The Conduct and Analysis of Sample Surveys [1] [2]

J. D. Petruccelli B. Nandram M. Chen

[1] Development of these materials was partially supported by the National Science Foundation's Division of Undergraduate Education under Grant DUE 9254087.

Contents

Preface

This document offers an introduction to the design and analysis of sample surveys. It assumes a basic knowledge of surveys, sampling, and statistical inference, as found, for example, in chapters 1-6 of Petruccelli, Nandram and Chen, *Applied Statistics for Engineers and Scientists*, 1999, Prentice Hall. Knowledge of simple linear regression, as found in chapter 7 of that text, is needed in the discussion of regression estimation. The reader will note that reference is made to these chapters in the narrative. This is because the material presented here consists of material omitted from the Petruccelli, Nandram and Chen book. In keeping with its origins, we present this material as chapter 16 of that text, complete with exercises, labs, and a mini-project. We also present it in the hopes it will be useful to those contemplating conduct of a sample survey.

Chapter 16

The Conduct and Analysis of Sample Surveys

"I am monarch of all I survey."

-William Cowper

16.1 Introduction

In Chapter 3 you learned the difference between a sampling study and an experiment. You also learned about some of the issues in sampling studies, and some of the terminology used to describe these issues. Before beginning this chapter, you should review the material on sampling studies from Chapter 3.

Sampling studies are used in many fields. Probably the most familiar examples of sampling studies are opinion polls, such as the Gallup poll. Sampling studies also occur frequently in industry. For example, a type of sampling called acceptance sampling is used to check the quality of incoming parts or materials, or of outgoing shipments of completed products.

In this chapter we present both theory and practical aspects of the design and conduct of sampling studies and of the analysis of data obtained from such studies. In particular, we consider a number of sampling designs, including simple random sampling, stratified random sampling, cluster sampling, systematic sampling, and double sampling. We describe simple random sampling and stratified random sampling in detail. Finally, details of how to plan and conduct a sampling study are presented.

In terms of the analysis of data obtained from a sampling study, we present basic results on unbiased estimation for a finite population total, mean and proportion under simple random sampling and stratified random sampling designs. We also present ratio and regression estimation.

When you complete this chapter, you will be able to plan, conduct and analyze a number of basic kinds of sampling studies.

Knowledge and Skills

By successfully completing this chapter, you will acquire the following knowledge and skills:

KNOWLEDGE

1. Various sampling schemes: simple random sampling with and without replacement, stratified random sampling, cluster sampling, multi-stage sampling, systematic sampling, and double sampling.
2. Unbiased estimation for population total, mean and proportion.
3. Ratio and regression estimation.
4. The steps in planning, conducting and analyzing a sampling study.

SKILLS

1. The ability to plan, conduct and analyze a sampling study based on simple random sampling, stratified random sampling or single-stage cluster sampling.
2. The ability to perform unbiased, ratio, and regression estimation.

16.2 Some Terminology

In this chapter we will address the problem of obtaining accurate information about a finite population: a population consisting of a finite number of units.

DEFINITIONS:

- Each element in a finite population will be called a **population unit**.
- The population of interest is called the **target population**.

EXAMPLE 16.1

For illustration, we consider a small target population: the population of all the automobile engines of a specified model manufactured on a particular day at a certain factory. To keep computations manageable, this population has size 8: it's a small factory! Before being shipped for installation in cars, the engines are test run. The quantity of interest for our purposes is the level of hydrocarbon emissions. The data for all eight engines in the target population are displayed in Table 16.1, where the hydrocarbon emissions are expressed as a percentage of the maximum value allowed under California law, the most stringent in the nation. A third variable that is included in the data set is a compliance variable. This variable indicates whether the engine is in compliance with the emission law. It takes on two values: 0 and 1. A ’0’ indicates the engine is not in compliance with the law; a ’1’ indicates that it is in compliance. Such a $0-1$ variable is called an **indicator variable**.

Hydrocarbon Emissions for a Population of Engines		
Engine Number	Emissions	Compliance
1	90	1
2	78	1
3	101	0
4	95	1
5	92	1
6	121	0
7	89	1
8	99	1

Table 16.1: *Hydrocarbon emission data*

16.3 Population Parameters

Let N denote the number of units in the target population. The units in the population are identified by **labels**. People are usually identified by their names or social security numbers but in sampling theory we identify them by the numbers 1, 2, 3, ..., N, which are the labels. The numbers 1-8 are the labels which identify the engines in the population in Table 16.1.

With each unit in the population is associated a value of interest for that unit. This value may be a measurement, such as the percentage of maximum legal hydrocarbon emissions from the engine example, or a categorical variable, such as the variable indicating whether the engine is in compliance with the emission standards. We will denote these values by subscripted upper case letters: for example, Y_i is the value corresponding to the i^{th} unit in the population. The entire sequence $Y_1, Y_2, ..., Y_N$ of Y-values in the population is considered a fixed parameter of the population.

There are several quantities in a finite population that may be of interest for an investigator. We will consider four of these:

SOME FINITE POPULATION PARAMETERS

- The **population mean** is

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N}.$$

- The **population variance** is

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N-1}.$$

- The **population standard deviation** is

$$\sigma = +\sqrt{\sigma^2}.$$

- The **population total** is

$$Y = \sum_{i=1}^{N} Y_i = N\mu.$$

- The **population proportion** is the proportion of units in the population having a certain characteristic. If each value Y_i is an indicator variable which takes on the value '1' when the i^{th} unit has that characteristic and 0 otherwise. then the sum

$$C = \sum_{i=1}^{N} Y_i$$

counts the number of units in the population having the characteristic. In this case the population mean

$$p = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{C}{N}$$

is the proportion of units in the population having the characteristic, and hence is the population proportion.

EXAMPLE 16.1, CONTINUED

To illustrate, consider the population data in Table 16.1. The population parameters for the emission measurements are:

$$\mu = (90 + 78 + \ldots + 89 + 99)/8 = 95.625,$$

$$\sigma^2 = ((90 - 95.625)^2 + (78 - 95.625)^2 + \ldots + (89 - 95.625)^2 + (99 - 95.625)^2)/7 = 154.839,$$

and

$$\sigma = +\sqrt{154.839} = 12.443.$$

The population proportion of engines in compliance is

$$p = (1 + 1 + 0 + 1 + 1 + 0 + 1 + 1)/8 = 0.75.$$

2

Sampling studies are conducted to learn about the target population. In particular, we will assume that the desired information consists of one or more of the population parameters described above. In practice, a researcher will have two ways to obtain the desired information: a census or a sample.

DEFINITIONS:

- A **census** consists of obtaining data from every unit in the target population.
- A **sample** is a subset of the population units. **Judgment samples** are selected by subjective decision of the planners of the sampling study. **Probability samples** are selected by some random mechanism.
- The sample is obtained from a listing or enumeration, called a **frame**, of the population units.
- A **sampling study** is an activity in which we obtain information about a population by obtaining information from a sample taken from that population. Sampling studies are also referred to as **sample surveys**.

Of course, a census will give the most complete information about a population. However, investigators often do not have the resources or time to conduct a census, so any information they obtain about the target population usually must come from a sample. Since it is selected by non-scientific methods, there is no scientific way to judge the accuracy or precision of information obtained from a judgment sample. On the other hand, we can use probability theory to judge both accuracy and precision of a probability sample. For this reason, we will consider only probability samples in this chapter. You should be skeptical of the results of any sampling study which does not use a probability sample.

EXAMPLE 16.1, CONTINUED

For the hydrocarbon emission example, the target population is the eight engines produced on the day in question. Let's assume that the quantities of interest are μ, the population mean emissions as a percent of the maximum allowed value, and p, the population proportion of engines in compliance with the emissions standards.

Investigators might take measurements on all eight engines, which would constitute a census. From the census measurements they can obtain μ and p exactly.

However, perhaps they can't afford to take measurements on all eight engines. Then their only option is to take a sample of the engines and obtain measurements on those engines in the sample. They will use those measurements to estimate μ and p.

To see how this works, we begin with the simplest probability sampling scheme, called appropriately enough, **simple random sampling**. 2

16.4 Simple Random Sampling

DEFINITION:

- Simple random sampling (**SRS**), also known as **random sampling without replacement**, is the sampling design in which n distinct units are selected from the N units in the population in such a way that every possible combination of n units is equally likely to be in the sample selected. Note that "distinct" means that no unit can appear more than once in a sample.
- The quantity $f = n/N$ is called the **sampling fraction** or **sampling proportion** of the SRS scheme.
- The quantity $1 - f = 1 - n/N$ is called the **finite population correction** (**FPC**).

In a SRS, the chance that the i^{th} unit in the population is included in the sample is the sampling fraction, $f = n/N$. This chance is the same for each unit in the population. Some other designs can give equal chance to the selection of a unit, but only with simple random sampling does each possible sample of n units have the same chance of being in the sample.

Selecting a Simple Random Sample

When we take a simple random sample, we are really sampling the labels, not the Y-values. In principle, a SRS may be selected by writing the numbers 1 through N on N pieces of paper, putting the pieces of paper in a hat, stirring them thoroughly, and, without looking, selecting n pieces of paper without replacing any. The sample consists of the set of population units whose labels correspond to the numbers selected. Alternatively, the labor of the selection process can be reduced by using a table of random numbers or a computer random number generator.

