

Appendix B.

DETERMINING ADEQUATE SAMPLE SIZE

As already stressed in this bulletin, sample selection is one of the most important determinants of the quality of survey results. One important component of sample selection is designing a procedure to insure random selection of respondents. Another key component is knowing approximately how large your sample must be to give accurate results from the survey.

Usually, the most critical data from a visitor survey are the expenditure figures. Generally, it is important to have a large enough sample so that the sampling error is kept to a minimum. Often we are interested in examining characteristics (such as expenditures) of a subgroup of the sample. The standard deviation on the mean expenditures for the subgroup may be much less than for the total sample, thus reducing the necessary sample size to achieve credible results. On the other hand, if the standard deviation for the subgroup is not much smaller than for the whole sample, then, the total sample size required may be much larger.

One equation for estimating sample size for a very large population (i.e., large total number of visitors) is:

$$n = 4X \text{ variance} / \text{limit squared}$$

where n is the sample size (Snedecor and Cochran). The variance is the standard deviation squared or the square of the difference between the observed value and the mean value for each observation summed and divided by one less than the number of observations. The limit is the amount we are willing to allow our sample mean to differ from the population mean at the 95 percent confidence level (Snedecor and Cochran).

Unless a previous study has been done and the variance on the key variable (typically per person per day expenditures) is known, variance can be approximated with a formula using the range of the key variable:

$$\text{estimated variance} = (1/16) (\text{range})^2$$

where the range is the difference between the highest and lowest value expected for the variable.

Let's go through some examples. In studies conducted in Arizona to date, we generally find that the standard deviation on mean expenditures, whether on a per person per day basis or on a total trip basis, are very large. In our example, let's suppose we know from a previous study that mean expenditures per person per day are \$50 with a standard deviation of \$50. The variance would then be the standard deviation squared or 2500. Multiplying this times four gives 10,000. Let's say that we will tolerate an estimate that is plus or minus \$5 (or 10%) of the real mean expenditure. Then, the limit squared is equal to 25. Our sample size is equal to 10,000/25= 400.

Usually, average expenditures are used to estimate total visitor expenditures for a region. Suppose that visitation to the region is known to be 200,000 visitor days. Total visitor expenditures would be 200,000 times \$50 or \$10,000,000. We know that this estimate is accurate approximately plus or minus 200,000 times \$5 or by \$1,000,000. Depending on how we use this information, this may or may not be accurate enough. If it is not, we will need to select a smaller limit, which will result in a larger sample size. Keep in mind that we typically do not know the actual number of visitors to the region.

So, our visitor number is an estimate too, which will increase the confidence interval or the amount our estimate may differ from the actual total expenditure.

As a second example, let's suppose that we do not have a previous study but we expect that total per person per day expenditures will be in the range of \$0 to \$200. Our estimate of variance is:

$$\text{estimated variance} = (1/16) (200-0)^2 = 2500$$

$$\text{sample size} = [4 (2500)] / 25 = 400.$$

This example uses the same limit or bound of \$5.00 as in the previous example.

These formulas work best when dealing with very large populations. If your population of visitors is small (less than a few thousand), then a more appropriate formula is:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} \quad \text{where } D = \frac{B^2}{4}$$

In this formula,

n is the sample size,

N is the estimated population size,

σ^2 is the variance,

B is the bound or limit on the error of the variable to be estimated

(Mendenhall

and Reinmuth). For other formulas for estimating sample size, read Chapter 12 in Ritchie and Goeldner.

