Sample Design

W. Penn Handwerker
University of Connecticut, Storrs, Connecticut, USA

Glossary

autocorrelation Correlation of case residuals, ordinarily among cases adjacent in time or space.
case A member of a population, the primary sampling unit on which one makes measurements.
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.
enumeration (listing) unit Spatially discrete sampling unit that 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.
point estimate One's single best guess about a parameter of interest.
population The set of cases to which one generalizes sample findings.
sampling distribution The frequency distribution of all possible statistics of a given kind (e.g., means) 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 of a variable calculated from sample data that constitutes a point estimate of a parameter.

Sample design refers to the means by which one selects 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, data are always collected at specific times and places about a specific set of cases (a sample) that 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 one's inferences are both accurate and appropriately precise.

Research Questions Can Target Variables or Cultures

Decisions about sample design depend on the 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?

Sampling and Inference for Variables

The set of cases, times, and places from which we sample is a population (or universe). Populations are characterized

by values called 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 1987–1988; 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 (e.g., the incidence of child abuse), an average (e.g., the average number of children born to women by age 50), or another univariate statistic. Answers to questions such as these ordinarily come from synchronic observational studies that employ a cross-sectional design. 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 past 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 such as 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 such as 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.

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 schemes. 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 another 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 cultures 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. However, 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.

Selection Criteria Provide the Ingredients for Sample Designs

Cases can be selected on the basis of one or more of six criteria:

1. Availability
2. Fulfilling a size quota
3. Random (or known probability) selection
4. Case characteristics
5. Presence in specific enumeration units
6. Presence along transects or at specific map coordinates

All samples that utilize random (or known probability) selection are called probability samples. If one does not employ random selection, one produces one of four different forms of nonprobability samples.

Nonprobability Samples

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.

Probability Samples

Probability samples are distinguished from nonprobability samples because the former exhibit known sampling distributions that warrant parameter estimation with classical statistical tests (e.g., chi-squared, t test, and F ratio). By convention, we identify parameters with Greek letters, such as β (beta), α (alpha), ε (epsilon), ρ (rho), and σ (sigma). Samples, in contrast, yield statistics. By convention, we identify statistics with Latin letters and words (e.g., b, median, percentage, and mean). Each statistic constitutes a point estimate of a parameter, which is one's 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 derive. Every sample yields different findings, and every statistic contains three sources of error (construct, measurement, and sampling). Construct error derives from trying to measure a construct that imperfectly fits the culture or cultures found in the population studied. Measurement error derives from imperfections in the means by which a value is assigned to an observation from a set of possible outcomes. To the extent to which significant construct and measurement errors can be ruled out, 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 exhibit a characteristic sampling distribution of errors around the parameter. Some statistics underestimate the parameter, whereas others overestimate the parameter.

Sampling errors may reflect chance or bias. Sampling errors that derive 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. The average amount of error in a sampling distribution is called the standard error rather than standard deviation to distinguish sampling distributions from the frequency distributions of the variables studied 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, and 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 (usually 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 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.

Simple Random Samples

Simple random samples (SRSs) constitute the reference standard against which all other samples are judged. The procedure for selecting a random sample requires 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 such as SYSTAT. One 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 important, even SRSs 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 increases as the ratio of sample size (n) to population size (N) increases.

Random Systematic Samples

Random systematic samples (RSSs) constitute a variation on SRSs in which random selection of a starting point is substituted for random selection of all cases. For example, to select an RSS 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 fifth case thereafter. To select an RSS 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 RSSs. 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). RSSs 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. Randomly 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.

Stratified, Cluster, Transect, and Case-Control Samples

All other probability samples incorporate SRSs or RSSs 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, and ethnicity) or combinations of characteristics (e.g., old and young women, and 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 SRSs or RSSs of enumeration units that contain sets of cases, such as households, hospitals, city blocks, buildings, files, file drawers, or census enumeration districts. Probability proportional to size 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 SRSs or RSSs 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, in which cases are characterized by specific health conditions not experienced by controls. However, 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.

How to Create Practical and 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.

Inferences about Variables

When one's 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. However, the primary types 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, SRSs 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 SRS. Even when SRSs 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 can be identified either within a spatially bounded region or by enumeration units. When the population occupies a spatially bounded region, samples based on transects or sets of map coordinates are efficient choices. When a comprehensive list of enumeration units can be assembled efficiently, cluster samples of one kind or another become both feasible and relatively cheap. However, both cluster samples and SRSs 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 one's ability to precisely identify a parameter, to detect differences in a parameter over time or space, or identify the parametric influence of one variable on another, if the effect is real rather than due to chance. Sample design determines the population to which one 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 requisite precisions.

