4-30-2010

Sampling

David Kline

Sampling

David Kline

Honors-Department of Mathematical Sciences

April 30, 2010

Table of Contents

Introduction

Data have become extremely important in the world today. Information is everywhere, and people are working to use it to their advantage, whether in business, politics, or science. For instance, political polls and market research surveys look to gain information about the people that they are serving. The amount of data that can be collected is incomprehensible. It is impossible to analyze all of the information that exists. That is where sampling comes into play. Sampling allows a researcher to select a part of a population to observe so that one may infer something about the whole population. It allows one to take the deluge of data and filter it down to a manageable amount by sampling. The sample design determines the quantity of information that is needed to estimate the population parameter of interest. (Scheaffer, Mendenhall, & Ott, 7) Also, since collecting data costs money, sampling tries to minimize costs, while maximizing the information from the sample. This implies a minimization of the variability. There is uncertainty that occurs in a sample since only a portion of the population is included. Sampling designs try to control and minimize the level of that uncertainty (i.e. variability). Then one can use the sample to make inferences about the larger population with a specified level of confidence. Sampling can be thought of as similar to experimental design. The main difference is that experimental design controls for factors and adjusts a certain variable for each group to compare. Sampling is simply an observation of something as it naturally occurs. There are no adjustments made to different groups. For example, a sample is conducted to estimate the deer population in two counties. Then the deer are observed according to the sampling design and the totals are calculated. It is an observation of how things are at the time of the survey. In an experimental design setting, a researcher would control what each observation would be. If one was interested in comparing fertilizers, the same type of seeds would be planted in the same soil

and the only thing that would change between observations would be the type of fertilizer. The experimenter controls exactly what information is to be provided. In the sampling setting, the information provided depends on the observation selected, instead of how the researcher constructed the experiment.

Some Basics

There are a few introductory terms that are necessary to define for use throughout the paper. An element is an object on which a measurement is taken. A collection of elements about which an inference will be made is a population. (Scheaffer, Mendenhall, & Ott, 8) For the purposes of generating a sample, the population must be completely defined before the data collection can begin. This means one must define specifically what is to be measured and from what elements the measurements are to be taken. Sampling units are non-overlapping collections of elements from the population so that the union of sampling units covers the entire population. (Scheaffer, Mendenhall, & Ott, 8) For example, a sampling unit could be an individual person or a city block. It need not be a singular item. Sampling units are usually easy to determine based on the setting, but sometimes can be more difficult to determine, especially in populations that are difficult to track, like those that might be used in a geological study. A list of the sampling units is called the frame. (Scheaffer, Mendenhall, & Ott, 9) Ideally, the frame would contain all of the sampling units of the population. In practice, it is not always possible to do this, but the hope is that the difference between the population and the frame is small enough that it does not dramatically impact the results. For instance, if a list of phone numbers is used as the frame, people with unlisted numbers would not be included, but there are not enough unlisted numbers for that to pose a major problem. The sample is the selection of sampling units that is selected

from the frame. (Scheaffer, Mendenhall, & Ott, 9) The focus of this paper will be to describe some of the various designs that exist for selecting the sample.

First, since sampling is often used in the context of a survey, it is important to understand some basics of survey design. There are eleven basic steps to design an effective survey. (Scheaffer, Mendenhall, & Ott, 37-38) The first step is for one to clearly and specifically state the objectives of the survey. It is important to be specific, so the intentions are understood by all involved. Along the same lines, the target population also must be specifically defined so the elements can be clearly identified. The selection of a frame is then needed to encompass the target population. At this point, all of the necessary background information is gathered and the appropriate sampling design can be selected, based on the criteria discussed later in the paper. It must then be determined how the measurements will be obtained and what will be used to obtain them. If a questionnaire is to be used, it is important that it is an unbiased and neutral questionnaire. The questions should not be leading or suggestive of a response. The questions and answer choices should also be ordered in a way that does not influence the subject being questioned. It is vitally important to use a questionnaire that is free from bias. The next step is to train workers to administer the survey. Workers need to be trained so that they are not leading in the way that questions are asked and also are able to gain the participation of the subject. Once the training has taken place, a pilot study or pretest should be run to check for any bias and allow the workers to practice administering the survey. Then it is time to collect and organize the data with some type of quality control to ensure accuracy. Finally the data are analyzed, and inferences are made about the target population.

