Madow. 2/ The conditions are usually satisfied with regard to estimates from sample surveys. As a rule of thumb the variance formula is usually accepted as satisfactory if the coefficient of variation of the variable in the denominator is less than 0.1; that is, if $\frac{\sigma_w}{\bar{w}} < 0.1$. In other words, this condition states that the coefficient of variation of the estimate in the denominator should be less than 10 percent. A larger coefficient of variation might be tolerable before becoming concerned about Equation (3.26) as an approximation.

The condition $\frac{\sigma_w}{\bar{w}} < 0.1$ is more stringent than necessary for regarding the bias of a ratio as negligible. With few exceptions in practice the bias of a ratio is ignored. Some of the logic for this will appear in the illustration below. To summarize, the conditions when Equations (3.25) and (3.26) are not good approximations are such that the ratio is likely to be of questionable value owing to large variance.

If u and w are linear combinations of random variables, the theory presented in previous sections applies to u and to w. Assuming u and w are estimates from a sample, to estimate $\text{Var}(\frac{u}{w})$ take into account the sample design and substitute in Equation (3.26) estimates of $\bar{u}$, $\bar{w}$, σ_u^2, σ_w^2, and ρ_{uw}. Ignore Equation (3.25) unless there is reason to believe the bias of the ratio might be important relative to its standard error.

It is of interest to note the similarity between Var(u-w) and $\text{Var}(\frac{u}{w})$. According to Theorem 3.5,

$$\text{Var}(u-w) = \sigma_u^2 + \sigma_w^2 - 2\rho_{uw}\sigma_u\sigma_w$$

2/ Hansen, Hurwitz, and Madow, Sample Survey Methods and Theory, Volume I, Chapter 4, John Wiley and Sons, 1953.

By definition the relative variance of an estimate is the variance of the estimate divided by the square of its expected value. Thus, in terms of the relative variance of a ratio, Equation (3.26) can be written

$$\text{Rel Var}(\frac{u}{w}) = \frac{\sigma_u^2}{\bar{u}^2} + \frac{\sigma_w^2}{\bar{w}^2} - 2\rho_{uw} \frac{\sigma_u \sigma_w}{\bar{u}\bar{w}}$$

The similarity is an aid to remembering the formula for $\text{Var}(\frac{u}{w})$.

Illustration 3.13. Suppose one has a simple random sample of n elements from a population of N. Let $\bar{x}$ and $\bar{y}$ be the sample means for characteristics X and Y. Then, $u = \bar{x}$, $w = \bar{y}$,

$$\sigma_u^2 = \frac{N-n}{N} \frac{S_X^2}{n} \quad \text{and} \quad \sigma_w^2 = \frac{N-n}{N} \frac{S_Y^2}{n}$$

Notice that the condition discussed above, $\frac{\sigma_w}{\bar{w}} < 0.1$, is satisfied if the sample is large enough so

$$\frac{N-n}{N} \frac{S_Y^2}{n\bar{Y}^2} < 0.1^2$$

Substituting in Equation (3.26) we obtain the following as the variance of the ratio:

$$\text{Var}(\frac{\bar{x}}{\bar{y}}) = (\frac{N-n}{N})(\frac{1}{n}) \frac{\bar{X}^2}{\bar{Y}^2} [\frac{S_X^2}{\bar{X}^2} + \frac{S_Y^2}{\bar{Y}^2} - \frac{2\rho_{XY}S_XS_Y}{\bar{X}\bar{Y}}]$$

The bias of $\frac{\bar{x}}{\bar{y}}$ as an estimate of $\frac{\bar{X}}{\bar{Y}}$ is given by the second term of Equation (3.25). For this illustration it becomes

$$(\frac{N-n}{N})(\frac{1}{n}) \frac{\bar{X}}{\bar{Y}} [\frac{S_Y^2}{\bar{Y}^2} - \frac{\rho_{XY}\sigma_X\sigma_Y}{\bar{X}\bar{Y}}]$$

As the size of the sample increases, the bias decreases as $\frac{1}{n}$ whereas the standard error of the ratio decreases at a slower rate, namely $\frac{1}{\sqrt{n}}$.

Thus, we need not be concerned about a possibility of the bias becoming important relative to sampling error as the size of the sample increases. A possible exception occurs when several ratios are combined. An example is stratified random sampling when many strata are involved and separate ratio estimates are made for the strata. This is discussed in the books on sampling.

3.9 CONDITIONAL EXPECTATION

The theory for conditional expectation and conditional variance of a random variable is a very important part of sampling theory, especially in the theory for multistage sampling. The theory will be discussed with reference to two-stage sampling.

The notation that will be used in this and the next section is as follows:

M is the number of psu's (primary sampling units) in the population.

m is the number of psu's in the sample.

N_i is the total number of elements in the i^{th} psu.

$N = \sum_{i}^{M} N_i$ is the total number of elements in the population.

n_i is the sample number of elements from the i^{th} psu.

$n = \sum_{i}^{m} n_i$ is the total number of elements in the sample.

$\bar{n} = \frac{n}{m}$

X_{ij} is the value of X for the j^{th} element in the i^{th} psu. It refers to an element in the population, that is, $j = 1,\ldots, N_i$, and $i = 1,\ldots, M$.

x_{ij} is the value of X for the j^{th} element in the sample from the i^{th} psu in the sample, that is, the indexes i and j refer to the set of psu's and elements in the sample.

$X_{i\cdot} = \sum_{j}^{N_i} X_{ij}$ is the population total for the i^{th} psu.

$\bar{X}_{i\cdot} = \frac{X_{i\cdot}}{N_i}$ is the average of X for all elements in the i^{th} psu.

