detect a real relationship (*power*) goes up, and the chances that you'll miss it (β)goes down. But how power goes up and β goes down varies dramatically with the size of the effect. If the real shared variance between variables is about .06 (a Pearson's ***r*** of .25), you'd miss it about half the time even with a sample of 100 cases. If the real shared variance between variables is about .25 (a Pearson's ***r*** of .50), you'd only miss it about 9% of the time with a sample of only 50 cases, and you'd only miss it 1% of the time with a sample of 75 cases. By contrast, if the real shared variance between variables is about .76 (a Pearson's ***r*** of .873), you could expect to miss it only about 3% of the time even with a sample of only 10 cases, and not at all with only 15 cases. Decisions about sample size ordinarily seek to be able to detect the smallest important effect 80% of the time or better. Power analyses may be conducted readily by specialty software or by power routines that come with the major statistical software packages.

INSERT FIGURES 1 & 2 ABOUT HERE

A useful balance of feasibility, cost, and power usually comes in the form of a multistage sample design. Table 1 shows a multistage design appropriate for a study of drug- and sex-related HIV risk behaviors among new Latino migrants to the United States from Mexico and Central America. The context for such a study illustrates many of the difficulties that sample designs must resolve.

INSERT TABLE 1 ABOUT HERE

First, the size of the population is unknown, which eliminates the choice of an SRS sample.

Second, who among the population engages in drug- or sex-related HIV risk behavior is unknown, which eliminates the choice of a case-control sample.

Third, the region in which migrants live is clearly delimited, but the target population of migrants may comprise a tiny proportion of the total number of people living within the region. Migrants frequently live in locations highly dispersed among the vast majority of the region's population; many may be effectively hidden from conventional enumeration units (like households) because they move residence frequently. When these conditions apply, transect or map coordinate samples would constitute costly sample design choices. Conventional ways of thinking about cluster sampling do not apply.

Fourth, it remains possible to assemble without undue cost a list of the unconventional enumeration units that would include even those migrants who otherwise remain hidden. These units might include street locations, farms, bars, community agencies, and churches, and significant time differences for each. If different kinds of locations and times attract cases with different characteristics, the comprehensive list of ennumeration units may be usefully stratified into different kinds of units based on those characteristics.

Fifth, Random Systematic Sampling (RSS) is easy to teach and easy to apply and does not require a comprehensive list of primary sampling units (cases). When cases are distributed randomly, RSS samples exhibit the same standard errors as SRS samples. Stratification of enumeration units by case characteristics orders the cases with regard to the variables studied. With ordered cases, RSS samples exhibit lower standard errors (greater power) than SRS samples. Further stratification on the basis of ethnicity and gender may or may not be cost-efficient relative to the gain in power it would yield. In the absence of stratification, explicit measurement of internal validity confounds implements a Post-Test Only Control Group research design that substitutes measurement for random assignment. RSS samples from each kind of enumeration unit, and RSS samples of cases from each randomly selected enumeration unit complete the multistage design.

An appropriate power analysis focuses on the objective of the proposed study to test hypotheses about circumstances that increase or decrease the likelihood of engaging in specific HIV risk behaviors. Given an alpha level of .05, a two-tailed test, and the assumptions of SRS sampling, the analysis would tell us the sample size necessary to detect effects of specific independent variables 80% of the time, or the power of tests based on different sample sizes. Table 2 shows how power would vary for the study in question with variation in sample size, the ratio of the reference and response groups, and with varying effect sizes, using binary independent variables. A sample size of up to 1,204 cases would be required to detect a 50% increase in the likelihood of a given risk behavior (or a 33% reduction in the likelihood of a given risk behavior) if the ratio of reference and response groups was 20/80. However, a 500 case sample would possess good-to-excellent power to detect an odds ratio equal to or greater than 1.76 (or equal to or less than 0.57) whether the ratio of reference and response groups approximates 60/40 or 80/20. A 600 case sample does not appreciably improve the power of these analyses. An argument that we could both anticipate effects of this size and that smaller effects would not be of clinical or substantive significance at the current time – or, not worth the expense of doubling the sample size -- warrants a total sample size of approximately 500 cases.

INSERT TABLE 2 ABOUT HERE

By employing random selection criteria and sample sizes determined by a power analysis, the sample design in Table 1 allows accurate and reasonably precise estimates of parameters bearing on drug- and sex-related HIV risk behavior for a specific population of Mexican and Central American migrants to the United States. That the total population of cases came to be explicitly known only during the course of case selection and data collection does not bear on the validity of inferences from the sample to the population.