In general, there are two types of errors that are associated with sampling. There are sampling errors and nonsampling errors. Sampling errors occur because only part of the

population is included in the sample, so errors occur because of the part of the population that is left out. (Thompson, 5) Sampling designs attempt to control and minimize sampling errors by using appropriate schemes for selecting from the population of interest. All other errors are referred to as nonsampling errors. Nonsampling errors include such things as nonresponse, measurement errors, and detectability problems. (Thompson, 5) Nonresponse occurs when a selected element of the population cannot or does not respond to the survey. It is not possible to substitute another element for the nonresponsive element due to the randomization involved in the original selection. It would damage the integrity of the results to allow a substitution. Sometimes measurements errors occur which are out of the control of the sampling design but affect the results. Detectability problems occur when it is difficult to generate a frame that covers the entire target population. This presents a problem since elements not included in the frame do not have a chance of selection. Nonsampling errors can be limited by training and planning in the initial stages of the study. Sampling errors can be minimized by the selection of an appropriate design.

Simple Random Sampling

-Without Replacement

The most basic sampling scheme is simple random sampling. Simple random sampling is a sampling design in which n distinct units are selected from the N units in the population in such a way that every possible combination of n units is equally likely to be sampled. (Thompson, 11) When one thinks of taking a random sample, they generally are thinking about simple random sampling. For example, if a community was interested in the average daily water consumption of households in the community, they would want to randomly select households to include in

the sample. Simple random sampling is the only sampling design where each sample has the same probability of selection. The probability of selection for an element under this design is $\pi_i = \frac{n}{N}$. There are various ways to generate a simple random sample. The most common would be to use a random number table or a random number generator. One would simply assign numbers to each of the sampling units and then generate random numbers in that range until the desired sample size is attained.

The three primary goals of conducting a sample are generally to estimate the population mean, the population total, or a population proportion. The actual values of the population mean and variance are given by $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$. (Thompson, 11) The population total is given by $\tau = \sum_{i=1}^{N} y_i$. These are the formulas that we would anticipate using for the population values. They are the common mean and variance formulas carried out for the whole population. The goal is to generate estimators that are unbiased or that minimize the amount of bias. For an estimator to be unbiased, the expected value of the estimator must equal the actual value of the parameter. So, in the case of the mean, it is unbiased if $E(\bar{y}) = \mu$. That is the average value of $\bar{y}$, when taken over all possible samples. The estimators for simple random sampling are all unbiased estimators. The estimator, $\bar{y}$, is considered to be a random variable whose outcome depends on which of the potential samples is selected. (Thompson, 17) The observations themselves are considered to be fixed. The following are estimators for simple random sampling without replacement.

Parameter: mean

Estimation: $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Estimation of variance: $\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$

(Scheaffer, Mendenhall, & Ott, 84)

Parameter: total

Estimation: $\hat{\tau} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i$

Estimation of variance: $\hat{V}(\hat{\tau}) = \hat{V}(N\bar{y}) = N^2\left(\frac{N-n}{N}\right)\frac{s^2}{n}$,

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, the usual formula for sample variance

(Scheaffer, Mendenhall, & Ott, 88)

Since the setting under consideration is without replacement, the sample is being taken from a finite population. When one samples without replacement, once a unit is selected it is removed from the frame and cannot be selected again. Since we have a finite population, it is necessary to correct our estimates to adjust for that fact. As can be seen above, each estimator of the variance is multiplied by the factor $\frac{N-n}{N}$. This factor is known as the finite population correction factor. It takes into account how much of the population is included in our sample relative to the size of the population. If the factor evaluates to be greater than 0.95, it can be ignored because the population is large enough, relative to the sample size, to be considered infinite. One can see that as $n \to N$ the factor goes to 0. This should make sense intuitively. If the sample includes all of the elements in the population, as in the case $n=N$, then there is no

uncertainty because the true value for each element of the population is known. Therefore, there would be no variance associated with that estimator. Likewise, as the sample size decreases, fewer elements of the population are included, leading to more uncertainty. So the variance is larger. The finite population correction adjusts the variance estimators for this fact in finite population settings.

