April 17, 2012

Chapter 2: Environmental Sampling

1 Introduction

This chapter discusses various means of obtaining observational data for environmental studies. That is, we have a population of interest that we want to study and the goal is to obtain a sample from the population. There exist several different types of sampling designs for obtaining data that are scientifically valid and which provide unbiased (or nearly unbiased) estimates of quantities of interest (e.g. means, totals, proportions). In addition, we can obtain standard errors of our estimates that indicate how precision of the estimates. Furthermore, one can save a lot of time, effort and money by using efficient sampling designs. Collecting data takes time and can be costly. Efficient sampling designs allow us to obtain the same level of information with smaller sample sizes compared to less efficient sampling designs. We shall illustrate these ideas with some simulation experiments using the software R.

One of the key points to understand is that a valid sampling plan is needed in order to obtain useful data. If the scientist simply goes out into the field and picks sites to sample with no plan ahead of time, then biases and other problems can lead to poor or worthless data.

Example: Estimate the number of trees in a forest with a particular disease. If the forest is large, then it may be impractical to examine every tree in the forest (e.g. it would take too much time or require a large team of researchers). Instead of examining all the trees, we can use ideas from sampling design. One idea is to divide the forest into plots of a particular size (say 1 acre) and then obtain a random sample of these plots. Next, count the number of diseased trees in each sampled plot and from these counts we can obtain an unbiased estimate of the total number of diseased trees and we can obtain a measure of precision of our estimated total.

Note that the term "*unbiased*" has a statistical definition: an estimator is unbiased if it does not systematically under, nor over-estimate, the true population value that is being estimated. If we are trying to estimate the total number of diseased trees in a forest, our estimated total is unbiased if it comes out exactly equal to the true total on average. What this means is that if we were to perform the sampling many times and compute a total estimate each time we take a sample, then the average of these estimated totals will equal the true total for the whole forest.

Also, in order to access the reliability of our estimates, we will often form *confidence intervals* that give an indication of the precision of our estimates. Confidence intervals can be thought of as giving a range of plausible values for the quantity being estimated. If one obtains a very small sample size, the resulting confidence interval will likely be very wide which is not very informative. However, as sample sizes get larger, confidence intervals get narrower leading to a more precise estimate of the parameter of interest.

Some of the most well-known sampling designs used in practice and discussed here are as follows:

- Simple Random Sampling
- Stratified Random Sampling
- Systematic Sampling
- Double Sampling
- Multistage Sampling

2 Terminology and Notation

First, we introduce some terminology and basic ideas.

Census: This occurs when one samples the entire population of interest.

The United States government tries to do this every 10 years. However, in practical problems, a true census is almost never possible.

In most practical problems, instead of obtaining a census, a **sample** is obtained by observing the population of interest, hopefully without disturbing the population. The sample will generally be a very tiny fraction of the whole population.

One must of course determine the population of interest – this is not always an easy problem. Also, the variable(s) of interest need to be decided upon.

Element: an object on which a measurement is taken.

Sampling Units: non-overlapping (usually) collections of elements from the population.

In some situations, it is easy to determine the sampling units (households, hospitals, plots, etc.) and in others there may not be well-defined sampling units (acre plots in a forest for example).

Example. Suppose we want to determine the concentration of a chemical in the soil at a site of interest. One way to do this is to subdivide the region into a grid. The sampling units then consist of the points making up the grid. The obvious question then becomes – how to determine **grid size**. One can think of the actual chemical concentration in the soil at the site varying over continuous spatial coordinates. Any grid that is used will provide a discrete approximation to the true soil contamination. Therefore, the finer the grid, the better the approximation to the truth.

Frame: A list of the sampling units.

Sample: A collection of sampling units from the frame.

Notation:

N	Number of Units in the Population
n	Sample size (number of units sampled)
y	Variable of interest.

Two Types of Errors.

- Sampling Errors – these result from the fact that we generally do not sample the entire population. For example, the sample mean will not equal the population mean. This statistical error is fine and expected. Statistical theory can be used to ascertain the degree of this error by way of standard error estimates.
- Non-Sampling Errors – this is a catchall phrase that corresponds to all errors other than sampling errors such as non-response and clerical errors. Sampling errors cannot be avoided (unless a census is taken). However, every effort should be made to avoid non-sampling errors by properly training those who do the sampling and carefully entering the data into a database etc.

3 Simple Random Sampling (SRS)

We begin with the simplest sampling design: the simple random sample.

Simple Random Sample: is the design where each subset of n units selected from the population of size N has the same chance (i.e. probability) of being selected.

Note: Under SRS each unit in the frame has the same chance of being selected in the sample. However, the converse is not true. That is, there are sampling plans where each unit has the same probability of selection into the sample, but it is not a SRS. The next example illustrates this.

Example: Suppose the frame for the population consists of sampling units labeled A, B, C, and D. Thus, $N = 4$ and we wish to obtain a sample of size $n = 2$. Then there are 6 possible random samples of size 2:

AB, AC, AD, BC, BD, CD

A simple random sample then requires that each of these 6 possible samples have an equal chance of being selected. In other words, the probability of obtaining anyone of these 6 samples is 1/6.

Now, if we only considered two possible samples: AB or CD, each with probability 1/2, then each sampling unit has a probability of 1/2 of being selected. But this is not a simple random sample.

Therefore, a simple random sample guarantees that each sampling unit has the same chance of being selected. On the other hand, a sampling plan where each unit has the same chance of being selected is not necessarily a simple random sample.

Question: How do we obtain a simple random sample? The answer is easy - simply label all the sampling units in the population as $1, 2, \ldots, N$ and then pick at random from this list a set of n numbers. This sampling is generally done *without replacement.* This is akin to putting the numbers 1 through N on a slip of paper, putting them in a hat and then random picking n slips of paper from the hat. Of course, actually writing numbers on a slip of paper and picking from a hat is quit tedious, especially if N is large. Instead, what is done in practice is to have a statistical or mathematical software package generate a random sample automatically. Many books make use of a table of random digits but these tables are rather archaic and it is suggested to simply use a computer for the task of choosing random samples.

In R, one can use the function "sample" to obtain a SRS. For example suppose the frame has $N = 100$ units and you want to select a SRS of $n = 5$ units. Then in R you can write the command

```
sample(1:100,5, replace=F),
```

which produces the sample: 46 23 20 93 78. If you run this again, you will get a completely different sample (with very high probability).

Note that the last part of the command says "replace=F" which means *sample without replacement.* This means that once a unit is selected, you do not return it to the mix when you go to pick your next sampling unit. That way, each unit can only be selected once. One could perform a sampling with replacement in which case the same unit can appear more than once in your sample. This is usually not preferred, although some sampling designs allow sampling with replacement.

3.1 Estimating the Population Mean

Let

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \text{and} \quad \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$$

denote the population mean and variance respectively. These population parameters are estimated by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

the sample mean and variance respectively. Using combinatorial counting techniques, it can be shown that the sample mean $\bar{y}$ is unbiased for μ. That is, the average value of $\bar{y}$ over all possible samples of size n is exactly equal to μ. Additionally, the sample variance s^2 is unbiased for σ^2.

Furthermore, using counting techniques, it also follows that

$$\text{var}(\bar{y}) = \{\sigma^2/n\}(1 - n/N).$$

The factor $(1 - n/N) = (N - n)/N$ is called the *finite population correction factor* which is approximately equal to 1 when n is a tiny fraction of N. The square-root of the variance of $\bar{y}$ is the **Standard error** of the sample mean. This is usually estimated by

$$\text{Estimated Standard Error of the mean:} = \frac{s}{\sqrt{n}}\sqrt{1 - n/N}.$$

Example: Consider two populations of sizes $N_1 = 1,000,000$ and $N_2 = 1000$. Suppose the variance of a variable y is the same for both populations. What will give a more accurate estimate of the mean of the population: a SRS of size 1000 from the first population or a SRS of size 30 from the second population? In the first case, 1000 out of a million is 1/1000th of the population. In the second case, 30/1000 is 3% of the population. Surprisingly, the sample from the larger population is more accurate.

Confidence Intervals. A $(1 - \alpha)100\%$ confidence interval for the population mean can be formed using the following formula:

$$\bar{y} \pm t_{\alpha/2,n-1}\widehat{SE}(\bar{y}) = \bar{y} \pm t_{\alpha/2,n-1}(s/\sqrt{n})\sqrt{1 - n/N},$$

where $t_{\alpha/2,n-1}$ is the $\alpha/2$ critical value of the t-distribution on $n-1$ degrees of freedom. This confidence interval is justified by applying a finite population version of the central limit theorem to the sample mean obtained from random sampling.

R-software for t-critical values: Traditionally statistics textbooks provided tables for looking up the t-critical value $t_{\alpha/2,n-1}$ needed in the confidence interval formula. These values can be easily obtained using the "qt" function in R where qt stands for quantile t-distribution. For instance, suppose a 95% confidence interval is needed and one has a sample size of $n = 50$. Then, since $0.95 = 1 - 0.05 = 1 - \alpha$, we have $\alpha/2 = 0.025$ and $1 - \alpha/2 = 0.975$. To get the t-critical value for the confidence interval formula, we can type "qt(0.975, 50-1)" in the R console window to get a value of $t_{\alpha/2,n-1} = 2.0096$. Quantiles of other well-known distributions can also be obtained (qnorm for the normal distribution for instance).