$\bar{X}_{\cdot\cdot} = \frac{\sum\sum_{ij}^{MN_i} X_{ij}}{N} = \frac{\sum_{i}^{M} X_{i\cdot}}{N}$ is the average of all N elements.

$\bar{X}_{\cdot} = \frac{\sum_{i}^{M} X_{i\cdot}}{M}$ is the average of the psu totals. Be sure to note the difference between $\bar{X}_{\cdot\cdot}$ and $\bar{X}_{\cdot}$.

$x_{i\cdot} = \sum_{j}^{n_i} x_{ij}$ is the sample total for the i^{th} psu in the sample.

$\bar{x}_{i\cdot} = \frac{x_{i\cdot}}{n_i}$ is the average for the n_i elements in the sample from the i^{th} psu.

$\bar{x}_{\cdot\cdot} = \frac{\sum\sum_{ij}^{mn_i} x_{ij}}{n}$ is the average for all elements in the sample.

Assume simple random sampling, equal probability of selection without replacement, at both stages. Consider the sample of n_i elements from the i^{th} psu. We know from Section 3.3 that $\bar{x}_{i\cdot}$ is an unbiased estimate of the psu mean $\bar{X}_{i\cdot}$; that is, $E(\bar{x}_{i\cdot}) = \bar{X}_{i\cdot}$ and for a fixed i (a specified psu) $EN_i\bar{x}_{i\cdot} = N_iE(\bar{x}_{i\cdot}) = N_i\bar{X}_{i\cdot} = X_{i\cdot}$. But, owing to the first stage of sampling,

$EN_i\bar{x}_i$ must be treated as a random variable. Hence, it is necessary to become involved with the expected value of an expected value.

First, consider X as a random variable, in the context of single-stage sampling, which could equal any one of the values X_{ij} in the population set of $N = \sum_{i}^{M} N_i$. Let P(ij) be the probability of selecting the j^{th} element in the i^{th} psu; that is, P(ij) is the probability of X being equal to X_{ij}. By definition

$$E(X) = \sum_{ij}^{MN_i} P(ij)X_{ij} \tag{3.27}$$

Now consider the selection of an element as a two-step procedure: (1) selected a psu with probability P(i), and (2) selected an element within the selected psu with probability P(j|i). In words, P(j|i) is the probability of selecting the j^{th} element in the i^{th} psu given that the i^{th} psu has already been selected. Thus, P(ij) = P(i)P(j|i). By substitution, Equation (3.27) becomes

$$E(X) = \sum_{ij}^{MN_i} P(i)P(j|i)X_{ij}$$

or

$$E(X) = \sum_{i}^{M} P(i) \sum_{j}^{N_i} P(j|i)X_{ij} \tag{3.28}$$

By definition, $\sum_{j}^{N_i} P(j|i)X_{ij}$ is the expected value of X for a fixed value of i. It is called"conditional expectation."

Let $E_2(X|i) = \sum_{j}^{N_i} P(j|i)X_{ij}$ where $E_2(X|i)$ is the form of notation we will be using to designate conditional expectation. To repeat, $E_2(X|i)$ means the expected value of X for a fixed i. The subscript 2 indicates

that the conditional expectation applies to the second stage of sampling. E_1 and E_2 will refer to expectation at the first and second stages, respectively.

Substituting $E_2(X|i)$ in Equation (3.28) we obtain

$$E(X) = \sum_{i}^{M} P(i)\ E_2(X|i) \tag{3.29}$$

There is one value of $E_2(X|i)$ for each of the M psu's. In fact $E_2(X|i)$ is a random variable where the probability of $E_2(X|i)$ is P(i). Thus the right-hand side of Equation (3.29) is, by definition, the expected value of $E_2(X|i)$. This leads to the following theorem:

Theorem 3.6. $E(X) = E_1E_2(X|i)$

Suppose $P(j|i) = \frac{1}{N_i}$ and $P(i) = \frac{1}{M}$. Then,

$$E_2(X|i) = \sum_{j}^{N_i}(\frac{1}{N_i})X_{ij} = \bar{X}_{i\cdot}$$

and $$E(X) = E_1(\bar{X}_{i\cdot}) = \sum_{i}^{M}(\frac{1}{M})(\bar{X}_{i\cdot}) = \frac{\Sigma\bar{X}_{i\cdot}}{M}.$$

In this case E(X) is an unweighted average of the psu averages. It is important to note that, if P(i) and P(j|i) are chosen in such a way that P(ij) is constant, every element has the same chance of selection. This point will be discussed later.

Theorem 3.3 dealt with the expected value of a linear combination of random variables. There is a corresponding theorem for conditional expectation. Assume the linear combination is

$$U = a_1u_1+\ldots+a_ku_k = \sum_{t=1}^{k} a_tu_t$$

where $a_1,\ldots,a_k$ are constants and $u_1,\ldots,u_k$ are random variables. Let $E(U|c_i)$ be the expected value of U under a specified condition, c_i, where c_i is one of the conditions out of a set of M conditions that could occur. The theorem on conditional expectation can then be stated symbolically as follows:

Theorem 3.7. $E(U|c_i) = a_1E(u_1|c_i) +\ldots+ a_kE(u_k|c_i)$

$$\text{or} \quad E(U|c_i) = \sum_t^k a_tE(u_t|c_i)$$

Compare Theorems 3.7 and 3.3 and note that Theorem 3.7 is like Theorem 3.3 except that conditional expectation is applied. Assume c is a random event and that the probability of the event c_i occurring is P(i). Then $E(U|c_i)$ is a random variable and by definition the expected value of $E(U|c_i)$ is $\sum_i^M P(i)E(U|c_i)$ which is E(U). Thus, we have the following theorem:

Theorem 3.8. The expected value of U is the expected value of the conditional expected value of U, which in symbols is written as follows:

$$E(U) = EE(U|c_i) \tag{3.30}$$

Substituting the value of $E(U|c_i)$ from Theorem 3.7 in Equation (3.30) we have

$$E(U) = E[a_1E(u_1|c_i)+\ldots+a_kE(u_k|c_i)] = E[\sum_t^k a_tE(u_t|c_i)] \tag{3.31}$$

Illustration 3.14. Assume two-stage sampling with simple random sampling at both stages. Let x', defined as follows, be the estimator of the population total:

$$x' = \frac{M}{m}\sum_i^m \frac{N_i}{n_i}\sum_j^{n_i} x_{ij} \tag{3.32}$$

Exercise 3.17. Examine the estimator, x', Equation (3.32). Express it in other forms that might help show its logical structure. For example, for a fixed i what is $\frac{N_i}{n_i} \sum_j^{n_i} x_{ij}$? Does it seem like a reasonable way of estimating the population total?

To display x' as a linear combination of random variables it is convenient to express it in the following form:

$$x' = [\frac{M}{m}\frac{N_1}{n_1} x_{11}+\ldots+ \frac{M}{m}\frac{N_1}{n_1} x_{1n_1}] +\ldots+ [\frac{M}{m}\frac{N_m}{n_m} x_{m1} +\ldots+ \frac{M}{m}\frac{N_m}{n_m} x_{mn_m}] \qquad (3.33)$$

Suppose we want to find the expected value of x' to determine whether it is equal to the population total. According to Theorem 3.8,

$$E(x') = E_1E_2\ (x'|i) \qquad (3.34)$$

$$E(x') = E_1E_2\{[\frac{M}{m} \sum_i^m \frac{N_i}{n_i} \sum_j^{n_i} x_{ij}]|i\} \qquad (3.35)$$

Equations (3.34) and (3.35) are obtained simply by substituting x' as the random variable in (3.30). The c_i now refers to any one of the m psu's in the sample. First we must solve the conditional expectation, $E_2(x'|i)$. Since $\frac{M}{m}$ and $\frac{N_i}{n_i}$ are constant with respect to the conditional expectation, and making use of Theorem 3.7, we can write

$$E_2(x'|i) = \frac{M}{m} \sum_i^m \frac{N_i}{n_i} \sum_j^{n_i} E_2(x_{ij}|i) \qquad (3.36)$$

We know for any given psu in the sample that x_{ij} is an element in a simple random sample from the psu and according to Section 3.3 its expected value is the psu mean, $\bar{X}_{i.}$. That is,

$$E_2(x_{ij}|i) = \bar{X}_{i.}$$

and $$\sum_{j}^{n_i} E_2(x_{ij}|i) = n_i \bar{X}_{i.} \tag{3.37}$$

Substituting the result from Equation (3.37) in Equation (3.36) gives

$$E_2(x'|i) = \frac{M}{m} \sum_{i}^{m} N_i \bar{X}_{i.} \tag{3.38}$$

Next we need to find the expected value of $E_2(x'|i)$. In Equation (3.38), N_i is a random variable, as well as $\bar{X}_{i.}$, associated with the first stage of sampling. Accordingly, we will take $X_{i.} = N_i \bar{X}_{i.}$ as the random variable which gives in lieu of Equation (3.38).

$$E_2(x'|i) = \frac{M}{m} \sum_{i}^{m} X_{i.}$$

Therefore,

$$E(x') = E_1[\frac{M}{m} \sum_{i}^{m} X_{i.}]$$

From Theorem 3.3

$$E_1[\frac{M}{m} \sum_{i}^{m} X_{i.}] = \frac{M}{m} \sum_{i}^{m} E_1(X_{i.})$$

Since

$$\sum_{i}^{m} E_1(X_{i.}) = m[\frac{\sum_{i}^{M} X_{i.}}{M}]$$

$$E_1[\frac{M}{m} \sum_{i}^{m} X_{i.}] = \sum_{i}^{M} X_{i.}$$

Therefore, $E(x') = \sum_{i}^{M} X_{i.} = X_{..}$ This shows that x' is an unbiased estimator of the population total.

3.10 CONDITIONAL VARIANCE

Conditional variance refers to the variance of a variable under a specified condition or limitation. It is related to conditional probability and to conditional expectation.

To find the variance of x' (See Equation (3.32) or (3.33)) the following important theorem will be used:

Theorem 3.9. The variance of x' is given by

$$V(x') = V_1E_2(x'|i) + E_1V_2(x'|i)$$

where V_1 is the variance for the first stage of sampling and V_2 is the "conditional" variance for the second stage.

We have discussed $E_2(x'|i)$ and noted there is one value of $E_2(x'|i)$ for each psu in the population. Hence $V_1E_2(x'|i)$ is simply the variance of the M values of $E_2(x'|i)$.

In Theorem 3.9 the conditional variance, $V_2(x'|i)$, by definition is

$$V_2(x'|i) = E_2\{[x'-E_2(x'|i)]^2 \,|i\}$$

To understand $V_2(x'|i)$ think of x' as a linear combination of random variables (see Equation (3.33)). Consider the variance of x' when i is held constant. All terms (random variables) in the linear combination are now constant except those originating from sampling within the i^{th} psu. Therefore, $V_2(x'|i)$ is associated with variation among elements in the i^{th} psu. $V_2(x'|i)$ is a random variable with M values in the set, one for each psu. Therefore, $E_1V_2(x'|i)$ by definition is

$$E_1V_2(x'|i) = \sum_{i}^{M} P(i)V_2(x'|i)$$

That is, $E_1V_2(x'|i)$ is an average of M values of $V_2(x'|i)$ weighted by P(i), the probability that the i^{th} psu had of being in the sample.