-With Replacement

Simple random sampling with replacement is another type of simple random sampling that can be used. This is considered to be the infinite population setting since the selected units are returned to the frame and are eligible to be selected again. Simple random sampling with replacement generates *n* selections that are independent. Each unit has the same probability of inclusion in the sample. Each possible sequence of length *n*, considering order and possible repeat elements, has equal probability of selection. A downside to with replacement sampling is that it is less efficient than without replacement since the same element can be selected multiple times, thus not providing as much information on the population. (Thompson, 19) Sampling with replacement is helpful if working in a very large setting or some other setting where it is easier to select the elements from the frame without having to delete selected items from the frame. The estimators of the mean and variance are similar to those from the without replacement setting, just lacking the finite population correction factor.

Parameter: mean

Estimation: $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$

Estimation of variance: $\hat{V}(\bar{y}_n) = \frac{s^2}{n}$,

where s^2 is as in the without replacement setting

(Thompson, 19-20)

There is another option for estimating the mean in simple random sampling with replacement. It uses the effective sample size. Effective sample size is the number of distinct units contained in the sample. (Thompson, 20) It is denoted by ν. It is also an unbiased estimator and will yield a variance less than that of $\bar{y}_n$.

Parameter: mean

Estimation: $\bar{y}_\nu = \frac{1}{\nu}\sum_{i=1}^{\nu} y_i$

(Thompson, 20)

Even with this improved estimator, simple random sampling with replacement is still less efficient than simple random sampling without replacement.

-Confidence Intervals, Sample Size, and Proportions

Often we are not simply interested in a point estimate for a parameter but rather an interval estimate. A confidence interval with a specified level of significance, $1-\alpha$, can be developed. The idea of a confidence interval is that $P(\mu \in I) = 1-\alpha$, where I is some interval which varies from one sample to the next. (Thompson, 29) The confidence interval is estimated by using the estimates for the mean and variance of the population and a critical value from the standard normal distribution or Student's t distribution if in a small sample setting. Normality of the estimation can be assumed as a consequence of the Central Limit Theorem.

Confidence Interval: $\bar{y} \pm z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\bar{y})}$

Bound on Error/Margin of Error: $b = z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\bar{y})}$

(Thompson, 30)

Often this will not be reported as an interval, but rather as the bound on error or margin of error depending on the preference of the statistician. They are all simply different names for the same concept. Often margin of error is reported as a percentage. It is helpful in practice because it provides a way to express variability in terms of the estimate, which is often more clear to whomever is trying to interpret the data.

Another issue that is of interest when working with samples is the sample size. The formula for the bound on error is necessary for this calculation. One sets what they want the bound, *b*, to be and the level of confidence, α, and then solve the formula for the sample size, *n*.

Sample size for estimating the mean: $$n = \frac{1}{\frac{b^2}{z^2_{\frac{\alpha}{2}}\sigma^2} + \frac{1}{N}}$$

Sample size for estimating totals: $$n = \frac{1}{\frac{b^2}{N^2 z^2_{\frac{\alpha}{2}}\sigma^2} + \frac{1}{N}}$$

(Thompson, 36)

Sometimes one might be interested in estimating a proportion of a population that has or does not have a certain feature. This can also be accomplished within the framework of the simple random sampling design. An example of a proportion setting is a political poll. The objective of the poll is to determine the percentage or proportion of public support for an issue or candidate. The formulas use the binomial notation of $\hat{p}$ and $\hat{q}$, where $\hat{q} = 1 - \hat{p}$. Also, in this setting y_i is a binary response variable, with potential values of 0 and 1, corresponding to whether or not the observation has the desired trait of interest.