3.2 Estimating a Population Total

Often, interest lies in estimating the population total, call it τ. For instance, in the diseased tree example, one may be interested in knowing how many trees have the disease. If the sampling unit is a square acre and the forest has $N = 1000$ acres, then $\tau = N\mu = 1000\mu$. Since μ is estimated by $\bar{y}$, we can estimate the population total by

$$\hat{\tau} = N\bar{y} \tag{1}$$

and the variance of this estimator is

$$\text{var}(\hat{\tau}) = \text{var}(N\bar{y}) = N^2\text{var}(\bar{y}) = N^2(1 - n/N)\sigma^2/n.$$

Confidence Interval for Population Total. A $(1 - \alpha)100\%$ confidence interval for the population total τ is given by

$$\hat{\tau} \pm t_{\alpha/2,n-1}(s/\sqrt{n})\sqrt{N(N-n)}.$$

Sample Size Requirements.

One of the crucial questions that has to be answered when planning a sampling design is to determine how large of a sample size n to obtain. This is true of experimental designs as well. If n is too small, then the standard errors of the estimates may be so large that the estimates are of little use. On the other hand, very large sample sizes will take a lot of time and perhaps lead to a large cost. The idea is to find a sample size that can achieve a desired level of accuracy or statistical power. A sample size calculation may reveal that in order to achieve the a desired level of accuracy, you may need a sample size that is beyond your resources. It is good to know this before embarking on a long data collection job that may end up being a waste of time.

When estimating a population quantity such as a mean or total, it is quite common to use a confidence interval. To determine an appropriate sample size when using a confidence interval, first specify the desired degree of confidence, denoted $1 - \alpha$. Typically 95% confidence intervals are used so that the confidence level is 0.95 in which case $\alpha = 0.05$ since $0.95 = 1 - \alpha$. Next you need to specify how wide you would like your interval. The half-width of the interval will be denoted d so that with a high degree of confidence we know our parameter lies within d units from the true population value of the parameter. Once α and d are selected, we can use the confidence interval formulas to solve for n, the sample size. The following two formulas give the (approximate) sample size required for the population mean and total:

$$\text{For the mean } \mu\text{:}\ \ n \geq \frac{N\sigma^2 z_{\alpha/2}^2}{\sigma^2 z_{\alpha/2}^2 + Nd^2},$$

and

$$\text{For the total } \tau\text{:}\ \ n \geq \frac{N^2\sigma^2 z_{\alpha/2}^2}{N\sigma^2 z_{\alpha/2}^2 + d^2},$$

where $z_{\alpha/2}$ is the standard normal critical value (for instance, if $\alpha = 0.05$, the $z_{0.025} = 1.96$). These two formulas are easily derived algebraically solving for n in the confidence interval formulas.

Note that these formulas require that we plug a value in for σ^2 which is unknown in practice. To overcome this problem, one can use an estimate of σ^2 from a previous study or a pilot study. Alternatively, one can use a reasonable range of values for the variable of interest to get an estimate of σ^2: $\sigma \approx \text{Range}/6$.

Example. Suppose a study is done to estimate the number of ash trees in a state forest consisting of $N = 3000$ acres. A sample of $n = 100$ one-acre plots are selected at random and the number of ash trees per selected acre are counted. Suppose the average number of trees per acre was found to be $\bar{y} = 5.6$ with standard deviation $s = 3.2$. Find a 95% confidence interval for the total number of ash trees in the state forest.

The estimated total is $\hat{\tau} = N\bar{y} = 3000(5.6) = 16800$ ash trees in the forest. The 95% confidence interval is

$$16800 \pm 1.96(3.2/\sqrt{100})\sqrt{3000(3000-100)} = 16800 \pm 1849.97.$$

A Note of Caution. The confidence interval formulas given above for the mean and total will be approximately valid if the sampling distribution of the sample mean and total are approximately normal. However, the approximate normality may not hold if the sample size is too small and/or if the distribution of the variable is strongly skewed. To illustrate the problem, consider the following illustration. Standardized tests are conducted every year in Ohio public schools. Let us take each county as a sampling unit. Then $N = 88$ for the eighty eight counties in Ohio.

In this example, data on the actual number of third grade students in each county is already known sampling is not needed. However, we will use this example to illustrate how things can go wrong if the underlying distribution is strongly non-normal. The table below gives the tabulation from a recent school year of the number of third grade students in each county. The central limit theorem is an asymptotic result meaning that the normal approximation the sampling distribution of our estimators gets better and better as the sample size gets larger and larger. For finite sample sizes, the confidence intervals above based on the normal approximation may not be very accurate if the underlying population deviates strongly from normality. Using the power of computer simulation, we can determine how a particular statistical procedure actually works. In this case, since we have data on the whole population, we know the true population total. We can take numerous SRS's from the population, compute a confidence interval for each sample and then simply check if the parameter of interest, in this case, the true total, is in each interval. If we are using a 95% confidence interval, then 95% of the simulated confidence intervals should contain the true total.

```
 1  Adams       359
 2  Allen      1296
 3  Ashlan      520
 4  Ashtab     1274
 5  Athens      580
 6  Auglaize    558
 7  Belmont     638
 8  Brown       679
 9  Butler     3980
```

```
10  Carrol     249
11  Champaign  549
12  Clark     1748
13  Clermo    2083
14  Clinton    586
15  Columb    1221
16  Coshocton  415
17  Crawford   522
18  Cuyahoga 14570
19  Darke      637
20  Defian     447
21  Delaware  1448
22  Erie      1012
23  Fairfield 1710
24  Fayett     373
25  Frankl   13440
26  Fulton     658
27  Gallia     389
28  Geauga     941
29  Greene    1550
30  Guerns     464
31  Hamilton  8250
32  Hancock    888
33  Hardin     448
34  Harris     209
35  Henry      346
36  Highland   601
37  Hockin     264
38  Holmes     380
39  Huron      867
40  Jackson    383
41  Jefferson  778
42  Knox       613
43  Lake      2499
44  Lawren     822
45  Lickin    1979
46  Logan      558
47  Lorain    3618
48  Lucas     4632
49  Madison    517
50  Mahoni    2608
51  Marion     824
52  Medina    2250
53  Meigs      264
54  Mercer     602
55  Miami     1192
56  Monroe     185
57  Montgo    5459
58  Morgan     178
59  Morrow     413
60  Muskin    1206
61  Noble      181
62  Ottawa     436
63  Pauldi     267
64  Perry      440
65  Pickaw     699
66  Pike       406
67  Portage   1812
68  Preble     572
69  Putnam     435
70  Richla    1473
71  Ross       893
72  Sandus     713
73  Scioto     849
74  Seneca     601
75  Shelby     684
76  Stark     4576
77  Summit    6205
78  Trumbu    2556
79  Tuscararawas 1117
80  Union      572
81  VanWert    289
82  Vinton     179
83  Warren    2404
84  Washington 784
85  Wayne     1279
86  Willia     499
87  Wood      1363
88  Wyando     247
```

Figure 1 shows the actual distribution of students for the $N = 88$ counties and we see a very strongly skewed distribution. The reason for the skewness is that most counties are rural with small populations and hence relatively small numbers of children. Counties encompassing urban areas have very large populations and hence large numbers of students.

To illustrate the sampling distribution of the estimated total $\hat{\tau}$ where

$$\hat{\tau} = N\bar{y},$$

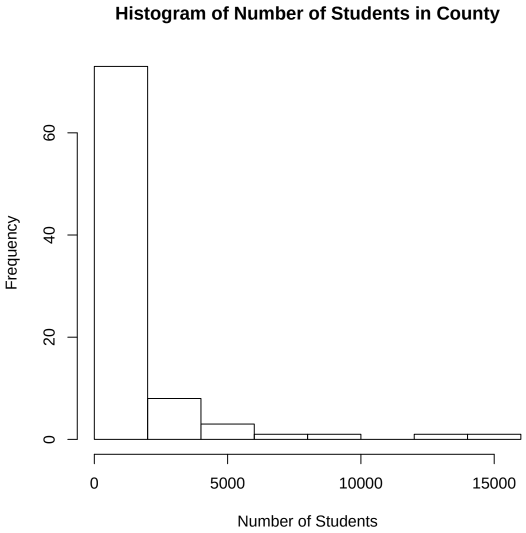

Figure 1: Actual distribution of student totals per county. Note that the distribution is very strongly skewed to the right.

10,000 samples of size n were obtained and for each sample, the total was estimated. The histograms show the sampling distribution for $\hat{\tau}$ for sample sizes of $n = 5, 25$, and 50 in Figure 2, Figure 3, and Figure 4 respectively. The long vertical line denotes the true total of $\tau = 131, 260$.

Clearly the sampling distribution of $\hat{\tau}$, the estimated total, is not nearly normal for $n = 5$. We see a bimodal distribution which results due to the presence of lightly populated and heavily populated counties.

Cochran (1977) gives the following rule of thumb for populations with positive skewness: the normal approximation will be reasonable provided the sample size n satisfies

$$n \geq 25G_1^2,$$

where G_1 is the population skewness,

$$G_1 = \sum_{i=1}^{N} (y_i - \mu)^3/(N\sigma^3).$$

For this particular example, we find

$$25G^2 = 357$$

which is much bigger than the entire number of sampling units (counties)!

