SAMPLING

By G. David Garson

North Carolina State University
School of Public And International Affairs

Table of Contents

Contact:

G. David Garson, President
Statistical Publishing Associates
274 Glenn Drive
Asheboro, NC 27205 USA

Email: gdavidgarson@gmail.com
Web: www.statisticalassociates.com

Overview

Sampling in conjunction with survey research is one of the most popular approaches to data collection in the social sciences. Sampling is choosing which subjects to measure in a research project. Usually the subjects are people, but subjects could also be objects, organizations, cities, or even nations. Sampling is associated with survey instruments but is independent of the specific measurement mode, which could also observation, lab instrumentation, audio or video recording, archival record, or other methods. Regardless, sampling will determine how much and how well the researcher may generalize the study's findings. A poor sample may well render findings meaningless.

Random sampling is the foundation assumption for much of inferential statistics and significance testing. Significance testing does not apply to enumerations, in which every case in the population of interest is included in the study. Most national and other large samples by governmental agencies and polling firms use complex sampling designs, such as multistage or stratified sampling. These complex designs mean that significance tests must be computed differently. That is, most significance tests in statistical software are somewhat inaccurate when applied to complex samples. Specialized software, discussed in the section on complex samples, should be used to obtain the highest level of accuracy.

See also the separate volume of the Statistical Associates "Blue Book" series dealing with survey research.

Key Concepts and Terms

Population

The population, also called the universe, is the set of people or entities to which findings are to be generalized. The population must be defined explicitly before a sample is taken. Care must be taken not to generalize beyond the population. Doing so is a common error in statistical writing.

Random sampling

Random sampling is data collection in which every person in the population has a chance of being selected which is known in advance. Normally this is an equal chance of being selected. If data are a random sample, the researcher must report not only the magnitude of relationships uncovered but also their significance level (the chance the findings are due to the chance of sampling). There are various types of random sampling, discussed below.

The sampling frame

The sampling frame is the list of ultimate sampling entities, which may be people, households, organizations, or other units of analysis. For example the list of registered students may be the sampling frame for a survey of the student body at a university. In multi-stage sampling, discussed below, there will be one sampling frame per stage (ex., in a study in which individuals are sampled from residences, residences sample from Census blocks, Census block sampled from Census tracts, Census tracts sampled from the state, and states sampled from the nation, there are five frames: the lists of individual residents, of residences, of Census blocks, of Census tracts, and of states.

The target population to which the researcher wishes to generalize is ideally the same as the sampling frame. In reality, the target population may be larger or smaller than the available sampling frame, in which case the researcher must address questions about the representativeness of the sampling frame. IF the sampling frame is smaller than the target population, there is undercoverage. If the sampling frame is larger than the target population, there is overcoverage. Either can involve sampling frame bias. Telephone directories are often used as sampling frames, for instance, but tend to under-represent the poor (who have fewer or no phones) and the wealthy (who have unlisted numbers). Random digit dialing (RDD) reaches unlisted numbers but not those with no phones, while overrepresenting households owning multiple phones.

Strata

Strata are grouping variables used in sampling. In "area sampling" the groups are geographic, such as Census blocks, Census tracts, metropolitan areas, and counties. A *primary sampling unit (PSU)* is the level of aggregation immediately

above the individual level, such as Census blocks for a sample of residents. However, strata may be non-geographic, such as gender or racial groups. Stratified sampling, discussed below, uses information on strata to increase the precision of the sample.

Significance

Significance is the percent chance that a relationship found in the data is just due to an unlucky sample, and if the researcher took another sample, no relationship might be found. That is, significance is the chance of a *Type I* error: the chance of concluding there is a relationship when there is not. Social scientists often use the .05 level as a cutoff: if there is 5% or less chance that a relationship is just due to chance, it is concluded that the relationship is real. More precisely, the researcher fails to accept the null hypothesis that the strength of the relationship is not different from zero.

Significance testing is not appropriate for non-random samples or for enumerations/censuses (discussed below). Researchers would like to make similar inferences for non-random samples, but that is impossible.

Confidence intervals

Confidence intervals are the "margin of error" commonly reported for opinion polls. Confidence intervals are directly related to coefficients of significance. For a given variable in a given sample, one could compute the standard error, which (assuming random sampling, a normal distribution, and assuming a .05 significance level cutoff for rejecting the null hypothesis) has a 95% confidence interval of approximately plus or minus 1.96 times the standard error. If a very large number of samples were taken, and a (possibly different) estimated mean and corresponding 95% confidence interval were constructed from each sample, then 95% of these confidence intervals would contain the true population value. The formula for calculating the confidence interval, significance levels, and standard errors, depends on the type of random sampling, whether simple or complex, discussed below.

