Practical Sampling

Sample Size

Contributors: Gary T. Henry
Editors: Gary T. Henry
Book Title: Practical Sampling
Chapter Title: "Sample Size"
Pub. Date: 1990
Access Date: October 14, 2013
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780803929593
Online ISBN: 9781412985451
DOI: http://dx.doi.org/10.4135/9781412985451.n7
Print pages: 117-129

http://dx.doi.org/10.4135/9781412985451.n7
[p. 117 ↓]

Sample Size

As noted earlier, the number of units for the sample is usually the first question addressed by a study team to the sampling consultant. Response to the question must await information on other design and sampling choices. Why is the sample size important for a study? Sample size is the most potent method of achieving estimates that are sufficiently precise and reliable for policy decisions or scientific inquiry. The impact of increasing sample size on the estimates of the sampling variability is shown in Figure 7.1. The downward sloping curve indicates that sampling variability decreases as the sample size increases. However, the gain in precision is greater for each unit increase in the smaller sample size range than the larger sample size range.

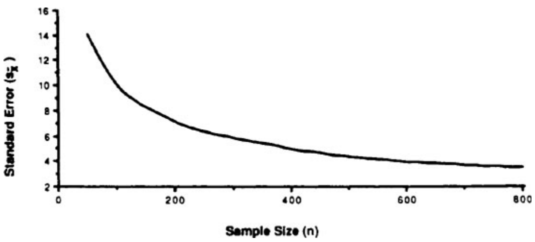
Increasing sample size obviously has a cost. Larger samples require more expenditures for collecting data, especially when interviews are being utilized; following up on nonresponse; and coding and analyzing data. When increasing the sample size is done at the expense of effort invested in follow-up of nonrespondents, for instance, the total error may rise due to non-sampling bias. Choice of a sample size cannot be considered in a vacuum. Once again, trade-offs in cost, total error, and other design choices must be considered.

To begin the process of making a sample size choice, a number of factors must be examined sequentially. Prior to beginning the process, the determination of the tolerable error of the estimates or power of the analysis must be made. Policy studies, though they serve different purposes than studies oriented toward testing theory, are subject to the same criteria. The determination of tolerable error or power needs for a policy study tends to be defined more by the use for the information in the particular situation at hand than by conventional standards.

For policy studies, it is useful to involve policymakers in making the determinations about the precision required of the data. Their responses can be elicited by posing "what if" type questions, such as, "Let's say the study estimates that 60% of the elderly are in need of services, but we are only reasonably confident that somewhere between 50% and 70% need services. Is the 20% gap too large for the information to be useful?" or alternatively,

[p. 118 ↓]

Figure 7.1. Relationship Between Sample Size and Standard Error



> You know that test scores of disadvantaged students are approximately at the 38th percentile on standardized tests. Considering the cost of this program and available alternatives, would you consider the program successful if scores were improved to the 40th percentile? How about the 43rd percentile?

Variations on these questions can be developed to fit the particular policy study. Cost estimates can be explicitly incorporated into the discussion to provide direct information about the relationship between cost and sampling variability.

The factors to be examined in the choice of sample size include:

- Efficient sample size
- Implications of the design for efficient sample size
- Implications of the sample size and design for subpopulation analysis
- Adjustments for ineligibles and nonresponse
- Expense of the design given the sample size

Credibility

Each of these factors is discussed below.

EFFICIENT SAMPLE SIZE

Efficient sample size is based on an estimate of the sample size required to limit sampling variability to the desired level. For a study that is essentially descriptive, the sampling variability is set in terms of the level of precision **[p. 119 ↓]** needed for the estimates. Analytical studies use the size of the effect that the study estimates should be able to detect. Generally, efficient sample size estimates assume a simple random sample design, although with more information for studies that are frequently repeated, design-specific estimates can be developed.

The computation of efficient sample sizes for descriptive studies begins with the tolerable error (*te*): the standard error times the *t*-value for the selected confidence level. The variance or standard deviation of the variable must be estimated, also. Usually, the standard deviation can be estimated from a previous study. Some adjustment may be necessary if the target population for the study is different than the target population for the previous study.

Another method for estimating the standard deviation is the use of a small pilot study. Sometimes as few as 50 cases can provide a useful estimate. It is best to select the cases randomly from the target population, but this option is not always available. Judgment must be applied to determine whether the target population estimate would be expected to differ from the pilot study population. If so, an adjustment should be made.