We can write down a formula for the number of possible SRSs drawn from a finite population consisting of N individuals. Using the factorial notation[1], the number of samples is

$$\left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N!}{n!(N-n)!}.$$

Thus, for our example n=2, N=8, and the number of possible samples is 28. For simple random sampling each sample of size n has the same chance of being taken. That is, each sample has chance of

$$\left(\begin{array}{c} N \\ n \end{array} \right)^{-1}$$

of being selected.

Estimation of a Population Mean and Variance

In this section, we will discuss estimation of the finite population mean and variance using sample data. You may want to review the material on normal theory estimation in Chapter 5 before continuing.

Suppose the sample data are the values $y_1, y_2, \ldots, y_n$.[2] Then a natural estimator of the population mean μ is the sample mean $\overline{y} = \sum_{i=1}^{n} y_i/n$, and a natural estimator of the population variance σ^2 is $s^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)$.

DEFINITION: An estimator is **unbiased** if its mean taken over all possible samples equals the population quantity being estimated.

[1] Recall that n factorial, denoted $n!$, is defined as the product $n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$.

[2] Notice that we use upper case letters to denote population values and lower case letters to denote sample values.

Both $\overline{y}$ and s^2 are unbiased estimators of μ and σ^2, respectively, since if we compute $\overline{y}$ and s^2 for each of the $\begin{pmatrix} N \\ n \end{pmatrix}$ possible samples of size n, and then take the mean of these (i.e. just average them), the results will equal μ and σ^2, respectively.

The variance of $\overline{y}$ is $\sigma^2_{\overline{y}} = (1-f)\sigma^2/n$, where $1-f$, you will recall, is the finite population correction. Compare this with the value σ^2/n for the variance of an infinite population. The variance is smaller in the finite population case because there are fewer possible samples to select. What do you think happens to f as the size of the population becomes infinite?

An unbiased estimator of $\sigma^2_{\overline{y}}$ is $\hat{\sigma}^2_{\overline{y}} = (1-f)s^2/n$.

For large n, a level L confidence interval for μ is given by

$$(\overline{y} - \hat{\sigma}_{\overline{y}} \cdot z_{\frac{1+L}{2}},\ \overline{y} + \hat{\sigma}_{\overline{y}} \cdot z_{\frac{1+L}{2}}),$$

where $z_{\frac{1+L}{2}}$ is the $\frac{1+L}{2}$ quantile of the $N(0,1)$ distribution.

These confidence intervals are what are being talked about in news reports with statements like "We are 95% confident that to within ±250 calories the mean daily food consumption of American poodles is 2100 calories per day."

Estimation of a Population Total

An unbiased estimate of the population total Y is

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

The variance of $\hat{Y}$ is $N(N-n)\sigma^2/n$, which is N^2 times the variance of $\overline{y}$.

For large n, a level L confidence interval for Y is given by

$$(\hat{Y} - \sqrt{N(N-n)s^2/n} \cdot z_{\frac{1+L}{2}},\ \hat{Y} + \sqrt{N(N-n)s^2/n} \cdot z_{\frac{1+L}{2}}).$$

Note that a level L confidence interval for Y can be obtained from a level L confidence interval for μ by multiplying the endpoints of the latter interval by N.

Estimation of a Population Proportion

Let C be the number of units in the population having a certain characteristic. Then $p = C/N$ is the population proportion with that characteristic. Now suppose a sample of size n is taken and that c of the n units in the sample have the characteristic. Then the sample proportion $\hat{p} = c/n$ is an unbiased estimator of p. The variance of $\hat{p}$ is $\sigma^2_{\hat{p}} = (1-f)p(1-p)/n$. The estimated variance of $\hat{p}$ is $\hat{\sigma}^2_{\hat{p}} = (1-f)\hat{p}(1-\hat{p})/n$. For large n, a level L confidence interval for p is given by

$$(\hat{p} - \hat{\sigma}_{\hat{p}} \cdot z_{\frac{1+L}{2}},\ \hat{p} + \hat{\sigma}_{\hat{p}} \cdot z_{\frac{1+L}{2}}).$$

RECAP: LARGE SAMPLE CONFIDENCE INTERVAL FORMULAS FOR SIMPLE RANDOM SAMPLING

- A large sample level L confidence interval for a population mean μ is

$$(\overline{y} - \hat{\sigma}_{\overline{y}} \cdot z_{\frac{1+L}{2}}, \overline{y} + \hat{\sigma}_{\overline{y}} \cdot z_{\frac{1+L}{2}}).$$

- A large sample level L confidence interval for a population total Y is

$$(\hat{Y} - \sqrt{N(N-n)s^2/n} \cdot z_{\frac{1+L}{2}},\ \hat{Y} + \sqrt{N(N-n)s^2/n} \cdot z_{\frac{1+L}{2}}).$$

- A large sample level L confidence interval for a population proportion p is

$$(\hat{p} - \hat{\sigma}_{\hat{p}} \cdot z_{\frac{1+L}{2}},\ \hat{p} + \hat{\sigma}_{\hat{p}} \cdot z_{\frac{1+L}{2}}).$$

EXAMPLE 16.1, CONTINUED

We now use the data from Table 16.1 to illustrate the computation and interpretation of these estimators and confidence intervals. Note that the sample sizes are too small to use the normal approximation we use here, so **these computations are for illustration only**.

Suppose we choose a SRS of size 3 consisting of engines 3, 7 and 8 from the population. The emission values are 90, 101 and 89 and the compliance values are 1, 0 and 1, respectively. We then compute the following quantities:

$$\begin{array}{lcrcr}
\overline{y} & = & (99+101+89)/3 & = & 289/3 \\
s^2 & = & ((99-96.33)^2 + (101-96.33)^2 + (89-96.33)^2)/2 & = & 124/3 \\
\hat{p} & = & (1+0+1)/3 & = & 2/3 \\
1-f & = & (1-3/8) & = & 5/8
\end{array}$$

We use these quantities now to compute the following:

- **Estimates of μ:** The point estimate is $\overline{y} = 289/3 = 96.33$. Noting that $z_{.025} = 1.96$, a 95% confidence interval for μ is

  $$289/3 \pm 1.96 \cdot \sqrt{(5/8)(124/3)/3} = (90.58, 102.08).$$

  The interpretation is that in repeated sampling form this population, 95% of all intervals computed in this way will actually contain μ. Note that this particular interval covers the true population mean 95.635.

- **Estimates of Y:** The point estimate is

  $$N\overline{y} = 8 \cdot 289/3 = 770.64,$$

  and a 95% confidence interval is

  $$(8 \cdot 90.58, 8 \cdot 102.08) = (724.64, 816.64).$$

- **Estimates of p:** A point estimate of the proportion of complying engines in the population is $\hat{p} = 2/3$. A 95% normal theory confidence interval for the proportion of complying engines is

  $$2/3 \pm 1.96\sqrt{(5/8)(2/3)(1-2/3)/3} = (0.25, 1.09) = (0.25, 1.00),$$

  where the upper limit of 1.09 is rounded to the nearest value that makes sense.

Determination of Sample Size

One consideration in designing a sampling study is the **precision** desired.

DEFINITION: **Precision** of an estimator is a measure of how variable that estimator is. Another equivalent way of expressing precision is the width of a level L confidence interval. For a given population and sampling method (e.g SRS), precision is a function of the size of the sample: the larger the sample, the greater the precision.

Usually in designing a sampling study, the desired precision is specified first, and the sample size needed to attain that precision computed. As an example, suppose a simple random sample will be used and it is desired to estimate a population proportion p to within d units with confidence level at least L. If we assume a large enough sample size (so the normal approximation can be used in computing the confidence interval), and that the number in the sample n is small relative to the number in the population N, then an estimate of sample size is

$$n_0 = (p(1-p) \cdot z^2_{\frac{1+L}{2}})/d^2 \tag{16.1}$$

If n_0/N is small, use $n = n_0$ (note that this is the usual formula for an infinite population). If not, use $n = n_0/(1 + n_0/N)$.

Of course, all this supposes we know p. If we don't, we can get an estimate from a pilot study (see Section 16.9). Or, since

$p(1-p) \leq 0.25$, we can use 0.25 in place of $p(1-p)$ in the above formulas to get a conservative (i.e. too large) sample size estimate.

There is an analogous formula when a simple random sample will be used and it is desired to estimate a population mean μ to within d units with confidence level at least L. If we assume a large enough sample size (so the normal approximation can be used in computing the confidence interval), and that the number in the sample n is small relative to the number in the population N, then an estimate of sample size is

$$n_0 = (\sigma^2 \cdot z^2_{\frac{1+L}{2}})/d^2. \tag{16.2}$$

If n_0/N is small, use $n = n_0$ (note that this is the usual formula for an infinite population). If not, use

$$n = n_0/(1 + n_0/N). \tag{16.3}$$

Again, this supposes we know σ^2. If we don't, we can get an estimate from a pilot study (see Section 16.9).

EXAMPLE 16.2

As an example of sample size calculations, consider the sample size necessary to estimate the proportion of Americans who prefer that Federal government cutbacks be returned to taxpayers through lower taxes instead of being used to reduce the Federal deficit. Assume that there are two hundred million adult Americans and that we will take a SRS (despite the fact that a SRS is not really practical). Suppose we want the estimate to be accurate to within 3 percentage points with 90% confidence. Finally, assume that we have no idea what the true population proportion p is.