Enumerations

Also called censuses, enumerations are collections of data from every person or entity in the population. If data are an enumeration, the researcher does not have to be concerned with significance, which is a form of estimate of sampling error. That is, if the significance level is .05, this means there is a 5% chance of getting an outcome as strong or stronger than the observed result if another random sample were taken. With enumerations, there is no possibility of "another random sample," so significance testing does not apply. Any relationship, no matter how small, is a true relationship (barring measurement error) for an enumeration. While some researchers nonetheless report significance levels for enumeration data, significance coefficients conflate sample size and strength of relationship and thus serve poorly when taken as measures of effect size.

The design effect

The design effect, D, is a coefficient which reflects how sampling design affects the computation of significance levels compared to simple random sampling. A design effect coefficient of 1.0 means the sampling design is equivalent to simple random sampling. A design effect greater than 1.0 means the sampling design reduces precision of estimate compared to simple random sampling (cluster sampling, for instance, reduces precision). A design effect less than 1.0 means the sampling design increases precision compared to simple random sampling (stratified sampling, for instance, increases precision). Unfortunately, most computer programs generate significance coefficients and confidence intervals based on the assumption of formulas for simple random sampling. Formulas for all types are found, for example, in Kalton (1983).

Types of non-random sampling

Overview

Non-random sampling is widely used as a case selection method in qualitative research, or for quantitative studies of an exploratory nature where random sampling is too costly, or where it is the only feasible alternative. Non-random samples are often “convenience samples,” using subjects at hand. Non-random

samples, like random samples, also raise the issue of whether the findings are merely an artifact of the chance of sampling or not, but the rules of statistical significance testing do not apply. That is, there is no statistical way to assess the validity of results of non-random samples.

Convenience sampling

Also called availability sampling or haphazard sampling, convenience samply occurs when the researcher selects subjects on the basis of availability. Examples include interviewing people who emerge from an event or location, interviewing a captive audience such as one's students, or mail-in surveys printed in magazines and newspapers.

Quota sampling

Quota sampling is availability sampling, but with the constraint that proportionality by strata be preserved. Thus the interviewer(s) will be told to interview so many white male Protestants, so many Hispanic female Catholics, and so on, with proportions determined based on the Census, to improve the representativeness of the sample.

Maximum variation sampling is a variant of quota sampling in which the researcher purposely and non-randomly tries to select a set of cases which exhibit maximal differences on variables of interest. Further variations include *extreme or deviant case sampling* or *typical case sampling*.

Expert sampling

Also called *judgment sampling*, expert sampling occurs when the researcher interviews a panel of individuals known to be expert in a field. Expertise in this context is defined as any special knowledge, not necessarily formal training. Depending on the topic of study, experts may be policy issue academics or devotees to a popular culture fad.

Chain referral sampling , also called *snowball sampling* or *network sampling*, is one method of selecting experts for expert sampling. The researcher starts with a subject who displays qualities of interest (ex., being a private militia member), then obtains referred subjects from the first subject, then additional referred

subjects from the second set, and so on. (Note chain referral sampling is different from chain sampling in quality control applications. Chain sampling in that sense refers to basing acceptance of a given production lot based not only on a sample of that lot, but also on samples of the previous 10 or so lots.)

Critical case sampling is a variant of expert sampling, in which the sample is a set of cases or individuals identified by experts as being particularly significant (ex., award winners in a given field).

Types of random sampling

Simple random sampling

Simple random sampling occurs when the researcher has a sampling frame (list) which approximates listing all members of the population, then draws from that list using a random number generator or possibly takes every nth subject (called interval sampling). Conventional significance testing is appropriate for simple random samples.

Simple random sampling is common when the sampling frame is small. Starting at an arbitrary point in a table of random numbers, such as found online (ex., www.random.org) or in most statistics books, sets of random numbers are read and associated with members of the sampling frame. The first three digits might be 712, for instance, and thus the 712th person in the sampling frame might be selected. If the sampling frame had over 999 people, then four-digit random sequences would be used, and so on. Selections outside the range of members of the frame would be ignored. Many statistical packages, such as SPSS, provide a built-in pseudo-random number generator, allowing researchers to request generation of a given number n random digits between 1 and N, where n is sample size and N is population size.

Simple random sampling with replacement

Simple random sampling with replacement (SWSWR) occurs when selections are replaced back into the sampling frame such that repeat selections are possible.

For a large sample, such repeats are rare and SWSWR becomes equivalent to simple random sampling without replacement, discussed below.

Simple random sampling without replacement

Simple random sampling without replacement (SRSWOR) does not allow the same random selection to be made more than once. Confidence intervals are slightly (usually trivially) smaller for SRSWOR samples compared to simple random samples. Most computer programs use formulas which assume SRSWOR sampling.

