Tools of the Trade:

STRATEGY FOR DETERMINING SAMPLE SIZE

One of the most frequently asked questions in survey statistics is, "What size sample do I need?". A sample that is too large may waste time, money, and resources. A sample that is too small may lead to inaccurate results. Most people expect a simple answer to this question but in reality the methods used to determine the sample size are influenced by a number of factors which include the precision, confidence level, expected eligibility rate, expected completion rate, expected variability of the results, survey design and the way the data will be analyzed. In the following paragraphs, a method to determine sample size will be presented using the Synar survey as an example. This method can be used for a number of different designs with minor modifications.

Before we begin, some background information about the Synar survey must be established. The survey uses a stratified and clustered design with the sampling units defined as retail outlets that sell cigarettes over-the-counter. The objective of the survey is to produce a statewide estimate for the rate that retail outlets sell cigarettes to minors. The sampling design is two stages with the first stage involving the selection of clusters within stratum using probability proportionate to size of the cluster. The second stage involves selecting 17 outlets from the sampled clusters. The survey protocol requires youth to enter the outlet and attempt to purchase cigarettes. The Synar survey is federally regulated and must adhere to federal requirements that aren't normally present in other surveys. Lastly, the survey does not analyze sub-populations. If sub-populations are the goal, more sample would have to be added.

Determining Sample Size
Step 1: Set precision and confidence levels.
Step 2: Calculate the effective sample size.
Step 3: Determine design effect and adjust sample size.
Step 4: Estimate eligibility & completion rates; adjust sample size.
Step 5: Adjust sample for variability and equal sample allocation.

Precision and Confidence Levels
The first step is to decide the level of precision and confidence to use for the survey. The confidence level refers to the probability value associated with a confidence interval. A confidence level is necessary when the results will be presented using confidence intervals. Common confidence levels are 90 or 95 percent for either a two-sided or one-sided interval. The confidence level is directly proportional to sample size, i.e. more sample is required for higher levels of confidence. The Synar survey is required by a federal regulation to use a one-sided 95% confidence interval.

Precision refers to the survey error. The amount of error that you are willing to allow must be set prior to the start of the survey. The precision level is inversely proportional to sample size. The lower the precision level is set, the more sample will be needed to reach that goal. Obviously survey error is undesirable and every survey should strive for as little error as possible when conducting the survey. While determining the sample size, the level should be set as low as possible without increasing the sample size to more than your resources will allow. A federal regulation requires the Synar error to be less than or equal to three percent.

Effective sample size
Calculate the effective sample size using the precision and confidence levels established in step 1. The effective sample size is the minimum sample size needed to meet the precision and confidence requirements under a simple random design. Ignoring the finite population correction factor, the

effective sample size equation is derived from the precision equation "$w = z(s.e.)$" and the standard error equation "$s.e. = \frac{\sqrt{p(1-p)}}{\sqrt{n_e}}$". The effective sample size equation is:

$$n_e = \left(\frac{z}{w}\right)^2 p(1-p) \qquad \textbf{(E1)}$$

Where *z* is the critical value of the standard normal distribution for a one-sided 95% confidence interval, *w* is the precision and *p* is the expected rate for the variable being measured.

For the Synar survey, *z* is 1.645 according to a critical value table and *w* is .03 since the precision is set at three percent. The expected sale rate must be estimated and should be set at a level that will produce the highest number of sample that can be afforded. It is better to estimate too high than to estimate too low. As displayed in the equation, the highest amount of sample occurs when the sale rate is 50 percent. The Synar survey isn't accepted if the statewide estimate is greater than 23 percent (20% + 3% error = 23%). Therefore, 23 percent was used for the value of *p* since it is the closest we can get to 50 percent. The effective sample size for the 2007 Synar survey is determined by substituting these numbers into **(E1)** and solving for n_e:

$$n_e = \left(\frac{z}{w}\right)^2 p(1-p) = \left(\frac{1.645}{.03}\right)^2 \times .23(1-.23) = 532.5 \approx 533$$