In order to get an idea of how well the 95% confidence interval procedure works for this data, we performed the sampling 10,000 times for various sample sizes and computed the percentage of intervals that contained the true population total. If the confidence procedure works correctly, the percentage of intervals containing the true population total should be approximately 95%. The results are given in the follow table:

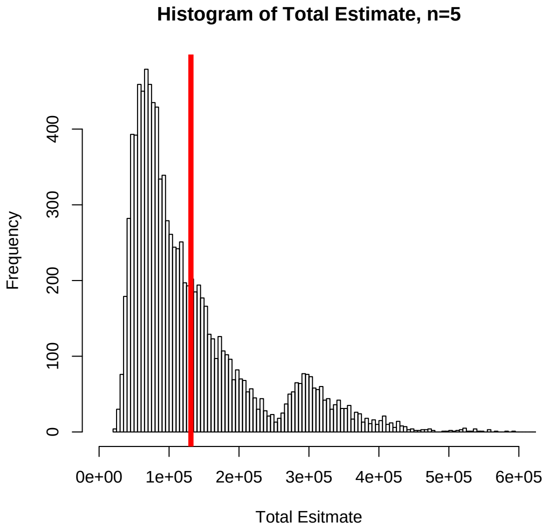

Figure 2: SRS of $n = 5$ for estimating the total number of students. The thick vertical line marks the true total.

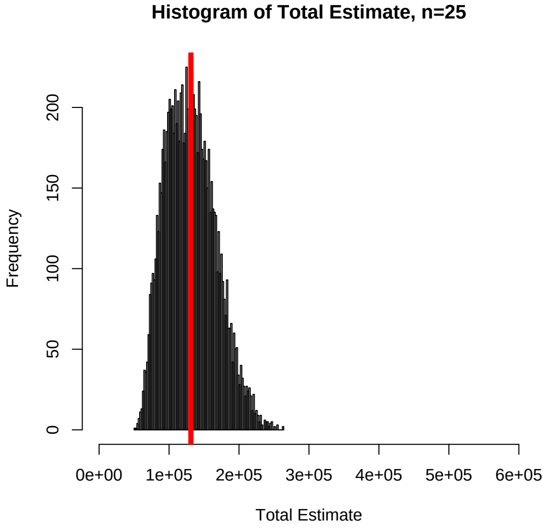

Figure 3: SRS of $n = 25$ for estimating the total number of students. The thick vertical line marks the true total.

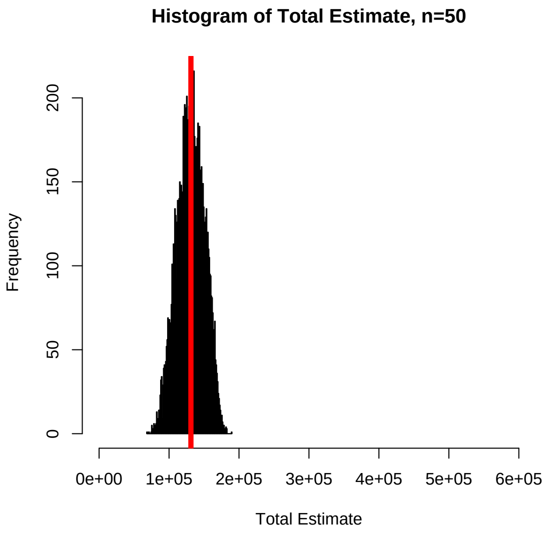

Figure 4: SRS of $n = 50$ for estimating the total number of students. The thick vertical line marks the true total.

Sample Size	Percentage
5	70%
10	74%
25	83%
50	89%

The simulation indicates that the true confidence level is quite a bit lower than the stated confidence level of 95%. For $n = 5$, only 70% of the 10,000 intervals contained the true population total.

This example illustrates that for a strongly non-normal population and relatively small sample sizes, the sample mean (and hence estimated total) will not be approximately normal and the confidence interval formulas given above may not valid. This example also illustrates the usefulness of computer simulations (run using R for this example) for checking if a statistical procedure will work properly or not.

3.3 Estimating a Population Proportion

Consider a situation where for each sampling unit we record a zero or a one indicating whether or not the sampling unit is of a particular type or not. A very common instance of this type of sampling is with opinion polls – do you or do you not support candidate X? Suppose you take a survey of plants and you note whether or not each plant has a particular disease. Interest in such a case focuses on the proportion of plants that have the disease. In this section we look at how to estimate the population proportion.

If we obtain a sample of size n from a population of size N, and each unit in the population either has or does not have a particular attribute of interest (e.g. disease or no disease), then the number of items in the sample that have the attribute is a random variable having a **hypergeometric distribution**. If N is considerably larger than n, then the hypergeometric distribution is approximated by the **binomial distribution**. We omit the details of these two probability distributions.

The data for experiments such as these looks like $y_1, y_2, \ldots, y_n$, where

$$y_i = \begin{cases} 1 & \text{if the } i\text{th unit has the attribute} \\ 0 & \text{if the } i\text{th unit does not have the attribute.} \end{cases}.$$

The population proportion is denoted by p and is given by

$$p = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

We can estimate p using the sample proportion $\hat{p}$ given by

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Note that in statistics, it is common to denote the estimator of a parameter such as p by $\hat{p}$ ("p"-hat). This goes for other parameters as well.

Using simple random sampling, one can show that

$$\mathrm{var}(\hat{p}) = (\frac{N-n}{N-1})\frac{p(1-p)}{n}. \tag{2}$$

An unbiased estimator of this variance is given by

$$\widehat{\mathrm{var}}(\hat{p}) = (\frac{N-n}{N})\frac{\hat{p}(1-\hat{p})}{n-1}. \tag{3}$$

An approximate $(1-\alpha)100\%$ confidence interval for the population proportion is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{(N-n)\hat{p}(1-\hat{p})}{N(n-1)}}.$$

This confidence interval is justified by assuming that the sample proportion behaves like a normal random variable which follows from the central limit theorem. The approximation is better when the true value of p is near $1/2$. If p is close to zero or one, the distribution of $\hat{p}$ tends to be skewed quite strongly unless the sample size is very large.

The sample size required to estimate p with confidence level $(1-\alpha)$ with half-width d is given by

$$n \geq \frac{z_{\alpha/2}^2 p(1-p)N}{z_{\alpha/2}^2 p(1-p) + d^2(N-1)}.$$

Note that this formula requires knowing p which is what we are trying to estimate! There are a couple ways around this problem. (1) Plug in $p = 1/2$ for p in the formula.

This will guarantee a larger than necessary sample size. (2) Use a rough estimate for p, perhaps based on a previous study.

A Common Misconception Related to Sampling. A few years ago, the town I live in wanted to do a survey of the town's residents. A controversy arose as to whether a sample should be taken or, if instead, a census should be done (i.e., send the survey to everyone in town). I addressed the town's council in favor of the sampling option. The person that spoke before me was also in favor of sampling and he said the sampling approach is valid *provided a fairly large proportion of the population is sampled.* However, this is wrong – precise results via sampling can be obtained even if the sample is a very tiny fraction of the whole population. To illustrate, consider two scenarios: (a) an opinion poll of $n = 1000$ people from a city of one million or (b) an opinion poll of $n = 100$ people from a city of ten thousand. Simple random sampling will guarantee both polls are accurate (unbiased) but which poll is more precise? In scenario (a), only 1/1000th of the population is polled ($n/N = 0.001$) whereas in scenario (b), the proportion of the population polled is 10-times greater at 1/100th of the population ($n/N = 0.01$). However, if the population proportion p is the same for both cities, the standard error of the estimator of p (i.e., the square root of the variance from equation (2)) is smaller for the large city than for the small city. For instance, if $p = 0.5$, then the standard error of $\hat{p}$ for the large city is 0.016 which is much smaller than the standard error of 0.050 for the small city. The main point here is that precise results can be obtained via sampling even if a very tiny fraction of the population is sampled. In most applications, N will be much larger than n in which case the finite population correction factor is negligible and the precision of the estimator is determined almost entirely by the sample size.

4 Stratified Random Sampling

Data is often expensive and time consuming to collect. Statistical ideas can be used to determine efficient sampling plans that will provide the same level of accuracy for estimating parameters with smaller sample sizes. The simple random sample works just fine, but we can often do better in terms of efficiency. There are numerous sampling designs that do a better job than simple random sampling. In this section we look at perhaps the most popular alternative to simple random sampling: Stratified Random Sampling. Stratified sampling can lead to dramatic improvements in precision over simple random sampling.

The idea is to partition the population into K different *strata*. Often the units within a strata will be more homogeneous. For stratified random sampling, one simply obtains a simple random sample in each strata. Then the values from each strata are combined to form an estimate for the overall population. Several decisions need to be made in order to implement stratified sampling, namely, how to define the strata and how many strata should be used. Often the population of interest will have naturally occurring strata. For instance, in a study of water quality, on could form strata corresponding to lakes in the region of interest. If a survey is to be conducted in a city, the different neighborhoods or subdivisions could define the strata. In other cases, it may not be clear how to define strata and there could many different possible choices for defining the strata. An optimal stratification corresponds to strata defined so that the corresponding estimates have the smallest variance possible. Once strata are defined, a decision has to be made as to how much of the overall sample will be allocated to each stratum.