Equal probability systematic sampling

Equal probability systematic sampling also involves the direct selection of subjects or other primary sampling units from the sampling frame. The researcher starts at a random point and selects every nth subject in the sampling frame. The random starting point equals the sampling interval, n, times a random number between 0 and 1, plus 1, rounded down. In systematic sampling there is the danger of *order bias*: the sampling frame list may arrange subjects in a pattern, and if the periodicity of systematic sampling matches the periodicity of that pattern, the result may be the systematic over- or under-representation of some stratum of the population. For instance, if street numbers are used, in some cities each block starts with a multiple of 100 and so houses with addresses of 100, 200, 300, etc., will be corner lots, which on average have more expensive houses, thus introducing bias. If, however, it can be assumed that the sampling frame list is randomly ordered, systematic sampling is mathematically equivalent to and equally precise as simple random sampling. If the list is stratified (ex., all females listed, then all males), systematic sampling is mathematically equivalent to stratified sampling and is more precise than simple random sampling.

Repeated systematic sampling

Repeated systematic sampling is a variant which seeks to avoid the possibility of systematic biases due to periodicity in the sampling frame. This is done by taking several smaller systematic samples, each with a different random starting point, rather than using one pass through the data as in ordinary systematic sampling. Repeated systematic sample has the side benefit that the variability in the

subsample means for a given variable is a measure of the variance of that estimate in the entire sample.

Stratified simple random sampling

Stratified simple random sampling is simple random sampling of individuals from each stratum of the population. For instance, in a study of college students, a simple random sample may be drawn from each class (freshman, sophomore, junior, senior) in proportion to class size. This guarantees the resulting sample will be proportionate to known sizes in the population. One may simultaneously stratify for additional variables, such as gender, with separate simple random samples of freshman women, freshmen men, sophomore women, etc. The finer the stratification, the more precision compared to unstratified simple random sampling. That is, confidence intervals will be narrower for stratified sampling than for simple random sampling of the same population. The more heterogenous the means of the strata are on a variable of interest, the more stratified sampling will provide a gain in precision compared to simple random sampling. Stratified sampling, therefore, is preferred to simple random sampling.

Multistage stratified random sampling

Multistage stratified random sampling occurs when the researcher draws simple random samples from successively more homogenous groups ("strata") until the individual subject level is reached. For instance, the researcher may stratify by industry groups, then sample companies within industries constraining the sample to stratified proportions; then randomly sample individual workers within companies. The purpose of multistage stratified random sampling is to increase research precision by ensuring that key populations of subjects are represented in the sample (ex., people in certain industry categories). The greater the heterogeneity of the strata and the finer the stratification (that is, the smaller the strata involved) depending on the topic of study., the more the precision of the results. At each stage, stratified sampling is used to further increase precision. Because the variance of individuals from their group mean in each strata is less than the population variance, standard errors are reduced. This means conventional significance tests, based on population variances, will be too conservative – there will be too many Type I errors, where the researcher wrongly accepts the null hypothesis.

Simple vs. multistage sampling

In simple stratified sampling, the primary sampling unit may be the Census block but there may be higher sampling units, such as metropolitan areas and counties, but the final sample of individuals will be proportionate to all strata (Census block, metropolitan area, county) and every combination of strata will have some representation in the sample. In multistage sampling, where counties are sampled within states, then metropolitan areas within counties, then Census blocks within metropolitan areas, then individuals within Census blocks, the final sample will be proportionate to all strata only if each of the hierarchical levels prior to the ultimate level is sampled according to the number of ultimate units (ex., individuals) it contains (this is called *probability proportional to size sampling* or PPS sampling) and not every combination of strata will be represented.

Where stratification reduces standard error, multistage sampling increases it. That is, there is an accumulation of chances of sampling error at each stage. Conventional significance tests will be too liberal – there will be too many Type II errors, where the researcher wrongly sustains an hypothesized relationship.

Overall, multi-stage or cluster sampling is usually less precise than simple random sampling, which in turn is less precise than one-stage stratified sampling. Since multistage sampling is the most prevalent form for large, national surveys, and since most computer programs use standard error algorithms based on the assumption of simple random samples, the standard errors reported in the literature often underestimate sampling error, leading to too many Type II errors (false positives). See the discussion below regarding estimation and software for complex samples.

Disproportionate stratified sampling

Disproportionate stratified sampling occurs when disproportionate numbers of subjects are drawn from some strata compared to others. A typical use of disproportionate stratified sampling is oversampling of certain subpopulations (ex., African Americans) to allow separate statistical analysis of adequate precision. When significance and confidence levels are computed for the entire sample under disproportionate stratified sampling, cases must be weighted to restore the original proportions. This weighting process generally reduces the precision benefits of stratified sampling. That is, disproportionate stratified

samples tend to be less precise than proportionate stratified samples of the same population. Standard error estimates based on disproportionate stratified samples may be either more or less precise than those based on simple random samples, whereas those based on proportionate stratified samples are more precise. The more heterogeneous the means of the strata are on a variable of interest, the more the positive effect on precision of stratification compared to the negative effect of weighting, and the more likely the disproportionate stratified sample will be more precise than one based on a simple random sample.

