

Last updated April 8, 2008

Chapter 2: Environmental Sampling

This chapter discusses means of obtaining data for environmental studies. Either the data will come from a planned experiment in the lab or from sampling done in the field. This chapter discusses several methodologies for obtaining data in a scientifically valid way via sampling.

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. How can we do this? One idea is to divide the forest into plots of size 1 acre say and then obtain a random sample of these acres. Count the number of diseased trees in each sampled acre. From this sample, we can use statistical principals to estimate the number of trees in the forest with the disease.

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

2.1 Introduction

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, 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.

2.2 Simple Random Sampling (SRS)

One of the simplest sampling designs available is 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: It is possible to have a sampling plan where each of the possible samples considered have the same probability of selection *but* the sampling plan is not a SRS.

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 $1, 2, \dots, N$ 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.

2.3 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 $(1 - n/N)$ 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.

2.4 Estimating a Population Total

Often, interest lies in estimating the population total, call it T_y . 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 $T_y = N\mu = 1000\mu$. Since μ is estimated by \bar{y} , we can estimate the population total by

$$t_y = N\bar{y} \tag{1}$$

and the variance of this estimator is

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

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

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

Sample Size Requirements.

When using a confidence interval to estimate μ or T_y , the total, we may require that our estimate lies within d units from the true population parameter. How large a sample size is required so that the half-width of the confidence interval is d ? The following two formulas give the (approximate) sample size required for the population mean and total:

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

and

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

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 $t_y = N\bar{y} = 3000(5.6) = 16800$ 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. Suppose a survey is to be conducted to estimate the total number of students in Ohio public schools suffering from asthma. Let us take each county as a sampling unit. Then $N = 88$ for the eighty eight counties in Ohio.

For the sake of illustration, suppose we know the number of students in each county suffering from asthma and that the data is given in the following table:

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 with asthma 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 with asthma. Counties encompassing urban areas have very large populations and hence large numbers of students with asthma.

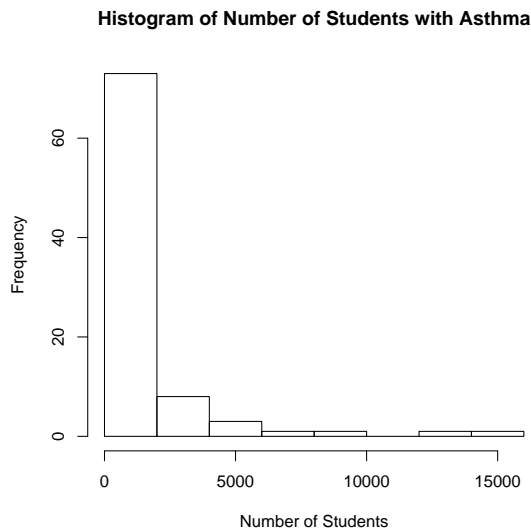


Figure 1: Actual distribution of student totals per county. Note that the distribution is very strongly skewed to the right.

To illustrate the sampling distribution of the estimated total t_y where

$$t_y = N\bar{y},$$

10,000 samples of size n were obtained and for each sample, the total was estimated. The histograms show the sampling distribution for t_y 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 $T = 131,260$.

Clearly the sampling distribution of t_y , 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,

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

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

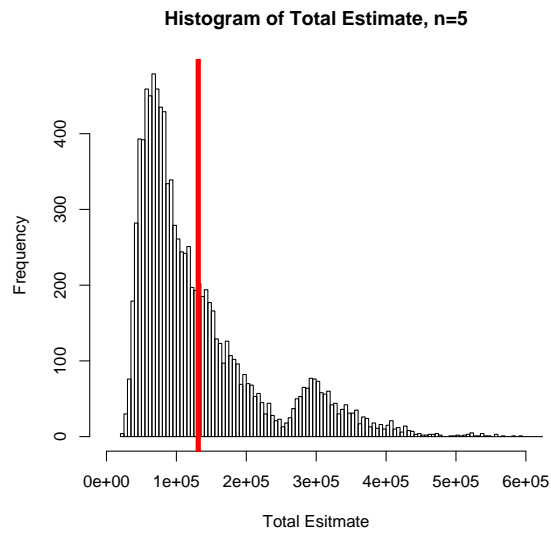


Figure 2: SRS of $n = 5$ for estimating the total number of students. The thick vertical line marks the true total.

