Social Studies 201
Notes for November 20, 2006

Sample size for estimation of a mean – Section 8.6, pp. 530-540.

These notes discuss how to determine the appropriate size for a sample when estimating a population mean, given a particular accuracy and confidence level.

By using the central limit theorem, prior to obtaining the sample, it is possible to specify the sample size required to achieve a given degree of accuracy for an estimate of the mean. A confidence level for the interval estimate must be specified and the researcher must have some knowledge of the variability of the population from which the sample is drawn. Since a larger sample size generally requires more time and effort, costs more, and may disturb the population, a researcher wishes to select the smallest possible sample size consistent with obtaining the required accuracy and confidence level. Such a sample must also be a random sample – other methods of sampling may be associated with different required sample sizes.

Since the required sample size is usually large, the central limit theorem can be used to describe the distribution of sample means. This provides the researcher with a way to determine the variability of sample means prior to obtaining the sample.

Notation. The required accuracy for the estimate is denoted by E, so that the interval constructed will be $\bar{X} \pm E$ after the sample data have been obtained. Note that this is an interval of $\pm E$ on either side of the sample mean $\bar{X}$, so the interval width is $W = 2E$. The letter E is used here to denote "error," that is, sampling error. This amount E is equivalent to the sampling error of the sample. As before, the confidence level is $C\%$, with the corresponding value from the normal table given the symbol Z_C. That is, $\pm Z_C$ are the Z-values such that $C\%$ of the distribution is between them.

Derivation of the formula for sample size

If the mean μ of a population is to be estimated, the central limit theorem describes the distribution of sample means, $\bar{X}$. The theorem states that when random samples are drawn from a population with mean μ and standard deviation σ,

$$\bar{X} \text{ is Nor } \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

when a random sample from the population has more than thirty cases. From this theorem, it can be seen that a larger n produces a smaller standard error (standard deviation of the mean, $\sigma/\sqrt{N}$) than does a smaller sample size. That is, there are different normal distributions for each different sample size. The method described here involves selecting a normal distribution that will result in an interval estimate of required accuracy E, that is, an interval of $\bar{X} \pm E$.

Since the sample means are normally distributed, for confidence level C, the corresponding Z-value from the normal distribution is Z_C. C% of the normal distribution lies within Z_C standard deviations of the true mean μ. Since one standard deviation for the distribution of sample means is $\sigma/\sqrt{n}$, this means that Z_C standard deviations amount to

$$Z_C \frac{\sigma}{\sqrt{n}}.$$

From this, the distribution of sample means (with mean μ and standard deviation $\sigma/\sqrt{n}$), C% of the area under the distribution is within the interval

$$\mu \pm Z_C \frac{\sigma}{\sqrt{n}}.$$

C% of the sample means $\bar{X}$ also lie within these limits.

Now if the interval estimate is to be accurate to within $\pm E$, this implies that the sample mean must be within E of the population mean μ. What we need to match the accuracy of the estimate with the interval of the previous paragraph is a sample size such that:

$$\mu \pm Z_C \frac{\sigma}{\sqrt{n}}$$

and

$$\mu \pm E$$

match. This occurs when:

$$E = Z_C \frac{\sigma}{\sqrt{n}}.$$

This can be obtained by solving this expression for n. Rearranging and solving this expression for n (see p. 533 of the text), gives

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

where Z, rather than Z_C is used. If a random sample of the size specified by this formula is obtained, then the confidence interval estimate obtained by the researcher should be of accuracy E, that is, it should be approximately $\bar{X} \pm E$.

In order to see that it is practical to use this formula, all the terms on the right side of this equation can be obtained prior to the sample being obtained. That is, the accuracy of the estimate desired, E, can be specified by the researcher prior to conducting the sample. The Z-value can be determined from the table of the normal distribution once a confidence level is given. Finally, a researcher has to have some estimate of the variability of the population from which the sample is to be drawn, so that σ can be specified in the formula. Some guidelines concerning this are contained later in these notes, and on pp. 538-9 of the text.

Example – sample size to estimate mean income

Use the data above concerning the incomes of Saskatchewan females aged 35-44 who were employed full-time, from Problem Set 5. For this group, find the sample size required to obtain an estimate of mean income correct to within one thousand dollars, 19 in 20 times.

For this group, also find the sample size required to determined the mean income correct to within two thousand dollars, with 90% confidence.

Answer

As with any problem of this sort, the first step is to be clear concerning what is to be estimated. In this case, the parameter to be estimated is μ, the mean income for all Saskatchewan females aged 35-44 employed full-time. Since this is an estimate of a mean, and since we expect the required sample size to be reasonably large, the central limit theorem can be used to describe the distribution of sample means. As shown above, from this, the required sample size is

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

where E is the accuracy required of the estimate.