Clustering and multilevel effects in multistage samples

Single stage cluster sampling is a synonym for stratified sampling. It occurs when the researcher draws simple random samples of individuals from certain aggregational units of interest (ex., from certain census block groups). *Multistage cluster sampling* is a synonym for multistage stratified sampling. When the strata are geographic units, this method is sometimes called *area sampling*. Note that some researchers reserve the term "cluster sampling" to refer to a two-stage process in which clusters are sampled, then individuals are sampled within the chosen clusters, and they reserve the term "multistage cluster sampling" for three-stage processes and higher.

Clustering, which is inherent in multistage sampling, often results in correlated observations, which in turn violates the assumption of independently sampled cases - an assumption of many statistical techniques. While it is common practice to treat data from multistage sampling as if it were randomly sampled data, this should be done only if intraclass correlation (ICC) is non-significant, meaning the grouping or clustering variable is independent of the dependent variable. I ICC is significant, some form of multilevel modeling (also known as linear mixed modeling or hierarchical linear modeling) is called for. Multilevel modeling is discussed in a separate Statistical Associates "Blue Book" volume.

Dealing with sampling problems

Dealing with mismatched sampling frames

Often, the sampling frame does not match the primary sampling unit. For instance, the sampling frame may be a list of residences, but the ultimate sampling units are individuals. The interviewer needs instruction on just which individual within a residence to select as a subject. This is handled by a *selection grid* (see Kish, 1965). The selection grid may be of a variety of forms. For instance, in a survey of taxpayers, for simplicity assume that a residence can have a maximum of three taxpayers. Let the researcher number the taxpayers from 1 to 3 by age and use the selection grid below:

		If no. of taxpayers in residence is		
		1	2	3
Survey form	% of all forms	Interview taxpayer #		
A	1/3	1	1	1
B	1/6	1	1	2
C	1/6	1	2	2
D	1/3	1	2	3

In this example, there are four different forms of the survey (A, B, C, D), printed and randomly distributed in the proportions in the table above. Each survey form corresponds to one of the rows in the table. For instance, in form B, if there were two taxpayers in the residence then the interviewer would interview taxpayer number 1. For forms C or D, taxpayer number 2 would be interviewed. If there is 1 taxpayer in the residence, that taxpayer (#1) gets interviewed regardless of form. If there are two taxpayers, each has an equal chance of being interviewed, and the same for three taxpayers.

Kalton (1983: 61) presents a similar table for the assumption of a maximum of four selectable subjects per residence. Analogous tables can be constructed for

any assumption. All such selection grid tables equalize the probability of any appropriate individual being chosen for inclusion in the sample.

A similar, simpler approach is the "last birthday" method, whereby the researcher asks the number of adults in the household, then interviews the sole adult in one-adult households; every other time interviews the adult with the most recent birthday in two-adult households, and every other time the other adult; every third time in three-adult households, etc. Asking about birthdays rather than ages may be less sensitive. Some evidence suggests that in telephone surveys, these methods may require additional screening time and interview time, and may generate too many callbacks or refusals. In face-to-face interviews there is greater subject tolerance for such screening questions.

Increasing the response rate

Response rates for face-to-face interviews are typically in the 75% range, and for mail surveys 10% is considered good in some marketing surveys. However, "good" depends of purpose and resources of the polling organization. In the federal government in the U.S., Office of Management and Budget minimal standards require a 75% expected response rate, and federal data collected with response rates under 50% should not be used except if a special OMB waiver is granted. The OMB guidelines point up the salient point, that high response rate, like random sampling, is essential to reliable statistical inference. Response rate can be increased by

- Having legitimating sponsorship for the survey, geared to sponsors who are highly-regarded in the community being surveyed
- Having a good, short explanation justifying the survey
- Notifying individuals in advance that a survey is coming
- Keeping the survey instrument short, and letting prospective subjects know this.
- Assure confidentiality and anonymity.
- Offer to reschedule to a time at the subject's convenience.
- Make call-backs (four is typical) where needed. In mail surveys, provide a new copy of the instrument each time.
- Start the instrument with non-threatening questions which arouse interest.
- Offer to provide final results later on.

- In mail surveys, use certified and express mail. In e-mail surveys, use priority e-mail. Postcard and e-mail reminders also help. In mail surveys, plan to wait about six weeks for all responses to drift in.
- Offer token remuneration (ex., $1 enclosed in mail surveys; $5 in face-to-face interviews; or use pens, keychains, calendars, or other tokens to make respondents feel like they are getting something in return). Linking the return of the survey to entrance into a lottery for a substantial prize (ex., $1,000) can also boost response rates.

Analyzing non-response

The researcher must assume that non-respondents differ significantly from respondents with regard to variables of interest. There are three general strategies for assessing the effect of non-response. These strategies are not mutually exclusive.