Three illustrations of the application of Theorem 3.9 will be given. In each case there will be five steps in finding the variance of x':

Step 1, find $E_2(x'|i)$

Step 2, find $V_1E_2(x'|i)$

Step 3, find $V_2(x'|i)$

Step 4, find $E_1V_2(x'|i)$

Step 5, combine results from Steps 2 and 4.

Illustration 3.15. This is a simple illustration, selected because we know what the answer is from previous discussion and a linear combination of random variables is not involved. Suppose x' in Theorem 3.9 is simply the random variable X where X has an equal probability of being any one of the X_{ij} values in the set of $N = \sum_i^M N_i$. We know that the variance of X can be expressed as follows:

$$V(x') = \frac{1}{N}\sum_{ij}^{MN_i}(X_{ij}-\bar{X}_{..})^2 \tag{3.39}$$

In the case of two-stage sampling an equivalent method of selecting a value of X is to select a psu first and then select an element within the psu, the condition being that $P(ij) = P(i)P(j|i) = \frac{1}{N}$. This condition is satisfied by letting $P(i) = \frac{N_i}{N}$ and $P(j|i) = \frac{1}{N_i}$. We now want to find V(X) by using Theorem 3.9 and check the result with Equation (3.39).

Step 1. From the random selection specifications we know that $E_2(x'|i) = \bar{X}_{i.}$. Therefore,

Step 2. $V_1E_2(x'|i) = V_1(\bar{X}_{i.})$

We know that $\bar{X}_{i.}$ is a random variable that has a probability of $\frac{N_i}{N}$ of being equal to the i^{th} value in the set $\bar{X}_1,\ldots,\bar{X}_M$. Therefore, by definition of the variance of a random variable,

$$V_1E(x'|i) = \sum_i^M \frac{N_i}{N}(\bar{X}_{i.}-\bar{X}_{..})^2 \tag{3.40}$$

where $\bar{X}_{..} = \sum_i^M \frac{N_i}{N}\bar{X}_{i.} = \frac{\sum_i^M X_{i.}}{N}$

Step 3. By definition

$$V_2(x'|i) = \sum_j^{N_i} \frac{1}{N_i} (X_{ij}-\bar{X}_{i.})^2$$

Step 4. Since each value of $V_2(x'|i)$ has a probability $\frac{N_i}{N}$

$$E_1V_2(x'|i) = \sum_i^M \frac{N_i}{N} \sum_j^{N_i} \frac{1}{N_i} (X_{ij}-\bar{X}_{i.})^2 \tag{3.41}$$

Step 5. From Equations (3.40) and (3.41) we obtain

$$V(x') = \frac{1}{N} [\sum_i^M N_i(\bar{X}_{i.}-\bar{X}_{..})^2 + \sum_i^M \sum_j^{N_i} (X_{ij}-\bar{X}_{i.})^2] \tag{3.42}$$

The fact that Equations (3.42) and (3.39) are the same is verified by Equation (1.10) in Chapter I.

Illustration 3.16. Find the variance of the estimator x' given by Equation (3.32) assuming simple random sampling at both stages of sampling.

Step 1. Theorem 3.7 is applicable. That is,

$$E_2(x'|i) = \sum_{ij}^{mn_i} E_2[\frac{M}{m} \frac{N_i}{n_i} x_{ij}|i]$$

which means "sum the conditional expected values of each of the n terms in Equation (3.33)."

With regard to any one of the terms in Equation (3.33), the conditional expectation is

$$E_2[\frac{M}{m} \frac{N_i}{n_i} x_{ij}|i] = \frac{M}{m} \frac{N_i}{n_i} E_2(x_{ij}|i) = \frac{M}{m} \frac{N_i}{n_i} \bar{X}_{i.} = \frac{M}{m} \frac{X_{i.}}{n_i}$$

Therefore

$$E_2(x'|i) = \sum_{ij}^{mn_i} \frac{M}{m} \frac{X_{i.}}{n_i} \tag{3.43}$$

With reference to Equation (3.43) and summing with respect to j, we have

$$\sum_{j}^{n_i} \frac{M}{m} \frac{X_{i\cdot}}{n_i} = \frac{M}{m} X_{i\cdot}$$

Hence Equation (3.43) becomes

$$E_2(x'|i) = \frac{M}{m} \sum_{i}^{m} X_{i\cdot} \tag{3.44}$$

Step 2. Find $V_1E_2(x'|i)$. This is simple because $\frac{\sum_{i}^{m} X_i}{m}$ in Equation (3.44) is the mean of a random sample of m from the set of psu totals $X_{1\cdot}, \ldots, X_{M\cdot}$. Therefore,

$$V_1E_2(x'|i) = M^2\left(\frac{M-m}{M-1}\right) \frac{\sigma^2_{b1}}{m} \tag{3.45}$$

where

$$\sigma^2_{b1} = \frac{\sum_{i}^{M} (X_{i\cdot} - \bar{X}_{\cdot})^2}{M} \quad \text{and} \quad \bar{X}_{\cdot} = \frac{\sum_{i}^{M} X_{i\cdot}}{M}$$

In the subscript to σ^2, the "b" indicates between psu variance and "1" distinguishes this variance from between psu variances in later illustrations.

