Sampling Methods and Practice

Richard L. Scheaffer
University of Florida

The topic of Sampling Methods and Practice fits well with that of Categorical Data Analysis. Indeed, most survey questionnaires produce categorical data by asking for Yes/No or Agree/Disagree responses. Typically, the reports on the surveys present proportions and percentages of the responses. In this section, we will consider the topic of Survey Sampling, its important features and appropriate techniques of analysis.

Sample Surveys and Experiments

A sample survey differs from an experiment in several important ways. A sample survey is characterized by

- a clearly specified population
- a sample selected by a random process from that population
- the goal of estimating some population parameters

An experiment is characterized by

- a treatment or treatments of interest
- some form of control, either a control group or another treatment
- randomized assignment of the experimental unit (subject) to a treatment
- the goal of establishing treatment differences, if they exist.

The goals of a sample survey and an experiment are very different. The role of randomization also differs. In both cases, without randomization there can be no inference. Without randomization, the researcher can only describe the observations and cannot generalize the results. In the sample survey, randomization is used to reduce bias and to allow the results of the sample to be generalized to the population from which the sample was drawn. In an experiment, randomization is used to balance the effects of confounding variables.

Some Terminology

Element: An element is an object on which a measurement is made. This could be a voter in a precinct, a product as it comes off the assembly line, or a plant in a field that has either bloomed or not.

Population: A population is a collection of elements about which we wish to make an inference. The population must be clearly defined before the sample is taken.

Sampling Units: Sampling units are nonoverlapping collections of elements from the population that cover the entire population. The sampling units partition the population of interest. The sampling units could be households or individual voters.

Frame: A frame is a list of sampling units.

Sample: A sample is a collection of sampling units drawn from a frame or frames. Data are obtained from the sample and are used to describe characteristics of the population.

Example 1 Suppose we are interested in what students in a particular high school think about the drilling for oil in our national wildlife preserves. The elements are the high school students and the population is the students who attend this high school. The sampling units could be the students as individuals with the frame the alphabetical listing of all students enrolled in the school. The sampling units could be homerooms, since each student has one and only one homeroom, and the frame the class list for homerooms.

Example 2 Suppose we are interested in what voters in a particular precinct think about the drilling for oil in our national wildlife preserves. The elements are the registered voters in the precinct. The population is the collection of registered voters. The sampling units will likely be households in which there may be several registered voters. The frame is a list of households in the precinct.

When the population is the residents of a city, the frame will commonly be the city phone book. However, not everyone in the city has their phone listed in the phone book. In this situation, the frame does not match the population. A survey conducted from the frame of the phone book would likely suffer from undercoverage bias.

Probability Samples

Sample designs that utilize planned randomness are called *probability samples*. The most fundamental probability sample is the *simple random sample*. In a simple random sample, a sample of n sampling units is selected in such a way that each sample of size n has the same chance of being selected. In practice, other more sophisticated probability sampling methods are commonly used, but most of the statistical theory for the introductory course in statistics is based on the simple random sample.

First, we will define a stratified random sample, a systematic sample, and a cluster sample.

Stratified Random Sample: A stratified random sample is one obtained be separating the population elements into non-overlapping groups, called strata, and then selecting a simple random sample from each stratum. (Scheaffer, Mendenhall, and Ott, *Elementary Survey Sampling, 5th edition*, page 125).

Systematic Sample: A systematic sample is obtained by randomly selecting at random one element from the first k elements in the frame and every k^{th} element thereafter. This is known as a 1-in-k systematic sample. (Scheaffer, Mendenhall, and Ott, *Elementary Survey Sampling, 5th edition*, page 252).

Cluster Sample: A cluster sample is a probability sample in which each sampling unit is a collection, or cluster, of elements. (Scheaffer, Mendenhall, and Ott, *Elementary Survey Sampling, 5th edition*, page 289).

Dick Scheaffer, in *Elementary Survey Sampling* (p. 407-408) gives and excellent overview and comparison of the different standard methods of conducting probability samples. We include this discussion with only slight modification.

COMPARISONS AMONG THE DESIGNS AND METHODS

Simple random sampling is the basic building block and point of reference for all other designs discussed in this text. However, few large-scale surveys use only simple random sampling, because other designs often provide greater accuracy or efficiency or both.

Stratified random sampling produces estimators with smaller variance than those from simple random sampling, for the same sample size, when the measurements under study are homogeneous within strata but the stratum means vary among themselves. The ideal situation for stratified random sampling is to have all measurements within any one stratum equal but have differences occurring as we move from stratum to stratum.

Systematic sampling is used most often simply as a convenience. It is relatively easy to carry out. But this form of sampling may actually be better than simple random sampling, in terms of bounds on the error of estimation, if the correlation between pairs of elements within the same systematic sample is negative. This situation will occur, for example, in periodic data if the systematic sample hits both the high points and the low points of the periodicities. If, in contrast, the systematic sample hits only the high points, the results are very poor. Populations that have a linear trend in the data or that have a periodic structure that is not completely understood may be better sampled by using a stratified design. Economic time series, for example, can be stratified by quarter or month, with a random sample selected from each stratum. The stratified and the systematic sample both force the sampling to be carried out along the whole set of data, but the stratified design offers more random selection and often produces a smaller bound on the error of estimation.

Cluster sampling is generally employed because of cost effectiveness or because no adequate frame for elements is available.

However, cluster sampling may be better than either simple or stratified random sampling if the measurements within clusters are heterogeneous and the cluster means are nearly equal. The ideal situation for cluster sampling is, then, to have each cluster contain measurements as different as possible but to have the cluster means equal. This condition is in contrast to that for stratified random sampling in which strata are to be homogeneous but stratum means are to differ.

Another way to contrast the last three designs is as follows. Suppose a population consists of $N = nk$ elements, which can be thought of as k systematic samples each of size n. The nk elements can be thought of as n clusters of size k, and the systematic sample merely selects one such cluster. In this case the clusters should be heterogeneous for optimal systematic sampling. By contrast, the nk elements can also be thought of as n strata of k elements each, and the systematic sample selects one element from each stratum. In this case the strata should be as homogeneous as possible, but the stratum means should differ as much as possible. This design is consistent with the cluster formulation of the problem and once again produces an optimal situation for systematic sampling. So we see that the three sampling designs are different, and yet they are consistent with one another with regard to basic principles.

The Need for Probability Samples

Consider the table shown below of the accuracy in the final Gallup Presidential Polls from 1936 to 1984.

Gallup Poll Accuracy

Year	Gallup Final Survey	Election Result	% Error
1936	55.7% Roosevelt	62.5% Roosevelt	6.8%
1940	52.0% Roosevelt	55.0% Roosevelt	3.0%
1944	51.5% Roosevelt	52.3% Roosevelt	0.8%
1948	44.5% Truman	49.9% Truman	5.4%
1952	51.0% Eisenhower	55.4% Eisenhower	4.4%
1956	59.5% Eisenhower	57.8% Eisenhower	1.7%
1960	51.0% Kennedy	50.1% Kennedy	0.9%
1964	64.0% Johnson	61.3% Johnson	2.7%
1968	43.0% Nixon	43.5% Nixon	0.5%
1972	62.0% Nixon	61.8% Nixon	0.2%
1976	48.0% Carter	50.0% Carter	2.0%
1980	47.0% Reagan	50.8% Reagan	3.8%
1984	59.0% Reagan	59.2% Reagan	0.2%

Source: G. Gallup, Jr. *The Gallup Poll, Public Opinion 1984*. Copyright © 1985, Scholarly Resources Inc., Wilmington, DE. From Scheaffer, Mendenhall, Ott, *Elementary Survey Sampling, 5th Edition*, Duxbury Press.

Prior to 1948, the Gallup Poll used a quota sampling technique, which is not a probability sample. They had sought to find a representative group that matched the demographics of the country. Although the resulting sample did accurately represent the demographics of the country, it incorrectly predicted that Dewey would beat Truman in the election. Quota sampling failed. The samples taken after 1948 were probability samples. Even though the number of people in the sample was smaller than for polls used prior to 1948, the errors are generally much smaller.

Sources of Errors in Surveys

Statistician Robert Gross of the University of Michigan has categorized the kinds of errors in surveys into errors of non-observation and errors of observation.

Errors of non-observation include sampling error, error in coverage, and errors due to non-response.

- Sampling error is the "natural" error that is a part of any sampling process. If the sampling process were repeated a number of times, the results would differ each time, producing a variation in the estimates of the population parameters.

- Coverage error results when the frame does not match the population. For example, if the frame is the town phone book, then people with unlisted numbers and those without phones will be missing from the frame.

- Non-response error is a result of elements in the frame that have died, moved away, refuse to participate, or otherwise are missing from the sample.

Errors of observation include interviewer error, respondent error, measurement error, and errors in data collection.

- Interviewer error is a result of the interaction between the interviewer and the subject being interviewed. Most people who agree to an interview do not want to appear disagreeable and will tend to side with the view apparently favored by the interviewer, especially on questions for which the respondent does not have a strong opinion. Reading a question with inappropriate emphasis or intonation can force a response in one direction or another. Interviewers of the same gender, racial, and ethnic groups as those being interviewed are, in general, slightly more successful.