Power varies with the risk of a type I or α error that one is willing to accept, sample size, and the size of the effect that one wants to be able to detect. The probability of making a type II or β error—of not detecting a real relationship between variables—is 1-power. For a fixed sample and effect size, when α is lowered, β simultaneously rises. When one wants to rigorously avoid concluding, for example, that traumatic stress in childhood influences the risk of depression in adulthood when, in fact, it does not, one might set α at 0.01. But, how power goes up and β goes down varies dramatically with the size of the effect.

Figure 1 illustrates the interdependencies between sample size, power, and effect size, when α is 0.01 (assuming the standard errors of SRSs). Figure 2 illustrates the interdependencies between sample size, β, and effect size, when α is 0.01 (assuming the standard errors of SRSs). As sample size increases, the ability to detect a real relationship (power) increases, and the possibility that it will not be detected (β) decreases. However, the way in which power increases and β decreases varies dramatically with the size of the effect. If the real shared variance between variables is approximately 0.06 (a Pearson's r of 0.25), it would be missed approximately half the time even with a sample of 100 cases. If the real shared variance between variables is approximately 0.25 (a Pearson's r of 0.50), it would be missed only approximately 9% of the time with a sample of only 50 cases and 1% of the time with a sample of 75 cases. In contrast, if the real shared variance between variables is approximately 0.76 (a Pearson's r of 0.873), one could expect to miss it only approximately 3% of the time even with a sample of 10 cases and not at all with 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 by specialty software or by power routines that come with the major statistical software packages.

A useful balance of feasibility, cost, and power usually comes in the form of a multistage sample design. Table I shows a multistage design appropriate for a study of drug- and sex-related HIV risk behaviors among recent 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.

First, the size of the population is unknown, which eliminates the choice of an SRS. Second, it is unknown who among the population engages in drug- or sex-related HIV risk behavior, 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 (e.g., households) because they change 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.

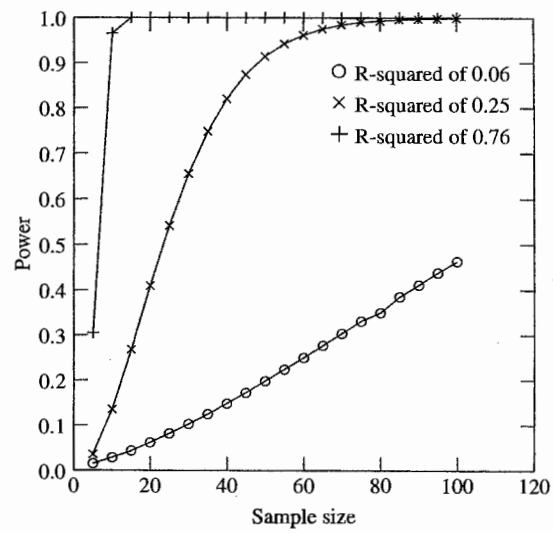

Figure 1 Relationship between power and sample size for effects of different sizes.

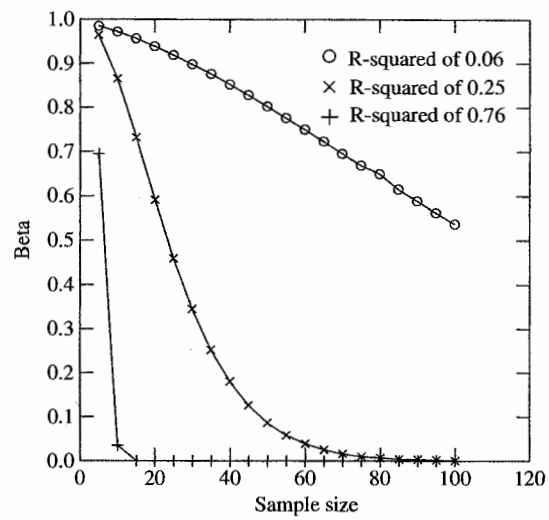

Figure 2 Relationship between beta and sample size for effects of different sizes.

Table I 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*
No. 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

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, churches, and significant time differences for each. If different types of locations and times attract cases with different characteristics, the comprehensive list of enumeration units may be usefully stratified into different kinds of units based on those characteristics.