Since we have no idea what the true population proportion p is, we will use the upper bound 0.25 in place of $p(1-p)$ in equation 16.1. The result is a first estimate ($z_{0.95} = 1.645$) of

$$n_0 = (0.25)(1.645)^2/(0.03)^2 = 752,$$

rounded upward. Since n_0 is small relative to $n = 200,000,000$, we may use 752 as the sample size.

If p is of moderate size, say between 0.3 and 0.7, the n_0 estimated using 0.25 for $p(1-p)$ will be reasonably close to the true n_0. However, it can overestimate appreciably if p is near 0 or 1. For example, suppose in the previous example we know that p is near 0.95. Then equation 16.1 yields

$$n_0 = (0.95)(0.05)(1.645)^2/(0.03)^2 = 143.$$

2

EXAMPLE 16.3

As a second example, suppose we are measuring the diameters of ball bearings in a shipment of 2500. We know that the variance of these measurements is always near 0.1 mm. We would like our estimate of mean diameter to be within 0.025 mm with 99% confidence. From equation 16.2 we see that

$$n_0 = (0.1)(2.58)^2/(0.025)^2 = 1066.$$

Since this is a substantial proportion of the population size of 2500, we use equation 16.3 to compute

$$n = 1066/(1 + 1066/2500) = 748.$$

Thus, the fact that we have a finite population means we need take a sample of only 748 ball bearings instead of 1066, a savings of 30%. 2

16.5 Stratified Sampling

DEFINITION:

- A **stratum** is a subgroup of the population.
- In **stratified sampling**, the population is divided into **strata** (the plural of stratum), and from each stratum a separate sample is drawn.

EXAMPLE 16.4

As an example of a population where stratification would be useful, consider the data set FIRMS, which contains the numbers of employees of 115 multinational corporations. The firms are arranged in two strata, the first containing the 21 firms with more than 100,000 employees, and the second the remaining 94 firms. These data are similar to populations of many types, in that some firms contribute a great deal to the total and display much greater variability than the remainder. These data will be investigated further in Lab 16.2. 2

Reasons for Stratifying

There are several reasons for choosing stratified sampling instead of a SRS:

1. If measurements on certain strata within the population are of particular interest, and a SRS will likely contain too few data points to obtain good results from these strata, stratification can ensure adequate sample sizes for the strata of interest.
2. If there is large variability between strata compared to the variability within strata, stratification will give more precise estimators.
3. Administrative convenience may dictate the use of strata. For example, national data may be best sampled by region if the data are collected by regional offices. In this case the regions are natural strata.
4. Conditions in a population may make different sampling methods appropriate for different strata. For example, in surveying businesses, it is not uncommon to sample the largest firms, of which there are a small number with high variability, at a higher rate and to sample the smaller firms, of which there are many with low variability, at a lower rate.

In this chapter, we will concentrate on the simplest kind of stratified sampling: that in which a simple random sample is taken in each stratum separately. Such a sampling scheme is called **stratified random sampling**.

Notation

Assume the population of N units is divided into K strata. We will consider stratified random sampling, in which a SRS is taken separately in each of the K strata. Table 16.2 displays the formulas used in describing stratified random sampling. Though the notation is messy, most formulas follow easily from those for SRS. The messiness comes from the subscript j that is used to keep track of the different strata.

From the formulas in Table 16.2, it follows that $N = \sum_{i=1}^{K} N_j$, (the number of units in the population equals the sum of the units in the K strata), that $n = \sum_{i=1}^{K} n_j$ (the total number of units in the sample equals the sum of the numbers of sample units in all the strata), and that $\mu = \sum_{j=1}^{K} N_j \mu_j / N$ (the overall population mean is a weighted average of the strata means).

Notation for Stratified Sampling: Population Quantities

Notation	Meaning
N_j	The number of population units in the j^{th} stratum.
$Y_{j1}, Y_{j2}, \ldots, Y_{jN_j}$	Population measurements in the j^{th} stratum.
$\mu_j = \sum_{i=1}^{N_j} Y_{ji}/N_j$	Population mean of measurements in the j^{th} stratum.
$\sigma_j^2 = \sum_{i=1}^{N_j} (Y_{ji} - \mu_j)^2/(N_j - 1)$	Population variance of measurements from the j^{th} stratum.
C_j	Population number of units in the j^{th} stratum having a certain characteristic.
$p_j = C_j/N_j$	Population proportion of units in the j^{th} stratum having a certain characteristic.

Notation for Stratified Sampling: Sample Quantities

Notation	Meaning
n_j	The number of sample units from the j^{th} stratum.
$1 - f_j = 1 - n_j/N_j$	The FPC in the j^{th} stratum.
$y_{j1}, y_{j2}, \ldots, y_{jn_j}$	Sample measurements from the j^{th} stratum.
$\overline{y}_j = \sum_{i=1}^{n_j} y_{ji}/n_j$	Sample mean of measurements from the j^{th} stratum.
$s_j^2 = \sum_{i=1}^{n_j} (y_{ji} - \overline{y}_j)^2/(n_j - 1)$	Sample variance of measurements from the j^{th} stratum.
$\hat{\sigma}^2_{\overline{y}_j} = (1 - f_j)s_j^2/n_j$	The estimated variance of $\overline{y}_j$.
c_j	Number of sample units from the j^{th} stratum having a certain characteristic.
$\hat{p}_j = c_j/n_j$	Proportion of sample units from the j^{th} stratum having a certain characteristic.

Table 16.2: *Notation for stratified sampling*

Estimation from a Stratified Sample

Estimation of the Population Mean

An unbiased estimator of μ is

$$\overline{y}_{st} = \sum_{j=1}^{K} N_j \overline{y}_j / N.$$

An unbiased estimator of its variance is

$$\hat{\sigma}^2_{\overline{y}_{st}} = \sum_{j=1}^{K} N_j^2 \hat{\sigma}^2_{\overline{y}_j} / N^2.$$

For large n, a level L confidence interval for μ is given by

$$(\overline{y}_{st} - \hat{\sigma}_{\overline{y}_{st}} \cdot z_{\frac{1+L}{2}},\ \overline{y}_{st} + \hat{\sigma}_{\overline{y}_{st}} \cdot z_{\frac{1+L}{2}}).$$

Estimation of the Population Total

An unbiased estimator of the population total Y is

$$\hat{Y}_{st} = N\overline{y}_{st}.$$

An unbiased estimator of the variance of $\hat{Y}_{st}$ is

$$\hat{\sigma}^2_{\hat{Y}_{st}} = N^2\hat{\sigma}^2_{\overline{y}_{st}} = \sum_{j=1}^{K} N_j^2\hat{\sigma}^2_{\overline{y}_j}.$$

For large n, a level L confidence interval for Y is given by

$$(\hat{Y}_{st} - \hat{\sigma}_{\hat{Y}_{st}} \cdot z_{\frac{1+L}{2}},\ \hat{Y}_{st} + \hat{\sigma}_{\hat{Y}_{st}} \cdot z_{\frac{1+L}{2}}).$$

Estimation of the Population Proportion

An unbiased estimator of the proportion p in the population having a certain characteristic is

$$\hat{p}_{st} = \sum_{j=1}^{K} N_j\hat{p}_j/N.$$

An unbiased estimator of its variance is

$$\hat{\sigma}^2_{\hat{p}_{st}} = \sum_{j=1}^{K} N_j^2\hat{\sigma}^2_{\hat{p}_j}/N^2,$$

where $\hat{\sigma}^2_{\hat{p}_j} = (1 - f_j)\hat{p}_j(1 - \hat{p}_j)/n_j$. For large n, a level L confidence interval for p is given by

$$(\hat{p}_{st} - \hat{\sigma}_{\hat{p}_{st}} \cdot z_{\frac{1+L}{2}},\ \hat{p}_{st} + \hat{\sigma}_{\hat{p}_{st}} \cdot z_{\frac{1+L}{2}}).$$

RECAP: LARGE SAMPLE CONFIDENCE INTERVAL FORMULAS
STRATIFIED RANDOM SAMPLING

- A large sample level L confidence interval for a population mean μ is

$$(\overline{y}_{st} - \hat{\sigma}_{\overline{y}_{st}} \cdot z_{\frac{1+L}{2}},\ \overline{y}_{st} + \hat{\sigma}_{\overline{y}_{st}} \cdot z_{\frac{1+L}{2}}).$$

- A large sample level L confidence interval for a population total Y is

$$(\hat{Y}_{st} - \hat{\sigma}_{\hat{Y}_{st}} \cdot z_{\frac{1+L}{2}},\ \hat{Y}_{st} + \hat{\sigma}_{\hat{Y}_{st}} \cdot z_{\frac{1+L}{2}}).$$

- A large sample level L confidence interval for a population proportion p is

$$(\hat{p}_{st} - \hat{\sigma}_{\hat{p}_{st}} \cdot z_{\frac{1+L}{2}},\ \hat{p}_{st} + \hat{\sigma}_{\hat{p}_{st}} \cdot z_{\frac{1+L}{2}}).$$

EXAMPLE 16.4, CONTINUED

Refer again to the data set FIRMS on multinational corporations described in Section 16.5. The quantity being measured is the number of employees. Suppose for these data we take a stratified random sample of 5 large and 10 small corporations. The data obtained are:

	Stratum 1	Stratum 2
	5206	102423
	15524	129434
	16835	138326
	22354	284000
	64604	332700
	15982	
	25198	
	28535	
	4304	
	35000	
$\overline{y}_j$:	23354.2	197376.6
s_j :	17380.8	103604.3

We want to estimate the total number of employees in the population of 115 firms. Notice the large difference in means and standard deviations which suggests that stratification may be effective.

