Sampling

by David A. Freedman
Department of Statistics
University of California
Berkeley, CA 94720

The basic idea in sampling is extrapolation from the part to the whole—from "the sample" to "the population." (The population is sometimes rather mysteriously called "the universe.") There is an immediate corollary: the sample must be chosen to fairly represent the population. Methods for choosing samples are called "designs." Good designs involve the use of probability methods, minimizing subjective judgment in the choice of units to survey. Samples drawn using probability methods are called "probability samples."

Bias is a serious problem in applied work; probability samples minimize bias. As it turns out, however, methods used to extrapolate from a probability sample to the population should take into account the method used to draw the sample; otherwise, bias may come in through the back door. The ideas will be illustrated for sampling people or business records, but apply more broadly. There are sample surveys of buildings, farms, law cases, schools, trees, trade union locals, and many other populations.

SAMPLE DESIGN

Probability samples should be distinguished from "samples of convenience" (also called "grab samples"). A typical sample of convenience comprises the investigator's students in an introductory course. A "mall sample" consists of the people willing to be interviewed on certain days at certain shopping centers. This too is a convenience sample. The reason for the nomenclature is apparent, and so is the downside: the sample may not represent any definable population larger than itself.

To draw a probability sample, we begin by identifying the population of interest. The next step is to create the "sampling frame," a list of units to be sampled. One easy design is "simple random sampling." For instance, to draw a simple random sample of 100 units, choose one unit at random from the frame; put this unit into the sample; choose another unit at random from the remaining ones in the frame; and so forth. Keep going until 100 units have been chosen. At each step along the way, all units in the pool have the same chance of being chosen.

Simple random sampling is often practical for a population of business records, even when that population is large. When it comes to people, especially when face-to-face interviews are to be conducted, simple random sampling is seldom feasible: where would we get the frame? More complex design are therefore needed. If, for instance, we wanted to sample people in a city, we could list all the blocks in the city to create the frame, draw a simple random sample of blocks, and interview all people in housing units in the selected blocks. This is a "cluster sample," the cluster being the block.

Notice that the population has to be defined rather carefully: it consists of the people living in housing units in the city, at the time the sample is taken. There are many variations. For example, one person in each household can be interviewed to get information on the whole household. Or, a person can be chosen at random within the household. The age of the respondent can be restricted; and so forth. If telephone interviews are to be conducted, "random digit dialing" often provides a reasonable approximation to simple random sampling—for the population with telephones.

CLASSIFICATION OF ERRORS

Since the sample is only part of the whole, extrapolation inevitably leads to errors. These are of two kinds: sampling error ("random error") and non-sampling error ("systematic error"). The latter is often called "bias," without connoting any prejudice. Sampling error results from the luck of the draw when choosing a sample: we get a few too many units of one kind, and not enough of another. The likely impact of sampling error is usually quantified using the "SE," or standard error. With probability samples, the SE can be estimated using (i) the sample design and (ii) the sample data.

As the "sample size" (the number of units in the sample) increases, the SE goes down, albeit rather slowly. If the population is relatively homogeneous, the SE will be small: the degree of heterogeneity can usually be estimated from sample data, using the standard deviation or some analogous statistic. Cluster samples—especially with large clusters—tend to have large SEs, although such designs are often cost-effective.

Non-sampling error is often the more serious problem in practical work, but it is harder to quantify and receives less attention than sampling error. Non-sampling error cannot be controlled by making the sample bigger. Indeed, bigger samples are harder to manage. Increasing the size of the sample—which is beneficial from the perspective of sampling error—may be counter-productive from the perspective of non-sampling

error. Non-sampling error itself can be broken down into three main categories: (i) selection bias, (ii) non-response bias, and (iii) response bias. We discuss these in turn.

(i) "Selection bias" is a systematic tendency to exclude one kind of unit or another from the sample. With a convenience sample, selection bias is a major issue. With a well-designed probability sample, selection bias is minimal. That is the chief advantage of probability samples.

(ii) Generally, the people who hang up on you are different from the ones who are willing to be interviewed. This difference exemplifies non-response bias. Extrapolation from respondents to non-respondents is problematic, due to non-response bias. If the response rate is high (most interviews are completed), non- response bias is minimal. If the response rate is low, non- response bias is a problem that needs to be considered. At the time of writing, U.S. government surveys that accept any respondent in the household have response rates over 95%. The best face-to-face research surveys in the U.S., interviewing a randomly-selected adult in a household, get response rates over 80%. The best telephone surveys get response rates approaching 60%. Many commercial surveys have much lower response rates, which is cause for concern.

(iii) Respondents can easily be lead to shade the truth, by interviewer attitudes, the precise wording of questions, or even the juxtaposition of one question with another. These are typical sources of response bias.

Sampling error is well-defined for probability samples. Can the concept be stretched to cover convenience samples? That is debatable (see below). Probability samples are expensive, but minimize selection bias, and provide a basis for estimating the likely impact of sampling error. Response bias and non-response bias affect probability samples as well as convenience samples.