A third method that obtains a rough estimate of the standard deviation can be obtained by dividing the range by four. If the highest and lowest value of a variable can be obtained from data or expert opinion, the estimate of the range can be plugged into the formula and the standard deviation estimated.

A final method of estimating the variance is often used when proportions are the statistics of interest. This method simply assumes the maximum variance that occurs when $p = .5$. The product of $p(1 - p)$ in this case is .25, the largest possible product of a proportion or a worst-case scenario.

The formulas for the efficient sample size for means and proportions are shown in Table 7.1 along with an example calculation for each.

In the first computation, the standard deviation is estimated to be 37.6. A tolerable error of 1.764 is used, which relates to a standard error of .9 units and a *t*-score of 1.96. The efficient sample size is 1,745 before the finite population correction (FPC) is applied, and 1,617 after. In the second example, the tolerable error of 2% yields a standard error of 1%. That is, the researchers will be 95% confident that the estimate of the proportion will fluctuate as much as 2% above or 2% below the true proportion. To obtain this level of precision, a sample size of 2,300 is needed after taking the FPC into account.

Even though these calculations are relatively straightforward, elementary statistics textbooks often present tables that give efficient sample sizes **[p. 120 ↓]** and associated standard errors. These tables usually are based on the assumption that an estimate of a proportion is the objective of the study and that the maximum assumption for the proportion ($p = .5$) is appropriate. Also, a finite population correction is not applied. These tables fit a very limited number of situations and should be used with caution.

TABLE 7.1 Efficient Sample Size for Means

$s = 37.6$ $N = 22{,}000$
$te = 1.764$ $t = 1.96$
$s_{\bar{x}} = te/t = .9$

$n' = (37.6)^2/(1.764/1.96)^2$
$n' = 1{,}745$
$n = 1{,}745/[1 + (1{,}745/22{,}000)]$
$n = 1{,}617$

where: s is standard deviation estimate
N is the population size
te is the tolerable error
t is the t-value for the desired confidence level
$s_{\bar{x}}$ is the allowable standard error
n' is the sample size without finite population correction
n is the sample size with the finite population correction

Efficient Sample Size for Proportions

$p = .6$ $1 - p = .4$
$te = .02$ $N = 1{,}100{,}000$
$s_p = .01$ $t = 1.96$

$n' = (.6)(.4)/(.02/1.96)^2$
$n' = 2{,}305$
$n = 2{,}305/(1 + 2{,}305/1{,}100{,}000)$
$n = 2{,}300$

where: p is a proportion of the sample
te is the tolerable error
N is the population size
s_p is the standard error of the proportion
n' is the sample size without the finite population correction
n is the sample size with the finite population correction

For analytical studies, efficient sample size calculations quickly exceed the capacity of this sampling text. Interested readers should begin with Lipsey (1989) in their quest for information on power analysis. An example of one study may adequately explain the concept of power analysis.

[p. 121 ↓]

Sample size for each group (total sample size = n × 2)	Sentence length differences (months)
n	$t = (\bar{x}_1 - \bar{x}_2) / (s_{\bar{x}\,2} + s_{\bar{x}\,2}{}^2)^{1/2}$
30	16
50	12
75	10
115	8
295	5

775	3
where: n is the sample size for each group	
x *B* is the mean sentence length for blacks	
x *W* is the mean sentence length for whites	

A study is to be undertaken to determine if differences exist between sentence lengths of whites and minorities convicted of the same crime. The study will analyze the differences between means of the samples of whites and blacks using the following formula:

$$s_{\bar{x}}^{2}$$

For this study it is determined that the researchers will risk an error in finding a difference only 5 times out of 100 if no difference exists (i.e., *t* = 1.96).

Furthermore, the standard deviation of sentence lengths for rape of 32.3 months for minorities and 29.3 months for whites can be assumed from an earlier study. Finally, equal sample sizes are assumed in the calculation ($n_1 = n_2$). Using these assumptions, a sample size for each sample of 30 cases is needed to detect a 16-month difference in sentence lengths. However, if the sentence length difference is expected to be only 3 months, each sample would need 775 observations to detect the difference. Sample sizes required for various expected differences are shown in Table 7.2.