- Respondent error is a result of the differing abilities of the respondents in a sample to answer correctly the questions asked. Most respondent errors are unintentional and are due to either recall bias (the respondent does not remember correctly) or prestige bias (the respondent exaggerates). At times, respondent error may be due to intentional deception (the respondent will not admit breaking a law or has a particular gripe against an agency).

- Measurement error occurs when inaccurate responses are caused by errors of definition in survey questions. For example, what does the term *unemployed* mean? Should the unemployed include those who have given up looking for work, teenagers who cannot find summer jobs, and those who lost part-time jobs? Does *education* include only formal schooling or technical training, on-the-job classes and summer institutes as well? Items to be measured must be precisely defined and be unambiguously measurable.

- Errors in data collection occur in all surveys. The most commonly used methods of data collection in sample surveys are personal interviews and telephone interviews. These methods, with appropriately trained interviewers and carefully planned callbacks, commonly achieve response rates of 60% to 75%. The procedure usually requires the interviewer to ask prepared questions and to record the respondent's answers.

  The primary advantage of these interviews is that people will usually respond when confronted in person. However, if the interviewers are not thoroughly trained, they may deviate from the required protocol, thus introducing a bias into the sample data. Any movement, facial expression, or statement by the interviewer can affect the response obtained. Errors in recording the response can also lead to erroneous results.

  A major problem with telephone surveys is the establishment of a frame that closely corresponds to the population. Telephone directories have many numbers that do not belong to households, and many households have unlisted numbers. A technique that avoids the problem of unlisted numbers is random digit dialing. In this method, a telephone exchange number (the first three digits of the seven-digit number) is selected, and then the last four digits are dialed randomly until a fixed number of households of a specified type are reached.

  A mailed questionnaire sent to a specific group of interested persons can achieve good results, but, response rates for this type of data collection are generally so low that all reported results are suspect. Nonresponse can be a problem in any form of data collection, but since we have the least contact with respondents in a mailed questionnaire, we frequently have the lowest rate of response. The low response rate can introduce a bias into the sample because the people who answer questionnaires may not be representative of the population of interest. To eliminate some of this bias, investigators frequently contact the nonrespondents through follow-up letters, telephone interviews, or personal interviews.

Steps in Planning a Survey

(modified from Scheaffer, et al. *Elementary Survey Sampling, 5th Ed*., 1996. p. 68-70)

1. *Statement of objectives*. State the objectives of the survey clearly and concisely and refer to these objectives regularly as the design and the implementation of the survey progress. Keep the objectives simple enough to be understood by those working on the survey and to be met successfully when the survey is completed.

2. *Target population.* Carefully define the population to be sampled. If adults are to be sampled, then define what is meant by *adult* (all those over the age of 18, for example) and state what group of adults are included (all permanent residents of a city, for example). Keep in mind that a sample must be selected from this population and define the population so that sample selection is possible.

3. *The frame.* Select the frame (or frames) so that the list of sampling units and the target population show close agreement. Keep in mind that multiple frames may make the sampling more efficient. For example, residents of a city can be sampled from a list of city blocks coupled with a list of residents within blocks.

4. *Sample design.* Choose the design of the sample, including the number of sample elements, so that the sample provides sufficient information for the objectives of the survey.

5. *Method of measurement.* Decide on the method of measurement, usually one or more of the following methods: personal interviews, telephone interviews, mailed questionnaires, or direct observations.

6. *Measurement instrument.* In conjunction with step *5*, carefully specify how and what measurements are to be obtained. If a questionnaire is to be used, plan the questions so that they minimize nonresponse and incorrect response bias.

7. *Selection and training* of *field-workers.* After the sampling plan is clearly and completely set up, someone must collect the data. Those collecting data, the field-workers, must be carefully taught what measurements to make and how to make them. Training is especially important if interviews, either personal or telephone, are used because the rate of response and the accuracy of responses are affected by the interviewer's personal style and tone of voice.

8. *The pretest.* Select a small sample for a pretest. The pretest is crucial because it allows you to field-test the questionnaire or other measurement device, to screen interviewers, and to check on the management of field operations. The results of the pretest usually suggest that some modifications must be made before a full-scale sampling is undertaken.

9. *Organization* of *fieldwork.* Plan the fieldwork in detail. Any large-scale survey involves numerous people working as interviewers, coordinators, or data managers. The various jobs should be carefully organized and lines of authority clearly established before the survey is begun.

10. *Organization of data management.* Outline how each piece of datum is to be handled for all stages of the survey. Large surveys generate huge amounts of data. Hence, a well-prepared data management plan is of the utmost importance. This plan should include the steps for processing data from the time a measurement is taken in the field until the final analysis is completed. A quality control scheme should also be included in

the plan in order to check for agreement between processed data and data gathered in the field.

11. *Data analysis.* Outline the analyses that are to be completed. Closely related to step 10, this step involves the detailed specification of what analyses are to be performed. It may also list the topics to be included in the final report.

12. *Final Report.* The final report should match the stated objectives in step 1. Considering the final report before the survey is conducted may be helpful in determining what items are to be measured in the survey.

13. *Recapitulation.* After the final report is completed, you should consider what changes should be made if/when the survey is repeated. Most surveys are conducted periodically. It is important to keep track of what went well and what difficulties occurred.

Simple Random Sampling

Suppose the observations $y_1, y_2, \ldots y_n$ are to be sampled from a population with mean m, standard deviation s, and size N in such a way that every possible sample of size n has an equal chance of being selected. Then the sample $y_1, y_2, \ldots y_n$ was selected in a simple random sample. If the sample mean is denoted by $\bar{y}$, then we have

$$E(\bar{y}) = m$$

and

$$V(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N-1}\right).$$

The term $\left(\frac{N-n}{N-1}\right)$ in the above expression is known as the finite population correction factor. For the sample variance s^2, it can be shown that

$$E(s^2) = \left(\frac{N}{N-1}\right)s^2.$$

When using s^2 as an estimate of s^2, we must adjust with $s^2 \approx \left(\frac{N-1}{N}\right)s^2$. Consequently, an unbiased estimator of the variance of the sample mean is given by

$$\hat{V}(\bar{y}) = \frac{\left(\frac{N-1}{N}\right)s^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{s^2}{n}\left(\frac{N-n}{N}\right).$$

As a rule of thumb, the correction factor $\left(\frac{N-n}{N}\right)$ can be ignored if it is greater than 0.9, or if the sample is less than 10% of the population.

As an example, consider the finite population composed of the $N=4$ elements $\{0, 2, 4, 6\}$. For this population $m=3$ and $s^2=5$. Simple random samples, without replacement, of size $n=2$ are selected from this population. All possible samples along with their summary statistics are listed below.

Sample	Probability	Mean	Variance
$\{0, 2\}$	1/6	1	2
$\{0, 4\}$	1/6	2	8
$\{0, 6\}$	1/6	3	18
$\{2, 4\}$	1/6	3	2
$\{2, 6\}$	1/6	4	8
$\{4, 6\}$	1/6	5	2

(1) The expected value of the sample means is

$$E(\bar{y}) = \sum_{i=1}^{6} \bar{y}_i \cdot p(\bar{y}_i) = \left(\frac{1}{6}\right)(1+2+3+3+4+5) = 3.$$

Notice that $E(\bar{y}) = m$.

(2) The variance of the sample means is

$$V(\bar{y}) = E(\bar{y}^2) - (E(\bar{y}))^2 = E(\bar{y}^2) - (3)^2. \quad \text{So}$$

$$E(\bar{y}^2) = \sum_{i=1}^{6} \bar{y}_i^2 \cdot p(\bar{y}_i^2) = \left(\frac{1}{6}\right)(1^2 + 2^2 + 3^2 + 3^2 + 4^2 + 5^2) = \frac{64}{6}$$

and

$$V(\bar{y}) = \frac{64}{6} - 9 = \frac{5}{3}$$

We see in this example that $V(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N-1}\right) = \left(\frac{5}{2}\right)\left(\frac{4-2}{4-1}\right) = \left(\frac{5}{2}\right)\left(\frac{2}{3}\right) = \frac{5}{3}$.

(3) The expected value of the sample variances is

$$E(s^2) = \sum_{i=1}^{6} s_i^2 \cdot p(s_i^2) = \left(\frac{1}{6}\right)(2+8+18+2+8+2) = \frac{20}{3}.$$

Again, we see that $E\left(s^2\right)=\left(\frac{N}{N-1}\right)s^2=\left(\frac{4}{3}\right)(5)=\frac{20}{3}$, as the theory states must be true.