B. Inferences About Cultures

When your research question calls for an answer in the form of a construct that summarizes behavioral and cognitive similarities among a set of people, accurate and precise answers depend on samples designed to actively search for cultural variation that comes from specific forms of variation in life experiences. When generalizing about cases rather than variables, the meaning of power changes, and sample size depends on the degree of similarity among cases.

For example, in ethnographic analyses, power refers to the reliability and validity of inferences about the content of the behavioral and cognitive similarities among cases (the culture or cultures they share). Important work by Susan Weller (1987) has shown that estimates of both the reliability and validity of those inferences come from the application of the Spearman-Brown Prophesy Formula to the average level of similarity among cases. If the average level of similarity is 0.50, 9 cases will yield a reliability coefficient of .90 and a validity coefficient of .95 Only 18 cases will yield a reliability coefficient of .95 and a validity coefficient of .97. If the average level of agreement is 0.60, only 12 cases are needed for the same level of reliability and validity. As the level of similarity rises to 0.70, 0.80, and 0.90, the number of cases (sample size, n) falls to 8, 6, and 3 cases respectively. At an average level of agreement of 0.90, 3 cases yield a reliability coefficient of .96 and a validity coefficient of .99.

Sample designs for ethnographic analysis thus differ in important respects from sample designs for variable analyses. They don't require large sample sizes and they don't depend on random selection. Useful sample designs for the study of cultures stratify the population by contrasting life experiences that may produce cultural differences, employ judgmental selection of key informants and critical cases, and select other cases based on their availability, either out of availability (convenience) or through a snowball procedure. Sample size for specific strata is set by quota, depending on the average level of agreement. Efficient sample designs track levels of agreement, and expand sample sizes and change stratification criteria consistent with levels of agreement and identified cultural boundaries.

The multistage sample design in Table 1 may be usefully employed for an ethnographic study of the same population and the same topic, with important changes in procedure:

(1) the list of enumeration units (Stage I) may be assembled in the process of conducting informal and semi-structured interviews or observations -- indeed, an ethnographic component to a research design allows one to assemble such lists for a later survey with which to make inferences about variable parameters, to assess the importance of stratifying such a list, and to assess and avoid construct errors that might otherwise find their way into a study's measuring instruments;
(2) purposive (judgmental) selection should substitute for RSS selection in Stage II;
(3) selection of cases from specific enumeration units in Stage III should employ a combination of purpose (judgment), availability (convenience), and snowball criteria, rather than RSS criteria;
(4) selection of cases must include selection of the social relations of those cases; and, most important,
(5) data collection
    (a) begins with the purposeful or convenient identification of cases (and their social relations) and
    (b) initiates an iterative process that results in the construction of the multiple stages shown in Table 1.

Informal and semi-structured interviews and observations are designed to actively search for sources of cultural difference. They elicit information on the adequacy of the initial stratification criteria. Identification of people who think and act differently leads to interviews with cases selected on the basis of new stratification criteria. Because people

construct cultures and make them evolve, valid, reliable generalization is restricted to the population which exhibits those specific life experiences and to the immediate future. This makes it particularly important for ethnographic studies to explicitly measure potentially pertinent life experience variables.

Different research goals require different stratification criteria. Demeaning remarks directed at and restricted opportunities provided for members of ethnic minorities (e.g., Native Americans) by members of a dominant ethnic majority constitute two forms of traumatic stress experienced in childhood that may exhibit dramatic effects on later behavior. Table 3 shows a stratified sampling design for a retrospective study of continuity and change in the meaning of social interaction between members of majority and minority ethnic groups. People in their 60s in 2000 can tell you what they remember about native-nonnative interaction in the 1960s, when they were in their 20s. People in their 40s in 2000 can tell you what they remember about native-nonnative interaction in the 1980s, when they were in their 20s. People in their 20s in 2000 can tell you what they remember about native-nonnative interaction at that historical period.

INSERT TABLE 3 ABOUT HERE

Table 4 shows a stratified sampling design for a case-control study, a design widely applicable to outcomes evaluation research. Evaluation outcomes research tests the efficacy of interventions designed to induce specific forms of cultural change. Judgments about the efficacy of interventions require information on whether or not or the degree to which people who started with one culture ended with another. Cultural differences between participants (cases) and nonparticipants (controls) that cannot be explained by other potential internal validity confounds, like gender and age, constitutes evidence of a successful intervention.