Parameter: proportion

Estimation: $\hat{p} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Estimation of variance: $\hat{V}(\hat{p}) = \left(\frac{N-n}{N}\right)\frac{\hat{p}\hat{q}}{n-1}$

(Scheaffer, Mendenhall, & Ott, 92)

The sample size computation for a proportion setting is the same as mentioned above where $\sigma^2 = pq$. If estimates for *p* and *q* are unavailable, then using *p=q=.5* will provide the most conservative sample size estimate. It does so because it is the maximum possible value that the variance can take on, thus giving a conservative result, since the actual variance can only be less than or equal to that value.

Ratio Estimation

Another type of estimator that is encountered in sampling problems is the ratio estimator. A ratio estimator sets up a ratio by using a variable that is highly correlated with the variable of interest in order to predict the desired outcome. Such a variable is called an auxiliary variable. A ratio estimator is especially helpful in situations where *N* is unknown or not cost effective to find, and it is not difficult to collect the auxiliary information. For example, the wholesale price paid for oranges in large shipments is based on the sugar content of the load. The exact sugar content can't be determined prior to the purchase and extraction of juice from the entire load. However, it is relatively easy to relate the weight of an individual orange with its sugar content. It is also easy to weigh the entire shipment. One can then set up a proportion to use the weight as an auxiliary variable. The proportion shows the ratio of the total weight to the individual weight is equal to the ratio of the total sugar content to the individual content. (Scheaffer, Mendenhall,

& Ott, 181-182) Sometimes an estimate of the ratio is desired and other times the estimation of the population mean or total are the desired estimators.

Parameter: ratio

Estimation: $r = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \dfrac{\bar{y}}{\bar{x}}$

Estimation of variance: $\hat{V}(r) = \left(\dfrac{N-n}{nN}\right)\left(\dfrac{1}{\mu_x^2}\right)\dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2$

(Scheaffer, Mendenhall, & Ott, 184)

The ratio estimator basically acts as a proportion to estimate the mean or total of a population.

$\dfrac{\mu_y}{\mu_x} \approx \dfrac{\bar{y}}{\bar{x}} \rightarrow \hat{\mu}_y = \dfrac{\bar{y}}{\bar{x}}\mu_x$ where x is the auxiliary variable

The logic would follow similarly if we were interested in estimating population totals. The variance estimate for the mean works out to be $\hat{V}(\hat{\mu}_y) = \mu_x^2 \hat{V}(r)$, and likewise, the variance estimate for the population total is $\hat{V}(\hat{\tau}_y) = \tau_x^2 \hat{V}(r)$. The ratio estimator provides a clever indirect alternative that can save one time and money depending on the setting of interest.

Stratified Random Sampling

There are many other sampling designs that are more efficient than simple random sampling in certain settings. Another type of sampling design is stratified random sampling. A stratified random sample is one obtained by separating the population elements into nonoverlapping groups, called strata, and then selecting a simple random sample from each stratum. (Scheaffer, Mendenhall, & Ott, 118) The goal when dividing the population into strata

is to make those elements classified to the same stratum as similar as possible. For example, if the question is one of a political nature, the strata could be based on political party affiliation.

There are several reasons why it would be beneficial to use a stratified design. Stratification can produce a smaller bound on error, or variance, of the estimate than a simple random sample of the same sample size. The more homogeneous the strata are within, the more likely it is that the bound on error would be smaller than that of a simple random sample. This is because the estimator for variance is made up of the within stratum variance terms. So when the within stratum variance is smaller, the variance estimate is smaller. Another benefit to stratifying is that it allows one to easily obtain estimates for population subgroups. Since the strata samples are independent, each can be treated as an individual simple random sample. Thus, it is very straight forward to obtain estimates by following the procedure for a simple random sample. This is not the case for all sampling designs. It can also be cheaper if the strata are grouped in a way that makes gathering the data convenient, such as geographical groupings.