For the first part of the example, the accuracy required is one thousand dollars so, in units of thousands of dollars, the accuracy required is $E = 1$. Since the requirement is that the estimate be accurate 19 in 20 times, this is equivalent to 95% confidence. As a result, $Z = 1.96$, since 95% of the area under a normal distribution lies between $Z = -1.96$ and $Z = +1.96$. While the true standard deviation of income for all female workers is not known, the sample in problem 5 of Problem Set 5 provides an idea of the variability of income. The standard deviation of $s = 20.7$ thousand dollars can be used as an estimate of σ.

From these values the determination of sample size is

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

$$n = \left(\frac{1.96 \times 20.7}{1.00}\right)^2$$

$$n = 40.752^2 = 1,646.09$$

or a sample size of $n = 1,647$. A random sample of $n = 1,647$ female full-time employees, each of age 35-44, should provide an estimate of mean income correct to within ±\$1 thousand dollars, 19 in 20 times, that is with probability 0.95 or 95% confidence.

For an interval estimate to be accurate to within two thousand dollars, or $E = \$2$, the same formula is used, but with $E = 2$ replacing $E = 1$. For 90% confidence level, $Z = 1.645$ and the required sample size is

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

$$n = \left(\frac{1.645 \times 20.7}{2}\right)^2$$

$$n = 17.026^2 = 289.876$$

or a sample size of $n = 290$. A random sample of $n = 290$ female employees should provide a sample mean $\bar{X}$ that differs from the population mean μ by no more than \$2,000, with 90% confidence.

Additional notes on sample size

1. **Round up**. In the above example, where there were decimals for the sample size n, these were always rounded up to the next integer when reporting the required sample size. In order to specify a large enough sample size, the answer should always be rounded up to the next integer.

2. **Units**. When using the formula above, make sure that E and the estimate of σ are in the same units. In the example above, everything was converted into dollars, to ensure consistency.

3. **Trade-off**. There is often a trade-off between the budget for a survey and the accuracy of the results. A larger sample size produces greater accuracy but this may cost much more and take much more time and effort. As a result, a researcher may not be able to obtain the sample size specified by the formula, and may have to live with the less accurate results from a sample size smaller than desired.

4. **Factors associated with larger required** n. A careful look at the formula
$$n = \left(\frac{Z\sigma}{E}\right)^2$$
shows that the required sample size, n, increases as Z increases, σ increases, and E decreases. This can be summarized as follows:

   (a) A larger confidence level, $C\%$, produces a larger Z-value and results in a larger required sample size.

   (b) A more variable population, with larger σ, means a larger sample size is required to achieve the given level of accuracy. In contrast, populations where members are similar to each other in the characteristics being examined, do not require such large sample sizes to achieve the required accuracy of estimate.

   (c) The greater the accuracy required, the smaller is the value of E, and the larger the required sample size.

5. **Population size not important**. The required sample size does not depend on the size of the population from which the sample is being drawn, unless the required sample size is a large proportion of the population. Suppose the above formula leads to a sample size of 200, but the population size is 10,000 people. Then a random sample from this population gives the accuracy required. If the size of the population is 100,000, or one million, the sample size is the same – a random sample with a size of $n = 200$ is required in each case.

   The only exception to this is when the population is relatively small. Say the population size is 1,000, so the sample size recommended is $200/1,000 = 0.2$, or 20%, of the population size. In this case, the required sample size may be reduced somewhat, since the sample size is a considerable portion of the total population. But if the sample size is less than, say, 5% of the population size, the above formula holds. The reason for this apparent paradox are within probability theory – consult a text on the mathematical principles involved in sampling if you are interested in this issue.

6. **Estimate of** σ. In order to calculate required sample size, some estimate of σ, the variability of the population, is required. Some methods

of obtaining a prior estimate of σ are as follows (see pp. 538-9 for a fuller discussion).

(a) **Small sample**. As in the above example, a researcher may have a small sample and, from this, an initial idea of the variability of the population from which a larger sample is to be drawn.

(b) **Other studies and other populations**. Other researchers may have obtained surveys from a population, or similar population, and the sample standard deviations from these surveys may provide a reasonable estimate of σ.

(c) **Range**. Recall that the standard deviation may be close to one-quarter of the range for a variable. This is not very exact but, in the absence of much knowledge of the variability of a population, may provide a quick and rough estimate of σ.

(d) **Sampling method**. It may be possible to devise a sampling method so the sample size can be increased later, after the initial sample is drawn. That is, conduct an initial random sample and if the results are not accurate enough, randomly select more cases.

Last edited November 22, 2006.