Social Studies 201
Notes for November 19, 2003

Determining sample size for estimation of a population proportion – Section 8.6.2, p. 541.

As indicated in the notes for November 17, when sample size is larger, the interval estimate is narrower and sampling error is reduced, compared with smaller sample size. This section of the notes outlines how to obtain the sample size required to estimate a population proportion for any specified sampling error and confidence level.

Notation. Let p represent the proportion of a population with a particular characteristic and q denote the proportion of the population not having this characteristic. Since members of the population must either have this characteristic or not, $p + q = 1$ and $q = 1 - p$.

Let the size of the sampling error be given the symbol E. That is, the C% confidence level will result in the interval estimates of $\hat{p} \pm E$ if the required sample size is obtained. And if the required sample size is obtained, C% of these intervals will contain the population proportion p.

Note that the units for E are proportions. For example, if the proportion of population members with a particular characteristic is to be estimated to within ±2 percentage points, the value of E will be 0.02. That is, the point estimate of p will be a proportion $\hat{p}$, and this will be accurate to within ± 0.02, so that the intervals will be $\hat{p} - 0.02$ to $\hat{p} + 0.02$.

Formula for deterining sample size

As with the interval estimates for a population proportion p, determining sample size begins by considering the sampling distribution of the sample proportion $\hat{p}$. Suppose that random samples of large sample size are taken from a population with a proportion p of members having a particular characteristic. The sample proportions $\hat{p}$ are normally distributed with mean p and standard deviation $\sqrt{pq/n}$. That is,

$$\hat{p} \text{ is Nor} \left(p, \sqrt{\frac{pq}{n}} \right) .$$

This is the case so long as n exceeds 5 divided by the smaller of p or q.

Larger sample sizes yield normal distributions of $\hat{p}$ that are more concentrated, smaller sample sizes yield normal distributions of $\hat{p}$ that are more dispersed. For any given confidence level C and associated Z-value, the aim is to find a distribution where the confidence interval estimates

$$\hat{p} \pm Z\sqrt{\frac{pq}{n}}$$

match the intervals associated with the specified sampling error E:

$$\hat{p} \pm E.$$

That is, the C% intervals are constructed so that they are $Z\sqrt{pq/n}$ on either side of $\hat{p}$. But the researcher specifies these are to be intervals of amount E on either side of $\hat{p}$. The desired error of estimate E and the confidence intervals are the same when a sample size is selected so that

$$E = Z\sqrt{\frac{pq}{n}}.$$

When this latter expression is solved for n, the required sample size is

$$n = \left(\frac{Z}{E}\right)^2 pq$$

This is the formula for the required sample size for a specified error of estimate E and for a Z-value associated with the specified confidence level.

The procedure for estimating sample size is to select a confidence level C and an error of estimate E that the researcher wishes to obtain. From the confidence level the Z-value can be determined from the table of the normal distribution. Using the above formula, the only other parts in question are the values of p and q. As stated earlier, when $p + q = 1$, the maximum value of the product of p and q occurs when $p = q = 0.5$. If a researcher wishes to determine a sample size that is sufficient to obtain sampling error E with confidence level C, then this is obtained when $p = q = 0.5$. In this circumstance, the formula for obtaining the required sample size becomes simply

$$n = \left(\frac{Z}{E}\right)^2 \times 0.25$$

since $pq = 0.5 \times 0.5 = 0.25$.

If a researcher has some knowledge that p and q are quite different than 0.5 each, then these alternate estimates for p and q can be used in the formula

$$n = \left(\frac{Z}{E}\right)^2 pq.$$

This will result in a smaller required sample size and it may be easier or less costly for the researcher to obtain this smaller sample. The concern a researcher might have though is that this smaller sample size may not be sufficient to produce intervals with the required error of estimate. Resulting interval estimates may be wider than desired.

Examples.

Suppose a researcher wishes to estimate the proportion of a population who support legalizing marijuana, correct to within (a) 5 percentage points, or (b) 2 percentage points, with probability 0.99. What are the required sample sizes?