Population comparison

Survey averages can be compared with known population averages, particularly feasible for demographic variables such as gender, age distribution, income, occupation, and other variables covered by the Census. In educational studies, population data may be available on the population distribution of gender, age, class, and test performance. The researcher seeks to identify variables where the sample mean deviates from the population mean, and speculates (preferably on the basis of prior literature) on the impact of such bias on the dependent variables of interest.

Intensive postsampling

An intensive effort may be made to interview a sample of non-respondents, using the postsample to assess nonresponse. Adding the postsample to the sample also reduces the bias of the sample from that estimated by this method, which is costly and therefore rare. If non-respondents are sampled, the researcher may also wish to ask instrumentation assessment questions to probe why non-response occurred in the initial sample, thus throwing light on the nature of non-response. If the non-respondent post-sample is by a different mode (ex., telephone vs. initial web survey), the researcher may wish to ask question about the initial modality (ex., does the non-responder have Internet access? an email

address? check his/her email regularly? have problems loading the survey web page? etc.)

Wave extrapolation

The researcher codes the initial response set and each of the four call-back response sets, computes the mean of key variables on each of the five sets, and then sees if consistent extrapolation is possible for any variable. For instance, each successive set might show fewer women respondents, suggesting the non-response set would be expected to have an even lower proportion of women yet. This method carries little marginal expense and therefore is highly recommended, particularly in supplementation to population comparison.

Imputing responses

Not all respondents answer all items in a questionnaire. Andrew Gelman, Gary King, and Chuanhai Liu (1998) have presented a method of analyzing a series of independent cross-sectional surveys in which some questions are not answered in some surveys and some respondents do not answer some of the questions posed. The method is also applicable to a single survey in which different questions are asked, or different sampling methods used, in different strata or clusters. The method involves multiply-imputing the missing items and questions by adding to existing methods of imputation designed for single surveys a hierarchical regression model that allows covariates at the individual and survey levels. Information from survey weights is exploited by including in the analysis the variables on which the weights were based, and then re-weighting individual responses (observed and imputed) to estimate population quantities. They also develop diagnostics for checking the fit of the imputation model based on comparing imputed to non-imputed data. They illustrate with the example that motivated this project --- a study of pre-election public opinion polls, in which not all the questions of interest are asked in all the surveys, so that it is infeasible to impute each survey separately. See Gelman, King, and Liu (1998).

Weighting

There are a variety of reasons for weighting and, indeed, weighting can be a multi-step process as the researcher sequentially adjusts for probability of selection, non-response, stratification, and other purposes. Some researchers

object to weighting on the grounds that it requires the artificial replication of cases, usually makes little or no difference to conclusions, and may convey a false confidence in the data while forestalling candid discussion of the biases of the sample. Supporters of weighting argue that weighting is a reasonable form of approximation to adjust an existing sample for known biases, and such correction is better than the alternative of no correction. The WEIGHT parameter in SAS and the WEIGHT BY command in SPSS are used for weighting, as discussed in their respective manuals.

Weighting for non-response. Sometimes it is possible to compare respondents and nonrespondents on the basis of gender or some other attribute. For instance, in a random sample of a city, census data gives a good estimate of the true proportion of males and females. In a mail survey, the true proportion of males and females may be estimated from first names. If in such situations one finds the observed distribution does not conform to the true population, one may wish to weight responses to adjust accordingly. For instance, if too few women are in the respondent pool, one might wish to weight their responses more than the male responses. For instance, if the true proportion by gender is 50-50, and if one got 40 females and 60 males, then one could weight each female response by 1.5. This, in effect, gives 60 females and 60 males. However, to avoid artificially increasing sample size from 100 to 120, one needs further weighting to scale back to 100. This could be achieved by further weighting both females and males by 5/6. This logic can be extended to the case of item non-responses for subjects who are in the respondent pool. Reliable non-response adjustment requires a high response rate (ex., > 70%).

Bourque and Clark (1992: 60) state, "It has been our experience that the use of weights does not substantially change estimates of the sample mean unless nonrespondents are appreciably different from respondents and there is a substantial proportion of nonrespondents."

Weighting for post-stratification adjustment. The same logic and same weighting strategy applies if the underrepresentation of a given strata (ex., black males age 18-25) is due to non-response or due to disproportionate stratified sampling. Either way the objective is to weight the existing cases in a way which increases the representation in the adjusted sample of the strata that are underrepresented in the raw data.

Weighting to account for the probability of selection. In a random sample of people, each individual should have an equal chance of being selected. However, the reality of the sampling procedure may be such that each household, not individual, has an equal chance. Individuals within households with more people have a lower chance of being selected: they are underrepresented and should be weighted more. A weighting adjustment can be made if the researcher thought to have an item asking for the number of eligible people in the household. Summing the responses to this item and dividing by the number of households surveyed, assuming one survey per household, gives the average number of individuals per household, say 2.5. The weight for any surveyed individual in the sample is then the number of people in that household divided by this average. For instance, if a given household had 5 eligible individuals, the weight for that case would be 5/2.5 = 2.