Step 3. Finding $V_2(x'|i)$, is more involved because the conditional variance of a linear combination of random variables must be derived. However, this is analogous to using Theorem 3.5 for finding the variance of a linear combination of random variables. Theorem 3.5 applies except that $V(u|i)$ replaces $V(u)$ and conditional variance and conditional covariance replace the variances and covariances in the formula for $V(u)$. As the solution proceeds, notice that the strategy is to shape the problem so previous results can be used.

Look at the estimator x', Equation (3.33), and determine whether any covariances exist. An element selected from one psu is independent of an

element selected from another; but within a psu the situation is the same as the one we had when finding the variance of the mean of a simple random sample. This suggests writing x' in terms of $\bar{x}_{i.}$ because the $\bar{x}_{i.}$'s are independent. Accordingly, we will start with

$$x' = \frac{M}{m} \sum_{i}^{m} N_i \bar{x}_{i.}$$

Hence

$$V_2(x'|i) = V_2\{[\frac{M}{m} \sum_{i}^{m} N_i \bar{x}_{i.}]|i\}$$

Since the $\bar{x}_{i.}$'s are independent

$$V_2(x'|i) = \frac{M^2}{m^2} \sum_{i}^{m} V_2(N_i \bar{x}_{i.}|i)$$

and since N_i is constant with regard to the conditional variance

$$V_2(x'|i) = \frac{M^2}{m^2} \sum_{i}^{m} N_i^2 V_2(\bar{x}_{i.}|i) \tag{3.46}$$

Since the sampling within each psu is simple random sampling

$$V_2(\bar{x}_{i.}|i) = (\frac{N_i - n_i}{N_i - 1}) \frac{\sigma_i^2}{n_i} \tag{3.47}$$

where

$$\sigma_i^2 = \sum_{j}^{N_i} \frac{1}{N_i} (X_{ij} - \bar{X}_{i.})^2$$

Step 4. After substituting the value of $V_2(\bar{x}_{i.}|i)$ in Equation (3.46), and then applying Theorem 3.3, we have

$$E_1 V_2(x'|i) = \frac{M^2}{m^2} \sum_{i}^{m} E_1[N_i^2 \frac{N_i - n_i}{N_i - 1} \frac{\sigma_i^2}{n_i}]$$

Since the first stage of sampling was simple random sampling and each psu had an equal chance of being in the sample,

$$E_1[N_i^2 \frac{N_i-n_i}{N_i-1} \frac{\sigma_i^2}{n_i}] = \frac{1}{M} \sum_i^M N_i^2 \frac{N_i-n_i}{N_i-1} \frac{\sigma_i^2}{n_i}$$

Hence

$$E_1V_2(x'|i) = \frac{M}{m} \sum_i^M N_i^2 \frac{N_i-n_i}{N_i-1} \frac{\sigma_i^2}{n_i} \tag{3.48}$$

Step 5. Combining Equation (3.48) and Equation (3.45) the answer is

$$V(x') = M^2 \frac{M-m}{M-1} \frac{\sigma_{b1}^2}{m} + \frac{M}{m} \sum_i^M N_i^2 \frac{N_i-n_i}{N_i-1} \frac{\sigma_i^2}{n_i} \tag{3.49}$$

Illustration 3.17. The sampling specifications are: (1) at the first stage select m psu's with replacement and probability $P(i) = \frac{N_i}{N}$, and (2) at the second stage a simple random sample of $\bar{n}$ elements is to be selected from each of the m psu's selected at the first stage. This will give a sample of $n = m\bar{n}$ elements. Find the variance of the sample estimate of the population total.

The estimator needs to be changed because the psu's are not selected with equal probability. Sample values need to be weighted by the reciprocals of their probabilities of selection if the estimator is to be unbiased. Let

$P'(ij)$ be the probability of element ij being in the sample,

$P'(i)$ be the relative frequency of the i^{th} psu being in a sample of m, and let

$P'(j|i)$ equal the conditional probability of element ij being in the sample given that the i^{th} psu is already in the sample.

Then

$$P'(ij) = P'(i)P'(j|i)$$

According to the sampling specifications $P'(i) = m \frac{N_i}{N}$. This probability was described as relative frequency because "probability of being

in a sample of m psu's" is subject to misinterpretation. The i^{th} psu can appear in a sample more than once and it is counted every time it appears. That is, if the i^{th} psu is selected more than once, a sample of $\bar{n}$ is selected within the i^{th} psu every time that it is selected. By substitution

$$P'(ij) = [m \frac{N_i}{N}] \frac{\bar{n}}{N_i} = \frac{m\bar{n}}{N} = \frac{n}{N} \tag{3.50}$$

Equation (3.50) means that every element has an equal probability of being in the sample. Consequently, the estimator is very simple,

$$x' = \frac{N}{m\bar{n}} \sum_{ij}^{m\bar{n}} x_{ij} \tag{3.51}$$

Exercise 3.18. Show that x', Equation (3.51), is an unbiased estimator of the population total.

In finding $V(x')$ our first step was to solve for $E_2(x'|i)$.