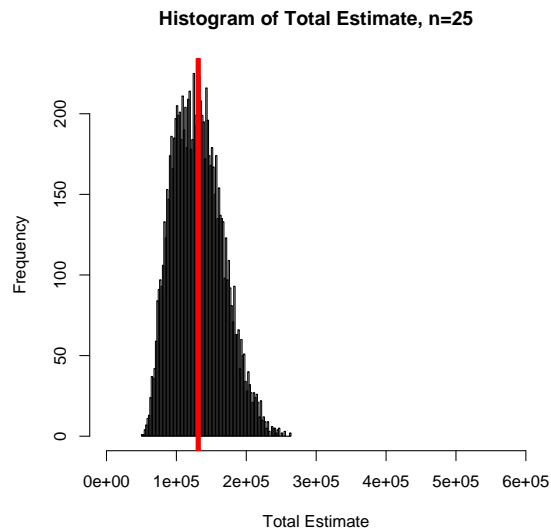


Figure 3: SRS of $n = 25$ for estimating the total number of students. The thick vertical line marks the true total.

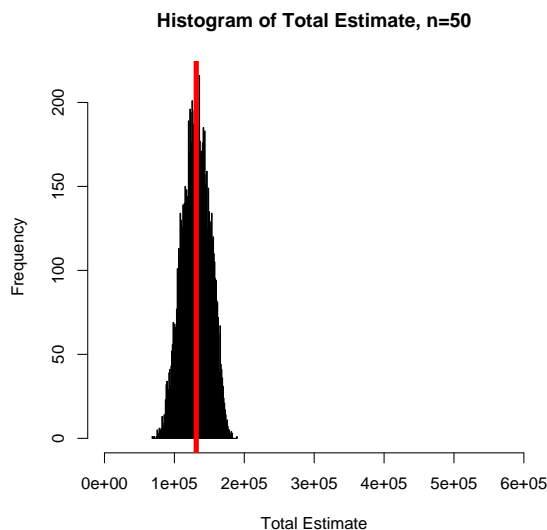


Figure 4: SRS of $n = 50$ for estimating the total number of students. The thick vertical line marks the true total.

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:

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.

Thus, 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 are not valid.

2.5 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, \dots, 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

$$\text{var}(\hat{p}) = \left(\frac{N-n}{N-1}\right) \frac{p(1-p)}{n}.$$

An unbiased estimator of this variance is given by

$$\hat{\text{var}}(\hat{p}) = \left(\frac{N-n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}.$$

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 guess for p , perhaps based on a previous study.

2.7 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.

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. Of course, the problem arises as to how many observations to allocate to each strata. Another issue is how to define the strata in the first place.

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.

Notation. Let N_i denote the size of the i th stratum for $i = 1, 2, \dots, K$, where K is the number of strata. Then the overall population size is

$$N = \sum_{i=1}^K N_i.$$

If we obtain a random of size n_i from the i th stratum, we can estimate the mean of the i th stratum, \bar{y}_i by simply averaging the data in the i th stratum. The estimated variance of \bar{y}_i is

$$(s_i^2/n_i)(1 - n_i/N_i),$$

where s_i^2 is the sample variance at the i th stratum.

The population mean is given by

$$\mu = \sum_{i=1}^K N_i \mu_i / N,$$

which can be estimated by

$$\bar{y}_s = \sum_{i=1}^K N_i \bar{y}_i / N,$$

with an estimated variance given by

$$\hat{\sigma}_{\bar{y}_s}^2 = \sum_{i=1}^K \left(\frac{N_i}{N}\right)^2 (s_i^2/n_i)(1 - n_i/N_i).$$

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

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

$$t_s = N\bar{y}_s$$

with estimated standard error

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

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

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

Example. A survey was done to estimate the average number of invasive honeysuckle plants per acre in a forest. The forest is partitioned into 158 acre plots. $N_1 = 86$ acres of the forest are new growth and $N_2 = 72$ acres are old growth. A sample of $n_1 = 14$ acres of new growth and $n_2 = 12$ acres of old growth forest were obtained yielding the following data:

New Growth	Old Growth
97 67 42 125	125 155 130 111
25 92 105 86	242 101 310 236
27 43 45 59	220 352 142 190
53 21	
$\bar{y}_1 = 63.36$	$\bar{y}_2 = 192.83$
$s_1 = 32.738$	$s_2 = 80.782$

The average number of plants per acre using the two-strata sampling is estimated to be:

$$\bar{y}_s = N_1\bar{y}_1/N + N_2\bar{y}_2/N = 86(63.36)/158 + 72(192.83)/158 = 122.36.$$

The standard error of this estimate is given by

$$\begin{aligned}\widehat{SE}(\bar{y}_s) &= \sqrt{(N_1/N)^2 s_1^2/n_1(1 - n_1/N_1) + (N_2/N)^2 s_2^2/n_2(1 - n_2/N_2)} \\ &= \sqrt{(86/158)^2 (32.738)^2/14(1 - 14/86) + (72/158)^2 (80.782)^2/12(1 - 12/72)} \\ &= 10.635.\end{aligned}$$

Thus, with 95% confidence, we estimate that the average number of honeysuckle per acre in the forest is

$$122.36 \pm 2(10.635) = 122.36 \pm 21.270 \text{ plants.}$$

It is interesting to note what would have happened if we had ignored the stratification and simply treated this as a simple random sample of size $n = n_1 + n_2 = 14 + 12 = 26$. The sample mean of all $n = 26$ acres is $\bar{y} = 123.12$ which is very close to the estimated mean found using the stratification formulas. The standard deviation for the $n = 26$ measurements is $s = 88.100$. The standard error of the mean using the simple random sampling formula is

$$\widehat{SE}(\bar{y}) = s/n(1 - n/N) = 88.100/26(1 - 26/158) = 15.792.$$

Thus, using a stratified sampling plan led to a much smaller standard error of the mean (10.635 compared to 15.792) than if we had just treated the data as a simple random sample. That is, the stratified design leads to a much more precise estimator of the mean. In addition, the stratification design allows us to obtain separate estimates of honeysuckle abundance in new and old growth parts of the forest.

2.8 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).

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 of elements in each stratum.
2. Variability in each strata, and
3. The cost 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, etc.). The optimal allocation of the total sample n to the i th stratum is to chose n_i proportional to

$$n_i \propto \frac{N_i \sigma_i}{\sqrt{c_i}},$$

where c_i is the cost for sampling a single unit from the i th 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, then the optimal allocation is given by

$$n_i \propto N_i \sigma_i,$$

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. This will be nearly optimal if the cost and variance at each stratum are nearly equal.

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

$$\hat{p}_s = \frac{1}{N} \sum_{i=1}^K N_i \hat{p}_i,$$

and the estimated variance of this estimator is given by

$$\widehat{\text{var}}(\hat{p}_s) = \frac{1}{N^2} \sum_{i=1}^K N_i (N_i - n_i) \hat{p}_i (1 - \hat{p}_i) / (n_i - 1).$$

2.9 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 k th 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, \dots, 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

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

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.

2.10 Other Design Strategies

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.

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 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 is often easy to compile a list of the primary sampling units (e.g. school districts).

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.

Estimation and standard error formulas for cluster sampling can be found in most textbooks on sampling (e.g. Scheaffer, Mendenhall, and Ott 1996).

Notation.

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

The population mean μ is estimated by

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

This estimator is a special case of a *ratio estimator* which we shall introduce a bit later. The estimated variance of \bar{y} is given by

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

where

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

and

$$\bar{M} = M/N,$$

the average size of a cluster for the population. Note that often in practice M and hence \bar{M} are unknown in which case \bar{M} can be estimated by

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

Estimating the Population Total in Cluster Sampling. An estimate of the population total in cluster sampling can be obtained in much the same way it was obtained in simple random sampling:

$$t_y = M\bar{y}.$$

The estimated variance of t_y is simply $M^2\widehat{\text{var}}(\bar{y})$. What is wrong with using this estimator of the population total? The problem is that it requires that we know M which is often unknown.