INSERT TABLE 4 ABOUT HERE

IV. HOW TO USE SAMPLE DESIGN TO AVOID SELECTION BIAS

Selection bias alters the population to which you may validly generalize. It may make it impossible to answer your research question. Unlike other aspects of sample design, the effects of selection bias do not vary with whether your research question calls for an analysis of variables or cases (ethnographic analysis).

Non-response cases and missing data constitute important and, perhaps, the most common sources of selection bias. The severity of these sources grows as the level of non-response and missing data grows. However, solutions to these problems come from how a survey is implemented (including recruitment, training and oversight of interviewers or observers), and from the design of data collection instruments. Solutions do not come from sample designs.

Sample design contributions to selection bias come from the inclusion or exclusion of important components of the population sampled. For example, health studies that draw clinic samples miss all the cases who don't attend the clinic in question or, more generally, don't seek care during the study. Studies of entrepreneurship that exclude failed

entrepreneurs can validly generalize only to successful examples. A study that seeks to evaluate trends will require a sample design that allows at least three data points, not just "before" and "after."

Final choices on a sample design must rest on careful examination of the possibility that a specific design might exclude an important subset of cases, and how that exclusion may affect your findings, your ability to generalize to your target population – even your ability to answer your original research question.

Bibliography

Berk, R. A. (1983). An Introduction to Sample Selection Bias in Sociological Data. *American Sociological Review* **48**, 386-398.

Bernard, H.R. (2001). *Research Methods in Anthropology.* 3rd Edition. AltaMira Publications, Walnut Creek, CA.

Buckland, S.T., D.R. Anderson, K.P. Burnham, and J.L. Laake. (1994). *Distance Sampling: Estimating Abundance of Biological Populations.* Chapman & Hall, NY.

Handwerker, W. P. (2001). *Quick Ethnography.* AltaMira Publications, Walnut Creek, CA.

Johnson, J. C. (1990). *Selecting ethnographic informants.* Qualitative Research Methods Series 22. Sage Publications, Newbury Park, CA..

Kish, L. (1963). *Survey Sampling.* John Wiley, NY.

Levy, P.S. and S. Lemeshow. (1999) *Sampling of Populations.* 3rd Edition. New York: Wiley.

Weller, S.C. (1987). Shared Knowledge, Intracultural Variation, and Knowledge Aggregation. *American Behavioral Scientist* **31**, 178-193.

TABLE 1

Multistage Sampling Design for a Cross-Sectional Observational Study of Drug- and Sex-related Risk Behaviors Among New Latino Migrants

STAGE I. Comprehensive List of Enumeration Units

STAGE Ib: Stratification of Enumeration Units

STAGE II: Random Systematic Sample of Each Kind of Enumeration Unit

STAGE IIb: Stratification by Ethnicity and Gender

Ethnicity	Mexican		Central American	
Gender	Men	Women	Men	Women
# of Interviews	Contingent on power analysis	Contingent on power analysis	Contingent on power analysis	Contingent on power analysis

STAGE III: Random Systematic Sample of Primary Sampling Units

TABLE 2

Power for Logistic Regression Tests with Varying Sample Size, Ratio of Reference to Response Group, and Size of Effect (Odds Ratio) with a Binary Independent Variable

Sample Split	Odds Ratio	Power (n=400)	Power (n=500)	Power (n=600)
80/20	2.07 or 0.48	82%	89%	94%
	1.76 or 0.57	62%	71%	78%
	1.50 or 0.67	37%	44%	50%
Sample size necessary to detect ands Rat Odio of 1.50 or 0.67 at a power of 80%: 1,204				
60/40	2.07 or 0.48	94%	97%	99%
	1.76 or 0.57	78%	86%	92%
	1.50 or 0.67	50%	59%	67%
Sample size necessary to detect an Odds Ratio of 1.50 or 0.67 at a power of 80%: 809				

TABLE 3
Stratified Sampling Design for a Retrospective Study

Aged 20 in Historical Period X	1960s				1980s				2000s			
Gender	Women		Men		Women		Men		Women		Men	
Native – non Native	N	nN	N	nN	N	nN	N	nN	N	nN	N	nN
# of Interviews Contingent on Average Level of Similarity	4-18	4-18	4-18	4-18	4-18	4-18	4-18	4-18	4-18	4-18	4-18	4-18