From this population, we will need the quantities $N = 115$, $N_1 = 94$ and $N_2 = 21$. From this sampling scheme, we need $1 - f_1 = 1 - 10/94 = 0.89$, and $1 - f_2 = 1 - 5/21 = 0.76$. From this sample, we will need the quantities

$$\hat{\sigma}^2_{\overline{y}_1} = (1 - f_1)s_1^2/n_1 = (0.89)(17380.8^2)/10 = 26886207,$$

and

$$\hat{\sigma}^2_{\overline{y}_2} = (1 - f_2)s_2^2/n_2 = (0.76)(103604.3^2)/5 = 1631545349.$$

Then

$$\hat{Y}_{st} = N_1\overline{y}_1 + N_2\overline{y}_2 = (94)(23354.2) + (21)(197376.6) = 6340203,$$

and

$$\hat{\sigma}_{\hat{Y}_{st}} = \sqrt{(94^2)(26886207) + (21^2)(1631545349)} = 978303.$$

So (assuming the normal approximation is valid for so small a sample) a 95% confidence interval for the population total is

$$6340203 \pm (1.96)(978303) = (4422728, 8257679).$$

Compare this with the true value of 7315991. 2

Sample Allocation

How to allocate the n observations in the strata is an important issue.

Optimal Allocation

It can be shown mathematically that when the cost of sampling is equal in each stratum, the allocation which minimizes the variance of $\overline{y}_{st}$ takes the sample size in the j^{th} stratum as close to

$$\frac{N_j\sigma_j}{\sum_{m=1}^{K} N_m\sigma_m}n$$

as possible. This formula also holds for estimating a population total.

However, if the cost of sampling is greatly different in the different strata, this optimal allocation is different. For example, if the cost per unit is c_j in stratum j, the optimal number in the j^{th} stratum is as close to

$$\frac{N_j\sigma_j/\sqrt{c_j}}{\sum_{m=1}^{K} N_m\sigma_m/\sqrt{c_m}}n$$

as possible. Similar formulas hold for estimating a population proportion.

The message is that in a given stratum, take a larger sample if:

1. the stratum is larger,

2. the stratum is more variable internally,

3. sampling is cheaper in the stratum.

Of course, the total number of units in the sample, n, must also be chosen. The choice of n will depend on such factors as the total available resources, and the desired precision of the estimator, either overall or within each stratum or both.

While it is good to know the formulas for optimal allocation, it is often difficult or impossible to use them in practice. For one thing, in order to use the formulas, we need to know each σ_j^2, which almost never happens. One solution to this problem is to do an initial pilot study (see Section 16.9), estimate the σ_j^2 from that study, and then sample according to the optimal allocation formulas using the estimated values.

But even if we do manage to estimate the σ_j^2, the optimal allocation formula applies only to a single measurement. Most sampling studies have many measurements, and each will have its own optimal allocation. When these conflict, it is not clear how to proceed.

For this reason optimal allocation is used less frequently than **proportional allocation**.

Proportional Allocation

In proportional allocation, the proportion of sample units from the j^{th} stratum is the same as the proportion of population units in the j^{th} stratum. That is, the number sampled from the j^{th} stratum is as close to $(N_j/N)n$ as possible.

It can be shown mathematically, that proportional allocation produces estimators with greater precision than a SRS, but less precision than optimal allocation.

Other Allocations

Other kinds of allocation may depend on the requirements of the study. For example, different precisions may be required for estimators in different strata individually, with yet another precision for estimators of population totals. These considerations will dictate the allocation of sample sizes to the different strata.

When Stratification Produces Large Gains in Precision

Stratification will yield large gains in precision when the following three conditions hold:

1. The population is composed of units varying widely in size.

2. The variables to be measured are closely related to the sizes of the units.

3. A good measure of size is available for setting up the strata.

Double Sampling

When a frame giving strata is not available, a technique called **double sampling** can be used. In double sampling for strata, units in an initial sample are classified into strata. The second sample is then selected from this initial sample by stratified sampling.

Post-Stratification

Sometimes a sample would benefit from stratification, but we are unable to come up with a frame for each stratum, and so are unable to perform stratified sampling. For example, we may want each stratum in a survey to consist of one type of religious affiliation, which is information we wouldn't know before an interview.

If we know how many units are in each stratum in the population, we can take a SRS of the entire population and stratify later, when we learn the stratum to which each unit belongs. In the example of religious affiliation, we would know the religion of a person after the interview.

This kind of stratification is called **post-stratification**. Post-stratified samples are analyzed in exactly the same way as stratified samples, and the estimators have the same properties (e.g. unbiasedness) provided all strata are represented in the sample. The difference between stratification and post-stratification is in how the sample is selected.

One final point: In order to use post-stratification we do not have to have planned to use it in the first place. If stratification was initially not done, by oversight or for whatever reason, and later it is found that stratification is useful, post-stratification can still be used.

16.6 Ratio and Regression Estimation

As a general rule, when solving a problem it is good practice to incorporate all the knowledge available about that problem into its solution. In statistical problems such knowledge often takes the form of observations on other variables related to the variable of interest.

EXAMPLE 16.6

Suppose we want to estimate the volume of lumber in harvestable trees in a certain forest. One approach would be to take a sample of the trees in the forest (how would YOU design the sampling scheme?), cut them down and measure the volume of lumber in them. Then we could estimate the total forest volume from the sample.

But consider another approach. Suppose the diameter of a tree a certain distance above the ground is closely related to the volume of lumber in the tree. Then it might make sense to proceed as follows:

1. For the original sample of trees, measure the diameter of each tree as well as its volume.

2. Use the data to quantify the relationship between tree diameter and tree volume.

3. Measure the diameters of a much larger sample of trees.

4. Use the relation between diameter and volume obtained in 2 and the measurements of the tree diameters obtained in 3 to give an improved estimate of the volume of trees in the forest.

In this approach, easily-obtained information, tree diameter, can help improve estimation of a quantity, tree volume, that is difficult to obtain. 2

Two kinds of estimators which use auxiliary information are **ratio** estimators and **regression** estimators. These estimators can improve the precision of estimation, in some cases substantially.

For both kinds of estimators, we assume the values of the variable of interest in the population are $Y_1, \ldots, Y_N$ and the values of an **auxiliary variable** are $X_1, \ldots, X_N$. The values observed in the sample are $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ respectively. The population totals of measurements on these two variables are Y and X and the sample totals are y and x. In the lumber example, the Y_i are the lumber volumes of the trees and the X_i are the tree diameters.

In the next two sections we will only consider estimation of the population total Y. Similar formulas apply to estimation of the population mean. We will also assume a SRS, though ratio and regression estimation can be easily adapted to stratified sampling.

Ratio Estimation

The ratio estimator of Y is

$$\hat{Y}_R = \frac{y}{x} X,$$

which has approximate variance

$$\tilde{\sigma}_R^2 = \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N}(Y_i - RX_i)^2}{N-1},$$

where $R = Y/X$. An estimator of the variance computable from the sample is

$$\hat{\hat{\sigma}}_R^2 = \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1},$$

where $\hat{R} = y/x$. For n large, an approximate level L confidence interval for Y is

$$(\hat{Y}_R - \hat{\hat{\sigma}}_R \cdot z_{\frac{1+L}{2}},\ \hat{Y}_R + \hat{\hat{\sigma}}_R \cdot z_{\frac{1+L}{2}}).$$

The ratio estimator is **not** unbiased, though unbiased versions can be constructed. Ratio estimators which use information from more than one auxiliary variable can also be constructed. Note also that ratio estimators require knowledge of X, the population total of the auxiliary variable. Ratio estimators are most useful when:

1. The relation between the X_i and Y_i is roughly linear with the line passing through the origin.

2. The variance of Y_i about this line is proportional to X_i.

Regression Estimation [3]

If the relation between the X_i and Y_i is roughly linear but the line doesn't pass through the origin, a regression estimator is preferable to a ratio estimator. The linear regression estimator is

$$\hat{Y}_{lr} = N\left[\overline{y} + \hat{\beta}_1(\overline{X} - \overline{x})\right],$$

where $\overline{y}$ is the sample mean of the variable of interest, $\overline{x}$ and $\overline{X}$ are the sample and population means of the auxiliary variable, and $\hat{\beta}_1$ is the estimated slope computed from the least squares regression of the y_i on the x_i (see Chapter 7 for details). For large samples, we may estimate the approximate variance of the regression estimator as

$$\hat{\sigma}_{lr}^2 = \frac{N^2(1-f)}{n}\text{MSE},$$

where MSE is the mean squared error resulting from the regression of the y_i on the x_i. For n large, an approximate level L confidence interval for Y is

$$(\hat{Y}_{lr} - \hat{\sigma} \cdot z_{\frac{1+L}{2}},\ \hat{Y}_{lr} + \hat{\sigma} \cdot z_{\frac{1+L}{2}}).$$

Like the ratio estimator, the regression estimator is **not** unbiased. Regression estimators which use information from more than one auxiliary variable can be constructed. Note also that regression estimators require knowledge of X.