Dealing with missing data

Non-response to individual survey items raises the question of whether to delete cases with missing data or to interpolate responses. Specialized software exists to aid in dealing with this problem. an example is the “Missing Values Analysis” module which can be added to base SPSS. This topic is dealt with in the separate Statistical Associates "Blue Book" volume on analysis of missing data

Pre-survey estimation of sample size

Sample size often needs to be estimated in the design phase of research. The size of the sample will need to be larger if one or more of the following applies:

- the weaker the relationships to be detected
- the more stringent the significance level to be applied
- the more control variables one will use
- the smaller the number of cases in the smallest class of any variable
- the greater the variance of one's variables.

Combinations of these factors create complexities which, in combination with lack of knowledge about the population to be sampled, usually make sample size estimation just that -- an arbitrary estimate. Note that needed sample size does not depend at all on the size of the population to be sampled. Even in the most complex analyses, samples over 1,500 are very rarely needed. Specialized

software, such as the “SamplePower” module which can be added to base SPSS, exists to help the researcher calculate needed sample size. This topic is dealt with in the separate Statistical Associates "Blue Book" volume on power analysis. Other such software as well as online sample size calculators may be found on the web. There are also rules-of-thumb and various manual methods.

Assumptions

Significance testing is only appropriate for random samples

Random sampling is assumed for inferential statistics (significance testing). "Inferential" refers to the fact that conclusions are drawn about relationships in the data based on inference from knowledge of the sampling distribution. Significance tests are based on a sampling theory which requires that every case have a chance of being selected known in advance of sample selection, usually an equal chance. Statistical inference assesses the significance of estimates made using random samples. For enumerations and censuses, such inference is not needed since estimates are exact. Sampling error is irrelevant and therefore inferential statistics dealing with sampling error are irrelevant. Significance tests are sometimes applied arbitrarily to non-random samples but there is no existing method of assessing the validity of such estimates, though analysis of non-response may shed some light. The following is typical of a disclaimer footnote in research based on a nonrandom sample:

"Because some authors (ex., Oakes, 1986) note the use of inferential statistics is warranted for nonprobability samples if the sample seems to represent the population, and in deference to the widespread social science practice of reporting significance levels for nonprobability samples as a convenient if arbitrary assessment criterion, significance levels have been reported in the tables included in this article." See Michael Oakes (1986). *Statistical inference: A commentary for social and behavioral sciences.* NY: Wiley.

Frequently Asked Questions

How can one estimate sample size in advance?

This is primarily discussed in the separate Statistical Associates "Blue Book" volume on power analysis. However, as mentioned above, sample size calculation depends on a number of complex factors. In practice, researchers use specialized software designed for these calculations. One rule of thumb is based on standard error as used in normal curve tests: Sample size = ss = $s*z/T)^2$, where s is the standard error of the variable with the largest variance (perhaps estimated in a pretest sample), z is the number of standard units corresponding to the desired proportion of cases (z = 1.96 for two-tailed tests at the .05 significance level), and T is the tolerated variation in the sample. For instance, this formula might be used to compute the necessary sample size such that the variable with the largest standard deviation will have a sample mean within t=2 years of the true population mean, with .05 significance.

The chi-square method is one possibility when estimating required sample size:

1. *Determine desired significance and difference levels.* The researcher must first select the desired level of significance (typically .05) and the smallest difference he or she wishes to be detected as significant. For instance, in a study of gender and presidential support, one might want a 10% gender difference to be found significant at the .05 level.
2. *Specifying expected and least-difference tables*. Researchers then must create two tables. This requires estimating the marginal frequencies (the number of men and women, and of presidential supporters and non-supporters, for example). Expected cell frequencies are then calculated as usual for chi-square. Then the researcher creates a least difference table as, for example, placing 10% more cases than expected on the diagonal (ex., on the male-non-supporters, female-supporters diagonal).
3. *Solving for n*. The researcher then solves for n, sample size, using the chi-square formula, which is chi-square = SUM{(Observed - Expected)^2/Expected}. For instance, in a 2-by-2 table, let the upper-left cell be .20n, the upper right .30n, the lower-left .20n, and the lower-right .30n. Let the least-difference cells be .25n, .25n, .15n, and .35n respectively. Degrees of freedom is (rows-1)*(columns-1), which is 1 for a 2-by-2 table.

For 1 degree of freedom, at the .05 significance level, the critical value of chi-square is 3.841. Therefore chi-square = 3.841 - SUM{[(.25n-.20n)^2/.20n] + [(.25n-.30n)^2/.30n] + [(.15n-.20n)^2/.20n] + [(.35n-.30n)^2/.30n]. Solving for n, n = 3.841/.0416 = 92.3. Therefore, a sample size of 93 is the minimum sample size needed to detect a 10% difference at the .05 significance level, by chi-square.