TRADING NON-RESPONDENTS FOR RESPONDENTS

Many surveys have a planned sample size: if a non-respondent is encountered, a respondent is substituted. That may be helpful in controlling sampling error, but makes no contribution whatsoever to reducing bias. If the survey is going to extrapolate from respondents to non-respondents, it is imperative to know how many non-respondents were encountered.

HOW BIG SHOULD THE SAMPLE BE?

There is no definitive statistical answer to this familiar question. Bigger samples have less sampling error. On the other hard, smaller samples

may be easier to manage, and have less non- sampling error. Bigger samples are more expensive than smaller ones: generally, resource constraints will determine the sample size. If a pilot study is done, it may be possible to judge the implications of sample size for accuracy of final estimates.

The size of the population is seldom a determining factor, provided the focus is on relative errors. For example, the percentage breakdown of the popular vote in a U.S. presidential election—with 200 million potential voters—can be estimated reasonably well by taking a sample of several thousand people. Of course, choosing a sample from 200 million people all across the U.S. is a lot more work than sampling from a population of 200,000 concentrated in Boise, Idaho.

STRATIFICATION AND WEIGHTS

Often, the sampling frame will be partitioned into groupings called "strata," with simple random samples drawn independently from each stratum. If the strata are relatively homogeneous, there is a gain in statistical efficiency. Other ideas of efficiency come into play as well. If we sample blocks in a city, some will be sparsely populated. To save interviewer time, it may be wise to sample such blocks at a lower rate than the densely-populated ones. If the objective is to study determinants of poverty, it may be advantageous to over-sample blocks in poorer neighborhoods.

If different strata are sampled at different rates, analytic procedures must take sampling rates into account. The "Horvitz-Thompson" estimator, for instance, weights each unit according to the inverse of its selection probability. This estimator is unbiased, although its variance may be high. Failure to use proper weights generally leads to bias, which may be large in some circumstances. (With convenience samples, there may not be a convincing way to control bias by using weights.) An estimator based on a complex design will often have a larger variance than the corresponding estimator based on a simple random sample of the same size: clustering is one reason, variation in weights is another. The ratio of the two variances is called "the design effect."

RATIO AND DIFFERENCE ESTIMATORS

Suppose we have to audit a large population of claims to determine their total audited value, which will be compared to the "book value." Auditing the whole population is too expensive, so we take a sample. A relatively large percentage of the value is likely to be in a small percentage of claims. Thus, we may over-sample the large claims and under-sample

the small ones, adjusting later by use of weights. For the moment, however, let us consider a simple random sample.

Suppose we take the ratio of the total audited value in the sample claims to the total book value, then multiply by the total book value of the population. This is a "ratio estimator" for the total audited value of all claims in the population. Ratio estimators are biased, because their denominators are random: but the bias can be estimated from the data, and is usually offset by a reduction in sampling variability. Ratio estimators are widely used.

Less familiar is the "difference estimator." In our claims example, we could take the difference between the audited value and book value for each sample claim. The sample average—dollars per claim—could then be multiplied by the total number of claims in the population. This estimator for the total difference between audited and book value is unbiased, and is often competitive with the ratio estimator.

Ratio estimators and the like depend on having additional information about the population being sampled. In our example, we need to know the number of claims in the population, and the book value for each; the audited value would be available only for the sample. For stratification, yet other information about the population would be needed. We might use the number of claims and their book value, for several different strata defined by size of claim. Stratification improves accuracy when there is relevant additional information about the population.

COMPUTING THE STANDARD ERROR

With simple random samples, the sample average is an unbiased estimate of the population average—assuming that response bias and nonresponse bias are negligible. The SE for the sample average is generally well approximated by the SD of the sample, divided by the square root of the sample size. With complex designs, there is no simple formula for variances; procedures like "the jackknife" may be used to get approximate variances. (The SE is the square root of the variance.) With non-linear statistics like ratio estimators, the "delta method" can be used.

THE SAMPLING DISTRIBUTION

We consider probability samples, setting aside response bias and nonresponse bias. An estimator takes different values for different samples ("sampling variability"); the probability of taking on any particular value can, at least in principle, be determined from the sample design. The

probability distribution for the estimator is its "sampling distribution." The expected value of the estimator is the center of its sampling distribution, and the SE is the spread. Technically, the "bias" in an estimator is the difference between its expected value and the true value of the estimand.

SOME EXAMPLES

In 1936, Franklin Delano Roosevelt ran for his second term, against Alf Landon. Most observers expected FDR to swamp Landon—but not the Literary Digest, which predicted that FDR would get only 43% of the popular vote. (In the election, FDR got 62%.) The Digest prediction was based on an enormous sample, with 2.4 million respondents. Sampling error was not the issue. The problem must then be non-sampling error, and to find its source, we need to consider how the sample was chosen.