Regression estimators are most useful when:

1. The relation between the X_i and Y_i is linear.

2. The variance of Y_i about this line is constant.

Double Sampling

As we have seen, both the ratio and regression estimation techniques rely on auxiliary information which may not be available at the time the sampling study is done. One way to obtain the auxiliary information is to do **double sampling**, a technique which you have already seen applied to stratified sampling.

In double sampling an initial sample is taken for obtaining auxiliary information only. Then a second sample is taken in which the variable of interest and the auxiliary variable are both observed.

In the lumber estimation example, we might take a relatively large first sample of the easily-measured tree diameter to estimate X, the total of the tree diameters in the forest. Then a smaller sample of both X_j and the harder to measure Y_j could provide a ratio or regression estimator.

[3] The material in this section assumes a knowledge of simple linear regression, as covered in Chapter 7.

EXAMPLE 16.6, CONTINUED

We illustrate ratio estimation by using the data found in the data set TREES. These data consist of measurements taken from 31 black cherry trees in the Allegheny National Forest in Pennsylvania. Among the variables recorded for these trees are the diameter in inches, taken at 4.5 feet above ground level, and the lumber volume, in cubic feet, obtained by cutting down the tree. It is desired to estimate the total lumber volume of mature black cherry trees in a sector of the forest.

Foresters used double sampling to obtain the quantities necessary to produce ratio estimators. To do this, they measured the diameters of all of the $N = 451$ mature black cherry trees in the sector (including the 31 in the data set), and obtained a total diameter $X = 5867.6$ inches for all 451 trees.

From the 31 trees in the data set, we have total volume $y = 935.3$ cubic feet and $x = 410.7$ inches. This yields $\hat{R} = 935.3/410.7 = 2.277$, and hence the estimator

$$\hat{Y}_R = (2.277)(5867.6) = 13360.5.$$

To compute the approximate variance of this estimate, we first find

$$\frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1} = 94.07.$$

The variance is then approximately

$$\tilde{\sigma}_R^2 = (451)^2(1 - 31/451)(94.07)/31 = 574798.$$

This is about one-third the estimated variance of 1651009 for the estimate $\hat{Y} = N\overline{y}$, which shows the value of incorporating the tree diameters into the estimation.

A level 0.99 confidence interval for the total of the tree volumes is given by

$$13360.5 \pm \sqrt{574798} \cdot 2.5758 = (11407.6,\ 15313.4).$$

The relation between diameter and volume for the 31 sample trees is, in fact, reasonably linear, but its intercept is not zero. As a result the regression estimator of total volume should perform better than the ratio estimator just computed. Exercise 16.14 gives those with a knowledge of simple linear regression a chance to see whether this proves true. 2

16.7 Some Other Commonly-Used Sampling Techniques

In this section we present a brief overview of some other commonly-used sampling techniques.

Cluster Sampling

In some applications the population units available for sampling consist of a group or **cluster** of smaller units called **elements**.

DEFINITION:

- **Cluster sampling** is a sampling scheme in which a sample of the clusters, instead of individual elements, is initially taken.
- If a census is then taken of each of the initially-sampled clusters, the procedure is called **single-stage cluster sampling**.
- If a sample is then taken from each of the initially-sampled clusters, the procedure is called **two-stage cluster sampling**.

EXAMPLE 16.7

Consider the problem of surveying all households in a city. Taking a SRS of all households could present difficulties. First, it might be too difficult or too expensive to get an accurate frame of all households. Second, it might prove too costly to try to physically cover the entire city as a SRS would require.

However, dividing the city into smaller clusters, such as city blocks, can help alleviate these problems. Suppose we use cluster sampling with city blocks as clusters. That is, at the first stage of sampling we choose a SRS of city blocks. This approach has two advantages over taking a random sample of households. First, it is likely that a frame of city blocks is easier to obtain and much shorter than a frame of all households. Second, by concentrating effort at a few locations (the selected blocks), resources used in traveling to the households are conserved.

If we conduct a census of all households in each block selected, we will be performing single-stage cluster sampling. If we conduct a SRS of the households in each block, we will be doing two-stage cluster sampling. 2

Single-Stage Cluster Sampling

For simplicity we will concentrate on estimating population totals of measurements, but similar results occur for population means of measurements and population proportions.

Suppose there are N clusters and that the number of elements in the j^{th} cluster is M_j. Let Y_{ji} denote the measurement value of the i^{th} element in the j^{th} cluster, and let $Y_j = \sum_{i=1}^{M_j} Y_{ji}$ denote the total for the j^{th} cluster. Let $Y = \sum_{j=1}^{N} Y_j$ denote the population total and $\overline{Y} = Y/N$ denote the population mean per cluster.

Suppose that n clusters are sampled and let $f = n/N$. For those clusters that are actually sampled, let $y_j = \sum_{i=1}^{M_j} Y_{ji}$ denote the total for the j^{th} cluster, and $\overline{y}_j = y_j/M_j$ the mean for the j^{th} cluster.

Then an unbiased estimator of Y is

$$\hat{Y} = \frac{N}{n}\sum_{j=1}^{n} y_j,$$

and its variance is

$$\sigma_{\hat{Y}}^2 = \frac{N^2(1-f)}{n}\frac{\sum_{j=1}^{N}(y_j - \overline{Y})^2}{N-1}.$$

As we will not know $\overline{Y}$, we may estimate it using $\overline{y}$, the mean per sample cluster, to obtain the estimated variance

$$\hat{\sigma}_{\hat{Y}}^2 = \frac{N^2(1-f)}{n}\frac{\sum_{j=1}^{n}(y_j - \overline{y})^2}{n-1}.$$

These estimators have the advantage that the M_j need not be known to compute them. However, the estimator $\hat{Y}$ is often found to have poor precision, particularly when the $\overline{y}_j$ vary little from cluster to cluster, and the M_j vary greatly.

Ratio Estimation We can often improve on the estimator $\hat{Y}$ if we know more about the M_j. For example, if we know the population total number of elements $M = \sum_{j=1}^{N} M_j$, then a **ratio estimator** for Y is

$$\hat{Y}_R = M\frac{\sum_{j=1}^{n} y_j}{\sum_{j=1}^{n} M_j}.$$

If N is large, this estimator has approximate variance

$$\tilde{\sigma}_R^2 = \frac{N^2(1-f)}{n}\frac{\sum_{j=1}^{N} M_j^2(\overline{y}_j - \overline{\overline{Y}})^2}{N-1},$$

where $\overline{\overline{Y}} = Y/M$ is the population total per element. In practice, we will not know $\overline{\overline{Y}}$, so we may estimate it with $\hat{Y}_R/M$ to obtain the estimate of $\tilde{\sigma}^2$,

$$\hat{\tilde{\sigma}}_R^2 = \frac{N^2(1-f)}{n}\frac{\sum_{j=1}^{n} M_j^2(\overline{y}_j - \hat{Y}_R/M)^2}{n-1}.$$

Note that the M_j need not be known ahead of time to compute $\hat{Y}_R$ (the M_j for the sampled units are needed, but these will be known once the units have been sampled), but the value of M is needed. The variance of $\hat{Y}_R$ is often much smaller than that of $\hat{Y}$.

EXAMPLE 16.7, CONTINUED

We will illustrate calculation of the unbiased and ratio estimators. Consider the problem of estimating the population of a mid-sized city. It is decided to conduct a single-stage cluster sampling. As a first stage in sampling, a simple random sample of size 100 will be obtained from the 1743 city blocks. The number of members in each household in each of the sampled city blocks will then be counted.

Data from ten of the blocks are shown in Table 16.3.

City Block	Number of Households	Population
1	40	106
2	31	96
3	27	74
4	24	54
5	11	23
.	.	.
.	.	.
.	.	.
96	24	60
97	48	115
98	8	18
99	26	78
100	25	78

Table 16.3: *City population data*

For these data, $N = 1743$, $n = 100$ and $\sum_{j=1}^{100} y_j = 9075$, so

$$\hat{Y} = \frac{1743}{100} 9075 = 158177.$$

In addition, $\sum_{j=1}^{100} (y_j - \overline{y})^2/99 = 1681.99$, so

$$\hat{\sigma}_{\hat{Y}}^2 = \frac{(1743)^2(1 - \frac{100}{1743})}{100} 1681.99 = 48167971.81,$$

and therefore $\hat{\sigma}_{\hat{Y}} = 6940.31$.

From past censuses, it is estimated that the total number of households in the city is $M = 62759$. From the data, $\sum_{j=1}^{n} M_j = 3445$.

The ratio estimate is then

$$\hat{Y}_R = 62759 \frac{9075}{3445} = 165323.$$

In addition,

$$\frac{\sum_{j=1}^{n} M_j^2 (\overline{y}_j - 165323/62759)^2}{n-1} = 80.87,$$

the estimated variance of $\hat{Y}_R$ is

$$\hat{\hat{\sigma}}_R^2 = \frac{(1743)^2(1 - \frac{100}{1743})}{100} 80.87 = 2315913.82.$$

This gives a standard deviation $\hat{\hat{\sigma}}_R = 1521.81$, a much smaller value than was obtained for the unbiased estimator. 2

PPS Sampling Another widely used technique is **sampling with probability proportional to size**, or **PPS**. For single-stage cluster sampling, this means that the chance of a cluster being included in the sample is proportional to M_j. As a model for how to choose such a sample, consider drawing numbered slips from a hat. A total of $M = \sum_{j=1}^{N} M_j$ slips are placed in the hat with those numbered 1 through M_1 corresponding to cluster 1, those numbered M_1+1 through M_1+M_2 corresponding to cluster 2, and so on. Then n slips are drawn from the hat **with replacement**. Drawing with replacement is necessary, when n exceeds 1, in order to keep the probability of selection proportional to cluster size.