However, as mentioned, these are simple rule-of-thumb methods and more satisfactory estimates require taking more complex factors into account.

How do I know if the distribution of my responses is what it should be, based on known population distributions?

The chi-square test of goodness of fit test addresses this question. This is discussed in the separate Statistical Associates "Blue Book" volume on significance testing, with a spreadsheet example.

What is adaptive sampling?

Adaptive sampling refers to strategies which change the sample during the process of sampling based on responses or observations to date. It is used particularly when that target is rare. For instance, in a study of immigrants, if a survey response revealed a rarity such as being a Vietnamese refugee airlifted out of Vietnam at the end of the Vietnamese War, the survey would add other residents of the neighborhood to the survey in an attempt to gain a larger sample and thus greater precision for that subgroup. Adaptive methods are discussed by Thompson (2012).

How is standard error computed for dichotomous responses?

For yes-no and other dichotomous items, where the object is to arrive at percentages which are within plus or minus a certain percentage (ex., 5%) of the unknown population percent, the formula is [P_y*P_n]/se, where se is the standard error term. The "se" term is estimated as (siglevel/z)-squared, where siglevel is our chosen significance level (.05) and z is the corresponding number of standard units in the normal curve (1.96 for the .05) level. Also P_y and P_n are the percentages estimated to answer yes and no on the dichotomy. If we lack any

knowledge at all, we estimate these percentages as both being .5. The formula then becomes .25/.0006507, or 384. That is, if we sample 384 persons, we can be confident that the percentages we arrive at on a dichotomous item will be within 5% of the true but unknown population percentage.

How do significance tests need to be adjusted for multi-stage, cluster, or other complex sampling designs, as compared to simple random samples?

This is a central focus of the reference by Lee, Forthofer, and Lorimar (1989). These authors set forth a number of different strategies for variance estimation for complex samples, including *replicated sampling*: Multiple independent subsamples are drawn using an identical sampling design such as repeated systematic sampling (discussed above). The sampling variance of the total sample is estimated from the variability of the subsample estimates. Let U be a parameter such as the response to a survey item. Each of t subsamples will generate an estimate of the mean of U, designated u_1, u_2, ..., u_t. One can also pool all the subsamples to get the sample mean, ü. One can then enter these values into a formula which gives the sampling variance of ü, whose square root is interpreted as an estimate of the standard error:

$$v(\ddot{u}) = SUM\ (u_i - \ddot{u})^2/t(t-1)$$

For descriptive statistics a minimum of 4 and a recommended number of 10 subsamples should be drawn. For multivariate analysis, the number of subsamples should be 20 or more. Note that the same logic and formula apply if U is a percentage, an odds ratio, an intercept, a regression coefficient, or any other parameter being estimated.

SPSS distributes an add-on module called *Complex Samples* for purposes of significance testing with multistage and other complex samples. In turn, it has the following sub-modules:

- Complex Samples Descriptives (CSDESCRIPTIVES)—
- Complex Sample Tabulate (CSTABULATE)—
- Complex Samples General Linear Models (CSGLM)—
- Complex Ordinals Selection (CSORDINAL)—

- Complex Samples Logistic Regression (CSLOGISTIC)—
- Complex Samples Cox Regression (CSCOXREG)

Other software handling estimation and variance estimation under stratified and other unequal probability sampling methods includes *SUDAAN* (from the Research Triangle Institute) and *VPLX* (from the Bureau of the Census).

Lee, Forthofer, and Lorimar (1989) also discuss three other methods of variance estimation: (1) *balanced repeated replication*, used primarily in paired selection designs; (2) *repeated replication through jackknifing*, which is based on pseudo-replication involving serially leaving cases out in successive subsamples ; and (3) the *Taylor series method*, which is less computationally intensive but cannot handle certain types of estimates, such as estimates of medians or percentiles. A replication method of dealing with complex samples is discussed by Brick et al. (2000).

It should be noted that in practice, most social science researchers utilize the default significance tests generated by SPSS and other leading statistical packages, whose defaults are based on the assumption of simple random sampling. This is justified on the argument that in most cases, substantive conclusions are not affected. However, ignoring complex sample design and using simple random sampling methods runs the risk of biased estimates (Landis et al., 1982; Kish 1992; Korn and Graubard, 1995).

What is resampling?

Discussed in a separate Statistical Associates "Blue Book" volume, resampling is an alternative inductive approach to significance testing. In this approach, repeated samples of n-1 size are drawn from a sample of n. Any statistic based on the repeated samples, such as a correlation coefficient, is thus replicated as many times as there are samples. The standard deviation of the coefficient may be computed based on the list of the coefficient derived from each sample. Based on the assumption of asymptotic normality, which is met by taking a large number of

samples, the probability (p) that the coefficient is not different from zero may be computed. This p value is reported as the significance of the coefficient.