The Digest mailed out 10 million questionnaires and got 2.4 million replies—leaving ample room for non-response bias. Moreover, the questionnaires were sent to people on mailing lists compiled from car ownership lists and telephone directories, among other sources. In 1936, cars and telephones were not as common as they are today, and the Digest mailing list was overloaded with people who could afford what were luxury goods in the depression era. That is selection bias.

We turn now to 1948, when the major polling organizations (including Gallup and Roper) tapped Dewey—rather than Truman—for the presidency. According to one celebrated headline,

DEWEY AS GOOD AS ELECTED, STATISTICS CONVINCE ROPER.

The samples were large—tens of thousands of respondents. The issue was non-sampling error, the problem being with the method used to choose the samples. That was "quota sampling." Interviewers were free to choose any subjects they liked, but certain numerical quotas were prescribed. For instance, one interviewer had to choose 7 men and 6 women; of the men, 4 had to be over 40 years of age; and so forth.

Quotas were set so that, in the aggregate, the sample closely resembled the population with respect to gender, age, and other control variables. But the issue was, who would vote for Dewey? Within each of the sample categories, some persons were more likely then others to vote Republican. No quota could be set on likely Republican voters, their number being unknown at the time of the survey. As it turns out, the interviewers preferred Republicans to Democrats—not only in 1948 but in all previous elections where the method had been used.

Interviewer preference for Republicans is another example of selection bias. In 1936, 1940, and 1948, Roosevelt won by substantial margins: selection bias in the polls did not affect predictions by enough to matter. But the 1948 election was a much closer contest, and selection bias tilted the balance in the polls. Quota sampling looks reasonable: it is still widely used. Since 1948, however, the advantages of probability sampling should be clear to all.

Our final example is a proposal to adjust the U.S. census. This is a complicated topic, but in brief, a special sample survey ("Post Enumeration Survey") is done after the census, to estimate error rates in the census. If error rates can be estimated with sufficient accuracy, they can be corrected. The Post Enumeration Survey is a stratified block cluster sample, along the lines described above. Sample sizes are huge (700,000 people in 2000), and sampling error is under reasonable control. Non-sampling error, however, remains a problem—relative to the small errors in the census that need to be fixed. For discussion from various perspectives, see Imber (2001). Also see Freedman and Wachter (2003).

SUPERPOPULATION MODELS

Samples of convenience are often analyzed as if they were simple random samples from some large, poorly-defined parent population. This unsupported assumption is sometimes called the "super-population model." The frequency with which the assumption has been made in the past does not provide any justification for making it again, and neither does the grandiloquent name. Assumptions have consequences, and should only be made after careful consideration: the problem of induction is unlikely to be solved by fiat. For discussion, see Berk and Freedman (1995).

An SE for a convenience sample is best viewed as a de minimis error estimate: if this were—contrary to fact—a simple random sample, the uncertainty due to randomness would be something like the SE. However, the calculation should not be allowed to divert attention from non-sampling error, which remains the primary concern. (The SE measures sampling error, and generally ignores bias.)

SOME PRACTICAL ADVICE

Survey research is not easy; helpful advice will be found in the references below. Much attention needs to be paid in the design phase. The research hypotheses should be defined, together with the target population. If people are to be interviewed, the interviewers need to be trained

and supervised. Quality control is essential. So is documentation. The survey instrument itself must be developed. Short, clear questions are needed; these should be worded so as to elicit truthful rather than pleasing answers. Doing one or more pilot studies is highly recommended. Non-response has to minimized; if non-response is appreciable, a sample of non-respondents should be interviewed. To the maximum extent feasible, probability methods should be used to draw the sample.

REFERENCES

Berk, R.A. and Freedman, D.A. (1995). Statistical assumptions as empirical commitments. In T.G. Blomberg and S. Cohen (Eds.), *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (pp. 245–258). New York: Aldine de Gruyter. 2nd ed. (2003) in press.

Cochran, W.G. (1977). *Sampling techniques*, 3rd ed. New York: John Wiley & Sons.

Diamond, S.S. (2000). Reference guide on survey research, in *Reference manual on scientific evidence*. 2nd ed. (pp. 229–276). Washington, D.C.: Federal Judicial Center.

Freedman, D.A., Pisani, R., and Purves, R.A. (1998). *Statistics*. 3rd ed. New York: W.W. Norton, Inc.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample survey methods and theory*. New York: John Wiley & Sons.

Freedman, D. A. and Wachter, K. W. (2003). On the likelihood of improving the accuracy of the census through statistical adjustment. In D. R. Goldstein (Ed.), *Science and Statistics: A Festscrift for Terry Speed*. Institute of Mathematical Statistics Monograph 40 pp. 197–230.

Imber, J.B. (Ed.). (2001). Symposium: Population politics. *Society*, 39, 3-53.

Kish, L. (1987). *Statistical design for research*. New York: John Wiley & Sons.

Sudman, S. (1976). *Applied sampling*. Academic Press.

Zeisel, H. and Kaye, D.H. (1997). *Prove it with figures*. New York: Springer.