There are three advantages to stratifying:

1. Parameter estimation can be more precise with stratification.
2. Sometimes stratifying reduces sampling cost, particularly if the strata are based on geographical considerations.
3. We can obtain separate estimates of parameters in each of the strata which may be of interest in of itself.

Examples.

- Estimate the mean PCB level in a particular species of fish. We could stratify the population of fish based on sex and also on the lakes the fish are living.
- Estimate the proportion of farms in Ohio that use a particular pesticide. We could stratify on the basis of the size of the farm (small, medium, large) and/or on geographical location etc.

These two examples illustrate a couple of points about stratification. Sometimes the units fall naturally into different stratum and sometimes they do not.

4.1 Notation and Estimators

Let N_h denote the size of the hth stratum for $h = 1, 2, \ldots, K$, where K is the number of strata. Then the overall population size is

$$N = \sum_{h=1}^{K} N_h.$$

If we obtain a random of size n_h from the hth stratum, we can estimate the mean of the ith stratum, $\bar{y}_h$ by simply averaging the data in the hth stratum. The estimated variance of $\bar{y}_h$ is

$$(s_h^2/n_h)(1 - n_h/N_h),$$

where s_h^2 is the sample variance at the hth stratum.

The population mean is given by

$$\mu = \sum_{h=1}^{K} N_h \mu_h / N,$$

which can be estimated by

$$\bar{y}_s = \sum_{h=1}^{K} N_h \bar{y}_h / N,$$

with an estimated variance given by

$$\hat{\sigma}^2_{\bar{y}_s} = \sum_{h=1}^{K} (\frac{N_h}{N})^2 (s_h^2/n_h)(1 - n_h/N_h).$$

The estimated standard error of $\bar{y}_s$, $S\hat{E}(\bar{y}_s)$ is the square root of this quantity.

Note that these last two estimators have sums involving N_h/N which is the proportion of the units in the hth stratum. Thus, these estimators can be regarded as *weighted averages* weighted by these proportions.

The population total $\tau = N\mu$ can be estimated using

$$\hat{\tau}_s = N\bar{y}_s$$

with estimated standard error

$$\widehat{SE}(\hat{\tau}_s) = N \cdot \widehat{SE}(\bar{y}_s)$$

Approximate $(1 - \alpha)100\%$ confidence intervals for the population mean and total using stratified random sampling are given by

$$\text{Population Mean: } \bar{y}_s \pm z_{\alpha/2} \widehat{SE}(\bar{y}_s),$$

and

$$\text{Population Total: } \hat{\tau}_s \pm z_{\alpha/2} \widehat{SE}(\hat{\tau}_s).$$

4.2 Post-Stratification

Sometimes the stratum to which a unit belongs is unknown until after the data is collected. For example, values such as age or sex which could be used to form stratum, but these values may not be known until individual units are sampled. The idea of post-stratification is to take a simple random sample first and then stratify the observations into strata after. Once this is done, the data can be treated as if it were a stratified random sample. One difference however is that in a post-stratification setting, the sample sizes at each stratum are not fixed ahead of time but are instead random quantities. This will cause a slight increase in the variability of the estimated mean (or total).

4.3 Allocation in Stratified Random Sampling

If a stratified sample of size n is to be obtained, the question arises as to how to allocate the sample to the different strata. In deciding the allocation, three factors need to be considered:

1. Total number N_h of elements in each stratum.
2. Variability σ_h^2 in each strata, and
3. The cost c_h of obtaining an observation from each stratum.

Intuitively, we would expect to allocate larger sample sizes to larger stratum and/or stratum with high variability. Surveys are often restricted by cost, so the cost may need to be considered. In some situations, the cost of sampling units at different strata could vary for various reasons (distance, terrain, water depth, etc.). One can show with some straightforward calculus that the optimal allocation of the total sample n to the hth stratum is to chose n_h proportional to

$$n_h \propto \frac{N_h \sigma_h}{\sqrt{c_h}},$$

where c_h is the cost for sampling a single unit from the hth stratum. Therefore, the i stratum will be allocated a larger sample size if its relative size or variance is big or its cost is low.

If the costs are the same per stratum or if cost information is unavailable or irrelevant, then the optimal allocation is given by

$$n_h \propto N_h \sigma_h,$$

which is known as *Neyman Allocation.*

A simple allocation formula is to use *proportional allocation* where the sample size allocated to each stratum is proportional to the size of the stratum, more specifically,

$$n_h = \frac{N_h}{N} n,$$

where n is the overall sample size. Proportional allocation is often nearly as good as optimal allocation if the costs and variances at the strata are nearly equal. Proportional allocation is a simple procedure that does not require knowledge of the within strata variances σ_h^2 which are usually unknown.

Example. (*open sturgeonsampling.r.*) In this example we use a computer simulation to compare SRS to stratified random sampling. Based on a study of sturgeon fish, data was simulated on mercury levels (mg/g) in the fish from three lakes. The goal is to estimate the mean mercury level in the fish from these lakes. The population size in the three lakes were set at $N_1 = 2000, N_2 = 3000$, and $N_3 = 4000$ and the population standard deviations were $\sigma_1 = 0.1642, \sigma_2 = 0.2087$, and $\sigma_3 = 0.1238$. Note that in this idealized world of simulation, we can know the truth since we have simulated the entire population. In practice, this information would not be available. However, the simulation will allow us to access how the different methods of estimation perform relative to each other. The population mean mercury level for this example is $\mu = 0.9482$. The goal is to find an estimator that comes closest to this parameter value on average. A sample of $n = 100$ fish are to be collected and the mercury levels measured in each. For SRS, one would randomly choose 100 fish from the total of $N = 2000 + 3000 + 4000 = 9000$.

The lengths (in cm) of the fish were also measured – we will use this information later when we discuss the ratio estimator.

For proportion allocation with stratified sampling, we would compute $n(N_1/N) = 100(2000/9000) = 22.22$ and round this to $n_1 = 22$. Similarly, from lake 2 we would compute $n(N_2/N) = 33.33$ and round this to get $n_2 = 33$, and then $n_3 = n - n_1 - n_2 = 45$.

For optimal allocation, one would compute

$$\{\frac{N_1\sigma_1}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3}\}n = 22.66$$

and then round this to $n_1 = 23$. Similarly, optimal allocation gives $n_2 = 43$, and $n_3 = 34$.

In order to compare how the three methods compare, we can take a large number of samples (since it is simulated data), compute our estimates for each sample and store the results. This is done in the R program "sturgeonsampling.r". In this program, 1000 samples were simulated and the mean mercury level was estimated using SRS and stratified sampling with proportion and optimal allocation. The average values of our estimators over all 1000 simulated samples are 0.9483, 0.9484, and 0.9487 which are all very close to the true mean of $\mu = 0.9482$. The standard deviation of these 1000 estimates are respectively $0.02021, 0.0166$, and 0.0156. As expected, stratified sampling with optimal allocation performs best – on average, over the 1000 simulated samples, this estimate has the smallest variance. However, stratified sampling with proportional allocation does almost as well as optimal allocation and SRS performs the worst. Figure 5 gives a visual representation of the performance of the three methods of estimation using nonparametric density estimates for the three methods. Note that the distributions for the stratified estimators are more tightly clustered around the true mean (indicated by the vertical line) than the SRS mean distribution.

4.4 Stratification for Estimating Proportions

A population proportion can be thought of as a population mean where the variable of interest takes only the values zero or one. Stratification can be used to estimate a proportion, just as it can be used to estimate a mean. The formula for the stratified estimate of a population proportion is given by

$$\hat{p}_s = \frac{1}{N}\sum_{h=1}^{K} N_h\hat{p}_h,$$

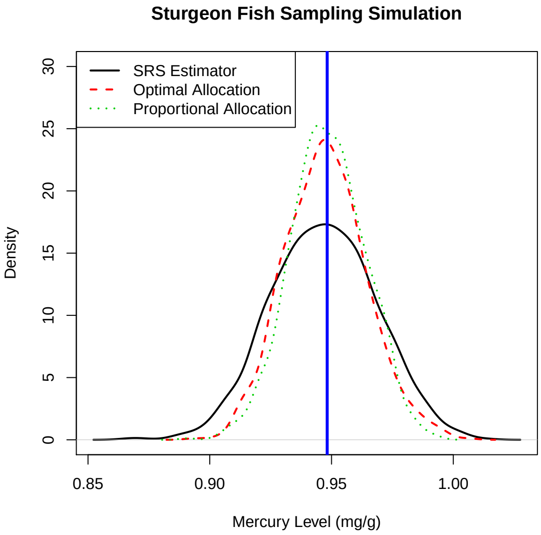

Figure 5: A comparison of SRS and stratified random sampling using proportional and optimal allocation for estimating the mean mercury levels in sturgeon fish from three lakes from simulated data.

and the estimated variance of this estimator is given by

$$\widehat{\text{var}}(\hat{p}_s) = \frac{1}{N^2}\sum_{h=1}^{K} N_h(N_h - n_h)\hat{p}_h(1-\hat{p}_h)/(n_h-1).$$

5 Systematic Sampling

Another sampling design that is often easy to implement is a systematic sample. The idea is to randomly choose a unit from the first k elements of the frame and then sample every kth unit thereafter. This is called a *one-in-k systematic sample.* A systematic sample is typically spread more evenly over the population of interest. This can be beneficial in some situations. In addition, a systematic sample may yield more precise estimators when the correlation between pairs of observations in the systematic sample is negative. However, if this correlation is positive, then the simple random sample will be more precise. We can use the same formulas for estimating the population mean and total as were used for a simple random sample. These estimators will be approximately unbiased for the population mean and variance. If the order of the units in the population are assumed to be arranged in a random order, then the variance of the sample mean from a systematic sample is the same of the variance from a simple random sample on average. In this case, the variance of $\bar{y}$ from a systematic sample can be estimated using the same formula as for a simple random sample: $(N-n)s^2/(Nn)$.