Estimation of a Population Mean

If we are interested in estimating a population mean from a simple random sample, we have

$$\hat{m}=\overline{y}=\frac{\sum_{i=1}^{n}y_i}{n}.$$

If we are interested in estimating a population variance from a simple random sample, we have

$$\hat{V}\left(\overline{y}\right)=\frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

where

$$s^2=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{n-1}.$$

The margin of error is 2 standard errors, so

$$2\sqrt{\hat{V}\left(\overline{y}\right)}=2\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}.$$

Estimation of a Population Proportion

If each observation in the sample is coded 1 for "success" and 0 for "failure", the sample mean becomes the sample proportion. In addition, we have

$$\frac{s^2}{n}=\frac{\hat{p}\left(1-\hat{p}\right)}{n-1},$$

where $\hat{p}$ denotes the sample proportion. To see this, recall that $s^2=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{n-1}$, so

$$\left(n-1\right)s^2=\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2=\sum_{i=1}^{n}\left(y_i^2-2y_i\overline{y}+\overline{y}^2\right)=\sum_{i=1}^{n}\left(y_i^2\right)-2\overline{y}\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}\overline{y}^2.$$

Since $\overline{y}=\frac{\sum_{i=1}^{n}y_i}{n}$, we have $n\,\overline{y}=\sum_{i=1}^{n}y_i$. Also, since each y_i is either 0 or 1, we have $\sum y_i^2=\sum y_i$ and $\overline{y}=\hat{p}$.

Then

$$\sum_{i=1}^{n}\left(y_i^2\right)-2\overline{y}\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}\overline{y}^2=\sum_{i=1}^{n}y_i-2n\overline{y}^2+n\overline{y}^2=n\overline{y}-n\overline{y}^2=n\hat{p}-n\hat{p}^2=n\hat{p}\left(1-\hat{p}\right).$$

So, we have $\left(n-1\right)s^2=n\hat{p}\left(1-\hat{p}\right)$ or equivalently, $\frac{s^2}{n}=\frac{\hat{p}\left(1-\hat{p}\right)}{n-1}$.

Using the formulas for the mean and the equality above, we can determine the estimator of the population proportion, of the variance of $\hat{p}$, and the margin of error for the proportion.

The estimator of the population proportion is $\hat{p}=\overline{y}=\frac{\sum_{i=1}^{n}y_i}{n}$.

The estimated variance of $\hat{p}$ is $\hat{V}\left(\hat{p}\right)=\frac{\hat{p}\left(1-\hat{p}\right)}{n-1}\left(\frac{N-n}{N}\right)$.

The margin of error of estimation is $2\sqrt{\hat{V}\left(\hat{p}\right)}=2\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n-1}\left(\frac{N-n}{N}\right)}$.

Estimating the Population Total

Finding an estimate of the population total is meaningless for an infinite population. However, for a finite population, the population total is a very important population parameter. For example, we may want to estimate the total yield of corn in Iowa, or the total number of apples in an orchard. If we know the population size N and the population mean m, then the total t is just $t=Nm$.

So, the estimator of the population total t is $\hat{t}=N\overline{y}=\frac{N\sum_{i=1}^{n}y_i}{n}$.

The estimated variance of t is $\hat{V}\left(\hat{t}\right)=\hat{V}\left(N\overline{y}\right)=N^2\cdot\hat{V}\left(\overline{y}\right)=N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)$.

Finally, the margin of error of estimation for t is

$$2\sqrt{\hat{V}\left(N\overline{y}\right)}=2\sqrt{N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)}=2Ns\sqrt{\frac{1}{n}-\frac{1}{N}}.$$

Sampling with Subsamples

Suppose you require several field workers to perform the sampling or the sampling takes place over several days. There will be variation in the measurements among the field workers or among the days of sampling. The population mean can be estimated using the subsample means of each of the field workers or for each of the days. This is not a stratified sample, but simply breaking up the sample into subsamples. This method of sampling was developed by Edward Deming.

The sample of size n is to be divided into k subsamples, with each subsample of size m. Let $\bar{y}_i$ denote the mean of the i^{th} subsample.

- The estimator of the population mean m is $\bar{y} = \frac{1}{k}\sum_{i=1}^{k} \bar{y}_i$, the average of the subsample means.
- The estimated variance of $\bar{y}$ is $\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{s_k^2}{k}$ where $s_k^2 = \frac{\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2}{k-1}$ and measures the variation among the subsample means.

Stratified Random Sampling

As described earlier, stratified random sampling produces estimators with smaller variance than those from simple random sampling, for the same sample size, when the measurements under study are homogeneous within strata but the stratum means vary among themselves. The ideal situation for stratified random sampling is to have all measurements within any one stratum equal but have differences occurring as we move from stratum to stratum. To create a stratified random sample, divide the population into subgroups so that every element of the population is in one and only one subgroup (non-overlapping, exhaustive subgroups). Then take a simple random sample within each subgroup.

The reasons one may choose to perform a stratified random sample are

(1) Possible reduction in the variation of the estimators (statistical reason)

(2) Administrative convenience and reduced cost of survey (practical reason)

(3) Estimates are often needed for the subgroups of the population

Stratification is a widely used technique as most large surveys have stratification incorporated into the design. Additionally, stratification is one of the basic principles of measuring quality and of quality control. (The noted statistician Edward Deming spent half of his life working in survey sampling and the other half in quality control.) Finally, stratification can substitute for direct control in observational studies.

A stratified sample cannot be a simple random sample. As an example, consider the population of 10 letters given below.

Simple Random Sample

Stratified Random Sample

Take a sample of size 4 from the population on the left. The probability that **A** is in the sample is $P(A)=\frac{4}{10}$. The probability of the sample ABCF (order does not matter) is $P(ABCF)=\frac{1}{\binom{10}{4}}$. In the stratified population on the right, in which two elements are taken from the first row and two from the second, the probability that **A** is in the sample is still $P(A)=\frac{4}{10}$. However, the probability of achieving the sample ABCF is $P(ABCF)=0$. Even though the probability of any single element being in the sample is the same, all samples of size 4 are not equally likely, and thus, this is not a simple random sample.

Stratification methods for the Gallop Poll and New York Times are presented below (quoted from Scheaffer, et al, *Elementary Survey Sampling, 5th Edition*, page 50-51):

The Gallup Poll

> Although most Gallup poll findings are based on telephone interviews, a significant proportion is based on interviews conducted in person in the home. The majority of the findings reported in Gallup Poll surveys is based on samples consisting of a minimum of 1,000 interviews. The total number, however, may exceed 1,000, or even 1,500, interviews, where the survey specifications call for reporting the responses of low-incident population groups such as young public-school parents or Hispanics.
>
> **Design of the Sample for Telephone Surveys**
>
> The findings from the telephone surveys are based on Gallup's standard national telephone samples, consisting of unclustered directory-assisted, random-digit telephone samples utilizing a proportionate, stratified sampling design. The random-digit aspect of the sample is used

to avoid "listing" bias. Numerous studies have shown that households with unlisted telephone numbers are different from listed households. "Unlistedness" is due to household mobility or to customer requests to prevent publication of the telephone number. To avoid this source of bias, a random-digit procedure designed to provide representation of both listed and unlisted (including not-yet-listed) numbers is used.

Telephone numbers for the continental United States are stratified into four regions of the country and, within each region, further arranged into three size-of-community strata. The sample of telephone numbers produced by the described method is representative of all telephone households within the continental United States.

Only working banks of telephone numbers are selected. Eliminating nonworking banks from the sample increases the likelihood that any sampled telephone number will be associated with a residence.

Within each contacted household, an interview is sought with the youngest man 18 years of age or older who is at home. If no man is home, an interview is sought with the oldest woman at home. This method of respondent selection within households produces an age distribution by sex that closely approximates the age distribution by sex of the total population.

Up to three calls are made to each selected telephone number to complete an interview. The time of day and the day of the week for callbacks are varied to maximize the chances of finding a respondent at home. All interviews are conducted on weekends or weekday evenings in order to contact potential respondents among the working population.

The final sample is weighted so that the distribution of the sample matches current estimates derived from the U.S. Census Bureau's Current Population Survey (CPS) for the adult population living in telephone households in the continental United States.

Design of the Sample for Personal Surveys

The design of the sample for personal (face-to-face) surveys is that of a replicated area of probability sample down to the block level in the case of urban areas and to segments of townships in the case of rural areas.

After stratifying the nation geographically and by size of community according to information derived from the most recent census, over 350 different sampling locations are selected on a mathematically random basis from within cities, towns, and counties that, in turn, have been selected on a mathematically random basis.

The interviewers are given no leeway in selecting the areas in which they are to conduct their interviews. Each interviewer is given a map on which a specific starting point is marked and is instructed to contact households according to a predetermined travel pattern. At each occupied dwelling unit, the interviewer selects respondents by following

a systematic procedure that is repeated until the assigned number of interviews has been completed.

The New York Times

The latest New York Times/CBS News Poll is based on telephone interviews conducted from Sept. 8 to 11 with 1,161 adults around the country, excluding Alaska and Hawaii.

The sample of telephone exchanges called was selected by a computer from a complete list of exchanges in the United States. The exchanges were chosen to assure that each region of the country was represented in proportion to its population. For each exchange, the telephone numbers were formed by random digits, thus permitting access to both listed and unlisted numbers. Within each household, one adult was designated by a random procedure to be the respondent for the survey.