The estimates clearly show that smaller expected effects require larger samples to detect the effect. This relationship generalizes to other analytical statistics. The specific formula for the calculation of power can become quite complex. Lipsey (1989) provides an excellent reference for guiding a researcher through the factors that affect the sensitivity of a design to detect relationships. A sampling expert is often required to assist in the computations when power is the overriding concern.

[p. 122 ↓]

IMPLICATIONS OF THE DESIGN FOR SAMPLE SIZE

While sample size has the most direct impact on the efficient sample size, the design also has an impact. The efficient sample size calculations assume simple random samples. If the sample design deviates from simple random sampling, the efficient sample size is likely to vary also. Sampling variability increases when cluster sampling is used; it decreases when stratification is used.

The design effect (deff) is a direct way of addressing the impact of design on sampling variability. The design effect can be multiplied by the expected sampling variance (

$$n' = \frac{n}{(e)(r)}$$

$$n' = \frac{1620}{(.95)(.85)}$$

$$n' = 2006$$

) in the calculation of an efficient sample size to adjust for the impact of the design. The design effect is the ratio of the sampling variance of the design to the sampling variance, assuming a simple random sample (Kish, 1965; Sudman, 1976). The square root of the design effect is used more often in practice to make it comparable to the standard error.

To incorporate the effect of the design into the calculation of the efficient sample size, information about the expected design effect is needed prior to the execution of the sample design. Naturally an estimate of the design effect is the best that a researcher can provide. For stratified samples, the design effect for means is likely to range from .5 to .95. The actual deff will depend on the number of strata and the correlation between the stratification variables and the variable studied.

Cluster samples can be expected to have a design effect greater than one. A common range would be 1.5 to 3.0. Obviously, a range of this size would have quite an impact on the efficient sample size. Determining the estimate of effect depends on characteristics of the particular design. The number of clusters, the homogeneity of the cluster members, and the use of stratification have an important bearing on the actual design effect.

Multistage samples, also, should be adjusted for design effects: "Sampling errors in multi-stage random samples are almost always larger than in unrestricted random sampling, and the effect of stratification at the first (and possibly later) stages is to reduce this excess but almost never to eliminate it" (Stuart, 1963, p. 89). Stuart offers a rough rule of a 1.25 to 1.50 increase in the sampling error. These numbers would be squared to multiple times the sample variance.

More recent work (Kish & Frankel, 1970; Frankel, 1971) gives the square root of the design effect for a variety of multistage samples and a variety of estimates. Two conclusions of Kish and Frankel are particularly relevant here: "Standard errors computed by machine programs, based on srs assumptions, were not gross underestimates [for multivariate analyses]"; **[p. 123 ↓]** and "Design effects were shown to be estimable and of appreciable magnitudes for standard errors of regression coefficients" (p. 1073).

TABLE 7.3 Sample Size and Subpopulation Analysis Analyzing 8 Equal-Size Districts

$n = 1{,}620$ $s = 37.6$ $s_{\bar{x}} = .9$

$n_d = 1{,}620/8 = 203$

$s_{d\bar{x}} = 37.6/(203)^{1/2} = 2.64$

where: n is entire sample size
n_d is sample size for each district
s is standard deviation estimate
$s_{\bar{x}}$ is standard error of the mean for full sample
$s_{d\bar{x}}$ is standard error for district subsample

In empirical investigations, the design effect was found to be larger for means than for regression coefficients. The rough guidance provided by Stuart (1963) proved reasonably accurate for the square root of the design effect for means. The square root of the design effect for regression coefficients tended to range from 1.06 to 1.30. Using the deff in formulas for efficient sample size can mean significant increases to the calculated sample size.

SUBPOPULATION ANALYSIS

Thus far, the consideration of sample size has assumed that the entire target population will be analyzed together. In many cases, subpopulations are of interest also. For example, in the frail elderly study a researcher may wish to single out females in the target population for a separate study. Another researcher may wish to analyze regions of the state separately. The subpopulation analyses have less precision than the analysis of the entire sample as a group. Fewer cases for the subpopulation increases sampling variability for the analysis by subpopulation, although smaller standard deviations for the subpopulation may offset the increase to some extent.