Step 1. By definition

$$E_2(x'|i) = E_2\{[\frac{N}{m\bar{n}} \sum_{ij}^{m\bar{n}} x_{ij}]|i\}$$

Since i is constant with regard to E_2,

$$E_2(x'|i) = \frac{N}{m\bar{n}} \sum_{ij}^{m\bar{n}} E_2(x_{ij}|i) \tag{3.52}$$

Proceeding from Equation (3.52) to the following result is left as an exercise:

$$E_2(x'|i) = \frac{N}{m} \sum_{i}^{m} \bar{X}_{i.} \tag{3.53}$$

Step 2. From Equation (3.53) we have

$$V_1E_2(x'|i) = V_1(\frac{N}{m} \sum_{i}^{m} \bar{X}_{i.})$$

Since the $\bar{X}_{i.}$'s are independent

$$V_1E_2(x'|i) = \frac{N^2}{m^2} \sum_{i}^{m} V_1(\bar{X}_{i.})$$

Because the first stage of sampling is sampling with probability proportional to N_i and with replacement,

$$V_1(\bar{X}_{i.}) = \sum_{i}^{M} \frac{N_i}{N} (\bar{X}_{i.}-\bar{X}_{..})^2 \tag{3.54}$$

Let

$$V_1(\bar{X}_{i.}) = \sigma_{b2}^2$$

Then

$$V_1E_2(x'|i) = \frac{N^2}{m^2} (m\sigma_{b2}^2) = \frac{N^2}{m} \sigma_{b2}^2 \tag{3.55}$$

Exercise 3.19. Prove that $E(\bar{X}_{i.}) = \bar{X}_{..}$ which shows that it is appropriate to use $\bar{X}_{..}$ in Equation (3.54).

Step 3. To find $V_2(x'|i)$, first write the estimator as

$$x' = \frac{N}{m} \sum_{i}^{m} \bar{x}_{i.} \tag{3.56}$$

Then, since the $\bar{x}_{i.}$'s are independent

$$V_2(x'|i) = \frac{N^2}{m^2} \sum_{i}^{m} V_2(\bar{x}_{i.})$$

and

$$V_2(\bar{x}_{i.}) = \frac{N_i-\bar{n}}{N_i-1} \frac{\sigma_i^2}{\bar{n}}$$

where

$$\sigma_i^2 = \sum_{j}^{N_i} \frac{1}{N_i} (X_{ij}-\bar{X}_{i.})^2$$

Therefore

$$V_2(x'|i) = \frac{N^2}{m^2} \sum_{i}^{m} \frac{N_i-\bar{n}}{N_i-1} \frac{\sigma_i^2}{\bar{n}}$$

Step 4.

$$E_1V_2(x'|i) = \frac{N^2}{m^2} \frac{1}{\bar{n}} \sum_{i}^{m} E_1\left(\frac{N_i-\bar{n}}{N_i-1} \sigma_i^2\right)$$

Since the probability of $V_2(x'|i)$ is $\frac{N_i}{N}$

$$E_1V_2(x'|i) = \frac{N^2}{m^2} \frac{1}{\bar{n}} \sum_{i}^{m} \left[\sum_{i}^{M} \frac{N_i}{N} \left(\frac{N_i-\bar{n}}{N_i-1}\right)\sigma_i^2\right]$$

which becomes

$$E_1V_2(x'|i) = \frac{N^2}{m\bar{n}} \sum_{i}^{M} \frac{N_i}{N} \left(\frac{N_i-\bar{n}}{N_i-1}\right)\sigma_i^2 \tag{3.57}$$

Step 5. Combining Equation (3.55) and Equation (3.57) we have the answer

$$V(x') = N^2 \left[\frac{\sigma_{b2}^2}{m} + \frac{1}{m\bar{n}} \sum \frac{N_i}{N} \left(\frac{N_i-\bar{n}}{N_i-1}\right)\sigma_i^2\right] \tag{3.58}$$

CHAPTER IV. THE DISTRIBUTION OF AN ESTIMATE

4.1 PROPERTIES OF SIMPLE RANDOM SAMPLES

The distribution of an estimate is a primary basis for judging the accuracy of an estimate from a sample survey. But an estimate is only one number. How can one number have a distribution? Actually, "distribution of an estimate" is a phrase that refers to the distribution of all possible estimates that might occur under repetition of a prescribed sampling plan and estimator (method of estimation). Thanks to theory and empirical testing of the theory, it is not necessary to generate physically the distribution of an estimate by selecting numerous samples and making an estimate from each. However, to have a tangible distribution of an estimate as a basis for discussion, an illustration has been prepared.

Illustration 4.1. Consider simple random samples of 4 from an assumed population of 8 elements. There are $\frac{N!}{n!(N-n)!} = \frac{8!}{4!4!} = 70$ possible samples. In Table 4.1, the sample values for all of the 70 possible samples of four are shown. The 70 samples were first listed in an orderly manner to facilitate getting all of them accurately recorded. The mean, $\bar{x}$, for each sample was computed and the samples were then arrayed according to the value of $\bar{x}$ for purposes of presentation in Table 4.1. The distribution of $\bar{x}$ is the 70 values of $\bar{x}$ shown in Table 4.1, including the fact that each of the 70 values of $\bar{x}$ has an equal probability of being the estimate. These 70 values have been arranged as a frequency distribution in Table 4.2.

As discussed previously, one of the properties of simple random sampling is that the sample average is an unbiased estimate of the population average; that is, $E(\bar{x}) = \bar{X}$. This means that the distribution of