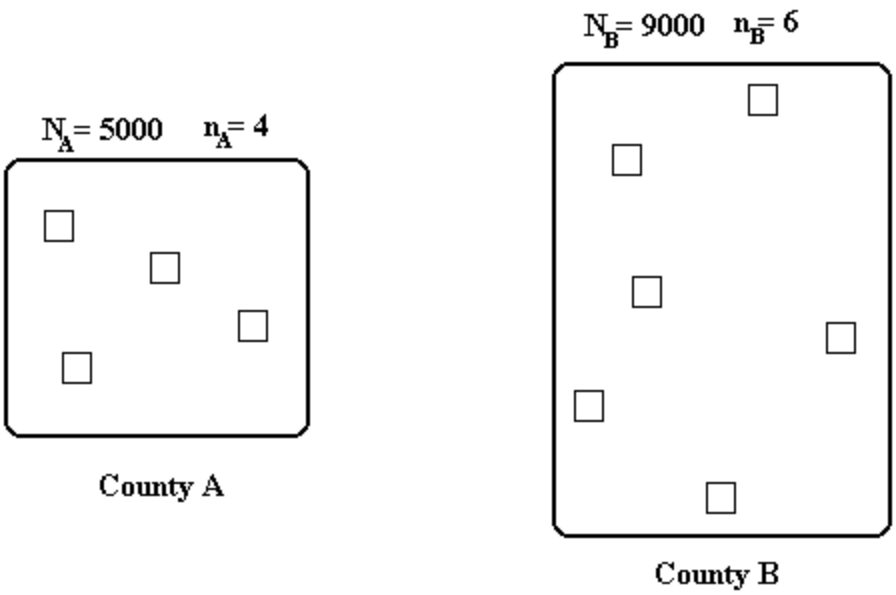
The results have been weighted to take account of the household size and the number of telephone lines into the residence, and to adjust for variations in the sample relating to region, race, sex, age and education.

In theory, in 19 cases out of 20 the results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all American adults. For smaller subgroups the potential sampling error is larger. For example, for blacks it is plus or minus 10 percentage points.

In addition to sampling error, the practical difficulties of conducting any survey of public opinion may introduce other sources of error into the poll. Variations in question wording or the order of questions, for example, can lead to somewhat different results.

Estimating the Population Mean in a Stratified Sample

Suppose we wish to estimate the yield of corn in two counties (A and B) in Iowa. County A has N_A acres of corn and County B has N_B acres of corn. Here, we are assuming that all N_i are sufficiently large so that the finite population correction factor can be ignored. The counties constitute two strata and we will take a simple random sample of size n_A from County A and n_B from County B, as described in the diagram below.



We want to estimate the total amount of corn for the two counties. If $\bar{y}_A$ is the mean yield of corn per acre for the 4 plots in County A and $\bar{y}_B$ is the mean yield of corn per acre for the 6 plots in County B, then

$$\hat{t} = N_A \bar{y}_A + N_B \bar{y}_B$$

is our estimate of the total amount of corn in the two counties.

Our estimate of the mean yield of corn per acre for the two counties is

$$\hat{m} = \frac{N_A \bar{y}_A + N_B \bar{y}_B}{N_A + N_B} = \frac{N_A}{N} \bar{y}_A + \frac{N_B}{N} \bar{y}_B ,$$

if we let $N = N_A + N_B$ be the total acreage for the two counties. This estimator can be written as a weighted average

$$\hat{m} = W_A \bar{y}_A + W_B \bar{y}_B \text{ with } W_A = \frac{N_A}{N} \text{ and } W_B = \frac{N_B}{N}$$

where the weights are the population proportions. The variance of $\hat{m}$ is easily computed

$$V(\hat{m}) = V(W_A \bar{y}_A + W_B \bar{y}_B) = W_A^2 V(\bar{y}_A) + W_B^2 V(\bar{y}_B) = W_A^2 \frac{s_A^2}{n_A} + W_B^2 \frac{s_B^2}{n_B} .$$

In general, if there are L strata of size N_i with $\sum_{i=1}^{L} N_i = N$ with samples of size n_i with $\sum_{i=1}^{L} n_i = n$ taken from each strata, respectively, then:

- the estimator of the total is $\hat{t} = \sum_{i=1}^{L} N_i \bar{y}_i$.

- the estimator of the mean is $\hat{m} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$ or $\hat{m} = \sum_{i=1}^{L} W_i \bar{y}_i$ with $W_i = \frac{N_i}{N}$ the population proportion.

We have our estimated mean

$$\bar{y} = \sum_{i=1}^{L} W_i \bar{y}_i \text{ , so } V(\bar{y}) = \sum_{i=1}^{L} W_i^2 V(\bar{y}_i) = \sum_{i=1}^{L} W_i^2 \frac{s_i^2}{n_i}.$$

This last expression can be rewritten using sample proportions as weights $w_i = \frac{n_i}{n}$. So,

$$V(\bar{y}) = \sum_{i=1}^{L} \frac{W_i^2 s_i^2}{n w_i}.$$

The Problems of Sample Size and Allocation

Suppose we want to estimate the mean yield of corn to within 100 bushels/acre. How can we use the equations above to determine the appropriate sample size *n* and the allocations n_i to produce an estimate accurate to a specified tolerance? We will, as usual, use $2\sqrt{V(\bar{y})} = B$ as our margin of error. We require values of *n* and n_i so that $V(\bar{y}) = \frac{B^2}{4} = D$ (called the dispersion). Then $D = \frac{1}{n}\left[\sum_{i=1}^{L} \frac{W_i^2 s_i^2}{w_i}\right]$ and consequently,

$$n = \frac{1}{D}\left[\sum_{i=1}^{L} \frac{W_i^2 s_i^2}{w_i}\right],$$

with $D = \frac{B^2}{4}$ when estimating m and $D = \frac{B^2}{4N^2}$ when estimating t.

We know that $W_i = \frac{N_i}{N}$ are population proportions. However, in order to find *n* we must know the weights w_i.

One method for determining the sample proportions w_i is to simply assign them the same values as the population proportions, so $w_i = W_i = \frac{N_i}{N}$. This method is particularly useful when the variances of the strata are similar.

Another standard procedure is to use the weights that minimize the variance. Consider the case when two strata are used. Then

$$V(\bar{y}) = \frac{W_1^2 s_1^2}{n_1} + \frac{W_2^2 s_2^2}{n_2} = \frac{k_1^2}{n_1} + \frac{k_2^2}{n - n_1} \text{ where } k_i^2 = W_i^2 s_i^2 \text{ is a constant.}$$

Now, to find the value of n_1 that minimizes $V(\bar{y})$, we use calculus. So,

$$\frac{d}{dn_1}\left(\frac{k_1^2}{n_1}+\frac{k_2^2}{n-n_1}\right)=\frac{-k_1^2}{n_1^2}+\frac{k_2^2}{(n-n_1)^2}=0.$$

Solving for n_1, we have

$$\frac{k_2^2}{(n-n_1)^2}=\frac{k_1^2}{n_1^2} \text{ or } \frac{n_1^2}{n_2^2}=\frac{k_1^2}{k_2^2}, \text{ so } \frac{n_1}{n_2}=\frac{k_1}{k_2}=\frac{W_1 s_1}{W_2 s_2}.$$

Then $n=n_1+n_2=n_1+\frac{k_2}{k_1}n_1=n_1\left(\frac{k_1+k_2}{k_1}\right)$. Solving for n_1, we have $n_1=n\left(\frac{k_1}{k_1+k_2}\right)$.

In general, we have $n_i=n\left(\frac{k_i}{\sum_{i=1}^{L}k_i}\right)=n\left(\frac{W_i s_i}{\sum_{i=1}^{L}W_i s_i}\right)$.

This last equation indicates that the allocation to region *i* will be large if $W_i=\frac{N_i}{N}$ is large, that is, if it contains a large portion of the population. This should make sense. It also indicates that the allocation to region *i* will be large if there is a lot of variability in the region. If there is little variation in the region, the allocation will be small, since a small sample will give the necessary information. As an extreme example, if there is no variation in a region, a single sample will tell you everything about the region. This optimal allocation was developed by the statistician Jerzy Neyman and is called the Neyman allocation.

Example 1. Consider the two counties A and B with $N_A=5000$ acres and $N_B=9000$ acres. Suppose we can approximate the variance of the yields for the two counties based on past performance as $s_A\approx 12$ bushels/acre and $s_B\approx 20$ bushels/acre. We want to estimate the mean yield in bushels per acre for the two counties with a margin of error of 5 bushels/acre. What are the values of *n*, n_A, and n_B if

a) we use proportional allocation
b) we allocate samples to minimize the variance (optimal allocation)

a) Here we have $\frac{n_A}{n_B}=\frac{N_A}{N_B}=\frac{5}{9}$. This means that $n_A=\frac{5}{14}n$ and $n_B=\frac{9}{14}n$ and $w_A=\frac{n_A}{n}=\frac{5}{14}$ with $w_B=\frac{9}{14}$. Using the formula derived above,

$$n=\frac{1}{D}\left[\sum_{i=1}^{L}\frac{W_i^2 s_i^2}{w_i}\right]=\frac{1}{D}\left[\frac{W_A^2 s_A^2}{w_A}+\frac{W_B^2 s_B^2}{w_B}\right],$$

we can find the appropriate values of n, n_A, and n_B. We know everything except D. To find D, we have $B = 5$, so $D = \frac{B^2}{4} = \frac{25}{4}$.