Answer. This is an estimate of a proportion – the proportion p of the population who support the legalization of marijuana. Since the sample size will likely be fairly large, it can be assumed that the sample proportions $\hat{p}$, of those who support legalization of marijuana, will be normally distributed. The distribution of the sample proportions

$$\hat{p} \text{ is Nor} \left(p, \sqrt{\frac{pq}{n}}\right).$$

The formula for sample size is

$$n = \left(\frac{Z}{E}\right)^2 pq$$

where $E = 0.05$ for part (a). The confidence level specified is 99% (0.99 probability) and the associated Z-value is 2.575. Letting $p = q = 0.5$, the required sample size is

$$n = \left(\frac{2.575}{0.05}\right)^2 0.5 \times 0.5 = (51.5)^2 \times 0.25 = 2,652.25 \times 0.25 = 663.1$$

The required sample size is 664.

For an accuracy of 2 percentage points, $E = 0.02$ and the required sample size is

$$n = \left(\frac{2.575}{0.02}\right)^2 0.5 \times 0.5 = (128.75)^2 \times 0.25 = 16,576.562 \times 0.25 = 4,144.1$$

or 4,145. This latter sample size is very large so it is unlikely that most research projects could obtain a sample with accuracy of ±2 percentage points with probability 0.99.

Conclusion. A few concluding points concerning the determination of sample size for estimation of a proportion are as follows.

1. The formula for determining sample size in the case of estimation of a proportion
$$n = \left(\frac{Z}{E}\right)^2 pq$$
has advantages over the formula for estimating a population mean in that the values of p and q can always be set to 0.5 each. This will always produce a sample size sufficient to produce the required accuracy E at whatever confidence level the researcher specifies. In the case of estimating the sample mean, the researcher needed some knowlege of the variability of the population being sampled – that is, an estimate of σ was required in order to determine sample size. In the case of a proportion, this is not necessary; a researcher can always use $p = q = 0.5$ and be sure this will produce a large enough sample size.

2. All of the above results apply to random sampling from a population. While researchers consider larger sample size to be better than smaller sample sizes, strictly speaking this may be the case only if the samples are random, or chosen using the principles of probability. If samples are judgment or snowball samples, large samples may not be all that much better than smaller samples.

   If other forms of probability samples are used, for example, cluster or stratified samples, formula such as that used in this section can be developed. But the formula in this section applies only to random sampling.

3. If a researcher considers the sample size too large when $p = q = 0.5$, different estimates of p and q can be used. In the example, if a researcher thinks that only 15% of the population oppose the legalization of marijuana, so that the researcher is willing to work with $\hat{p} = 0.85$ and $\hat{q} = 0.15$ when estimating $\sqrt{pq/n}$, the required sample size for (b) would be

$$n = \left(\frac{2.575}{0.02}\right)^2 0.85 \times 0.15 = (128.75)^2 \times 0.1275$$

$$n = 16,576.562 \times 0.1275 = 2,113.5$$

   or 2,114. This is much less than the earlier sample size of $n = 4,145$. The only danger here is that if the proportions supporting or opposing legalization of marijuana are closer to 0.5 than 0.85 and 0.15, then this sample size may produce a confidence interval estimate that has a sampling error greater than 0.02.

4. Given that $p = q = 0.5$ can always be used in order to determine sample size, it is possible to construct tables of required sample size for different confidence levels C and accuracy of estimate E. Table 8.8, p. 544 of the text is reproduced here as Table 1. Using $p = q = 0.5$ and the above formula, you should be able to calculate all the sample sizes in this table.

Table 1: Sample Sizes for a Proportion, Common Levels of Accuracy and Confidence

Level of	Confidence Level		
Accuracy (E)	90%	95%	99%
0.05	271	385	664
0.04	423	601	1,037
0.03	752	1,068	1,842
0.02	1,692	2,401	4,145
0.01	6,766	9,604	16,577

From Table 1, note that as the researcher is more demanding in terms of accuracy (smaller E), required sample size is greater. Similarly, as a researcher is more demanding in terms of requiring greater confidence that the intervals will contain the mean, sample size is again increased. In practice, the actual sample size selected is likely to be informed by the considerations of this section, but may depend more on the budget and time available for the researcher. With limited budget and time for a survey, a researcher may just have to live with the lesser accuracy associated with a smaller sample size.

Next section: Hypothesis testing – chapters 9 and 10.