Confidence Levels and Sample Size

A statistical confidence level is defined using a statement like: Identify the average number of trips made by households of the specified type to within plus or minus five percent at a 95 percent confidence level. If we calculate from the survey data that a household type makes an average of 10.0 person trips per day, this statement would imply that the "true value" ranges between 9.5 and 10.5 trips per day (i.e., plus or minus five percent) 95 percent of the time.

The number of observations that are needed to make a statement such as this depends on the statistical variability of the population group being measured. If the number of trips or mode preferences of a particular population group are similar, the number of observations that is required to calculate the trip rate or mode share will be smaller. If the population group includes people with widely different behavior, the number of observations that is required to calculate average statistics for this group will be larger.

The basic parameter used to determine the minimum number of observations (sample size) required for a given level of statistical reliability is called the Coefficient of Variation (CV). The Coefficient of Variation is the Standard Deviation divided by the Mean of a particular measure for a specific population group. Mean is the statistical term for average defined as

$$\text{Mean} = (\sum x_i) / n$$

where "x" is the value of each observation and "n" is the population size.

Standard Deviation is the parameter that measures the variability of individual observations around the mean. It is calculated as:

$$\text{Standard Deviation} = ((\sum (x_i - \text{mean})^2) / n$$

Or, in English, the sum of the squared deviations of each observation from the mean divided by the population size.

Given these statistics, the Coefficient of Variation is:

$$\text{Coefficient of Variation} = \text{Standard Deviation} / \text{Mean}.$$

If you know the Coefficient of Variation, the sample size is defined by the following equation:

$$\text{Sample size} = CV^2 * CF^2 / \text{Error}^2$$

where:

CV is the Coefficient of Variation
Error is the maximum error in the estimate (expressed as 0.10 = 10% error)
CF is the confidence factor based on the target confidence level
CF = 1.645 for a 90 percent confidence level
CF = 1.960 for a 95 percent confidence level

CF = 2.576 for a 99 percent confidence level

(Note: CF is derived from the standard normal distribution.)

By using the formula above and known parameters for your population, you can determine the minimum sample size required to achieve a given confidence level and accuracy. The trick, of course, is getting the known parameters, because if you knew them you would not need to do the survey.

A common solution to this shortcoming is to borrow or approximate the parameters from previous surveys or similar regions. The following table includes data from a travel survey conducted in the San Francisco Bay area. The table shows the range of sample size values that are required to accurately estimate various trip rate statistics. If you are interested in estimating the total trip rate per household at 95 percent confidence and plus or minus 10 percent error, you only need to survey 277 households in the San Francisco region. If, on the other hand, you are interested in estimating the transit trip rate per household at 95 percent confidence and plus or minus 2 percent error, you would need to survey 61,236 households.

Trip Rate	San Francisco Travel Survey			Sample Size Required 95% Confidence by Percent Error		
	Mean	S.D.	C.V.	10%	5%	2%
Total Trips / HH	8.713	7.399	0.849	277	1,108	6,926
Vehicle Trips / HH	5.231	5.009	0.958	352	1,409	8,806
Transit Trips / HH	0.558	1.409	2.525	2,449	9,798	61,236
HBW / HH	1.890	1.883	0.996	381	1,525	9,533
HBSH / HH	2.274	2.778	1.222	573	2,293	14,333
HBSR / HH	1.262	2.034	1.612	998	3,992	24,948
HBSK / HH	0.952	1.883	1.978	1,503	6,012	37,573
NHB / HH	2.335	3.351	1.435	791	3,165	19,780

Notice that there is no explicit consideration of the population size in the sample size calculation. If the Coefficient of Variation is the same, the population could be 10 million or 100,000, and the sample size would be the same. On the other hand, the Coefficient of Variation is likely to be larger for more diverse population groups which in turn leads to larger sample sizes. As the example above demonstrates, however, the overall accuracy of the Coefficient of Variation calculation can have a huge impact on the sample size. This suggests that borrowing statistics from other regions or previous years can introduce significant errors into the sample size calculations and the resulting precision of the estimates can differ greatly from what you intended.