Now,

$$n = \frac{4}{25}\left[\frac{\left(\frac{5}{14}\right)^2 (12)^2}{\left(\frac{5}{14}\right)} + \frac{\left(\frac{9}{14}\right)^2 (20)^2}{\left(\frac{9}{14}\right)}\right] \approx 50$$

So proportional allocation gives $n = 50$, $n_A = \left(\frac{5}{14}\right)50 \approx 18$ and $n_B = \left(\frac{9}{14}\right)50 \approx 32$.

b) Optimal allocation requires that

$$n_A = n\left(\frac{W_A s_A}{W_A s_A + W_B s_B}\right) = (n)\frac{\left(\frac{5}{14}\right)(12)}{\left(\frac{5}{14}\right)(12) + \left(\frac{9}{14}\right)(20)} = \frac{1}{4}n$$

and

$$n_B = n\left(\frac{W_B s_B}{W_A s_A + W_B s_B}\right) = (n)\frac{\left(\frac{9}{14}\right)(20)}{\left(\frac{5}{14}\right)(12) + \left(\frac{9}{14}\right)(20)} = \frac{3}{4}n.$$

As before,

$$n = \frac{1}{D}\left[\frac{W_A^2 s_A^2}{w_A} + \frac{W_B^2 s_B^2}{w_B}\right],$$

and so,

$$n = \frac{4}{25}\left[\frac{\left(\frac{5}{14}\right)^2 (12)^2}{\left(\frac{1}{4}\right)} + \frac{\left(\frac{9}{14}\right)^2 (20)^2}{\left(\frac{3}{4}\right)}\right] \approx 47$$

So proportional allocation gives $n = 47$, $n_A = \left(\frac{1}{4}\right)47 \approx 12$ and $n_B = \left(\frac{3}{4}\right)47 \approx 35$.

Notice that, although fewer samples were needed, more samples came from County B, since it had both greater variation and was a larger proportion of the population.

Considering Cost and Finite Population Factor

The equations developed in this section become somewhat more complex if the finite population correction factor must be included in the calculations. In this case, we have

$$n = \frac{\sum_{i=1}^{L} N_i^2 \frac{s_i^2}{w_i}}{N^2 D + \sum_{i=1}^{L} N_i s_i^2}$$

with $D = \frac{B^2}{4}$ when estimating m and $D = \frac{B^2}{4N^2}$ when estimating t.

The approximate allocation that minimizes total cost for a fixed variance, or minimizes variance for a fixed costs (c_i) is

$$n_i = n\left(\frac{\frac{N_i s_i}{\sqrt{c_i}}}{\sum_{k=1}^{L} \frac{N_k s_k}{\sqrt{c_k}}} \right).$$

Note that n_i is directly proportional to N_i and s_i and inversely proportional to $\sqrt{c_i}$.

Also note that if all c_i are equal, the allocation is Neyman's optimal allocation presented earlier.

Comparison of Stratified Random Sampling to Simple Random Sampling

Stratification usually produces gains in precision, especially if the stratification is accomplished through a variable correlated with the response. We would like to stratify when the strata are homogeneous and different, that is, we have

1) low variation in the strata
2) differing means among the strata.

The following comparisons apply for situations in which the N_i are all relatively large, so we can replace $\frac{1}{N_i - 1}$ with $\frac{1}{N_i}$. Here we use $f = \frac{n}{N}$ and $W_i = \frac{N_i}{N}$.

The variance of a SRS, denoted V_{SRS}, compared to the variance of a proportional allocation, denoted V_{prop} is described in the equation

$$V_{SRS} - V_{prop} = \frac{1 - f}{n} \sum_i W_i \left(\bar{Y}_i - \bar{Y} \right)^2.$$

From this equation, we see that the proportional allocation will be useful (produce a smaller variance than SRS) when there is a large difference in the means for the different strata.

The variance of proportional allocation compared to the variance of an optimal Neyman allocation, denoted V_{opt} is described in the equation

$$V_{prop} - V_{opt} = \frac{1}{n}\sum_i W_i \left(S_i - \overline{S}\right)^2 ,$$

where S_i is a measure of the random variation of the population strata and $\overline{S} = \sum_i W_i S_i$. From this equation, we see that the optimal allocation is an improvement over proportional allocation when there is a large difference in the variation among the strata.

In summary, one should attempt to construct strata so that the strata means differ. If strata variances do not differ much, use proportional allocation. If strata variances differ greatly, use optimum Neyman allocation.

A Word on Post Stratification

At times, we wish to stratify a sample after a simple random sample has been taken. For example, suppose you wish to stratify on gender based on a telephone poll, where you can't know the gender of the respondent until after the SRS is taken. What penalty do we pay if we decide to stratify after selecting a simple random sample? It is possible to show that the estimated variance, $\hat{V}_p\left(\overline{y}\right)$, is given by

$$\hat{V}_p\left(\overline{y}\right) = \left(\frac{N-n}{Nn}\right)\sum_{i=1}^{L} W_i\, s_i^2 + \frac{1}{n^2}\sum_{i=1}^{L}\left(1-W_i\right)s_i^2 .$$

The first term is what you would expect from a stratified sample mean using proportional allocation, so the second term is the price paid for stratifying after the fact. Notice that the term $\frac{1}{n^2}$ reduces the penalty as n increases. Post-stratification produces good results when *n* is large and all n_i are large as well.

Ratio Estimation

Ratio estimation is an important issue in cluster sampling. We will develop the principles of ratio estimation and then proceed to cluster sampling.

How do you determine the mpg for your car? One way would be to note the miles driven and the number of gallons of gas used each time you fill up the gas tank. This will produce a set of ordered pairs, each of which can be used to estimate your mpg. What is the best estimate you can make from this information?

miles	y_1	y_2	y_3	$\cdots$	y_n
gallons	x_1	x_2	x_3	$\cdots$	x_n

We can compute all n ratios $\frac{y_i}{x_i}$ and find the average value $\frac{1}{n}\sum\left(\frac{y_i}{x_i}\right)$. Unfortunately, $E\left(\frac{y_i}{x_i}\right) \neq \frac{m_y}{m_x}$. Each division of $\frac{y_i}{x_i}$ produces some bias, so we want to perform as few divisions as possible.

The best estimator of the population ratio $R = \frac{m_y}{m_x}$ is $r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\overline{y}}{\overline{x}}$.

The estimated variance of r can be approximated by

$$\hat{V}(r) = \hat{V}\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right) = \left(\frac{N-n}{N}\right)\left(\frac{1}{m_x^2}\right)\left(\frac{s_r^2}{n}\right),$$

where $s_r^2 = \frac{\sum_{i=1}^{n}(y_i - r x_i)^2}{n-1}$. The estimated variance of r is similar to the formula for the variance of a sample mean, but has the additional $\left(\frac{1}{m_x^2}\right)$ term. The value of s_r^2 is similar to the variance of residuals.

If we plot the ordered pairs (x_i, y_i), we are comparing these points to the line $y = r\,x$.

Our estimate of the ratio r allows us to make estimates of the population mean, $\hat{m}_y$, and the population total, $\hat{t}_y$. If $\frac{m_y}{m_x}$ is estimated by $\frac{\overline{y}}{\overline{x}}$, then we should be able to estimate m_y with

$$\hat{m}_y = \frac{\overline{y}}{\overline{x}} m_x = r\, m_x.$$

The estimated variance of m_y is

$$\hat{V}(\hat{m}_y) = m_x^2 \hat{V}(r) = \left(\frac{N-n}{N}\right)\frac{s_r^2}{n}.$$

Similarly, the ratio estimator of the population total, t_y, is

$$\hat{t}_y = \frac{\overline{y}}{\overline{x}} t_x = r t_x .$$

The estimated variance of t_y is

$$\hat{V}\left(\hat{t}_y\right) = t_x^2 \hat{V}(r) = t_x^2 \left(\frac{N-n}{N}\right)\left(\frac{1}{m_x^2}\right)\frac{s_r^2}{n} .$$

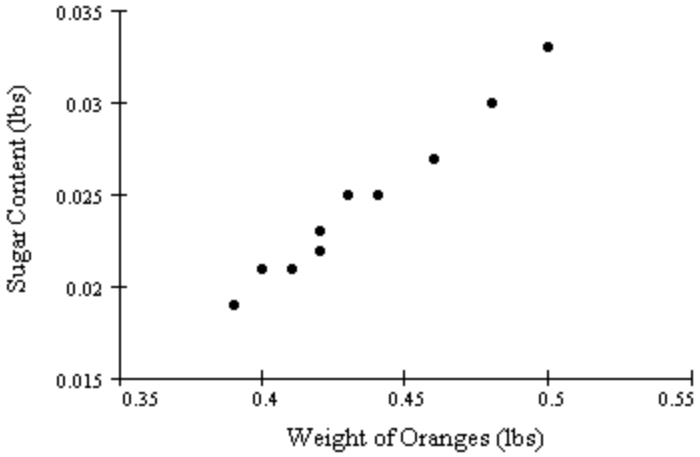
Note that we do not need to know t_x or N to estimate m_y when using the ratio procedure. However, we must know m_x.