An unbiased estimator of Y based on PPS sampling is

$$\hat{Y}_{pps} = \frac{M}{n} \sum_{j=1}^{n} \overline{y}_j,$$

and its variance is

$$\sigma^2_{pps} = \frac{M}{n} \sum_{j=1}^{N} M_j(\overline{y}_j - \overline{\overline{Y}})^2.$$

An estimated variance for $\hat{Y}_{pps}$ is

$$\hat{\sigma}^2_{pps} = \frac{M^2}{n(n-1)} \sum_{j=1}^{n} (\overline{y}_j - \frac{\hat{Y}_{pps}}{M})^2.$$

PPES Sampling One problem with PPS sampling is that we have to know all the M_j to design the sampling scheme. If we don't know the M_j, but can estimate them in a reasonable way, then we can use the estimates in place of the M_j in the above scheme. In the city example, we may have estimates of the number of households in each city block from the last census that we can use to estimate the M_j.

To see how this works in practice, suppose we estimate M_j by $\hat{M}_j$ based on information obtained prior to sampling. Let $\hat{M} = \sum_{j=1}^{N} \hat{M}_j$, and let $z_j = \hat{M}_j / \hat{M}$. Then if we draw the n clusters with replacement with probabilities proportional to the $\hat{M}_j$, an unbiased estimator for Y is the **probability proportional to estimated size (PPES)** estimator

$$\hat{Y}_{ppes} = \frac{1}{n} \sum_{j=1}^{n} \frac{y_j}{z_j},$$

which has variance

$$\sigma^2_{ppes} = \frac{1}{n} \sum_{j=1}^{N} z_j \left(\frac{y_j}{z_j} - Y \right)^2.$$

An unbiased estimator of this variance is

$$\hat{\sigma}^2_{ppes} = \frac{1}{n(n-1)} \sum_{j=1}^{n} \left(\frac{y_j}{z_j} - \hat{Y}_{ppes} \right)^2.$$

When the sample size is sufficiently large, normal theory confidence intervals can be computed from these estimators in the usual way.

EXAMPLE 16.8

We will illustrate PPES estimation with a small fabricated example. Suppose in a very small town all residents live in 8 blocks. The town wishes to estimate the population based on a survey of 4 blocks, which is all they can afford to do. They have an estimate of the number of households in each of the 8 blocks, so it is decided to do conduct a PPES sampling. To do the sampling, the cumulative sum of the estimated numbers of households in the 8 blocks was formed and a range of values was assigned based on this cumulative sum. Table 16.4 shows the estimated numbers of households, the cumulative sums

Block	Estimated Number of Households, $\hat{M}_i$	Cumulative Sum	Range	z_j
1	6	6	$1-6$	0.023
2	17	23	$7-23$	0.064
3	21	44	$24-44$	0.079
4	53	97	$45-97$	0.200
5	46	143	$98-143$	0.174
6	33	176	$144-176$	0.125
7	32	208	$177-208$	0.121
8	57	265	$209-265$	0.215

Table 16.4: *Estimated numbers of households, the cumulative sums and the ranges used in Example 16.8*

and the ranges constructed. Based on this information, $\hat{M} = 265$, and the $z_j = \hat{M}_j/\hat{M}$ are shown in Table 16.4.

Four random numbers between 1 and 255, inclusive were obtained: 9, 60, 90 and 251. The blocks whose ranges contained the random numbers were chosen for the sample. Thus, the sample consisted of blocks 2, 4, 4, and 8. Notice that sampling with replacement means that we allow a block to appear more than once in the sample.

The samplers interviewed all households in blocks 2, 4 and 8, and found the respective block populations to be 36, 147 and 151. The PPES estimator was obtained as

$$\hat{Y}_{ppes} = \frac{1}{4}\left[\frac{36}{0.064} + \frac{147}{0.200} + \frac{147}{0.200} + \frac{151}{0.215}\right] = 683.71.$$

Computation of the estimated variance, $\hat{\sigma}^2_{ppes}$, is left as an exercise. 2

Two- and Multi-Stage Sampling

If in one-stage cluster sampling we select a sample from each cluster instead of doing a census on that cluster, we are performing two-stage sampling. In Example 16.7, two-stage sampling results if after selecting a sample of city blocks, we then select a sample of households from each block for interviews. The formulas for estimation in two-stage sampling are more complex versions of those for single-stage cluster sampling because of the extra level of sampling. We will not consider them here.

Two-stage sampling can be extended to multi-stage sampling. Suppose that instead of doing a survey only on one city, we wanted to do a nationwide survey. We would want to do some type of cluster sampling nationwide for the same reasons as doing it in one city. One way to proceed is to sample clusters within clusters (within clusters, within clusters,..., etc). So if we want to do a nationwide survey of households, we might take a random sample of counties nationwide, then within the selected counties a sample of cities, then within the selected cities a sample of households. This would be a three-stage sampling procedure.

Systematic Sampling

Systematic sampling refers to sampling in a well-defined non-random manner. The simplest example is the **every k^{th} sample**. In this scheme, the units in the population are assigned the labels $1, \ldots, N$, just as in a SRS. Then one label is drawn at random from the first k labels and every k^{th} label after that is drawn. The units whose labels are drawn constitute the sample. For a sample of size n, choose $k = [N/n]$, (that is, the greatest integer less than or equal to N/n).

As an example, suppose we draw a sample of size 2 from the population in Table 16.1. Then we would take $k = 4$ and draw a number at random from the set {1,2,3,4}. Say we draw 3. Next we would choose label $3 + k = 3 + 4 = 7$, so the sample would be engines 3 and 7.

Systematic sampling has connections to both stratified random sampling and cluster sampling:

- Systematic sampling stratifies the population into strata consisting of the first k, the second k, etc. labels. One observation is sampled from each stratum. The difference in sampling is that in systematic sampling the labels are sampled from the same relative location in each stratum while in stratified random sampling they are sampled at random. This results in the systematic sample being more evenly spread out over the population, which sometimes leads to it being more precise than stratified random sampling.

  In the engine example, there would be two strata consisting of engines 1-4 and 5-8.

- Suppose $N = nk$ and consider the population as being divided into k clusters with the first cluster consisting of labels $1, k+1, 2k+1, \ldots, (n-1)k+1$, the second cluster consisting of labels $2, k+2, 2k+2, \ldots, (n-1)k+2$, and so on. The systematic sample is then a single-stage cluster sample with $N = k$ clusters in the population and $n = 1$ cluster in the sample. If $N \neq nk$ the clusters have unequal sizes.

  In the engine example, there would be 4 clusters: {1,5}, {2,6}, {3,7}, and {4,8}. The sample consists of one of these clusters selected at random.

A major advantage of systematic sampling schemes is the ease with which they may be implemented. A major disadvantage is the risk of poor precision and misleading results when unsuspected periodicity exists in the data. As an example of this, consider the extreme case in which $N = 40, k = 4$ and the value of the unit in the population with label i is $Y_i = sin(\pi i/2)$. Then the population total is 0. The sample totals will be either 0 (if the first label is 2 or 4), 10 (if the first label is 1) or -10 (if the first label is 3). If a SRS or stratified random sample is used, the sample totals will seldom be so extreme. This means the variance of the systematic sample will be unusually large.

A systematic sample may be treated as a SRS when the labels are or may be considered to be in random order. In other settings (stratified systematic samples or systematic multi-stage sampling, for example) the treatment of systematic samples is more complicated and will not be dealt with here.

Combining Types of Sampling

Types of sampling can be combined to obtain a sampling scheme that meets the needs of a specific application. For example, in the three-stage sampling scheme described on page 26, we might want to stratify the counties according to their primary economic base (industrial, agricultural or high technology), then sample different size counties by PPS or PPES sampling.

16.8 Steps in Designing a Sampling Study

What Information is Required?

At the initial design stages, discussion should focus on the goals of the study and what information is needed to meet those goals.

What Are the Relevant Target Populations?

It is important to define precisely all target populations (including subpopulations) for the study. It is also important to specify the sampling units. For example, if doing a survey of spending habits, do you want to measure individual or household spending? If the latter, how is "household" defined?

For these populations find out if a frame is available. If there is a frame, what information does it contain, and what is its cost?

What Are the Variables of Interest?

Make sure that what is being measured will provide the required information. In a survey this means asking the right questions of the right persons in the right way. For example, whom do you ask about teen age drug use? Parents? School officials? Teen agers? When you decide whom to ask, how do you phrase potentially incriminating or embarrassing questions to obtain accurate answers?

Match the variables selected with the goals of the study to make sure that all variables of interest are included. Establish the precision required for each variable.

How Will the Information be Obtained from the Sample Units?

For example, will a survey be conducted by mail, phone or in-person? What will be done about missing data or nonresponse?

Type of Sampling

The type of sampling will depend on the target populations, the types of frames, if any, that are available, and the goals of the study. Some rules of thumb are:

1. **Give special treatment to unusual population units.** For example, in a study of business firms, you may want to do a census of the very largest firms and sample the rest.