An alternative to estimating the variability is to consider the order of the observations in the systematic sample: $y_1, y_2, \ldots, y_n$ and then note that for consecutive neighboring points y_i and y_{i-1}, we have $E[(y_i - y_{i-1})^2] = 2\sigma^2$ assuming that neighboring points are independent. From this, it follows that

$$s_L^2 = 0.5 \sum_{i=2}^{n} (y_i - y_{i-1})^2/(n-1)$$

can be used to estimate the variance and therefore the standard error of the mean $\bar{y}$ can be estimated using

$$\widehat{\mathrm{SE}}(\bar{y}) = s_L/\sqrt{n}.$$

If the population has some periodic variation, then the systematic sampling approach may lead to poor estimates. Suppose you decide to use a systematic sample to monitor river water and you plan on obtaining samples every seventh day (a 1-in-7 systematic sample). Then this sampling plan reduces to taking a sample of water on the same day of the week for a number of weeks. If a plant upstream discharges waste on a particular day of the week, then the systematic sample may very likely produce a poor estimate of a population mean.

Systematic sampling can be used to estimate proportions as well as means and totals.

Systematic sampling can be used in conjunction with stratified random sampling. The idea is to stratify the population based on some criterion and then obtain a systematic sample within each stratum.

There are many different sampling designs used in practice and the choice will often be dictated by the type of survey that is required. We have discussed simple random sampling, stratified random sampling and systematic sampling. Now we briefly discuss a few other well-known sampling methodologies.

6 Cluster Sampling

The situation for cluster sampling is that the population consists of groups of units that are close in some sense (clusters). These groups are known as *primary units.* The idea of cluster sampling is to obtain a simple random sample of primary units and then to sample *every* unit within the cluster.

For example, suppose a survey of schools in the state is to be conducted to study the prevalence of lead paint. One could obtain a simple random sample of schools throughout the state but this could become quite complicated and lead to high costs due to a lot of travel. Instead, one could treat school districts as clusters and obtain a simple random sample of school districts. Once an investigator is in a particular school district, she could sample every school in the district.

A rule of thumb for determining appropriate clusters is that the number of elements in a cluster should be small (e.g. schools per district) relative to the population size and the number of clusters should be large. Note that one of the difficulties in sampling is obtaining a frame. Cluster sampling often makes this task much easier since it if often easy to compile a list of the primary sampling units (e.g. school districts).

Question: How is systematic sampling a special case of cluster sampling?

Cluster sampling is often less efficient than simple random sampling because units within a cluster often tend to be similar. Thus, if we sample every unit within a cluster, we are in a sense obtaining redundant information. However, if the cost of sampling an entire cluster is not too high, then cluster sampling becomes appealing for the sake of convenience. Note that we can increase the efficiency of cluster sampling by increasing the variability within clusters. That is, when deciding on how to form clusters, say over a spatial region, one could choose clusters that are long and thin as opposed to square or circular so that there will be more variability within each cluster.

Notation.

$$\begin{aligned}
N &= \text{The number of clusters}\\
n &= \text{Number of clusters selected in a simple random sample}\\
M_i &= \text{Number of elements in cluster } i\\
M &= \sum_{i=1}^{N} M_i = \text{Total number of elements in the population}\\
y_i &= \text{The total of all observations in the } i\text{th cluster}
\end{aligned}$$

There are two possibilities for estimating the population mean using a cluster sampling approach. A straightforward approach to estimating the population total using cluster sampling is to simply regard the y_i as a SRS from the set of primary units. Thus, the estimated population total is given by

$$\hat{\tau} = N\bar{y},$$

where

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

is the average of the cluster totals for the sampled clusters. The estimated variance of $N\bar{y}$ is

$$\widehat{\text{var}}(N\bar{y}) = N(N-n)s^2/n,$$

where

$$s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1).$$

$\hat{\tau} = N\bar{y}$ is an unbiased estimator of the population total, but because it does not use the information on the cluster sizes (e.g. the M_i's), the variance of $N\bar{y}$ tends to be bigger than the ratio estimator to be described next. The mean per primary unit can be estimated by simply $\bar{y}$ and the mean per secondary unit is estimated by $\hat{\mu} = \bar{y}/M$

The Ratio Estimator: A second approach to estimating the total is to use a *ratio* estimator:

$$\hat{\tau}_r = rM$$

where the ratio r is given by

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}.$$

The variance of $\hat{\tau}_r$ can be estimated using

$$\widehat{\text{var}}(\hat{\tau}_r) = \{\frac{N(N-n)}{n}\}s_r^2,$$

where

$$s_r^2 = \sum_{i=1}^{n}(y_i - rM_i)^2/(n-1).$$

Note that the ratio r can be used as an estimator of the mean per secondary unit. One problem with the ratio estimator is that it requires knowledge of M which is often not available.

7 Multistage Sampling

Multistage sampling is similar to cluster sampling. The idea is to determine a set of clusters (i.e. primary units). The first stage is to obtain a simple random sample of these clusters. The second stage is to obtain a simple random sample of units from each of the selected clusters. In cluster sampling, one would sample every unit within the cluster. However, for multistage sampling, only a sample of units within the selected clusters is obtained. In the school lead sampling, if the number of schools in districts is large, then multistage sampling may be preferred over cluster sampling. The variability of the estimators in two-stage sampling have an extra variance component compared to cluster sampling because only a sample is obtained in each cluster as opposed to conducting a census within each cluster. That is, because not all secondary units within a cluster are sampled, as they are in cluster sampling, the estimates from each cluster have variability associated with them. Multistage sampling differs from stratified sampling in that only a sample of clusters (or strata) are obtained. In stratified sampling, every cluster would be sampled.

Of course, multistage sampling can be generalized to any number of stages. Suppose you want to survey farms in the country. The first stage may be to randomly select a sample of states. In the second stage, select a sample of counties from each of the selected states. Finally, obtain a SRS of farms in each county.

Consider a two-stage sampling using the same notation as in cluster sampling. At the first stage, a SRS of n clusters (or primary units) is obtained. From each of these selected clusters, a SRS of m_i secondary units is selected from the ith cluster, for $i = 1, \ldots, n$. Let y_{ij} denote the response for the jth secondary unit in the ith primary unit. An unbiased estimate for the total in the ith primary unit is given by

$$\hat{y}_i = \frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{ij} = M_i\bar{y}_i.$$

The population total can then be estimated using

$$\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n}\hat{y}_i.$$

The variance of this estimated total will consist of two components due to sampling primary units and due to sampling secondary units within primary units. In particular, the variance of $\hat{\tau}$ in a two-stage sampling can be estimated using the following formula:

$$\widehat{\mathrm{var}}(\hat{\tau}) = N(N-n)s_b^2/n + \frac{N}{n}\sum_{i=1}^{n} M_i(M_i - m_i)s_i^2/m_i,$$

where s_b^2 is the estimated variance of the totals between clusters and is given by

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \hat{\mu}_1)^2,$$

with

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i,$$

is the estimated mean per primary unit. The within cluster variance is estimated by

$$s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2.$$

8 Summary of Classic Sampling Designs

In this section, we point out that stratified and cluster sampling are both special cases of two-stage sampling. The variance formulas for the estimated totals for these three sampling designs are:

$$\begin{aligned}
\text{Stratified} &: \mathrm{var}(\hat{\tau}) = \sum_{h=1}^{L} N_h(N_h - n_h)\sigma_h^2/n_h \\
\text{Cluster} &: \mathrm{var}(\hat{\tau}) = N(N-n)\sigma^2/n \\
\text{Two–stage} &: \mathrm{var}(\hat{\tau}) = N(N-n)\sigma_b^2/n + \frac{N}{n}\sum_{i=1}^{N} M_i(M_i - m_i)\sigma_i^2/m_i
\end{aligned}$$

where $\sigma_b^2 = \sum_{i=1}^{N}(y_i - \mu_1)^2/(N-1)$.

- Stratified sampling is a special case of two-stage sampling where for stratified sampling, in the first stage every cluster (or stratum) is sampled, as opposed to two-stage sampling where only a SRS of clusters is obtained. In this case, the first term in the variance formula for two-stage clustering is zero since $n = N$ and the N_h and n_h in the stratified formula correspond to the M_i and the m_i in the two-stage formula.

- Cluster sampling is a special case of two-stage sampling where, in the second stage, all secondary units in selected clusters are sampled. Thus, $m_i = M_i$ in each selected cluster and the second term in the two-stage variance formula goes to zero.