The impact of subpopulation analysis is shown in an example where program analysts are to estimate the length of time in weeks that cases have been open in a social service agency (Table 7.3). The efficient, full sample size is approximately 1,620. Conducting a subpopulation analysis of eight districts, where the districts are approximately equal size, would yield district subsamples of 203 units. The standard error of 2.64 for the district sub-samples compares with a standard error of .9 for the entire sample. This standard error is 2.9 times the standard error for the total sample. The total size of the 95% confidence interval for the districts will be 10.3 (2.64 × **[p. 124 ↓]** 1.96 × 2). If district estimates are important, the researchers must consider whether this confidence interval, ±5 weeks, is sufficiently precise for the purpose of the study. If not, the researcher must consider the cost of the total sample size necessary to increase the precision of each district to the tolerable error level.

ADJUSTMENTS FOR INELIGIBLES AND NONRESPONSE

In choosing the size of the sample, the researcher must remember that the precision of the sample is estimated by the number of target population members for whom data are actually obtained. Two reasons for not obtaining usable information for some of the sample selected are:

- Ineligibles in the sampling frame
- Nonresponse

Ineligibles include those listed on the sampling frame that are not members of the target population. Ineligibles contribute to increased sampling variability by lowering the actual sample size. For example, in the North Carolina Citizen Survey a resident of Virginia working and paying taxes in North Carolina could conceivably be included on taxpayer rolls. This individual would be ineligible for a poll of residents. Analogously, using random digit dialing for a special population survey will result in calls to many residences that do not contain a member of the target population and are screened out of the sample. The Florida study is an example.

Nonresponse occurs for a variety of reasons, including inability to contact the respondent and refusal to respond. Nonresponse can create non-sampling bias in the sample, because a portion of the population is under-represented in the sample. Evaluation of potential nonresponse bias will be examined in the next chapter. Here an adjustment to the efficient sample size is offered to compensate for the impact of nonresponse on sampling variability.

Impacts from ineligibles and nonresponse can be compensated for by dividing the efficient sample size by the proportion of eligibles times the proportion of respondents. An efficient sample size of 1,620 will be adjusted to an initial sample size of approximately 2,006 when .95 of the sampling frame is estimated to be eligible and .85 of the sample are expected to respond.

[p. 125 ↓]

$$\bar{x} = (n_r/n)(\bar{x}_r) + (n_n/n)(\bar{x}_n)$$

where n' is the adjusted sample size,

n is the efficient sample size,

e is the proportion of eligibles on the list, and

r is the proportion of respondents expected.

The higher the response rate the fewer the initial contacts that have to be made as a result of adjusting the sample size. Extensive follow-up procedures, while costly, have cost savings resulting from smaller initial sample sizes that can partially offset the additional costs.

EXPENSE

The cost of the data collection can be reasonably examined at this point. The examination should include costs arising from:

Except for the first two, these costs vary by the size of the sample selected or the number of responses obtained. Follow-up procedures are extremely important in the cost calculation. Investing in follow-up procedures can reduce the size of the sample selected by increasing the response rate, reduce costs associated with initial contacts, and eliminate the costs of nonresponse bias evaluation by reducing potential nonsampling bias. If, for example, **[p. 126 ↓]** the response rate in the example cited above could be increased from 85% to 95%, the adjusted sample size could be reduced from 2006 to 1795. The expenses associated with attempting to contact the additional

211 sample units, acquiring addresses or phone numbers, mailing questionnaires, calling for appointments, and so on, may be more than an intensive follow-up and reduce nonsampling bias.

CREDIBILITY

An efficient sample size is not always a credible sample size. Often users of information mistrust sample information that they perceive to be based on too few cases. Alternatively, the audience for the information may have a conception of differences across regions or unique local conditions that requires allocation of a larger sample.