Alternatively, if we do not know M , we could estimate the population total using

$$N\bar{y}_t,$$

where

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

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

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

where

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

$N\bar{y}_t$ 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 $N\bar{y}_t$ tends to be bigger than the variance of t_y .

Example. Roberts et al (2004) used a cluster sampling approach to estimate the number of additional deaths in Iraq that resulted due to the Iraq war that started in 2003. From this article, it was widely reported that the number of Iraqi's killed from the war (so far) is 100,000. Their estimate of Iraqi deaths due to the war was 98,000 (not including Falluja which had a very high number of deaths). A 95% confidence interval for this total was given as (8000, 194000). 33 clusters were sampled based on Governorates and 30 households were interviewed in each cluster. The 33 clusters were sampled using a systematic sampling approach. Additional details can be found in the article.

Question: How is a cluster sample different from a stratified sample?

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. Multistage sampling differs from stratified sampling in that only a sample of clusters 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 lakes 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, sample lakes in each county.

Composite sampling – mixing 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.

Ranked set sampling – 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>).

Ranked set sampling 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 fluorescent 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.

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

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.

2.11 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 x is the variable of interest and for each unit, there is another (auxiliary) variable u available. If u is correlated with x , then measurements on u provide information on x . Typically in practice, measurements on the variable u 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 x . 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.

A common use of ratio estimation is in situations where u is an earlier measurement taken on the population and x 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 $(u_1, x_1), \dots, (u_n, x_n)$. We can compute the means of the two variables \bar{x} and \bar{u} and form their ratio:

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

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

$$\frac{\mu_x}{\mu_u} \approx \frac{\bar{x}}{\bar{u}},$$

in which case

$$\mu_x \approx r\mu_u.$$

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

$$\bar{x}_{\text{ratio}} = r\mu_u,$$

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

$$t_x = rN\mu_u.$$

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_x using \bar{x} and for this particular sample, \bar{x} underestimates the true mean μ_x . Then the corresponding mean of u will also tend to underestimate μ_u for this sample if x and u are positively correlated. In other words, μ_u/\bar{u} will be greater than one. The ratio estimator of μ_x is

$$\bar{x}_{\text{ratio}} = r\mu_u = \bar{x}\left(\frac{\mu_u}{\bar{u}}\right).$$

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

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

$$E[\bar{x}_{\text{ratio}}] = \mu_x - \text{cov}(r, \bar{x}).$$

However, the variability of the ratio estimator often tends to be smaller than the variability of the usual estimator of \bar{x} indicating that it may still be preferable.

An estimate of the variance of the ratio estimator \bar{x}_{ratio} is given by the following formula:

$$\widehat{\text{var}}(\bar{x}_{\text{ratio}}) = (1 - n/N) \sum_{i=1}^n (x_i - ru_i)^2 / [n(n-1)]. \quad (2)$$

By the central limit theorem applied to the ratio estimator, \bar{x}_{ratio} 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 $\sigma_x/\mu_x < 0.10$. 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

$$\bar{x}_{\text{ratio}} \pm z_{\alpha/2} \widehat{se}(\bar{x}_{\text{ratio}}),$$

where $\widehat{se}(\bar{x}_{\text{ratio}})$ is the square-root of the estimated variance of the ratio estimator in (2).

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

$$t_x \pm z_{\alpha/2} \widehat{se}(t_x),$$

where

$$\widehat{se}(t_x) = N \widehat{se}(\bar{x}_{\text{ratio}}).$$

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{x} or the ratio estimator. If the correlation between x and u 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 x and u exceeds 0.5. There is a theoretical justification for this given in Cochran (1977, page 157) based on assuming the coefficient of variation for x and u are approximately equal.

