Sample Size

In: Introduction to Survey Sampling

By: Graham Kalton

Pub. Date: 2011
Access Date: September 17, 2018
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780803921269
Online ISBN: 9781412984683
DOI: http://dx.doi.org/10.4135/9781412984683
Print pages: 82-84

Sample Size

One of the first questions that arises in sample design is, “What sample size is needed?” The discussion of this question has been left until now because it depends on several aspects of the preceding material.

To describe the basic ideas, consider a simple example of a face-to-face interview survey that is to be conducted to estimate the percentage of a city's population of 15,000 adults who say they would make use of a new library if one were built. To determine an appropriate sample size, it is first necessary to specify the degree of precision required for the estimator. This is no easy task, and initially the degree of precision required is often overstated. Suppose, for instance, the initial specification calls for an estimator that is within 2% of the population percentage with 95% probability; in other words, the 95% confidence interval should be the sample percentage plus or minus 2%. This specification thus requires that 1.96 SE(p) = 2%, where p is the sample percentage. Assuming initially the use of simple random sampling, and ignoring the fpc term, SE(p) = $\sqrt{PQ/n'}$, where P is the population percentage, Q = 100 − P, and n' is the initial estimate of the sample size. Thus $1.96\sqrt{PQ/n'} = 2$ or $n' = 1.96^2PQ/2^2$. In order to determine n', a value is needed for P. Since PQ is largest at P = Q = 50%, a conservative choice is to set P equal to a percentage as close to 50% as is likely to occur. Suppose that P is thought to lie between 15% and 35%; then the conservative choice is P = 35%. With this choice, n' = 2185. If this initial sample size were small compared with the population size, so that the fpc term could be ignored, it would be the required sample size. In the present case, however, the fpc term should not be neglected. A revised estimate of the sample size to take account of the fpc term is obtained, with N = 15,000, as

$$n = Nn'/(N + n') = 1907$$

The above calculation assumes simple random sampling, and a modification is needed for other sample designs. The modification consists of multiplying the SRS sample size by the design effect for the survey estimator under the complex design. If a list of the city's adults were available in the above example, then an unclustered proportionate stratified sample might well be used. In this case, a somewhat smaller sample might suffice because of the gains in precision arising from the stratification. As has been noted earlier, however, the gains from proportionate stratification are generally small when estimating a percentage, so the reduction in sample size will often be modest. Say, in the present case, that the design effect for the sample percentage with a proportionate stratified design is predicted to be 0.97. Then the required sample size for an unclustered proportionate stratified design to give a confidence interval of within ± 2% is 0.97 × 1907 = 1850.

If no list of the city's adults or dwellings is available, area sampling may be needed, perhaps first sampling city blocks, then listing dwellings within blocks, sampling dwellings in selected blocks, and finally sampling one (or more) adults from each selected dwelling. Stratification and PPS selection would almost certainly be used in such a design. Suppose that with a stratified multistage design in which an average of ten adults are to be sampled from each PSU (block), the design effect is predicted to be about 1.3. Then the required sample for this design would be 1.3 × 1907 = 2479.

Another factor that needs to be included in the calculation of sample size is nonresponse. Suppose that the response rate is predicted to be 75%. Then the selected sample size needed to generate the achieved sample of 2479 adults with the multistage design has to be set at 2479/0.75 = 3305. Of course, this adjustment serves only to produce the desired sample size; it does not address the problem of nonresponse bias.

Having reached this point, the researcher may decide to review the initial specification of precision to see if it can be relaxed. Suppose that, on reflection, a confidence interval of ± 3% is deemed acceptable. Then the selected sample size can be substantially reduced to 1581. In practice, the level of precision required for an estimator is seldom cast in concrete. In consequence, the sample size is usually determined from a rough-and-ready assessment of survey costs relative to the level of precision that will result. It should be noted that the selected sample size depends on predictions of a number of quantities, such as the percentage of the population who say they would use the library, the design effect, and the nonresponse rate. Errors in predicting these quantities cause the survey estimator to have a level of precision different from that specified, but that is the only adverse effect; the estimator remains a reasonable estimator of the population parameter.

Having fixed the required sample size, the next step is to determine the sampling fraction to be used. If the sample is to be drawn from the list of the city's 15,000 adults, consideration will need to be given to the possibility of blanks (deaths and movers out of the area) and foreign elements on the list as well as to the consequences of any linking procedure that might be employed in dealing with missing elements. If, say, 4% of the listings are blanks and no linking is employed, the sampling fraction will need to be set at 2479/(0.96 × 15,000) = 0.172, or 1 in 5.81, to yield a sample of 2479. In practice, this sampling fraction may then be rounded for convenience, perhaps to 1 in 5.8 or even 1 in 6, to yield expected samples of 2483 or 2400, respectively.

In the case of the multistage area sample, the sample design calls for a sample of dwellings, with one adult sampled per dwelling. Suppose that at a recent census the city contained 6500

occupied dwellings. This figure should first be updated to correct for changes that have occurred since the census date, and also adjusted for any differences between the survey and census definitions of the city boundaries. Suppose that, as a result of these adjustments, the current number of occupied dwellings in the city is estimated to be 6750. In addition to these adjustments, allowance also needs to be made for the fact that the survey's sampling operations will probably fail to attain as complete a coverage of the city's dwellings as the census enumeration; the coverage rate for the sample might, say, be estimated as 95% of that of the census. Using this coverage rate, a sampling fraction of 2479/(0.95 × 6750) = 0.3866, or 1 in 2.59 dwellings, is needed to give the desired sample of 2479 adults. As before, the sampling rate may be rounded for convenience to 1 in 2.6, accepting a marginally smaller expected sample size (2466) for the use of a simpler rate.

While the above example has served to bring out a number of the issues involved in choice of sample size, it is nevertheless an oversimplified representation. In practice surveys are multipurpose, with a substantial number of estimators needing to be considered. Moreover, these estimators are required not only for the total sample but also for a wide range of subclasses, perhaps for different regions of the country, for people in different age groups or different educational levels, and so on. A major reason for the large samples typical of many surveys is the need to provide adequate precision for subclass estimators and for differences between subclass estimators. Larger samples permit finer divisions of the sample for subclass analysis, and, in general, the larger the sample the more detailed the analysis that can be conducted. The choice of sample size often depends on an assessment of the costs of increasing the sample compared with the possible benefits of more detailed analyses.

http://dx.doi.org/10.4135/9781412984683.n11