Design Effect

Since most surveys are not truly random, the sample size must be adjusted to account for deviations from a simple random design. These deviations are known as the design effect (DEFF). The higher the design effect, the larger the sample size needs to be. The DEFF is calculated by dividing the variance of the survey with the complex design by the variance of a simple random sample of the same size. Since the DEFF cannot be calculated before the survey is conducted, the DEFF has to be estimated by gathering information from surveys with similar designs and calculating their DEFFs. With this information and the knowledge of your own survey's circumstances, make an informed estimate for the DEFF. The DEFF equation is:

$$n_t = \mathrm{Deff_h} \times n_e \qquad \textbf{(E2)}$$

Where n_e is the effective sample size and $Deff_h$ is the highest DEFF of the last 3 surveys. Solving **(E2)** with $Deff_h$ =1.29, the increase sample size for design effect is:

$$n_t = \mathrm{Deff_h} \times n_e = 1.29 \times 533 = 687.57 \approx 688$$

Eligibility and Completion Rates

The sample size is adjusted to account for ineligible sample and non-completions. The eligibility and completion rates have to be estimated. The eligibility rate is the percent of the sample that is eligible and the completion rate is the percent of eligible sample completed. The best estimates come from previous versions of the same survey or other surveys with similar designs. If this information is available it should be used. If past surveys aren't available, then make your best guess. It is better to underestimate these rates than to overestimate them. If you assume your completion/eligibility rates will be high, you will need less sample but if they actually turn out to be low, you won't meet your precision goal because the sample will be too small.

For the Synar survey, the rates of the past 3 surveys were reviewed and 68 percent was used for the eligibility rate and 80 percent for the completion rate. The equation used to increase the sample to account for ineligible sample and non-completions is:

$$n_{ec} = \frac{n_t}{r_l r_c} \qquad \textbf{(E3)}$$

Where r_l = lowest eligibility rate of historical Synar surveys of a similar design (2004-2006) and r_c = 80 percent, the lowest completion rate allowed by the Federal government. Solving **(E3)** the sample size is increased to:

$$n_{ec} = \frac{n_t}{r_l r_c} = \frac{688}{(.68)(.80)} = 1264.7 \approx 1265$$

Original Sample Size

Lastly, the sample size needs to be adjusted for variability and equal sample allocation. This is the last step and the result is called the original sample size. Variability refers to the variability between clusters. One method to test variability is to simulate survey results repeatedly, using a computer programming language such as SAS (Statistical Analysis System), and study the results. SAS programs were used to test the variability for the Synar survey. If SAS isn't available, manually simulate the results a few times and examine them. It is recommended to add to the sample if you can afford it.

If the survey is clustered, it will have to be adjusted to fit the design so it can be equally distributed among the clusters. Consider a clustered design with 17 samples per cluster. The total sample size cannot be 1,265 because 1,265 cannot be divided by 17 evenly. Therefore, the total sample has to be increased so each cluster has 17 samples.

After testing, it was determined that more sample was needed for the Synar Survey. The amount of sample was determined by viewing the program results and spreading it out among the clusters and random areas of the design. n_A = 287 sampling units were added. The original sample size is determined using the following equation:

$$n_o = n_{ec} + n_A \qquad \textbf{(E4)}$$

By solving this equation, it was determined that the 2007 Synar survey needed the following number of sample size:

$$n_o = n_{ec} + n_A = 1265 + 287 = 1552$$

Final Equation

Although the above equations can be combined into one equation, each of the variables will still have to be determined individually. Combining the equations gives:

$$n_o = \left(\frac{Deff_h}{r_l r_c}\right) n_e + n_A \qquad \textbf{(E5)}$$

Solving the equation, we obtain a number that differs slightly due to rounding error.

$$n_o = \left(\frac{Deff_h}{r_l r_c}\right) n_e + n_A = \left(\frac{1.29}{(.68)(.80)}\right) 533 + 287 \approx 1551$$