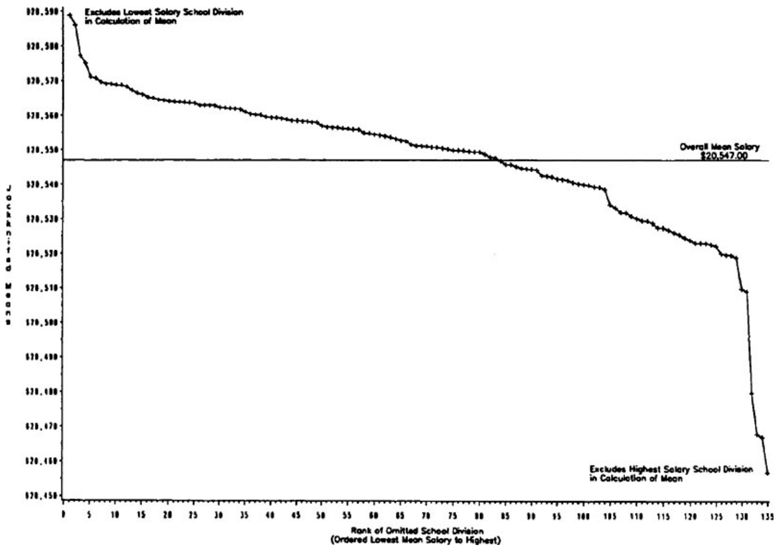
Sometimes these concerns arise as a function of population size. Computation of an efficient sample size, as shown, has little to do with the size of the population. In fact, the population size is used only in the finite population correction factor. Yet the perception exists that sample size should be a percentage of the population size—often 10% is the figure used. This perception is not accurate. Sample sizes of 1,500 to 2,500 used in general population polls and voter surveys are common, although on occasion the question arises, how can 1,500 individuals speak for citizens in this country? Media use and accuracy of the polls have overcome much skepticism, perhaps too much.

Smaller sample sizes used for medical research and other studies of special populations are sometimes viewed with incredulity. Skepticism about sample credibility is exacerbated by departures from proportional allocation of the sampling units in the random selection process. Lack of proportional distribution of the sample, or in a more extreme case lack of representation of some legislative districts, may be grounds for dismissing the sample information in legislative policy-making. The fallacy of overreliance on the sample proportions mirroring population proportions is discussed Chapter 8.

For example, in an evaluation where a sample of 60 licensed homes for adults was selected for inspection and data collection, program administrators refused to accept the results. After a census inspection effort of over 400 homes, percentages of homes with problems differed by no more than 3% across several variables being observed.

Attacks on credibility of the sample cannot be eliminated in studies that have policy impacts. Prior planning and attention to factors that may serve to undermine sample credibility may thwart undue attacks. The researcher **[p. 127 ↓] [p. 128 ↓]** should ask questions of the audience for whom the results are intended, pertaining to sample characteristics arising from the design. Expressed concerns, such as sample size by region, may enable the researcher to alter the design to accommodate potential criticisms.

Figure 7.2. Sensitivity Curve for Means Using Jackknife Method



SAMPLING SMALL POPULATIONS

Frequently researchers confront the question of how large a sample to select from small populations. Populations such as counties, probation officers, or students enrolled in Latin courses are sometimes too large to allow data collection on the entire population due to resource constraints. However, potential sample sizes appear too small to produce reliable results. In these cases it is often necessary to consider certainty selections for target population members that must be represented for the sample to be credible. Results must be appropriately weighted to account for the probability of selection.

Another technique that can be used with small samples involves the analysis of outliers in the data. Small samples are particularly vulnerable to outliers. Outliers can unduly influence the estimates produced by the sample. The difficult question is when is an observation an outlier and when is it a reasonable representation of the population? Techniques for outlier identification and statistics that provide reliable, robust estimates have been developed in recent years (Andrews, Bickel, Hampel, Huber, Rogers, & Tukey, 1972; Barnett & Lewis, 1984).

One technique, jackknifing, is extremely useful and simple (Efron, 1982). Jackknifing involves the iterative removal of a single observation and calculation of the sample estimates, until each observation has been removed from the calculation one time. The estimates can then be ordered and plotted to show the sensitivity of the estimate to any single observation. Figure 7.2 shows the sensitivity of the mean for one sample. The sensitivity curve exhibits a range of 128 units, indicating a strong influence of the extreme values on the data.

SUMMARY

The selection of the sample size depends upon the amount of sampling variability that can be tolerated, the variability of important variables, the design effect, need for subpopulation analysis, ineligibles and nonresponse, cost, and credibility factors. These factors should be considered in combination when evaluating sample design alternatives. The formulas provided in this chapter allow the researcher to estimate the sample size needed when considering these factors.

http://dx.doi.org/10.4135/9781412985451.n7