2. **Sample homogeneous groups lightly and nonhomogeneous groups heavily.**

3. **Spread the sample out.** In general, all other things being equal, it is better to sample 10% of the cities in all states rather than all cities in 10% of the states. While it will often prove too costly to follow this advice exactly, follow it to the extent that you can.

4. **For unequally sized groups use PPS sampling.**

Determine Sample Size

Once the above issues have been settled sample sizes can be determined subject to budget and precision constraints (which are always in conflict).

16.9 Some Steps in Conducting a Sampling Study

Some (though assuredly not all) steps you may want to consider in conducting a sampling study follow.

Develop the Operational Plan

The operational plan lays out the practical details of how the study will be conducted. A flow chart detailing the steps that need to be done, who will do them and the schedule for doing them will facilitate the plan.

Prepare the Initial Design of the Study

Follow the steps outlined in the previous section.

Develop the Data Collection Instruments

In a survey this means questionnaire preparation and instruction and training of interviewers. In other sampling studies it means developing the sampling instructions for technicians, data collection sheets, etc.

Conduct a Pilot Study

The importance of this cannot be over-emphasized. A pilot study is a small sampling study conducted before the main study. A pilot study will enable you to:

1. Get the bugs out of the whole operation. By conducting a pilot study, you will find out which parts of the operational plan work and which don't. You can then revise the plan as necessary for the main study.

2. Obtain valuable information on population parameters. For example, you may obtain estimates of population variances needed to determine sample sizes.

The pilot study should be small enough to allow adequate resources for the main study, but large enough to yield useful information.

Revise as Necessary

Using the results of the pilot study, you should now revise the study design, the study instruments and the operational plan.

Conduct and Analyze the Main Study

Now is the time when all the planning pays off. Once the data are obtained, they need to be entered into the computer and analyzed. Then conclusions are drawn. The work culminates in a final report.

16.10 Binomial and Hypergeometric Distribution Models (Optional)

This section draws a comparison between sampling from an infinite population (or equivalently sampling with replacement from a finite population) and sampling without replacement from a finite population. The discussion will lead you to an understanding of where the FPC comes from (though it will be given by a slightly different formula).

First, we make a fundamental observation with regards to the binomial model. This is really an illustration of the basic difference between sampling from a finite population and an infinite population (i.e., the difference between N finite and N infinite).

Consider N balls in an urn with M red balls and $N - M$ white ones. We take n balls at random without replacement (i.e., once a ball is chosen it is not put back in the urn for subsequent draws). What is the distribution model for Y, the number of red balls obtained? If N is infinite, and the proportion of red balls in the urn is p, then $Y \sim b(n,p)$ (i.e., Y follows the binomial distribution model).

In terms of the discussion in Chapter 4, if N is infinite, the number of red balls in n draws from the urn without replacement satisfies the conditions of a binomial experiment:

1. There are n independent trials (i.e. draws).

2. At each trial there are two possible outcomes, "success" (here, a red ball) and "failure" (here, a white ball).

3. The chance of a success is p at each draw.

It is item 3 that is crucial here. The reason the chance of a success is p at each draw is that the proportion of red balls in the urn is p at each draw. And the reason this proportion remains the same as more draws are taken is that N, the number of balls in the urn is infinite, so drawing one (or n) balls does not alter the proportion.

Recall from Chapter 4 that the $b(n,p)$ distribution has probability mass function

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

where $y = 0, 1, 2, \ldots, n$. Recall also the interpretation of $p(y)$ as the proportion of all samples of size n from the population which have $Y = y$ red balls. Finally, recall from Chapter 4 that the mean of the $b(n, p)$ distribution model is np and the variance is $np(1 - p)$.

If N is finite, the $b(n, p)$ model is incorrect. This is because item **3** above is no longer satisfied. To see this, suppose there are $N = 5$ balls in the urn and that 3 of them are red. On the first draw, the chance of drawing a red ball is 3/5. On the second draw, however, it is either 2/4 (if the first draw resulted in a red ball) or 3/4 (if the first draw resulted in a white ball).

In the case of finite N, the number of red balls, Y, obtained in n draws without replacement from an urn containing M red balls and $N - M$ white balls follows the **hypergeometric distribution model**, and we write $Y \sim H(N, M, n)$. The proportion of all samples of size n from the population which have $Y = y$ red balls is

$$p(y) = \frac{\begin{pmatrix} M \\ y \end{pmatrix}\begin{pmatrix} N - M \\ n - y \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

where $\max(0, n - (N - M)) \leq y \leq \min(n, M)$. Here, max stands for maximum and min stands for minimum.

Consider the $H(N, M, n)$ distribution model, and let $p = M/N$, the initial proportion of red balls in the urn. For this $H(N, M, n)$ distribution model, the mean is np just as it is for the $b(n, p)$ distribution model. However, in contrast to the $np(1 - p)$ for the $b(n, p)$ model, the variance of this $H(N, M, n)$ model is

$$\frac{(N - n)}{(N - 1)} np(1 - p).$$

The quantity

$$f = \frac{N - n}{N - 1},$$

called the **finite population correction**, is the factor by which the variability in sampling from a finite population is smaller than the variability in sampling from an infinite population. What becomes of the finite population correction and the variance of the $H(N, M, n)$ model as the number of balls in the urn, N, goes to infinity?

Discussion Questions

1. Give the meaning of the following terms:

   a. Target population

   b. Population units

   c. Census

   d. Sample, judgment sample, probability sample

   e. Frame

   f. Sampling study, sample survey

   g. Simple random sample

   h. Unbiased estimator

2. Tell how to compute the following, and what each means:

   a. Population total, mean, variance, standard deviation, proportion

   b. Sampling fraction, finite population correction

3. Tell how to estimate the population total, mean, variance, and proportion in large simple random samples.

4. Tell how to determine the size of a SRS need to attain a precision d with confidence L.
5. Explain the ideas behind stratified random sampling.
6. Tell how to estimate the population total, mean, variance, and proportion in large stratified random samples.
7. What is optimal allocation? Proportional allocation?
8. When does stratification produce large gains in precision?
9. Explain post-stratification.
10. What is an auxiliary variable?
11. Explain ratio and regression estimation.
12. Explain cluster sampling.
13. What are PPS and PPES?
14. What is multi-stage sampling?
15. What is systematic sampling? How is it connected to stratified sampling and cluster sampling?
16. What is double sampling, and when is it used?
17. Describe the steps in designing a sampling study.
18. Describe the steps in conducting a sampling study.

Exercises

16.1. The student body of a small technical institute consists of 2600 undergraduates. Assume you are a student at this institute.

   a. A simple random sample of size 260 is to be taken. Can you tell what your chance is of being included in this sample? If so, what is it? If not, why not?

   b. A stratified random sample of size 260 is to be taken. Can you tell what your chance is of being included in this sample? If so, what is it? If not, why not?

16.2. A SRS of size n was taken from the 2600 undergraduates at the technical institute mentioned in exercise 16.1. It was desired to estimate the proportion p of the undergraduates who favor the building of a student center whose operation will be partially supported by an increase in student fees. The investigators have no knowledge of what p is. What is the smallest sample size they can take in order to be certain of estimating to within 3 percentage points with 95% confidence?

16.3. The investigators mentioned in exercise 16.2 who were conducting the survey to estimate the proportion p of the 2600 undergraduates who favor building a student center thought that seniors, who would graduate soon, would have a different view than freshmen. So they decided to do a stratified random sample with classes (freshmen, sophomore, junior and senior) as the strata. Assume the numbers in the freshmen, sophomore, junior and senior classes are 600, 650, 650 and 700 respectively, that the numbers sampled were 60, 75, 85 and 65 respectively, and that the numbers favoring the center were 55, 50, 57 and 30 respectively.

   a. Obtain an unbiased estimate of the proportion of freshmen who favor a student center. Construct and interpret a 90% confidence interval for the proportion of freshmen who favor a student center.

   b. Do the same for the seniors.

c. Obtain $\hat{p}_{st}$. Explain what it means here. Construct and interpret a 90% confidence interval, based on $\hat{p}_{st}$, for the proportion of undergraduates who favor a student center.

16.4. Table 16.5 displays the heights in inches of the members of the Cleaver family. Money is a little tight with the Cleavers this year, so they can afford physicals for only three family members. Therefore Doctor Kildare, their family doctor, selects a SRS of size three to receive physicals. At the physical the sample of Cleavers learn their official heights. Ward is anxious to estimate the mean family height from this sample.

a. What is the population here?

b. What are μ and σ^2?

c. By writing out all samples of size three from the population, show that $\overline{y}$ is an unbiased estimator of μ.

Family Member	Height
Ward	71
June	64
Wally	65
Beaver	55

Table 16.5: *Heights of the Cleaver Family*

16.5. The town in which I live has 2,500 voters. It is desired to take a SRS from these voters in order to estimate the proportion who will vote in the next election.

a. How large a sample would I have to take in order to be assured that with 95% confidence the estimate will be accurate to within ±3%?

b. An SRS of size 1000 is actually taken and 850 of the 1000 voters questioned said that they intend to vote in the next election. Estimate the true proportion p of voters who will vote in the next election. Calculate and interpret a 95% confidence interval for p.

16.6. A population consists of 4 households. A SRS of size 3 is to be taken and the respondents asked how many members are in their household. The population units are numbered 1 to 4 and the household sizes are 1, 5, 2, and 6 respectively. You know that the sample mean is an unbiased estimator of the population mean household size. By writing out all samples of size 3 and computing the median of each, verify that the sample median is an unbiased estimator of the population median.