Example. A study of acid rain was undertaken by examining samples of water in 32 lakes in 1977 (Mohn and Volden 1985). In 1976, the pH was measured in the population of all $N = 68$ lakes which gave a mean value of $\mu_u = 5.715$ in 1976. Figure 5 shows a scatterplot of the pH values from the sample of $n = 32$ lakes in 1977. The goal is to estimate the mean pH level μ_x for all $N = 68$ lakes for 1977. The data for the $n = 32$ lakes are given in the following table:

1976 1977

 4.32 4.23
 4.97 4.74
 4.58 4.55
 4.72 4.81
 4.53 4.70
 4.96 5.35
 5.31 5.14
 5.42 5.15
 4.87 4.76
 5.87 5.95
 6.27 6.28
 6.67 6.44
 5.38 5.32
 5.41 5.94
 5.60 6.10
 4.93 4.94
 5.60 5.69
 6.72 6.59

5.97 6.02
 4.68 4.72
 6.23 6.34
 6.15 6.23
 4.82 4.77
 5.42 4.82
 5.31 5.77
 6.26 5.03
 5.99 6.10
 4.88 4.99
 4.60 4.88
 4.85 4.65
 5.97 5.82
 6.05 5.97

The sample means for the $n = 32$ lakes are

$$\bar{x} = 5.3997 \text{ and } \bar{u} = 5.4159,$$

which gives an estimated ratio of

$$r = \frac{\bar{x}}{\bar{u}} = \frac{5.3997}{5.4159} = 0.9970.$$

The ratio estimator of μ_x , the average pH in the 68 lakes is

$$\bar{x}_{\text{ratio}} = r\mu_u = (0.9970)(5.715) = 5.6979,$$

which is higher than the simple estimate of $\bar{x} = 5.3997$. Therefore, the ratio estimate takes the usual estimate of 5.3997 and scales it up by a factor of $\mu_U/\bar{u} = 5.715/5.4159 = 1.0552$. The sample correlation between pH in 1976 and 1977 for the 32 lakes is 0.883 which indicates that the ratio estimator will be more efficient than the usual simple random sample estimator of the mean. The estimated coefficient of variation for 1976 and 1977 are respectively 0.1234 and 0.1244. Although the coefficient of variation for 1977 exceeds our rule of thumb value of 0.10, it does not exceed it by much.

The estimated variance for the ratio estimator can be computed as

$$\widehat{\text{var}}(\bar{x}_{\text{ratio}}) = (1-n/N) \sum_{i=1}^{32} (x_i - 0.9970u_i)^2 / [32(31)] = (1-32/68)(3.2473) / [32(31)] = 0.0017.$$

The standard error of \bar{x}_{ratio} is obtained by taking the square root of this quantity which gives $\widehat{\text{se}}(\bar{x}_{\text{ratio}}) = \sqrt{0.0017} = 0.0412$. A 95% confidence interval for μ_x is

$$5.6979 \pm 1.96(0.0412) = 5.6979 \pm 0.0808.$$

Note that if we had just used the sample mean to estimate the population mean (obtaining $\bar{x} = 5.3997$), the associated standard error would be

$$\widehat{\text{se}}(\bar{x}) = (s/\sqrt{n})\sqrt{1-n/N} = (0.6716/\sqrt{32})\sqrt{1-32/68} = 0.0864$$