The idea behind stratified sampling is the *stratification principal* where the goal is to choose strata that are as homogeneous as possible. In this case, the within stratum variances will be small making stratified sampling more efficient than SRS. On the other hand, with cluster sampling, every unit within a selected cluster is sampled. Therefore, to optimize cluster sampling, one would want to choose clusters that are heterogeneous as possible. If all the units within a cluster are very similar to each other, then sampling every unit within the cluster is not very efficient. Thus, the goals of choosing strata in stratified sampling and choosing clusters in cluster sampling are different from one another.

9 Composite Sampling

In some applications we may obtain several samples from different units, such as blood samples from many people. Analyzing the samples may be time consuming and expensive. The idea of composite sampling is to mix samples that were obtained near each other to save on the cost of analyzing the sample. For example, consider the problem of testing blood to determine the proportion of people with syphilis. Initially, take one drop from each blood sample, mix these drops, and test the mixture for syphilis. If the test is negative, then syphilis is not present in any of the blood samples. However, if the test is positive, then the individual samples need to be tested. On average, the expected number of tests using composite sampling is much less than the number of samples present.

10 Ranked Set Sampling

This method is often used to save time and money for analyzing samples. The following example will help illustrate the procedure (this example is taken from: http://www.kent.ac.uk/IMS/personal/msr/rss1.html).

Example. The goal is to estimate the average amount of spray deposit on apple tree leaves. The sampling units are the leaves of the tree. Accurately computing the deposit density from the spray is time consuming: it requires an image analysis of the leaf to obtain a total pixel grey-scale value which is then divided by the leaf area. Suppose a sample of size $n = 5$ is to be obtained. The basic idea of ranked set sampling is to obtain a random sample of five leaves and *rank* them from highest to lowest spray deposit density. Pick the leaf with the highest spray concentration and accurately measure this concentration. Ranked set sampling requires that the observations can be quickly ranked. In this example, ranking the observations can be done if leaves are sprayed with a flourescent dye and examining them visually under ultraviolet light. Next, randomly pick five more leaves, rank them and then measure the spray density on the *second* highest leaf. Again, randomly pick five leaves, rank them and perform the measurement on the third highest leaf. Repeat this to get the fourth and fifth measurements. We can think of the data in the following illustration – each row corresponds to five sampled leaves. In the first row, the largest value is denoted by $x_{1(1)}$ and in the second row, the second largest value is denoted by $x_{2(2)}$, and so on.

$$\begin{array}{ccccccc}
x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & : & x_{1(1)} \\
x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & : & x_{2(2)} \\
x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & : & x_{3(3)} \\
x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & : & x_{4(4)} \\
x_{51} & x_{52} & x_{53} & x_{54} & x_{55} & : & x_{5(5)}
\end{array}$$

An unbiased estimator of the mean is given by the ranked set mean estimator:

$$\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x_{i(i)}.$$

It can be shown that the ranked set sample mean is more efficient than the simple random sample mean, i.e. the variance of $\bar{\bar{x}}$ is less than the variance of the sample mean from an ordinary simple random sample. In fact, the increased efficiency of ranked set sampling can be quite substantial. Of course if errors are likely when ranking the observations in each row above, then the efficiency of the ranked set sampling will decrease.

11 Ratio Estimation

A ratio estimator was introduced in the section on cluster sampling. In this section, we provide the general framework for ratio estimation. It is quite common that we will obtain auxiliary information on the units in our sample. In such cases, it makes good sense to use the information in this auxiliary information to improve the estimates of the parameters of interest, particularly if the auxiliary information provides information on the variable of interest.

Suppose y is the variable of interest and for each unit, there is another (auxiliary) variable x available. If x is correlated with y, then measurements on x provide information on y. Typically in practice, measurements on the variable x will be easier and/or less expensive to obtain and then we can use this information to get a more precise estimator for the mean or total of y. For instance, suppose we want to estimate the mean number of European corn bore egg masses on corn stalks. It is time consuming to inspect each and every leaf of the plant for corn borers. We could do this on a sample of plants. However, it is relatively easy to count the number of leaves on each given stalk of corn. It seems plausible that the number of egg masses on a plant will be correlated with the number of leaves on the plant. Returning to our sturgeon fish simulation, measuring the level of mercury in a fish is time consuming. However, one would expect the length of the fish to be positively correlated with the level of mercury since longer fish will tend to be older and will have bio-accumulated more mercury than shorter fish. The length of a fish is very easy to measure. Using the auxiliary information provided by the lengths of the fish can potentially lead to much more efficient estimators of means and totals.

A common use of ratio estimation is in situations where x is an earlier measurement taken on the population and y represents the current measurement. In these situations, we can use information from the previous measurements to help in the estimation of the current mean or total.

Suppose we obtain a sample of pairs $(x_1, y_1), \ldots, (x_n, y_n)$. We can compute the means of the two variables $\bar{y}$ and $\bar{x}$ and form their ratio:

$$r = \frac{\bar{y}}{\bar{x}}.$$

Letting μ_y and μ_x denote the population means of y and x respectively, then we would expect that

$$\frac{\mu_y}{\mu_x} \approx \frac{\bar{y}}{\bar{x}},$$

in which case

$$\mu_y \approx r\mu_x.$$

Using this relationship, we can define the ratio estimator of mean μ_y as

$$\bar{y}_r = r\mu_x, \tag{4}$$

and if N is the total population size, then the ratio estimator of the total τ_r is

$$\hat{\tau}_r = Nr\mu_x.$$

What is the intuition behind the ratio estimator? If the estimated ratio remains fairly constant regardless of the sample obtained, then there will be little variability in the estimated ratio and hence little variability in the estimated mean using the ratio estimator for the mean (or total).

Another way of thinking of the ratio estimator is as follows: suppose one obtains a sample and estimates μ_y using $\bar{y}$ and for this particular sample, $\bar{y}$ underestimates the true mean μ_y. Then the corresponding mean of x will also tend to underestimate μ_x for this sample if y and x are positively correlated. In other words, $\mu_u/\bar{x}$ will be greater than one. The ratio estimator of μ_y is

$$\bar{y}_r = r\mu_x = \bar{y}(\frac{\mu_x}{\bar{x}}).$$

From this relationship, we see that the ratio estimator takes the usual estimator $\bar{y}$ and scales it upwards by a factor of $\mu_x/\bar{x}$ which will help correct the under-estimation of $\bar{y}$.

There is a problem with the ratio estimator: it is biased. In other words, the ratio estimator of μ_y does not come out to μ_y on average. One can show that

$$E[\bar{y}_r] = \mu_y - \text{cov}(r, \bar{y}).$$

However, the variability of the ratio estimator can be considerably smaller than the variability of the usual estimator of $\bar{y}$ indicating that the ratio estimator may be preferred.

An estimate of the variance of the ratio estimator $\bar{y}_r$ is given by the following formula:

$$\widehat{\text{var}}(\bar{y}_r) = (\frac{N-n}{N})\sum_{i=1}^{n}(y_i - rx_i)^2/[n(n-1)]. \tag{5}$$

By the central limit theorem applied to the ratio estimator, $\bar{y}_r$ follows an approximate normal distribution for large sample sizes. In order to guarantee a good approximation, a rule of thumb in practice is to have $n \geq 30$ and the coefficient of variation $\sigma_y/\mu_y < 0.10$. If the coefficient of variation is large, then the variability of ratio estimator tends to be large as well.

An approximate confidence interval for the population mean using the ratio estimator is

$$\bar{y}_r \pm z_{\alpha/2}\widehat{se}(\bar{y}_r),$$

where $\widehat{se}(\bar{y}_r)$ is the square-root of the estimated variance of the ratio estimator in (5).

An approximate confidence interval for the population total using the ratio estimator is given by

$$\hat{\tau}_r \pm z_{\alpha/2}\widehat{se}(\hat{\tau}_r),$$

where

$$\widehat{se}(\hat{\tau}_r) = N\widehat{se}(\bar{y}_r).$$

When estimating the mean or total of a population when an auxiliary variable is available, one needs to decide between using the usual estimator $\bar{y}$ or the ratio estimator. If the correlation between y and x is substantial, then it seems that using the ratio estimator should be preferred. A rough rule of thumb in this regard is to use the ratio estimator when the correlation between y and x exceeds 0.5. There is a theoretical justification for this given in Cochran (1977, page 157) based on assuming the coefficient of variation for y and x are approximately equal.

Example. We return to the sturgeon fish simulation example illustrated above using the R-program *sturgeonsampling.r.* In this case, we will use the length of the fish as the auxiliary variable and ignore the different strata. Figure 6 shows a scatterplot of mercury levels versus length for the entire population from this simulated data set. Note that there is a very strong positive association between the two variables and in fact, the correlation between these two variables is $\rho = 0.81$. Once again, we shall use a sample size of $n = 100$ for illustration.

Once again, we will use simulation to compare the SRS estimator and the ratio estimator by simulating 1000 samples, each of size $n = 100$, and computing these estimators for each sample. Figure 7 shows the results of the simulation. Note that the sampling distribution for the ratio estimator (indicated by the dashed nonparametric density estimate curve) appears centered over the true mean so the bias is quite small for the ratio estimator in this example. In addition, the ratio estimator distribution is much tighter about the true mean indicating that it performs much better than the SRS estimator.