Since the strata samples are themselves simple random samples, the estimators for the stratified sample make use of this property. They essentially combine the estimators for each of the strata to generate the estimators for the population.

Parameter: mean

Estimation: $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$

Estimation of variance: $\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$

(Scheaffer, Mendenhall, & Ott, 121)

Parameter: total

Estimation: $\hat{\tau} = N\bar{y}_{st} = \sum_{i=1}^{L} N_i \bar{y}_i$

Estimation of variance: $\hat{V}(\hat{\tau}) = N^2 \hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \left(\frac{s_i^2}{n_i} \right)$

(Scheaffer, Mendenhall, & Ott, 125)

Parameter: proportion

Estimation: $\hat{p}_{st} = \frac{1}{N} \sum_{i=1}^{L} N_i \hat{p}_i$

Estimation of variance: $\hat{V}(\hat{p}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \left(\frac{\hat{p}_i \hat{q}_i}{n_i - 1} \right)$,

where L = number of strata

N_i = number of sampling units in stratum i

N = number of sampling units in the population

s_i^2 = sample variance of stratum i

(Scheaffer, Mendenhall, & Ott, 137)

-Sample Size and Allocation

There is another issue to consider when using stratified sampling. The sample size cannot simply be calculated as in simple random sampling. There are two considerations that must be made in regard to sample size. The total sample size and the sample size within each stratum must be calculated. This is known as allocation of the sample. Allocation depends on a value known as the allocation fraction, a_i. In the simplest setting, the allocation fraction is constant for each stratum which creates equal sample sizes for each stratum. In a slightly more complicated setting, stratum sample sizes can be proportionally allocated by using an allocation

fraction of $\frac{N_i}{N}$. The allocation fraction is simply multiplied by the total sample size to determine the allocation for the strata.

Allocation with allocation fraction: $n_i = na_i$

(Scheaffer, Mendenhall, & Ott, 126)

However, there are other factors to consider when determining the allocation that complicate matters. There are three important factors to consider in the allocation process. The total number of elements in each stratum influences the allocation. For instance, a large stratum should have a large sample size allocated to it. The variability of observations within each stratum also needs to be considered. If there is a lot of variation, then a larger allocation is required and vice versa. The final factor is cost. If it is more expensive to get observations from one stratum, then the allocation will be smaller. The optimal allocation takes all of these factors into account when calculating the allocation. The total sample size formula and the allocation fraction are complicated by the consideration of these factors. The formulas below are for estimating the population mean. If one is estimating the population total, the formulas would have to be slightly modified by multiplying $z^2_{\frac{\alpha}{2}}$ by N^2 in the denominator.

Total sample size: $$n = \frac{\left(\sum_{k=1}^{L} \frac{N_k \sigma_k}{\sqrt{c_k}}\right)\left(\sum_{i=1}^{L} N_i \sigma_i \sqrt{c_i}\right)}{N^2 \left(\frac{b^2}{z^2_{\frac{\alpha}{2}}}\right) + \sum_{i=1}^{L} N_i \sigma_i^2}$$

Allocation: $$n_i = n\left(\frac{\frac{N_i \sigma_i}{\sqrt{c_i}}}{\sum_{k=1}^{L} \frac{N_k \sigma_k}{\sqrt{c_k}}}\right),$$

where c_i = cost per observation for stratum i

(Scheaffer, Mendenhall, & Ott, 130)

It is worth noting that the allocation for a stratum is directly proportional to the total size and variance of the stratum and inversely proportional to the cost per observation. The allocation of the sample size is a critical step in the process of a stratified sample. Since the estimators are formed by combining estimators from the strata, the accuracy of the results from each stratum affects the overall accuracy of the estimators.