Table 4.1--Samples of four elements from a population of eight 1/

Sample number	Values of x_i	$\bar{x}$	s^2	Sample number	Values of x_i	$\bar{x}$	s^2
1c	2,1,6,4	3.25	4.917	36s	1,6,8,9	6.00	12.667
2	2,1,4,7	3.50	7.000	37s	1,4,8,11	6.00	19.333
3	2,1,4,8	3.75	9.583	38s	2,6,8,9	6.25	9.583
4	2,1,6,7	4.00	8.667	39s	2,4,8,11	6.25	16.250
5	2,1,4,9	4.00	12.667	40s	1,6,7,11	6.25	16.917
6	2,1,6,8	4.25	10.917	41s	1,4,11,9	6.25	20.917
7	2,1,6,9	4.50	13.667	42	1,7,8,9	6.25	12.917
8	2,1,4,11	4.50	20.333	43cs	6,4,7,8	6.25	2.917
9cs	2,1,7,8	4.50	12.333	44s	2,6,7,11	6.50	13.667
10	1,6,4,7	4.50	7.000	45s	2,4,11,9	6.50	17.667
11s	2,1.7,9	4.75	14.917	46	2,7,8,9	6.50	9.667
12	2,6,4,7	4.75	4.917	47s	1,6,8,11	6.50	17.667
13	1,6,4,8	4.75	8.917	48s	6,4,7,9	6.50	4.333
14	2,1,6,11	5.00	20.667	49s	2,6,8,11	6.75	14.250
15s	2,1,8,9	5.00	16.667	50s	1,6,11,9	6.75	18.917
16	2,6,4,8	5.00	6.667	51	1,7,8,11	6.75	17.583
17	1,6,4,9	5.00	11.337	52s	6,4,8,9	6.75	4.917
18s	1,4,7,8	5.00	10.000	53s	2,6,11,9	7.00	15.333
19s	2,1,7,11	5.25	21.583	54	2,7,8,11	7.00	14.000
20	2,6,4,9	5.25	8.917	55	1,7,11,9	7.00	18.667
21s	2,4,7,8	5.25	7.583	56s	6,4,7,11	7.00	8.667
22s	1,4,7,9	5.25	12.250	57	4,7,8,9	7.00	4.667
23s	2,1,8,11	5.50	23.000	58	2,7,11,9	7.25	14.917
24s	2,4,7,9	5.50	9.667	59	1,8,11,9	7.25	18.917
25	1,6,4,11	5.50	17.667	60s	6,4,8,11	7.25	8.917
26s	1,6,7,8	5.50	9.667	61	2,8,11,9	7.50	15.000
27s	1,4,8,9	5.50	13.667	62cs	6,4,11,9	7.50	9.667
28cs	2,1,11,9	5.75	24.917	63	6,7,8,9	7.50	1.667
29	2,6,4,11	5.75	14.917	64	4,7,8,11	7.50	8.333
30s	2,6,7,8	5.75	6.917	65	4,7,11,9	7.75	8.917
31s	2,4,8,9	5.75	10.917	66	6,7,8,11	8.00	4.667
32s	1,6,7,9	5.75	11.583	67	4,8,11,9	8.00	8.667
33s	1,4,7,11	5.75	18.250	68	6,7,11,9	8.25	4.917
34s	2,6,7,9	6.00	8.667	69	6,8,11,9	8.50	4.333
35s	2,4,7,11	6.00	15.333	70c	7,8,11,9	8.75	2.917

1/ Values of X for the population of eight elements are $X_1 = 2$, $X_2 = 1$, $X_3 = 6$, $X_4 = 4$, $X_5 = 7$, $X_6 = 8$, $X_7 = 11$, $X_8 = 9$; $\bar{X} = 6.00$; and

$$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{N-1} = 12.$$

Table 4.2--Sampling distribution of $\bar{x}$

$\bar{x}$	Relative frequency of $\bar{x}$		
	Simple random sampling Illustration 4.1	Cluster sampling Illustration 4.2	Stratified random sampling Illustration 4.2
3.25	1	1	
3.50	1		
3.75	1		
4.00	2		
4.25	1		
4.50	4	1	1
4.75	3		1
5.00	5		2
5.25	4		3
5.50	5		4
5.75	6	1	5
6.00	4		4
6.25	6	1	5
6.50	5		4
6.75	4		3
7.00	5		2
7.25	3		1
7.50	4	1	1
7.75	1		
8.00	2		
8.25	1		
8.50	1		
8.75	1	1	
Total	70	6	36
Expected value of $\bar{x}$	6.00	6.00	6.00
Variance of $\bar{x}$	1.50	3.29	0.49

$\bar{x}$ is centered on $\bar{X}$. If the theory is correct, the average of $\bar{x}$ for the 70 samples, which are equally likely to occur, should be equal to the population average, 6.00. The average of the 70 samples does equal 6.00.

From the theory of expected values, we also know that the variance of $\bar{x}$ is given by

$$S_{\bar{x}}^2 = \frac{N-n}{N}\frac{S^2}{n} \tag{4.1}$$

where

$$S^2 = \frac{\sum_{i}^{N}(X_i-\bar{X})^2}{N-1}$$

With reference to Illustration 4.1 and Table 4.1, $S^2 = 12.00$ and $S_{\bar{x}}^2 = \frac{8-4}{8}\frac{12}{4} = 1.5$. The formula (4.1) can be verified by computing the variance among the 70 values of $\bar{x}$ as follows:

$$\frac{(3.25-6.00)^2 + (3.50-6.00)^2 + \ldots + (8.75-6.00)^2}{70} = 1.5$$

Since S^2 is a population parameter, it is usually unknown. Fortunately, as discussed in Chapter 3, $E(s^2) = S^2$ where

$$s^2 = \frac{\sum_{i}^{n}(x_i-\bar{x})^2}{n-1}$$

In Table 4.1, the value of s^2 is shown for each of the 70 samples. The average of the 70 values of s^2 is equal to S^2. The fact that $E(s^2) = S^2$ is another important property of simple random samples. In practice s^2 is used as an estimate of S^2. That is,

$$s_{\bar{x}}^2 = \frac{N-n}{N}\frac{s^2}{n}$$

is an unbiased estimate of the variance of $\bar{x}$.