TABLE 4
Stratified Sampling Design for a Case-Control Study

Intervention	Cases				Controls			
Gender	Women		Men		Women		Men	
Age	<20	>20	<20	>20	<20	>20	<20	>20
# of Interviews Contingent on Average Level of Similarity	4-18	4-18	4-18	4-18	4-18	4-18	4-18	4-18

Fig. 1 Relationship between Power and Sample Size for Effects of Different Sizes

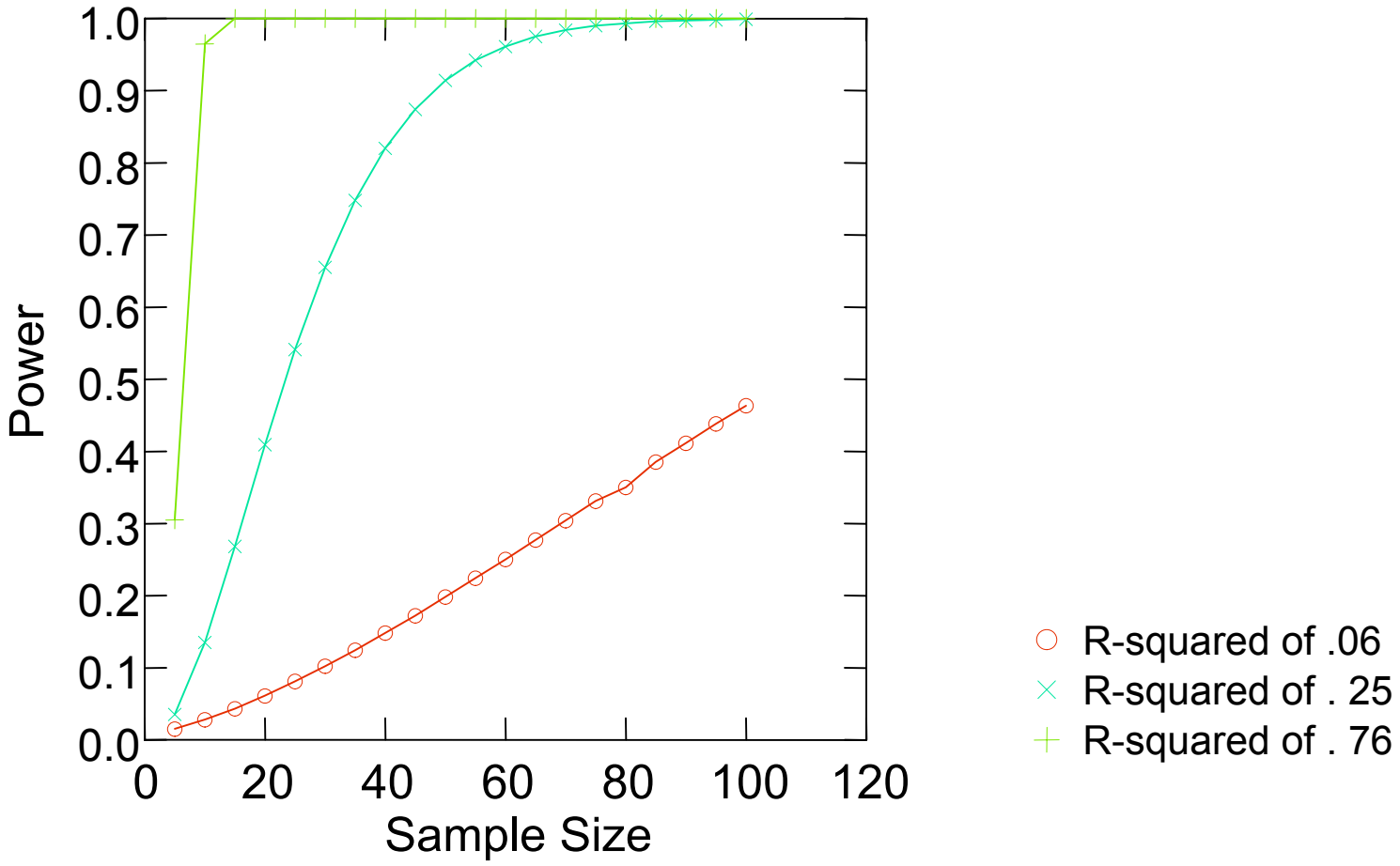

Fig. 2 Relationship between Beta and Sample Size for Effects of Different Sizes