Note that resampling changes the meaning of significance. In the usual meaning, significance is the chance of getting a result as strong or stronger than the observed result if another random sample is taken from the population. In resampling, significance is the chance of getting a result as strong or stronger than the observed result if another n-1 sample is taken from the current dataset (not from the population).

On the one hand, resampling is a distribution-free method of computing significance, which can be applied to any dataset. On the other hand, significance computed in this manner cannot justify generalizations to the population, only to the data at hand.

Bibliography

Bourque, Linda B. and Virginia A. Clark (1992). *Processing data: The survey example*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No. 85.

Brick, J. Michael; Morganstein, David; & Valliant, Richard (2000). *Analysis of complex sample data using replication*. Westat, Inc: http://www.westat.com/westat/pdf/wesvar/acs-replication.pdf.

Brogan, D. (1998). Software for sample survey data, misuse of standard packages. Pp. p. 4167-4174. in P. Armitage and T. Colton, eds., *Encyclopedia of Biostatistics, Volume 5*. New York: Wiley.

Czaja, Ron and Johnny Blair (2005). Designing surveys: A guide to decisions and procedures. Second ed.. Thousand Oaks, CA: Sage Publications.

Deming, William Edwards (2010). *Some theory of sampling* (Dover Books on Mathematics). Dover Publications.

Fink, Arlene (2002). *How to sample in surveys, Vol. 7*. Thousand Oaks, CA: Sage Publications.

Goldstein, H. (1995). Multilevel statistical models. London, Edward Arnold; New York, Halstead Press.

Gurr, Ted (1972). *Polimetrics*. Englewood Cliffs, NJ: Prentice-Hall.

Henry, Gary T. (1990). *Practical sampling*. Thousand Oaks, CA: Sage Publications. A very accessible introductory treatment.

Kalton, Graham (1983*). Introduction to survey sampling*. Quantitative Applications in the Social Sciences Series, No. 35. Thousand Oaks, CA: Sage Publications.

Kalton, Graham and D. Kasprzyk (1982). Imputing for missing survey responses. Proceedings of the Section on Survey Research Methods, American Statistical Association: 22-31.

Landis, Richard J., Lepkowski, James M., Eklund, Stephen A., and Kish, L. (1965). *Survey sampling.* NY: John Wiley. Cited with regard to selection grids.

Lee, Eun Sol, Ronald N. Forthofer, and Ronald J. Lorimor (1989). *Analyzing complex survey data.* Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No. 71.

Lohr, Sharon (2009). *Sampling: Design and analysis* . NY: Duxbury Press.

Kish, L. (1992). Weighting for unequal Pi, *Journal of Official Statistics*, 8, 183-200.

Korn, E.L. and Graubard, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49, 291-295.

Stehouwer, Sharon A., (1982). A Statistical Methodology for Analyzing Data From a Complex Survey: The First National Health and Nutrition Examination Survey, Vital and Health Statistics, National Center for Health Statistics, Series 2-No. 92, DHHS Pub. No. 82-1366.

Thompson, Steven K. (2012). *Sampling* (Wiley Series in Probability and Statistics). NY: Wiley.

White, Paul (2006). *Basic sampling*. Music Sales America.

Statistical Associates Publishing

Blue Book Series

Association, Measures of
Assumptions, Testing of
Canonical Correlation
Case Studies
Cluster Analysis
Content Analysis
Correlation
Correlation, Partial
Correspondence Analysis
Cox Regression
Crosstabulation
Curve Estimation
Data Distributions and Random Numbers
Data Imputation/Missing Values
Data Levels
Delphi Method
Discriminant Function Analysis
Ethnographic Research
Evaluation Research
Event History Analysis
Factor Analysis
Focus Groups
Game Theory
Generalized Linear Models/Generalized Estimating Equations
GLM (Multivariate), MANOVA, and MANCOVA
GLM (Univariate), ANOVA, and ANCOVA
GLM Repeated Measures
Grounded Theory
Hierarchical Linear Modeling/Multilevel Analysis/Linear Mixed Models
Kaplan-Meier Survival Analysis
Latent Class Analysis
Life Tables

Logistic Regression
Log-linear Models,
Longitudinal Analysis
Multidimensional Scaling
Multiple Regression
Narrative Analysis
Network Analysis
Nonlinear Regression
Ordinal Regression
Partial Least Squares Regression
Participant Observation
Path Analysis
Power Analysis
Probability
Probit and Logit Response Models
Reliability Analysis
Resampling
Research Designs
Sampling
Scales and Standard Measures
Significance Testing
Structural Equation Modeling
Survey Research
Time Series Analysis
Two-Stage Least Squares Regression
Validity
Weighted Least Squares Regression

Statistical Associates Publishing
http://www.statisticalassociates.com
sa.publishers@gmail.com