Example (Adapted from Scheaffer, et al, *Elementary Survey Sampling, 5th Edition*, page 205-206):

In Florida, orange farmers are paid according to the sugar content in their oranges. How much should a farmer be paid for a truckload of oranges? A sample is taken, and the total amount of sugar in the truckload can estimated using the ratio method.

Suppose 10 oranges were selected at random from the truckload to be tested for sugar content. The truck was weighed loaded and unloaded to determine the weight of the oranges. In this case, there were 1800 pounds of oranges. Larger oranges have more sugar, so we want to know the sugar content per pound for the truckload and use this to estimate the total sugar content of the load.

Orange	1	2	3	4	5	6	7	8	9	10
Sugar Content (lbs)	0.021	0.030	0.025	0.022	0.033	0.027	0.019	0.021	0.023	0.025
Wt of Orange (lbs)	0.40	0.48	0.43	0.42	0.50	0.46	0.39	0.41	0.42	0.44



The scatterplot above shows a strong linear relationship between the two variables, so a ratio estimate is appropriate. Using the formula $\hat{t}_y = \frac{\overline{y}}{\overline{x}} t_x = r t_x$ we estimate

$$\hat{t}_y = \frac{0.0246}{0.4350}(1800) = (0.05655)(1800) = 101.8 \text{ pounds}$$

of sugar in the truckload. A bound on the error of estimation can be found as well. We have $\hat{V}(\hat{t}_y) = t_x^2 \hat{V}(r) = t_x^2 \left(\frac{N-n}{N}\right)\left(\frac{1}{m_x^2}\right)\frac{s_r^2}{n}$, but in this case, we know neither N nor m_x. Since N is large (a truckload of oranges will be at least 4,000 oranges), so the finite population correction $\left(\frac{N-n}{N}\right)$ is essentially 1. We will use $\bar{x}$ as an estimate of m_x. With these modifications, we can compute

$$2\sqrt{\hat{V}(\hat{t}_y)} = 2\sqrt{t_x^2\left(\frac{1}{\bar{x}^2}\right)\frac{s_r^2}{n}} = 2\sqrt{(1800)^2\left(\frac{1}{0.435^2}\right)\frac{0.0024^2}{10}} = 6.3$$

Our estimate of the total sugar content of the truckload of oranges is 101.8 ± 6.3 pounds.

If the population size N is know, we could also use the estimator $N\bar{y}$ instead of $r\,t_x$ to estimate the total. Generally, the estimator $r\,t_x$ has a smaller variance than $N\bar{y}$ when there is a strong positive correlation between x and y. As a rule of thumb, if $r > \frac{1}{2}$, the ratio estimate should be used. This decrease in variance results from taking advantage of the additional information provided by the subsidiary variable x in our calculations with the ratio estimation.

Relative Efficiency of Estimators

Suppose there are two unbiased (or nearly unbiased) estimators, E_1 and E_2, for the same parameter. The relative efficiency of the two estimators is measured by the ratio of the reciprocals of their variances. That is,

$$RE\left(\frac{E_1}{E_2}\right) = \frac{V(E_2)}{V(E_1)}.$$

If $RE\left(\frac{E_1}{E_2}\right) > 1$, estimator E_1 will be more efficient. If the sample sizes are the same, the variance of E_1 will be smaller. Another way to view this is that estimator E_1 will produce the same variance as E_2 with a smaller sample size.

We can compute the relative efficiency of m_y and $\bar{y}$. Here, we have

$$\widehat{RE}\left(\frac{\hat{m}_y}{\bar{y}}\right)=\frac{V(\bar{y})}{V(\hat{m}_y)}=\frac{s_y^2}{s_r^2}.$$

Both variances have the same values of N and n, so the finite population correction factor divides out. The variance of $\hat{m}_y$ can be re-written in terms of the predicted correlation $\hat{r}$ so that

$$\widehat{RE}\left(\frac{\hat{m}_y}{\bar{y}}\right)=\frac{s_y^2}{s_y^2+r^2s_x^2-2r\hat{r}\,s_xs_y}.$$

If $\widehat{RE}\left(\frac{\hat{m}_y}{\bar{y}}\right)>1$ then $\hat{m}_y$ is a more efficient estimator. To determine when $\widehat{RE}\left(\frac{\hat{m}_y}{\bar{y}}\right)>1$, we consider $\frac{s_y^2}{s_y^2+r^2s_x^2-2r\hat{r}\,s_xs_y}>1$. Then

$$s_y^2>s_y^2+r^2s_x^2-2r\hat{r}\,s_xs_y,$$

or

$$2\hat{r}\,s_xs_y>rs_x^2.$$

If $r>0$, then

$$\hat{r}>\frac{r\,s_x^2}{2s_xs_y}=\frac{1}{2}\left(\frac{s_x/\bar{x}}{s_y/\bar{y}}\right).$$

As is often the case in ratio estimation, $\frac{s_x}{\bar{x}}\approx\frac{s_y}{\bar{y}}$, we see that $\hat{m}_y$ is a more efficient estimator than $\bar{y}$ when $\hat{r}>\frac{1}{2}$.

Cluster Sampling

Sometimes it is impossible to develop a frame for the elements that we would like to sample. We might be able to develop a frame for clusters of elements, though, such as city blocks rather than households or clinics rather than patients. If each element within a sampled cluster is measured, the result is a **single-stage cluster sample**. A cluster sample is a probability sample in which each sampling unit is a collection, or cluster, of elements. Cluster sampling is less costly than simple or stratified random sampling if the cost of obtaining a frame that lists all population elements is very high or if the cost of obtaining observations increases as the distance separating the elements increases.

To illustrate, suppose we wish to estimate the average income per household in a large city. If we use simple random sampling, we will need a frame listing all households

(elements) in the city, which would be difficult and costly to obtain. We cannot avoid this problem by using stratified random sampling because a frame is still required for each stratum in the population. Rather than draw a simple random sample of *elements*, we could divide the city into regions such as blocks (or clusters of elements) and select a simple random sample of blocks from the population. This task is easily accomplished by using a frame that lists all city blocks. Then the income of every household within each sampled block could be measured.

Cluster sampling is an effective design for obtaining a specified amount of information at minimum cost under the following conditions:

1. A good frame listing population elements either is not available or is very costly to obtain, while a frame listing clusters is easily obtained.

2. The cost of obtaining observations increases as the distance separating the elements increases.

Elements other than people are often sampled in clusters. An automobile forms a nice cluster of four tires for studies of tire wear and safety. A circuit board manufactured for a computer forms a cluster of semiconductors for testing. An orange tree forms a cluster of oranges for investigating an insect infestation. A plot in a forest contains a cluster of trees for estimating timber volume or proportions of diseased trees.

Notice the main difference between the optimal construction of strata and the construction of clusters. Strata are to be as homogeneous (alike) as possible within, but one stratum should differ as much as possible from another with respect to the characteristic being measured. Clusters, on the other hand, should be as heterogeneous (different) as possible within, and one cluster should look very much like another in order for the economic advantages of cluster sampling to pay off.

Estimation of a Population Mean and Total

Cluster sampling is simple random sampling with each sampling unit containing a collection or cluster of elements. Hence, the estimators of the population mean m and total t are similar to those for simple random sampling. In particular, the sample mean $\bar{y}$ is a good estimator of the population mean m.

The following notation is used in this section:

N = the number of clusters in the population

n = the number of clusters selected in a simple random sample

m_i = the number of elements in cluster i, $i = 1, \ldots, N$

$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i =$ the average cluster size for the sample

M $= \sum_{i=1}^{n} m_i$ = the number of elements in the population

$\bar{M} = \frac{M}{N}$ = the average cluster size for the population

y_i = the total of all observations in the *i*th cluster

y_{ij} = the measure for the *j*th element in the *i*th cluster

The estimator of the population mean m is the sample mean $\bar{y}$, which is given by

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}.$$

Since both y_i and m_i are random variables, $\bar{y}$ is a ratio estimator, so the formulas developed earlier will apply. We simply replace x_i with m_i.

The estimated variance of $\bar{y}$ is

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\bar{M}^2}\right)\left(\frac{s_r^2}{n}\right)$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2}{n-1}.$$

If $\bar{M}$ is unknown, it can be estimated by $\bar{m}$. This estimated variance is biased and will be a good estimate of $V(\bar{y})$ only if *n* is large. A rule of thumb is to require $n \geq 20$. The bias disappears if all m_i are equal.

Example 8.2 (Scheaffer, et al, page 294)

A city is to be divided into 415 clusters. Twenty-five of the clusters will be sampled, and interviews are conducted at every household in each of the 25 blocks sampled. The data on incomes are presented in the table below. Use the data to estimate the per-capita income in the city and place a bound on the error of estimation.

