If one is estimating a population mean (numerical average) the formula for sample size is very similar. In the case of average number of physician visits per year, if we wanted to estimate the population average within plus or minus .5 visits (which would be within 10 percent of a mean of 5.0) with 95 percent confidence, we would solve for N in the equation: $1.96\ SD_p/\sqrt{N} = .5$. The problem, of course, is that we must have an estimate of the standard deviation in the population (SD_p) before the sample is drawn in order to solve for N, and this is not always available. As a result, sample size is often computed using the formula for a proportion, frequently using the conservative estimate of $p = .5$ unless better information is available. Also, it may be possible to estimate the population standard deviation based on expert knowledge of the range, mode, and other characteristics of the distribution (6). After determining a sample size based on formulas such as these, it is usually a good idea to increase the sample size as much as possible to allow for nonresponse and other potential problems. Again, it is assumed in the calculations above that a simple random sample will be drawn.

If it has been determined that a sample size of 500, for example, is required then the process of drawing a simple random sample may be relatively easy. First we must have a list of each individual (person, household, etc.) in the population and be sure that each individual is listed only once. This may be impossible in some cases, for example if a sample of all nonwhites in Wake County were desired, then cluster or area sampling might be required. To the degree that persons omitted from a list differ from the rest of the population with respect to the characteristics being studied, a biased sample will be obtained. Telephone directories, for example, may be biased in that lower-income groups are likely to be underrepresented; also, persons with unlisted numbers may be an unusual population group. In some cases it may be desirable to redefine the population to conform to the list that is available. Assuming, however, that a complete list has been obtained, one would first associate a number with each position on the list. If there were 10,000 individuals on the list, then for a sample size of 500 one would randomly choose 500 numbers between 1 and 10,000 and the individuals corresponding to these numbers would fall in the sample. If numbers are chosen that have already been selected, these repetitions are omitted until we have 500 individuals in the sample. Random numbers may be chosen from a random number table, or some calculators and most computers will generate random numbers.

If a list is extremely long, or if a sample is to be drawn from administrative records or files, **systematic sampling** may be much easier than the procedure described above. In systematic sampling, instead of using a table of random numbers one simply goes down a list taking every *kth* individual, starting with a randomly selected case among the first *k* individuals. Thus, if one wanted to select a sample of 180 persons from a list of 3600, we would take every twentieth in the list. The first choice, however, must be determined randomly. If the number 5 were selected between 1 and 20, the sample would consist of individuals numbered 5, 25, 45, 65, etc. If the ordering used in compiling the list or in a set of records can be considered to be essentially random with respect to the variables being measured, then the standard error for a systematic sample will be equivalent to that for a simple random sample. Many lists or files are alphabetical in order, but in most cases alphabetical ordering is irrelevant to the variables being studied. Systematic sampling may result in a standard error larger than that for a simple random sample of the same size if the individuals have been ordered so that a trend occurs or if the list has some periodic or cyclical characteristic that corresponds to the sampling fraction. (7)

Synopsis

The concept of random errors of measurement has been presented, in the context of sampling from a population. Formulas were presented for calculating these errors of measurement in terms of the standard error of estimates of means and proportions. Procedures for selecting simple random and systematic samples were also described. Part 2 of this statistical primer will discuss errors in measures based on a "complete count" and present formulas for calculating standard errors of simple rates and of the difference between two rates.

This paper only touches on the basics of sampling and measurement error, and the reader interested in computing standard errors for more complex measures or in carrying out more complex sample designs should consult the references listed here or a sampling specialist. The simple examples presented here are not always encountered in practice. Frequently one will study more than one variable at a time in a sampling project, and this may complicate the picture considerably. It is usually best to draw a sample size that will allow adequate estimation of the variable with the largest standard error.

It is hoped that this overview will provide the reader with enough background to understand the general problems involved in a sampling project and to seek further help and consultation as necessary.