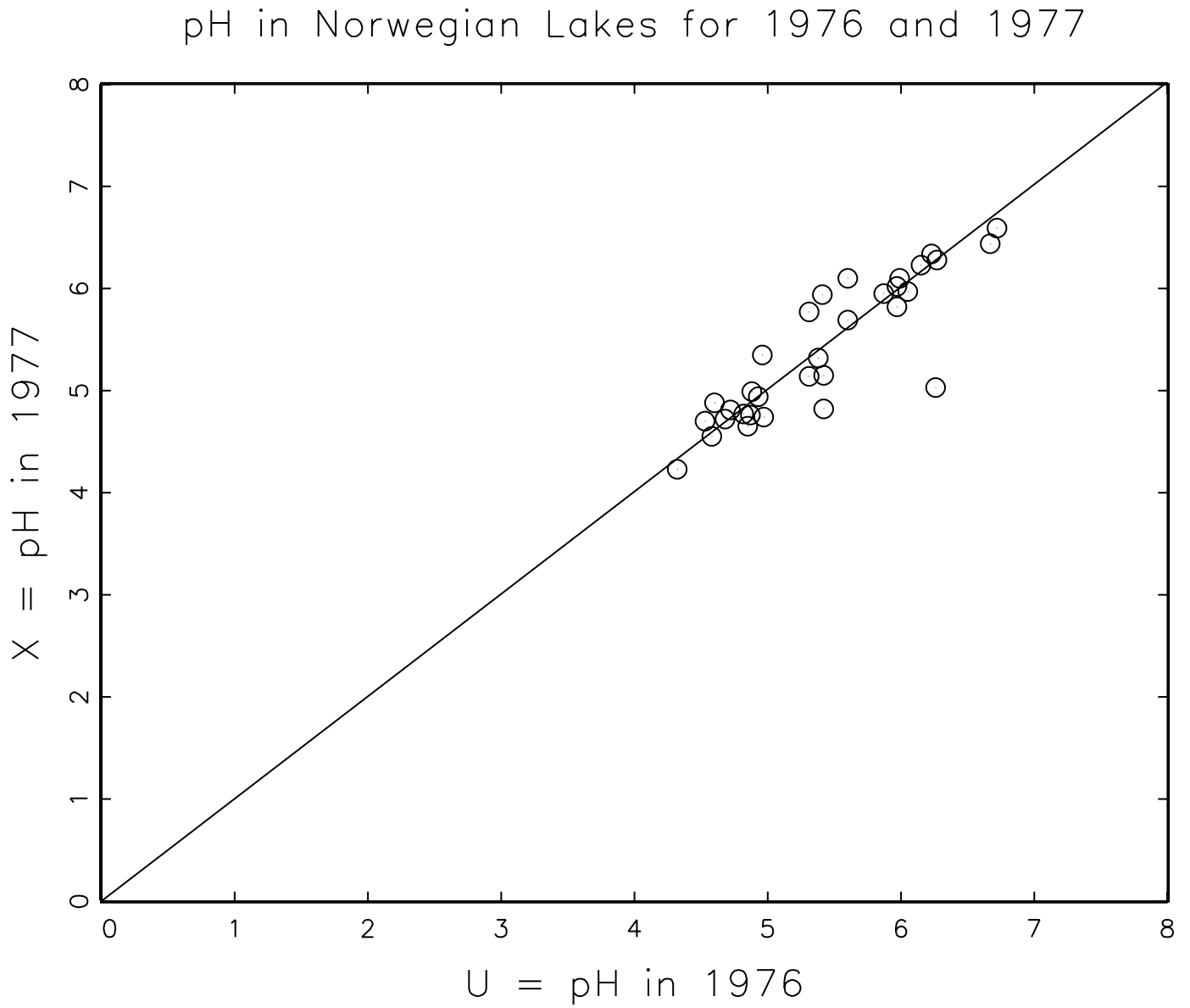


Figure 5: Scatterplot of pH in 1977 versus 1976 at 32 Norwegian lakes. These 32 lakes are a subset of all 68 lakes.

which is more than twice the standard error of the ratio estimator. This indicates that the ratio estimator is a more efficient estimator of the population mean.

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 will be required for the same level of precision.

Regression Estimation

Note that the line in Figure 5 appears to go through the origin which stands to reason if the relationship $x = ru$ is approximately valid. There exist other examples where an auxiliary variable is available and the relationship between x and u 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 x and u :

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

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.

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

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

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

$$\hat{\mu}_L = \bar{x} + \hat{\beta}_1(\mu_u - \bar{u}).$$

Thus, the regression estimator takes the usual estimator \bar{x} of the mean and adjusts it by adding $\hat{\beta}_1(\mu_u - \bar{u})$.

- 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.

2.12 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_u 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 x 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 u . From this preliminary sample, we can get an estimate of μ_u using

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

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:

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

Then, the population total for x can be estimated using

$$t_x = r\hat{\mu}'_u.$$

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

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

where

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

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.

2.14 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.

Let p_i denote the probability that the i th unit will be selected.

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

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

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 i th unit:

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

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 i th 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

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

The population mean can be estimated using

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

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.

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:

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

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

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

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

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

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

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

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).

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 $D = (\# \text{ of animal/unit area})$. 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 i th 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.

2.15 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.