Cluster i	Number of Residents, m_i	Total income per cluster, y_i
1	8	$96,000
2	12	121,000
3	4	42,000
4	5	65,000
5	6	52,000
6	6	40,000
7	7	75,000
8	5	65,000
9	8	45,000
10	3	50000
11	2	85,000
12	6	43.000
13	5	54,000

Cluster i	Number of Residents, m_i	Total income per cluster, y_i
14	10	$49,000
15	9	53,000
16	3	50,000
17	6	32,000
18	5	22,000
19	5	45,000
20	4	37,000
21	6	51,000
22	8	30,000
23	7	39,000
24	3	47,000
25	8	41,000

Here we have $\sum_{i=1}^{n} m_i = 151$, $\sum_{i=1}^{n} y_i = 1{,}329{,}000$, and $s_r = 25{,}189$.

Solution

The best estimate of the population mean m is $\bar{y} = \frac{\$1{,}329{,}000}{151} = \$8801$. The estimate of per capita income is $8801.

Since M is not known, $\bar{M}$ must be estimated by $\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n} = \frac{151}{25} = 6.04$. Since there were at total of 415 clusters, $N = 415$. So,

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\bar{M}^2}\right)\left(\frac{s_r^2}{n}\right) = \left(\frac{415-25}{415}\right)\left(\frac{1}{6.04^2}\right)\left(\frac{25189^2}{25}\right) = 653{,}785$$

Thus, the estimate of m with a bound on the error of estimation is given by

$$\bar{y} \pm 2\sqrt{\hat{V}(\bar{y})} = 8801 \pm 2\sqrt{653{,}785} = 8801 \pm 1617$$

The best estimate of the average per-capita income is $8801, and the error of estimation should be less than $1617 with probability close to 0.95. This bound on the

error of estimation is rather large; it could be reduced by sampling more clusters and, consequently, increasing the sample size.

Comparing Cluster Sampling and Stratified Sampling

It is advantageous to use a cluster sample when the individual clusters contain as much within cluster variability as possible, but the clusters themselves are as similar as possible. This can be seen in the computation of the variation,

$$s_r^2 = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y} m_i \right)^2}{n-1} = \frac{\sum_{i=1}^{n} m_i^2 \left(\bar{y}_i - \bar{y} \right)^2}{n-1},$$

which will be small when the $\bar{y}_i$'s are similar in value. For cluster sampling, the differences are found within the clusters and the similarity between the clusters.

It is advantageous to use stratified sampling when elements within each strata are as similar as possible, but the strata themselves are as different as possible. Here, the differences are found between the strata and the similarity within the strata. Two examples will help illustrate this distinction.

Example 1 Suppose you want to take a sample of a large high school and you must use classes to accomplish your sampling. In this school, students are randomly assigned to homerooms, so each homeroom has a mixture of students from all grade-levels (Freshman-Senior). Also, in this school, the study halls are grade-level specific, so all of the students in a large study hall are from the same grade. If you believe that students in the different grade-levels will have different responses, you want to be assured that each grade-level is represented in the sample.

You could perform a cluster sample by selecting *n* homerooms at random and surveying everyone in those homerooms. You would not use the homerooms as strata, since there would be no advantage over a simple random sample.

You could perform a stratified sample using study halls as your strata. Randomly select *k* students from study halls for each grade-level. Study halls would make a poor cluster, since the responses from all of the students are expected to be similar.

Example 2 We would like to estimate the number of diseased trees in the forest represented below. The diseased trees are indicated with a D, while the trees free of disease are represented by F. Consider the rows and columns of the grid.

(a) If a cluster sample is used, should the rows or columns be used as a cluster?

(b) If a stratified sample is used, should the rows or columns be used as strata?

Row	C1	C2	C3	C4	C5
1	F	F	F	D	D
2	F	F	D	D	D
3	F	F	F	F	F
4	F	F	D	F	D
5	F	F	F	F	D
6	D	F	D	F	F
7	F	F	D	F	D
8	F	D	D	F	D
9	F	F	F	D	D
10	F	F	F	D	D
11	F	F	F	D	F
12	F	D	D	D	D
13	F	D	F	D	D
14	F	F	F	D	D
15	F	D	F	D	D
16	F	F	D	D	D
17	F	F	D	D	D
18	F	F	F	D	D
19	F	F	D	D	D
20	F	F	F	F	F
21	D	F	F	D	F
22	F	D	F	F	D
23	F	F	D	D	F
24	F	F	F	D	D
25	F	F	F	D	D
26	F	D	F	F	D
27	F	F	D	F	D
28	D	F	F	F	D
29	F	F	F	F	D
30	F	F	D	D	D

It appears that there are more diseased trees in the right-most columns, however, there does not appear to be a difference among the rows. If we wanted a sample of size 25, we could obviously select a simple random sample, but we might miss the concentration of diseased trees in C4 and C5 just by chance. We want to insure that C4 and C5 show up in the sample. We have two choices:

- For a cluster sample, we should use the rows as clusters. We could select 5 rows at random, and consider every tree in each of those clusters (rows).

- For a stratified sample, we could use the columns as strata. We would select 5 elements from each of the 5 strata (columns) to consider.

Systematic Sampling

Suppose the population elements are on a list or come to the investigator sequentially. It is convenient to find a starting point near the beginning of the list and then sample every k^{th} element thereafter. If the starting point is random, this is called a 1-in-k systematic sample.

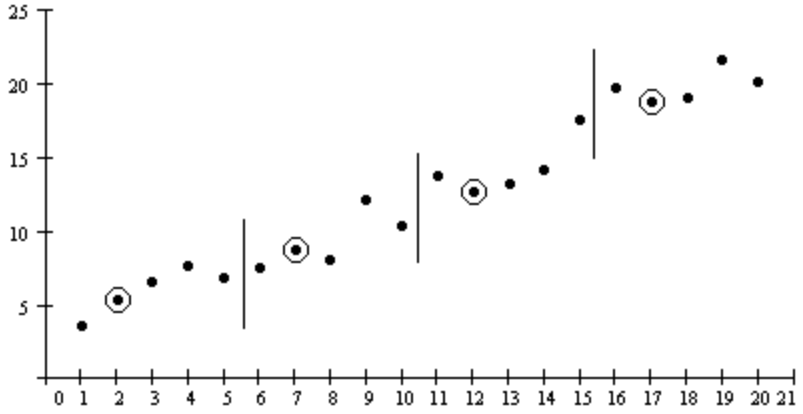
If the population elements are in random order, systematic sampling is equivalent to simple random sampling. If the population elements have trends or periodicities, systematic sampling may be better or worse than simple random sampling depending on how information on population structure is used. Many estimators of variance have been proposed to handle various population structures.

Repeated Systematic Sampling

In the 1-in-k systematic sample, there is only one randomization, which limits the analysis. The randomness in the systematic sample can be improved by choosing more than one random start. For example, instead of selecting a random number between 1 and 4 to start and then picking every 4^{th} element, you could select 2 numbers at random between 1 and 8, and then selecting those elements in each group of 8.

Relationship to Stratified and Cluster Sampling

Recall that if the elements are in random order, we have no problem with systematic sampling. If there is some structure to the data, as shown below, we can compare systematic sampling to stratified and cluster samples.



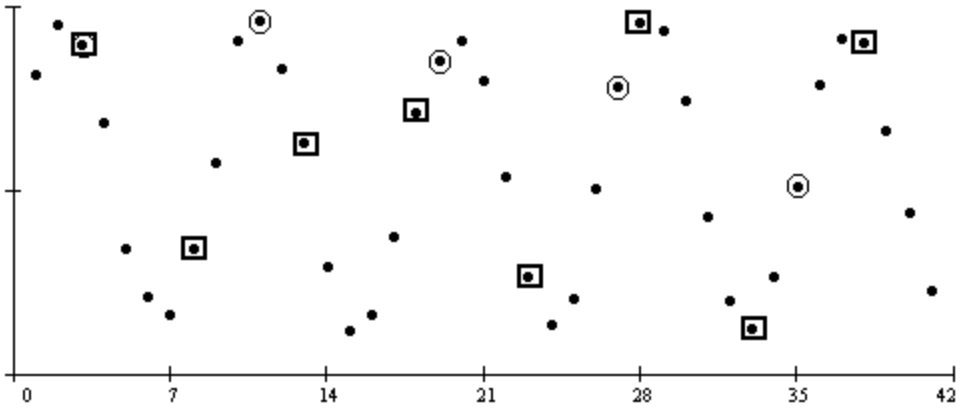
Systematic sampling is closely related to

- stratified sampling with one sample element per stratum
- cluster sampling with the sample consisting of a single cluster

As a stratified sample, we think of having 4 different strata, each with 5 elements. The elements of the strata are similar and the means of the strata are different, so this fits the requirements for a stratified sample. We take one element from each stratum (in this illustration, the second in each stratum). We have lost some randomness, since the second item is taken from all strata rather than a random element from each stratum.

As a cluster sample, we think of the 5 possible clusters. Cluster 1 contains all of the first elements, cluster 2 (the one selected) contains all the second elements, etc. Here we have surveyed all elements in one cluster (cluster 2). In this case, the clusters contain as much variation as possible with similar means, so the cluster process is appropriate. Since we have only one cluster, we have no estimate of the variance. A repeated systematic sample (taking clusters 2 and 5, for example) would eliminate this difficulty.