Just for fun, the simulation was repeated except this time, the sample size for the ratio estimator was cut in half to $n = 50$ although a sample size of 100 was used for the

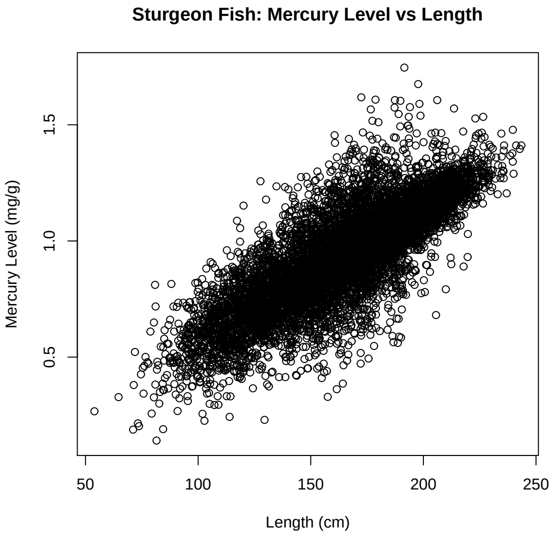

Figure 6: Mercury level versus length in the simulated Sturgeon fish example. Note the strong positive association between the two variables.

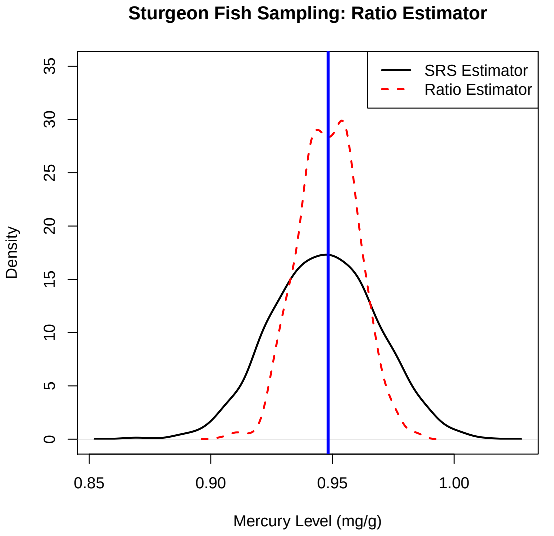

Figure 7: Sturgeon Fish Example: Ratio Estimator vs SRS estimator based on a simulation of 1000 data sets, each of sample size $n = 100$.

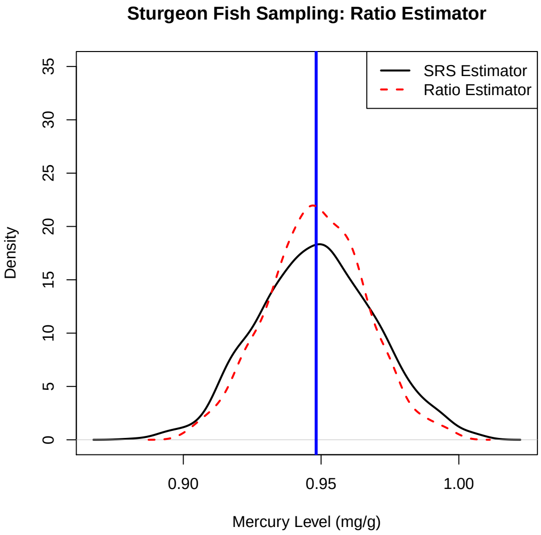

Figure 8: Sturgeon Fish Example: Ratio Estimator vs SRS estimator based on a simulation of 1000 data sets. In this simulation, a sample size of only 50 was used for the ratio estimator compared to a sample size of 100 using SRS.

usual SRS estimator. Figure 8 shows the results of this simulation and remarkably, the ratio estimator still outperforms the usual SRS estimator even though it is using only half the amount of data. That is, one can cut their sampling efforts in half (except for measuring the lengths of the fish) and still outperform the SRS estimator.

There exist sample size formulas for estimating means and totals using a ratio estimator which can be found in most textbooks on sampling. Note that if ratio estimation is more efficient than the usual simple random sample estimate, then smaller sample sizes can be obtained without sacrificing precision.

12 Regression Estimation

Ratio estimation does well when the regression line for y on x goes through the origin which stands to reason if the relationship $y = rx$ is approximately valid. There exist other examples where an auxiliary variable is available and the relationship between y and x is linear, but the line does not necessarily go through the origin. In these situations, it makes sense to utilize the information in the auxiliary variable using a simple linear regression relation between y and x:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where β_0 and β_1 are the intercept and slope of the line and ϵ is a random error to account for the fact that the sample points will not all lie exactly on a line. (See the next chapter on further details on regression analysis.)

Let $\hat{\beta}_1$ denote the usual least-squares estimator of the slope. Then the estimated regression line is given by

$$\hat{y} = \bar{y} + \hat{\beta}_1(x - \bar{x}).$$

Additionally, the least-squares regression line always passes through the mean $(\bar{x}, \bar{y})$. This suggest the following least-square regression estimator of the mean of y, denoted $\hat{\mu}_L$:

$$\hat{\mu}_L = \bar{y} + \hat{\beta}_1(\mu_x - \bar{x}). \tag{6}$$

Thus, the regression estimator takes the usual estimator $\bar{y}$ of the mean and adjusts it by adding $\hat{\beta}_1(\mu_x - \bar{x})$. In contrast, the ratio estimator takes the usual SRS estimator and adjusts it by a multiplicative factor as in (4).

- Typically the ratio estimator is preferred over the regression estimator for smaller sample sizes.
- Ratio and regression estimation can be used in conjunction with other types of sampling such as stratified sampling.

13 Double Sampling

Double sampling (also known as 2-phase sampling) is similar to ratio estimation in that it uses information from an auxiliary variable. For ratio estimation, it was assumed that the population mean μ_x was known for the auxiliary variable, but this may not always be the case.

The basic idea of double sampling is to first take a large preliminary sample and measure the auxiliary variable. It is assumed that the auxiliary variable will be easy and/or inexpensive to measure and that it will be correlated with the variable of interest. Then another sample (often a sub-sample of the first sample) is obtained where the variable y of interest is measured.

Some examples of easy-to-measure auxiliary variables are

- Examine aerial photographs of sampling units to get rough counts of trees, animals etc.
- Published data from past surveys.
- A quick computer search of files using a keyword for example.

In order to perform a double sampling, one first obtains a preliminary sample of size n' say and measures the variable x. From this preliminary sample, we can get an estimate of μ_x using

$$\hat{\mu}'_x = \sum_{i=1}^{n'} x'_i/n'.$$

Then one obtains the usual sample of size n, perhaps as a sub-sample of the preliminary sampled units. From this sample, we can compute the ratio as in a ratio sample:

$$r = \frac{\bar{y}}{\bar{x}}.$$

Then, the population total for y can be estimated using

$$\hat{\tau} = r\hat{\mu}'_x.$$

The variance for the estimated total using double sampling is more complicated than the variance of the ratio estimator because we have an extra source of variability with double sampling – namely the variability associated with the preliminary sample. The estimated variance of the double sampling total estimator is given by

$$\widehat{var}(\hat{\tau}) = N(N-n')s^2/n' + \frac{N^2(n'-n)}{nn'}s_r^2,$$

where

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

Notice that if $n' = N$, that is if the preliminary sample is of the entire population (i.e. a census), then the first term in this variance formula becomes zero and we end up with the same formula as the ratio estimator variance.

14 Unequal Probability Sampling

The sampling procedures discussed up to this point involve simple random sampling of sampling units in which case each unit has the same chance of being selected for the sample. Even with sampling designs more complicated than simple random sampling, such as stratified random sampling, a simple random sample was obtained in each stratum. In many situations, a simple random sample is either not possible or not preferable.

In *line-intercept* sampling for example, a line is more likely to intercept larger units than smaller units. If we divide an area into plots of sampling units, the plots may not all have the same size. In these cases, the probability of the unit to be selected into the sample will depend on the size of the unit. This is sometimes known as *probability proportional to size* estimation.

In what follows, let p_i denote the probability that the ith unit will be selected. For unequal probability sampling, we do not require all the p_i's to be equal.

14.1 Hansen-Hurwitz Estimator

Suppose sampling is done with replacement. Recall that when using simple random sampling, the population total is estimated by $\hat{\tau} = N\bar{y}$. We can rewrite this as

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} y_i/(1/N).$$

If we are sampling with replacement when each unit has the same chance of being selected, then the probability that a unit is selected at any given draw is $1/N$. For the Hansen-Hurwitz estimator, we simply replace the $1/N$ by p_i for the ith unit:

$$\hat{\tau}_{HH} = \frac{1}{n}\sum_{i=1}^{n} y_i/p_i \quad \text{(Hansen-Hurwitz estimation of total)}$$

14.2 Horvitz-Thompson Estimator

Sampling with replacement is not done often in practice as in the case of the Hansen-Hurwitz estimator. With the Horvitz-Thompson estimator, the sampling can be done either with or without replacement. We shall consider the case when the sampling is done without replacement. Let π_i denote the probability the ith sampling unit is selected in the sample. (Note that if all units have the same chance of being selected and we sample without replacement, then $\pi_i = n/N$. Can you explain why?)