16.7. Estimate the total number of individuals in all households in the population of the previous problem, by computing a 95% confidence interval based on a sample consisting of units 1, 2 and 4. Assume the normal theory interval will be adequate despite the small sample size.

16.8. A salary survey of recent (i.e. over the past five years) management graduates of four local four year colleges was recently conducted. A simple random sample was taken at each of the colleges and the resulting data were obtained:

Measure	College 1	College 2	College 3	College 4
Number in population	720	550	490	390
Number in sample	50	40	30	20
Mean salary (in $10,000)	3.2	2.8	3.7	2.4
Sample SD	.6	.8	1.1	.5

a. Find a 99% confidence interval for the mean salary for all recent management graduates from college 1.

b. Find an estimate of the mean salary of all recent management graduates of the four colleges using an unbiased estimation procedure.

c. Find a 90% confidence interval for the mean salary of all recent management graduates of the four colleges. Interpret this interval.

16.9. Each year the Internal Revenue Service Auditing Division samples tax returns for a full audit. One of the aims of these audits is to estimate the dollar amount of shortfall in the tax system.

A simple random sample of size 200 from all tax returns at a regional IRS office is taken and the returns in the sample are subjected to a complete audit. The table below summarizes the data. The data in the table are divided into four post-strata based on declared adjusted gross income. The totals in each stratum for the variable of interest, tax shortfall, are given. Based on these numbers, do you think post-stratification will greatly increase precision? Would you recommend that the IRS continue taking simple random samples in the future?

Declared Income ($ 1000s)	Tax Shortfall ($ 1000s)	Number (n_i)
Over 200	525	15
100 to 200	481	29
50 to 100	217	48
Under 50	86	108

16.10. Design a sampling scheme for IRS audits. Explain why yours is a better choice than other possibilities.

16.11. Consider the problem of sampling trees in a certain forest to estimate the volume of lumber in harvestable trees in that forest. Devise a workable sampling scheme for this task. State what information your scheme requires. Justify your choice by explaining why it is more workable than other methods.

16.12. A survey of student attitudes is to be conducted at a well-known college. Because of limited resources, the investigators decided to obtain a frame of all classes and then select a SRS of these classes. In each selected class a census was taken. Is this single stage cluster sampling? Why or why not?

16.13. A national political survey seeks to estimate the proportion p of all eligible voters who favor a certain government policy. A SRS is to be used. Suppose that there are 175 million eligible voters and it is desired to estimate p to within ± 0.03 with 95% confidence. Can you tell exactly how large the sample must be? If not, can you give an upper bound on the sample size needed?

16.14. (For students who have a knowledge of simple linear regression.) Obtain the regression estimate and its estimated variance for the tree data in Example 16.6. Compare the results with those for ratio estimation and for estimation if the tree diameters are ignored.

16.15. Compute $\hat{\sigma}^2_{ppes}$ for the data in Example 16.8.

Mini-Project: Do Your Own Survey

Purpose

Your group's task in this project is to plan, conduct and analyze your own survey.

Process

Before designing and conducting your survey, submit a short (one page or less) proposal for your instructor's approval. The proposal should state what question(s) you want answered, what kind of instrument you will use, what kind of sampling design you will use, and what analysis you intend to conduct. Your instructor will discuss the proposal with your group, and may suggest modifications.

After your group's proposal has been approved, proceed to collect the data and use any of the techniques discussed in this chapter that are appropriate.

Lab 16.1: Explore Simple Random Sampling

Introduction

By having you look at the nuts and bolts of the sampling process for a simple example, this lab demonstrates the notions involved in sampling and estimation for simple random samples from a finite population.

Objectives

To give you an understanding of how sampling and estimation work for simple random samples from a finite population.

Lab Materials

None needed.

Experimental Procedure

The first two columns of Table 16.6 show all $\binom{8}{2} = 28$ possible samples of size two from the population in Table 16.1 of Chapter 16. The third column shows the measured emissions for each unit in the sample, and the fourth and fifth columns show the sample mean and variance for the measured emissions in each sample. The sixth column shows the values of the compliance variable for each unit in the sample, and the seventh column displays the sample proportion of units in compliance.

Under simple random sampling, each of these 28 possible samples has equal chance of being chosen. These data are found in the data set EMISS1.

Unbiasedness

Compute the mean of YBAR. You will find that it equals 95.625, which is the value of μ, the population mean. In other words, the mean of the sample means for all possible samples equals the population mean. This shows that $\overline{y}$ is an unbiased estimator of μ.

How would you check that for these data, s^2 is an unbiased estimator of σ^2? That $\hat{p}$ is an unbiased estimator of p? Do so now for both of these

Finite Population Correction

Now compute the variance of YBAR in the data set EMISS1. However, when computing this variance, do not use the formula

$$\frac{1}{28-1}\sum_{i=1}^{28}(\overline{y}_i - \mu)^2,$$

which is the default in SAS/INSIGHT, but rather the formula

$$\frac{1}{28}\sum_{i=1}^{28}(\overline{y}_i - \mu)^2.$$

To do this, choose *Analyze:Distribution (Y)* from the menu bar of the data window. From the dialog window, choose YBAR as the Y variable, then click on the "Method" button, and choose 'N' instead of 'DF' as the variance divisor.

The variance you obtain for YBAR is 58.0647. Verify that this is $0.75 \times (154.839)/2$ (i.e. $((1-n/N) \times \sigma^2/n)$. This shows the validity of the formula for the variance of $\overline{y}$ involving the FPC.

Sample	Engines	Emissions		$\overline{y}$	$s^2_{\overline{y}}$	Compliance	$\hat{p}$
Number	in Sample	in Sample				in Sample	
1	2 1	78	90	84.0	72.0	1 1	1.0
2	3 1	101	90	95.5	60.5	0 1	0.5
3	3 2	101	78	89.5	264.5	0 1	0.5
4	4 1	95	90	92.5	12.5	1 1	1.0
5	4 2	95	78	86.5	144.5	1 1	1.0
6	4 3	95	101	98.0	18.0	1 0	0.5
7	5 1	92	90	91.0	2.0	1 1	1.0
8	5 2	92	78	85.0	98.0	1 1	1.0
9	5 3	92	101	96.5	40.5	1 0	0.5
10	5 4	92	95	93.5	4.5	1 1	1.0
11	6 1	121	90	105.5	480.5	0 1	0.5
12	6 2	121	78	99.5	924.5	0 1	0.5
13	6 3	121	101	111.0	200.0	0 0	0.0
14	6 4	121	95	108.0	338.0	0 1	0.5
15	6 5	121	92	106.5	420.5	0 1	0.5
16	7 1	89	90	89.5	0.5	1 1	1.0
17	7 2	89	78	83.5	60.5	1 1	1.0
18	7 3	89	101	95.0	72.0	0 1	0.5
19	7 4	89	95	92.0	18.0	1 1	1.0
20	7 5	89	92	90.5	4.5	1 1	1.0
21	7 6	89	121	105.0	512.0	1 0	0.5
22	8 1	99	90	94.5	40.5	1 1	1.0
23	8 2	99	78	88.5	220.5	1 1	1.0
24	8 3	99	101	100.0	2.0	1 0	0.5
25	8 4	99	95	97.0	8.0	1 1	1.0
26	8 5	99	92	95.5	24.5	1 1	1.0
27	8 6	99	121	110.0	242.0	1 0	0.5
28	8 7	99	89	94.0	50.0	1 1	1.0

Table 16.6: *All samples of size 2 from emission population*

Lab 16.2: Explore Stratification

Introduction

By having you look at the nuts and bolts of the sampling process for a simple example, this lab demonstrates the notions involved in sampling and estimation for stratified random samples from a finite population. It also demonstrates the gains that may be obtained from stratification, and compares optimal and proportional allocation.

Objectives

To give you an understanding of how sampling and estimation work for stratified random samples from a finite population. To demonstrate the gains that may be obtained from stratification. To compare optimal and proportional allocation.

Lab Materials

None needed.

Experimental Procedure

Look at the data set FIRMS. As explained in the chapter, these are the number of employees of 115 multinational corporations. The firms are arranged in two strata, the first containing the 94 firms with fewer than 100,000 employees, and the second the remaining 21 firms.

Assume that these 115 firms are the population about which inference is to be made. Assume also that $10,000 is available to sample from this population, and that you are to spend as close to the full $10,000 as possible. Your goal is to use a sample to estimate the total number of employees in the firms in the population. Begin by assuming that regardless of the size of the firm, it costs $400 to sample a single firm.

1. Calculate N, n, f, Y μ, and σ. Also, for $i = 1, 2$ calculate N_i, Y_i, μ_i, and σ_i.

2. Calculate n_i and $f_i, i = 1, 2$ when sampling under optimal allocation. Do the same when sampling under proportional allocation.

3. Calculate the standard deviation of $\hat{Y}$ under SRS, optimal allocation and proportional allocation. Which type of estimation is most precise? Which is least precise? What are their relative magnitudes? (Note: in computing the variance of $\hat{Y}_{st}$ use the actual population variances instead of variances estimated from a sample).

4. Suppose the cost of sampling a large firm is $900 and the cost of sampling a small firm is $100. Find n and the optimal allocation. How does it differ from the case in which all firms cost the same amount to sample?