Systematic Sampling

The next two designs seem quite different but are actually very similar. Systematic sampling is when a sample is obtained by randomly selecting one element from the first *k* elements in the frame and every *kth* element thereafter. (Scheaffer, Mendenhall, & Ott, 232) A cluster sample is a probability sample in which each sampling unit is a collection or cluster of elements, where the selection of a cluster means that all of the elements of the cluster are included in the sample. (Scheaffer, Mendenhall, & Ott, 266) These two designs are similar because in systematic sampling, one basically forms *k* clusters systematically and then selects which of the *k* to include randomly. In cluster sampling, the clusters are generally formed in a way that is convenient for the situation and then randomly selected for the sample. We will look at systematic sampling first.

There are several reasons why someone would choose to use a systematic sample. The main reason is that it is easier to execute in the field than other samples and does not require a good frame. It is easier because rather than randomly selecting all of the units for the sample, it is only necessary to select a *k* and a starting point. A random starting point is selected and then

every k elements thereafter are selected for the sample. It is a very straightforward method for selecting a sample. For example, a systematic sample would be easier to execute for a factory trying to evaluate the quality of an assembly line. It also is a good design for mobile populations, such as shoppers in a store or passengers on a bus. It should be noted that to attain a sample of n elements, one must choose k such that $k \leq \frac{N}{n}$. Systematic sampling uses the same estimators as simple random sampling, since they are approximately equivalent. It is not possible to calculate an unbiased estimator for the variance of a systematic sample because of the possibility of correlation between the elements in the sample, which introduces bias. The variance estimate is nearly unbiased if the sample is taken from a random population where the correlation does not exist. The reason for the correlation is that the population could have some periodic properties. So the elements of the population could have values that tend to cycle upward and downward in specific patterns. If k was improperly chosen, it could reasonably select either all of the high or low values in the population and produce and estimate that is way off of the actual value. In a situation like this, one should change starting points after taking a certain number of observations. This would help to ensure that different points of the population cycle were being sampled. There is a type of population that systematic sampling is more efficient in producing estimates. Systematic samples are good to use in situations where the population is ordered. It is good for ordered populations because it ensures that values throughout the spectrum of possibilities are selected. If it were a simple random sample, there would be no guarantee that the sample would cover the whole range of values. Another reason to use systematic samples is that it provides more information per unit cost than simple random sampling for populations such as those that are ordered. A systematic sample can be a relatively easy way to select a sample as long as the distribution of the population is considered.

Cluster Sampling

Now we will take a look at cluster sampling. Cluster sampling is an effective design when the population has the following qualities. It is helpful to use in situations where a good frame of population elements cannot be formed, but a frame of clusters is easy to find. It is also a good design to use if the clusters are geographically grouped since it is cheaper to obtain observations that involve less travel between them. An important consideration for cluster sampling is the make up of the individual clusters. Within each cluster, there should be as much difference as possible between the elements. Ideally, the cluster would exhibit the same trends and behavior of the population as a whole. Often, groupings like city blocks are used as clusters since a frame would be easy to generate and observations would be geographically close, reducing costs. Each of the clusters should also be similar to one another. Each cluster needs to be similar so when a cluster is selected it serves as an adequate representation of the population. Unlike stratified sampling, where each stratum is sampled, cluster sampling selects clusters and takes all of the elements contained in the selected clusters for inclusion in the sample. So, not all of the clusters are represented in the sample. The estimators for cluster sampling are similar to those for simple random sampling and also make use of ratio estimation. It is a ratio estimator because the response variable, y, and the size of the cluster, m, are random. The cluster size is random because it depends upon which cluster is selected. There are no set or predetermined cluster sizes, so it must be considered as random. The cluster size is viewed as the auxiliary information, using the vocabulary of the ratio estimator. The variance estimator is biased, but improves with $n>20$ and is unbiased if the clusters are of equal size.