Fifth, RSS is easy to apply and does not require a comprehensive list of primary sampling units (cases). When cases are distributed randomly, RSSs exhibit the same standard errors as SRSs. Stratification of enumeration units by case characteristics orders the cases with regard to the variables studied. With ordered cases, RSSs exhibit lower standard errors (greater power) than SRSs. 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 posttest-only control group research design that substitutes for random assignment the explicit measurement of internal validity confounds. RSSs from each kind of enumeration unit and RSSs 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 0.05, a two-tailed test, and the assumptions of SRS, the analysis would indicate 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 II shows how power would vary for the study in question with variation in sample size, the ratio of the reference and response groups, and varying effect sizes using binary independent variables. A sample size of up to 1204 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 have good to excellent power to detect an odds ratio ≥ 1.76 (or ≤ 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.

Table II 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[a]	37	44	50
60/40	2.07 or 0.48	94	97	99
	1.76 or 0.57	78	86	92
	1.50 or 0.67[b]	50	59	67

[a] Sample size necessary to detect an odds ratio of 1.50 or 0.67 at a power of 80% is 204.

[b] Sample size necessary to detect an odds ratio of 1.50 or 0.67 at a power of 80% is 809.

By employing random selection criteria and sample sizes determined by a power analysis, the sample design in Table I 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.

Inferences about Cultures

When a 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 has shown that estimates of both the reliability and the 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 0.90 and a validity coefficient of 0.95. Only 18 cases will yield a reliability coefficient of 0.95 and a validity coefficient of 0.97. 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 increases to 0.70, 0.80, and 0.90, the number of cases (sample size) declines to 8, 6, and 3 cases, respectively. At an average level of agreement of 0.90, 3 cases yield a reliability coefficient of 0.96 and a validity coefficient of 0.99.

Sample designs for ethnographic analysis thus differ in important respects from sample designs for variable analyses. They do not require large sample sizes, and they do not 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 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 I may be usefully employed for an ethnographic study of the same population and the same topic, with the following important changes in procedure:

1. The list of enumeration units (stage I) may be assembled in the process of conducting informal and semistructured 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.
5. Data collection (i) begins with the purposeful or convenient identification of cases (and their social relations) and (ii) initiates an iterative process that results in the construction of the multiple stages shown in Table I.

Informal and semistructured 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 that 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 III shows a stratified sampling design for a retrospective study of continuity and change in

Table III Stratified Sampling Design for a Retrospective Study

	Age 20 in historical period X											
	1960s				*1980s*				*2000s*			
Gender	Women		Men		Women		Men		Women		Men	
Native (N)–nonnative (nN)	N	nN	N	nN	N	nN	N	nN	N	nN	N	nN
No. 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 IV Stratified Sampling Design for a Case–Control Study

	Cases				*Controls*			
Gender	Women		Men		Women		Men	
Age	<20	>20	<20	>20	<20	>20	<20	>20
No. of interviews contingent on average level of similarity	4–18	4–18	4–18	4–18	4–18	4–18	4–18	4–18

the meaning of social interaction between members of majority and minority ethnic groups. People in their 60s in 2000 can tell what they remember about native–nonnative interaction in the 1960s, when they were in their 20s. People in their 40s in 2000 can tell what they remember about native–nonnative interaction in the 1980s, when they were in their 20s. People in their 20s in 2000 can tell what they remember about native–nonnative interaction during that historical period.

Table IV 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 and 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, such as gender and age, constitute evidence of a successful intervention.

How to Use Sample Design to Avoid Selection Bias

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

Nonresponse cases and missing data constitute important and, perhaps, the most common sources of selection bias. The severity of these sources increases as the level of nonresponse and missing data increases. 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 on clinic samples miss all the cases who do not attend the clinic in question or, more generally, do not 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 be based on careful examination of the possibility that a specific design might exclude an important subset of cases and how that exclusion may affect the findings, one's ability to generalize to one's target population, and even one's ability to answer the original research question.

See Also the Following Articles

Clustering • Confidence Intervals • Data Distribution and Cataloging • Ethnography • Randomization • Spatial Autocorrelation

Further Reading

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *Am. Soc. Rev.* **48,** 386–398.

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

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* 2nd Ed. Erlbaum, Hillsdale, NJ.

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

Kish, L. (1963). *Survey Sampling.* Wiley, New York.

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

Schulz, K. F., and Grimes, D. A. (2002). Case-control studies: Research in reverse. *Lancet* **359,** 431–434.

Sudman, S. (1976). *Applied Sampling.* Academic Press, New York.

Weller, S. C. (1987). Shared knowledge, intracultural variation, and knowledge aggregation. *Am. Behav. Sci.* **31,** 178–193.