If the structure of the data is periodic, it is important that the systematic sample not mimic the periodic behavior. In the diagram below, the circles begin at the 3rd element and select every 8th element. Since this matches closely the period of data, we select only values in the upper range. If we begin at the 3rd element and select every 5th element, we are able to capture data across the full range.



Estimating the Size of the Population

In the preceding sections, we estimated means, totals, and proportions, assuming that the population size was either known or so large that it could be ignored if not expressly needed to calculate an estimate. Frequently, however, the population size is not known and is important to the goals of the study. In fact, in some studies, estimation of the population size is the main goal. The maintenance of wildlife populations depends crucially on accurate estimates of population sizes.

Direct Sampling

One common method for estimating the size of a wildlife population is *direct sampling*. This procedure entails drawing a random sample from a wildlife population of interest, tagging each animal sampled, and returning the tagged animals to the population.

At a later date, another random sample of a fixed size n is drawn from the same population, and the number s of tagged animals is observed. If N represents the total population size, t represents the number of animals tagged in the initial sample, and p represents the proportion of tagged animals in the population, then $\frac{t}{N} = p$. Also, we expect to find approximately the same proportion (p) of the sample of size n tagged as well. So $p \approx \frac{s}{n}$. This gives us a way to estimate the size of the population N, since $\frac{s}{n} \approx \frac{t}{N}$. Solving for N provides an estimator for N,

$$\hat{N} = \frac{nt}{s}.$$

The approximate estimated variance of $\hat{N}$ is

$$\hat{V}\left(\hat{N}\right) = \frac{t^2 n(n-s)}{s^3}.$$

Notice that we have serious problems when s is zero, and a large variance when s is small.

As an example, suppose we initially capture and tag 200 fish in a lake. Later, we capture 100 fish, of which 32 were tagged. So $t = 200$, $n = 100$, and $s = 32$. Then our estimate of N is

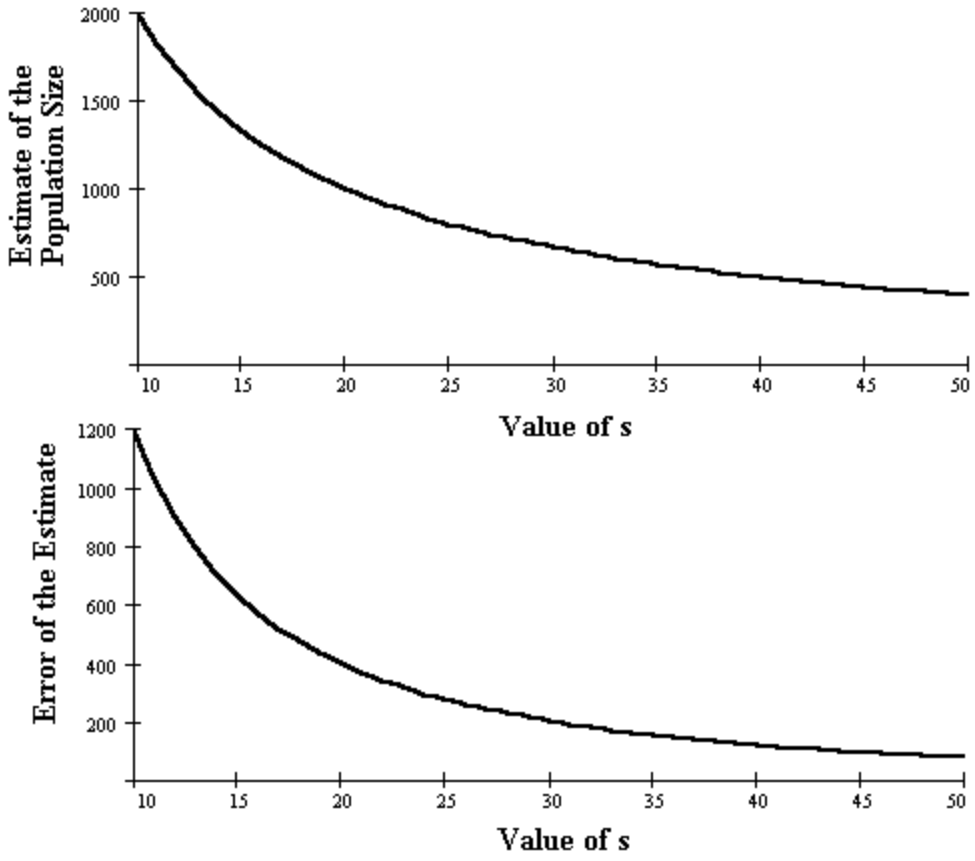
$$\hat{N} = \frac{nt}{s} = \frac{100(200)}{32} = 625 \text{ fish.}$$

Also, we approximate the variance with

$$\hat{V}\left(\hat{N}\right) = \frac{t^2 n(n-s)}{s^3} = \frac{(200)^2 (100)(100-32)}{32^3} = 8301.$$

The margin of error is $2\sqrt{\hat{V}\left(\hat{N}\right)} = 2\sqrt{8301} = 182$. Our estimate of the number of fish in the Lake is between 443 and 807 fish.

The graphs below illustrate how sensitive are both the estimate of N and the margin of error when s is small. If s is less than 4, the error of the estimate is larger than the estimate for these values of n and t.



Inverse Sampling

We can get around the problem of having a small value of s by sampling until we have a pre-specified value of s. For example, we could fish until we have caught 50 of the tagged fish. This technique is called *inverse sampling*. That is, we sample until a fixed number of tagged animals, *s*, is observed. Using this procedure, we can also obtain an estimate of *N*, the total population size by computing $N = \frac{nt}{s}$. This is the same computation as before, only *s* is fixed and *n* is random. This changes the variance of $\hat{N}$. The estimated variance of $\hat{N}$ is

$$V\left(\hat{N}\right) = \frac{t^2 n(n-s)}{s^2(s+1)}.$$

This variance estimate is almost the same as before, but it can be considered as a function of *n*, rather than *s*. We no longer have to worry about a small *s*, but this procedure may take much longer and be more expensive, since we do not know how long we need to continue the recapturing process before the pre-set value of *s* is achieved.

Example Consider our earlier example in which we initially captured and tagged 200 fish in a lake. Later, we fished until we had captured 50 fish that had previously been tagged. This required us to catch 162 fish. So $t = 200$, $n = 162$, and $s = 50$. Then our estimate of *N* is

$$\hat{N} = \frac{nt}{s} = \frac{162(200)}{50} = 648 \text{ fish.}$$

We approximate the variance with

$$\hat{V}\left(\hat{N}\right) = \frac{t^2 n(n-s)}{s^2(s+1)} = \frac{(200)^2(162)(162-50)}{50^2(51)} = 5692.$$

The margin of error is $2\sqrt{\hat{V}\left(\hat{N}\right)} = 2\sqrt{5692} = 151$. Our margin of error is smaller since we forced a larger value of s, but it required more resources to catch the extra 62 fish.

Another method for computing the estimated interval is to find a confidence interval on the proportion $\hat{p} = \frac{s}{n}$ and use it to create the interval for N algebraically. In our first example, we have $\frac{s}{n} = \frac{32}{100} = 0.32$. A 95% confidence interval for this proportion is

$$0.32 \pm \frac{1.96\sqrt{0.32(0.68)}}{\sqrt{100}} = 0.32 \pm 0.09 \text{ or (0.23, 0.41).}$$

Now, we have $\hat{N} = \frac{n}{s}t$, so our estimate is $\hat{N} \approx \frac{t}{\left(\frac{s}{n}\right)} = \frac{200}{0.32} = 625$ fish. An interval estimate can be derived using the two extremes of the interval (0.23, 0.41). So $\hat{N} \approx \frac{t}{\left(\frac{s}{n}\right)} = \frac{200}{0.23} = 870$ and $\hat{N} \approx \frac{t}{\left(\frac{s}{n}\right)} = \frac{200}{0.41} = 488$. Our estimate then is between 488 and 870 fish. Notice that the point estimate 625 is not in the center of the interval (488, 870).

Experimental Design for Capture-Recapture

There are two factors, *t* and *n*, that influence the variability of the estimate of N when using capture/recapture. A common question about capture recapture is, "Is it better to mark more fish initially or is it better to take a larger sample in the recapture phase?" The question is really about where to put your energy and resources. The recapture phase can be repeated many times and the resulting estimates of *N* compared, perhaps in a stem-leaf plot. One could then vary the number of tagged animals and repeat the process to see how the variability of the estimates depends on *t*. One could also vary the size of the second capture to see how the variability of the estimates depends on *n*.

Below are box-plots comparing 100 estimates of $N = 1000$ using either $t = 100$ or $t = 200$ and either $n = 60$ or $n = 120$. From the boxplots you can see that changing *t*

from 100 to 200 has a greater effect on reducing the variability than does changing *n* from 60 to 120. Greater benefits are achieved when more effort is put on the initial sample to be tagged. Note that when both *t* and *n* are small, small values of *s* were more often generated producing large estimates of *N*.