The estimator of the population total is given by

$$\hat{\tau}_{HT} = \sum_{i=1}^{n} y_i/\pi_i \quad \text{(Horvitz-Thompson Estimator).}$$

The population mean can be estimated using

$$\hat{\mu}_{HT} = \hat{\tau}_{HT}/N.$$

assuming the n units selected are all distinct (this will not necessarily be the case when sampling with replacement). The variance formula for the Horvitz-Thompson estimator is quite complicated and involves probabilities of the form π_{ij} which denotes the probability that units i and j are both selected. Recent research into simpler variance formulas that do not require knowing the π_{ij} has been published, see for example Berger (2004). If sampling is done proportional to size and size of units vary, then the π_{ij} will vary in value as well.

15 Detectability

In some sampling cases, the elements may be difficult to detect within the sampling units. This may be the case in certain wildlife populations (e.g. fish, birds, etc.). If one is obtaining a simple random sample from a population of N units, then whether or not an animal in the unit is detected may not be certain, but instead a probability is associated with the chance the animal is detected. A non-animal example could occur when soil samples are assessed for a particular contaminant, some of the material may be missed due to sparsity of the contaminant.

Definition. The probability that an object in a selected unit is observed is termed its *detectability*.

For the sake of discussion, we shall refer to the objects as "animals." The following is some notation:

$$\begin{aligned} y &= \text{\# of animals observed} \\ \tau &= \text{total \# of animals} \\ p &= \text{probability an animal is observed.} \end{aligned}$$

If we assume independence between observations and a constant detectability probability p throughout a region, then

$$Y \sim \text{Binomial}(\tau, p),$$

that is, Y, the number of animals observed follows a binomial distribution on τ trials and success probability p. Therefore, the expected value of Y is

$$E[Y] = \tau p,$$

which indicates that we can estimate the total number of animals by solving for τ and using an estimate for the mean:

$$\hat{\tau} = y/p.$$

The variance of the binomial random variable Y is $\tau p(1-p)$ and thus

$$\text{var}(\hat{\tau}) = \frac{\tau p(1-p)}{p^2} = \frac{\tau(1-p)}{p},$$

which can be estimated by substituting $\hat{\tau}$ for τ to get

$$\widehat{\text{var}}(\hat{\tau}) = \frac{y(1-p)}{p^2}.$$

Notice that if the probability p of detection is small, then this variance becomes large.

If the area of the region of interest is A, then we can define the animal *density* as

$$D = \tau/A,$$

the number of animals per unit area. An estimate for the density then is

$$\hat{D} = \frac{y}{pA},$$

which has an estimated variance of

$$\widehat{\text{var}}(\hat{D}) = \frac{y}{A^2}(\frac{1-p}{p^2}).$$

These formulas require that we know the value of p but this is typically not the case in practice.

The question arises as to how to estimate p. Methods such as double sampling, capture–recapture or line transects can be used to estimate p. One way to estimate

p is to select n sampling units and let y_i denote the number of animals detected in the ith unit using the standard sampling technique. Then do an intensive search of each of these sampling units and let u_i denote the actual number of animals at the ith unit. Then an estimate of p is obtained by computing

$$\hat{p} = \frac{\bar{y}}{\bar{u}}.$$

The variance of this estimator can be estimated using ideas from ratio estimation.

If p has to be estimated, then the previous estimate of the population total τ can now be given as

$$\hat{\tau} = \frac{y}{\hat{p}}.$$

Since we now have the random $\hat{p}$ in the denominator instead of a fixed p, the variance of the estimated total increases by an extra term. An approximate formula for the variance of this estimated total can be derived using a Taylor series approximation to the ratio

$$\text{var}(\hat{\tau}) = \tau(\frac{1-p}{p}) + \frac{\tau^2}{p^2}\text{var}(\hat{p}).$$

In the formulas above, we have let y denote the number of animals observed from our sample. The value of y obtained depends on the sampling design used. For instance, if a simple random sample was used, then the estimate of the total was found to be $N\bar{y}$ assuming all animals could be detected. If p is the probability of detection, then the estimate of the total becomes

$$\hat{\tau} = N\bar{y}/p.$$

We can replace p by $\hat{p}$ in this formula when p needs to be estimated. The variance formula approximations become quite complicated in this case (e.g. see Thompson 1992).

16 Line Transect Method

In this section we give a brief introduction to some of the basic ideas of line transect sampling. The basic idea of the line transect method of sampling is for the observer to move along a selected line in the area of interest and note the location of animals (or plants) along the line and the distance from the line. The goal of the line transect method is to estimate the animal density D = (# of animal/unit area). Then the total number of animals can be found by computing

$$\tau = DA,$$

where A is the area of the region of interest. The observer will obtain a random sample of line transects. Let y_i denote the number of animals detected along the ith transect.

The Narrow Strip Method: Choose a strip of length L and let w_0 denote the distance to the left and right of the line where the observer will observe the animals – w_0 is called the half-width. A simple estimate of the density along the strip is

$$\frac{\text{Number of animals in the strip}}{\text{Area of the strip}} = \frac{y}{2w_0L}.$$

The narrow strip method assumes that animals anywhere in the strip are just as likely to be observed as anywhere else in the strip. However, a more realistic scenario is that the detectability decreases with the distance from the transect.

Instead of using the narrow strip method then, the data can be used to estimate a detectability function where the probability of detection drops off with the distance from the line transect. A couple popular parametric choices for the detectability functions are given by the exponential function and the half-normal function:

$$\begin{aligned} g(x) &= e^{-x/w} \text{ Exponential Function} \\ g(x) &= e^{-\pi x^2/(4w^2)} \text{ Half-Normal Function,} \end{aligned}$$

where w is a parameter typically estimated using maximum likelihood and x is the distance from the line. Instead of specifying a parametric form for the detection function (e.g. exponential and half-normal), nonparametric detection functions can be estimated using *kernel* methods.

For line transect sampling, more than one transect is obtained. One can obtain a simple random sample of transects. This is usually accomplished by drawing a line along one edge of the region and then selecting n points at random along this line. Then the transects are perpendicular lines extending from this baseline into the region at the n points. Note that biases can occur for transects that occur near the boundary of the region (e.g. there may be few animals along the boundary – there are ways of dealing with this that we will not go into here). If the region has an irregular shape, then the lengths L_i of the n transects will have varying lengths and therefore the lengths are random variables.

Instead of taking a simple random sample of transects, one could instead obtain a systematic sample of transects. This will help guarantee a more even coverage of the region.

Also, transect lines can also be selected with probability proportional to the length of the transect. The probability proportional to length selection can be accomplished by selected n points at random from the entire two-dimensional region and then select transects based on perpendicular lines that go through these selected points from the baseline.

17 The Data Quality Objectives Process

The collection of data can be time consuming and expensive. Therefore, it is very important to plan matters very carefully before undertaking a survey or experiment.

If too small a sample size is used, then there may not be enough information to make the resulting statistical analysis useful. For instance, confidence intervals may be too wide to be of any use or a statistical test may yield insignificant results even if there is a real effect. On the other hand, one does not want to unnecessarily expend too much money and resources obtaining more data than what is necessary in order to make a decision.

The U.S. Environmental Protection Agency (EPA) developed the *Data Quality Objectives (DQO)* to ensure the data collection process will be successful. Details can be found on the web at http://www.epa.gov/quality/qs-docs/g4-final.pdf.

The steps of the DPO can be summarized as following:

1. State the problem: describe the problem, review prior work, and understand important factors.
2. Identify the decision: what questions need to be answered?
3. Identify the inputs to the decision: determine what data is needed to answer questions.
4. Define the boundaries of the study: time periods and spatial areas to which the decisions will apply. Determine when and where data is to be gathered.
5. Develop a decision rule: define the parameter(s) of interest, specify action limits,
6. Specify tolerable limits on decision errors: this often involves issues of type I and type II probabilities in hypothesis testing.
7. Optimize the design for obtaining data: consider a variety of designs and attempt to determine which design will be the most resource-efficient.

This process may very well end up being an iterative process. Not only will later steps depend on the earlier steps but the later steps may make it necessary to rethink earlier steps as the process evolves. For instance, one may initially set unrealistic error bounds (type I and/or II) and then come to realize that these constraints would make the project go way over budget.

References

Berger, Y. G. (2004), "A Simple Variance Estimator for Unequal Probability Sampling without Replacement," *Journal of Applied Statistics*, **31**, 305–315.

Cochran, W. G. (1977), *Sampling Techniques*, 3rd edition, Wiley, New York.

Mohn, E. and Volden, R. (1985) "Acid precipitation: effects on small lake chemistry," in *Data Analysis in Real Life Environment: Ins and Outs of Solving Problems*, (Eds J. F. Marcotorchino, J. M. Proth and J. Janssen), pp. 191-196, Elsevier, Amsterdam.

Roberts, L., Lafta, R., Garfield, R., Khudhairi, J., Burnham, G., (2004), "Mortality before and after the 2003 invasion of Iraq: cluster sample survey," *The Lancet*, **364**, 1857-1864.

Scheaffer, R., Mendenhall, W. and Ott, R. (1996), *Elementary Survey Sampling*, 5th edition, New York: Duxbury Press.

Thompson, S. K. (1992), *Sampling*, New York: Wiley.