Parameter: mean

Estimation: $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$

Estimation of variance: $\hat{V}(\bar{y}) = \left(\dfrac{N-n}{Nn\overline{M}^2}\right)\dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2$

(Scheaffer, Mendenhall, & Ott, 269)

Parameter: total

Estimation: $\hat{\tau} = M\bar{y}$

Estimation of variance: $\hat{V}(\hat{\tau}) = M^2\hat{V}(\bar{y})$

(Scheaffer, Mendenhall, & Ott, 273)

Parameter: proportion

Estimation: $\hat{p} = \dfrac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i}$

Estimation of variance: $\hat{V}(\hat{p}) = \left(\dfrac{N-n}{Nn\overline{M}^2}\right)\dfrac{\sum_{i=1}^{n}(a_i - \hat{p}m_i)^2}{n-1}$,

where N = number of clusters in the population

n = number of clusters selected in simple random sample

m_i = number of elements in cluster i

M = number of elements in the population

$\overline{M} = \dfrac{M}{N}$ = average cluster size for the population

y_i = total of all observations in the i^{th} cluster

a_i = number of elements in cluster i that possess a characteristic of interest

(Scheaffer, Mendenhall, & Ott, 282)

It is also worth noting that there is an alternative formula for calculating the population total that does not depend on knowing the total number of elements in the population. It is however a less precise estimator and can have a larger variance than the estimator above if there is a lot of variation in cluster sizes.

Parameter: total

Estimation: $\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} y_i$

Estimation of variance: $\hat{V}(\hat{\tau}) = N^2\left(\frac{N-n}{nN}\right)\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{\sum_{i=1}^{n} y_i}{n}\right)^2$

(Scheaffer, Mendenhall, & Ott, 273)

The formula to calculate the sample size is similar to the ones that we have previously examined. Again, only the formula for the sample size for estimating the mean will be presented.

Sample size: $n = \frac{N\sigma_r^2}{N\left(\frac{b^2\overline{M}^2}{z_{\frac{\alpha}{2}}^2}\right) + \sigma_r^2}$

(Scheaffer, Mendenhall, & Ott, 279)

-Two-Stage Cluster Sampling

Sometimes it is not necessary to include all of the elements in a cluster in the sample. In such cases, it is beneficial to use two-stage cluster sampling. Two-stage cluster sampling is a

sample obtained by first selecting a probability sample of clusters and then selecting a probability sample of elements from each sampled cluster. (Scheaffer, Mendenhall, & Ott, 304) There are different approaches that can be taken depending on the setting. For instance, one could sample a few elements from many clusters or many elements from a few clusters. The sample size from each cluster can also depend on the size of the cluster, with larger clusters having more elements sampled. It can be customized depending on the properties of the population of interest. Basically, a cluster sampling procedure is followed and then a simple random sample is drawn from each cluster that is selected. The estimators must now consider both parts of the sampling procedure. They must consider the random selection of the cluster and also the random selection of elements within each selected cluster. This point is quite clear when examining the formula for the variance estimate. It has two parts, one for each stage of the selection procedure.

Parameter: mean

Estimation: $\hat{\mu} = \frac{1}{\overline{M}n}\sum_{i=1}^{n} M_i \bar{y}_i$

Estimation of variance:

$$\hat{V}(\hat{\mu}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n\overline{M}^2}\right)s_b^2 + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right)$$

(Scheaffer, Mendenhall, & Ott, 306-307)

Parameter: total

Estimation: $\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} M_i \bar{y}_i$

Estimation of variance: $\hat{V}(\hat{\mu}) = \left(\frac{N-n}{N}\right)\left(\frac{N^2}{n}\right)s_b^2 + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right)$,

where $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\bar{y}_i - \overline{M}\hat{\mu}\right)^2$ and $s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2$

$\bar{y}_i$ = sample mean for i^{th} cluster

(Scheaffer, Mendenhall, & Ott, 309)

Two stage cluster sampling can be very beneficial in practice. It allows for the utilization of clusters but does not require the inclusion of every element in the cluster. This can help to reduce the cost of the research and also make the sample of a more manageable size for the analyst to compute estimates. The downside to the added convenience is that the formulas are much more complicated. Two-stage cluster sampling also illustrates how sampling theory has evolved beyond the basics to facilitate different settings and goals.

Capture-Recapture Sampling

Another example of a technique designed for a totally different setting is the capture-recapture sampling design which is designed to estimate the size of a population. The goal of capture-recapture sampling is to estimate the total of a mobile population, for example wildlife. There are two main methods for carrying out a capture-recapture study. First, we will consider the direct sampling approach. The direct sampling approach involves drawing a random sample of size *t* from the population and tagging or marking them in some way, then releasing them back into the population. Then, at a later time, a second sample of size *n* is taken noting the number of elements that were tagged, *s*. In order to come up with an estimator, *s* must be greater than zero. So, at least one element from the first sample must appear in the second sample. The random variable in this setting is *s*, the number of tagged elements in the second sample. The estimator is formed much like the ratio estimator for the mean. It is based on the proportionality of the ratio of the recaptured elements and the second sample size to the first sample size and the

total. The estimator for the total population is not unbiased and has a tendency to overestimate the population size.

Parameter: total

Estimation: $\frac{s}{n} \approx \frac{t}{\tau} \rightarrow \hat{\tau} = \frac{nt}{s}$

Estimation of variance: $\hat{V}(\hat{\tau}) = \frac{t^2 n(n-s)}{s^3}$

(Scheaffer, Mendenhall, & Ott, 329)

There is another estimator for direct sampling that was developed by Chapman and is nearly unbiased. Again, the trade off is a slightly more complex set of formulas.

Parameter: total

Estimation: $\hat{\tau}_c = \frac{(t+1)(n+1)}{s+1} - 1$

Estimation of variance: $\hat{V}(\hat{\tau}_c) = \frac{(t+1)(n+1)(t-s)(n-s)}{(s+1)^2(s+2)}$

(Scheaffer, Mendenhall, & Ott, 330)

The other method of capture-recapture sampling is called inverse sampling. Inverse sampling is very similar to direct sampling except the second sample continues until a pre-specified *s* is obtained. So in this setting, the random variable becomes *n*, since *s* is a set value that must be reached. Inverse sampling can be more precise than direct sampling if *n* is small relative to the population size, *N*. The estimators for inverse sampling are unbiased. Obviously, *s* must be greater than zero to calculate estimates.

Parameter: total

Estimation: $\hat{\tau} = \frac{nt}{s}$

Estimation of variance: $\hat{V}(\hat{\tau}) = \frac{t^2 n(n-s)}{s^2(s+1)}$

(Scheaffer, Mendenhall, & Ott, 331)

Much like cluster sampling, capture-recapture sampling can be done in multiple stages to improve the estimates. In multi-stage capture-recapture, the number of tagged units is tracked at each stage. For example, the second stage would note the number of tagged units captured and tag all of the untagged units captured. It would then be noted at each stage how many of the tagged and untagged are captured. Again, though, as the accuracy of the method improves, the calculations become more complex. Multi-stage capture-recapture sampling uses the product of hypergeometric distributions and maximum likelihood to estimate the population total. (Thompson, 242)

Conclusion

There are many other specialized sampling designs that have been created for specific purposes. The area of sampling is still rather young, with less than a century of intense study. Researchers are continuing to adapt designs to the needs of their studies and will continue to build on what is currently known. Sampling is a very important area of statistics since it is the foundation of a lot of research work. Without sampling, it would be very difficult to calculate reliable estimates for many important studies. For instance, sampling is crucial in creating some of the key economic indicators that can have a lot of influence with investors. Sampling will continue to grow from the basics and become increasingly complex in the future. It is, however, an area of which researchers of all disciplines should at least gain a broad understanding and is definitely worthwhile to study.

References

Scheaffer, R. L., Mendenhall, W., & Ott, L. (2006). *Elementary survey sampling* (6th ed.). Southbank, Vic. ; Belmont, CA: Thomson Brooks/Cole.

Thompson, S. K. (2002). *Sampling* (2nd ed.). New York: Wiley.