To recapitulate, we have just verified three important properties of simple random samples:

(1) $E(\bar{x}) = \bar{X}$

(2) $S_{\bar{x}} = \sqrt{\frac{N-n}{N}} \quad \frac{S}{\sqrt{n}}$

(3) $E(s^2) = S^2$

The standard error of $\bar{x}$, namely $S_{\bar{x}}$, is a measure of how much $\bar{x}$ varies under repeated sampling from $\bar{X}$. Incidentally, notice that Equation (4.1) shows how the variance of $\bar{x}$ is related to the size of the sample. Now we need to consider the form or shape of the distribution of $\bar{x}$.

Definition 4.1. The distribution of an estimate is often called the sampling distribution. It refers to the distribution of all possible values of an estimate that could occur under a prescribed sampling plan.

4.2 SHAPE OF THE SAMPLING DISTRIBUTION

For random sampling there is a large volume of literature on the distribution of an estimate which we will not attempt to review. In practice, the distribution is generally accepted as being normal (See Figure 4.1) unless the sample size is "small." The theory and empirical tests show that the distribution of an estimate approaches the normal distribution rapidly as the size of the sample increases. The closeness of the distribution of an estimate to the normal distribution depends on: (1) the distribution of X (i.e., the shape of the frequency distribution of the values of X in the population being sampled), (2) the form of the estimator, (3) the sample design, and (4) the sample size. It is not possible to give a few simple, exact guidelines for deciding when the degree of approximation is good enough. In practice, it is generally a matter of working as though the distribution of an estimate is normal but being mindful of the possibility that the distribution might differ

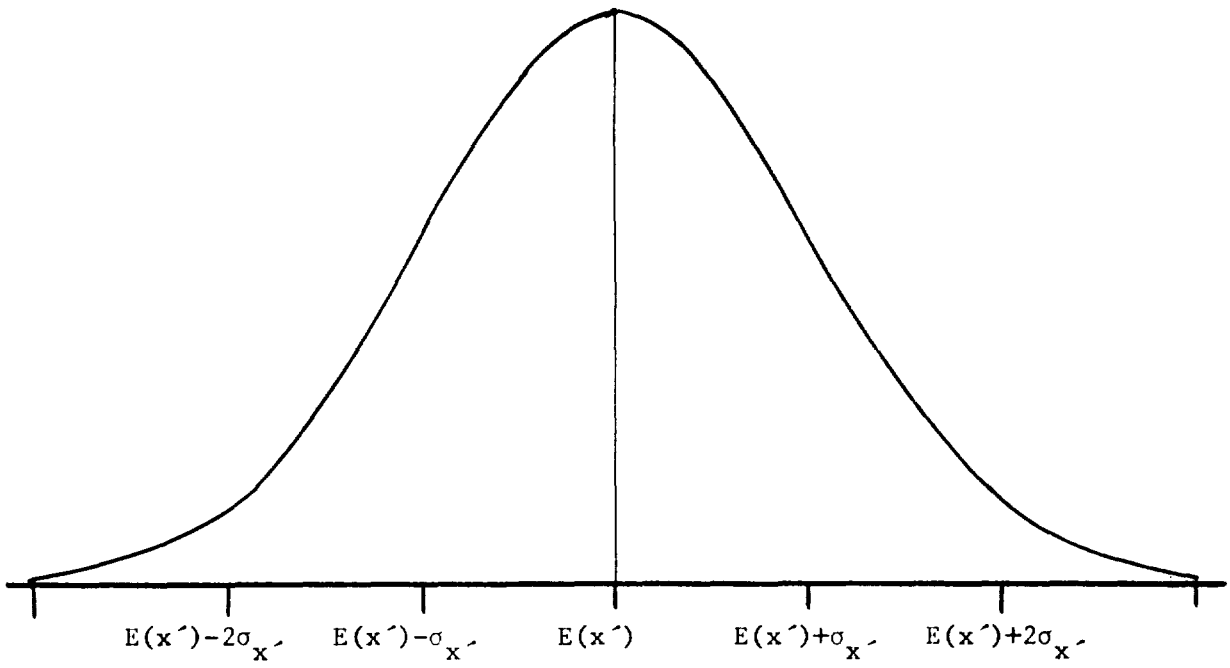


Figure 4.1--Distribution of an estimate (normal distribution)

considerably from normal when the sample is very small and the population distribution is highly skewed. 3/

It is very fortunate that the sampling distribution is approximately normal as it gives a basis for probability statements about the precision of an estimate. As notation, x' will be the general expression for any estimate, and $\sigma_{x'}$ is the standard error of x'.

Figure 4.1 is a graphical representation of the sampling distribution of an estimate. It is the normal distribution. In the mathematical equation for the normal distribution of a variable there are two parameters: the average value of the variable, and the standard error of the variable.

3/ For a good discussion of the distribution of a sample estimate, see Vol. I, Chapter 1, Hansen, Hurwitz, and Madow. Sample Survey Methods and Theory, John Wiley and Sons, 1953.

Suppose x' is an estimate from a probability sample. The characteristics of the sampling distribution of x' are specified by three things: (1) the expected value of x', $E(x')$, which is the mean of the distribution; (2) the standard error of x', $\sigma_{x'}$, and (3) the assumption that the distribution is normal. If x' is normally distributed, two-thirds of the values that x' could equal are between $[E(x') - \sigma_{x'}]$ and $[E(x') + \sigma_{x'}]$, 95 percent of the possible values of x' are between $[E(x') - 2\sigma_{x'}]$ and $[E(x') + 2\sigma_{x'}]$, and 99.7 percent of the estimates are within $3\sigma_{x'}$ from $E(x')$.

Exercise 4.1. With reference to Illustration 4.1, find $E(\bar{x}) - \sigma_{\bar{x}}$ and $E(\bar{x}) + \sigma_{\bar{x}}$. Refer to Table 4.2 and find the proportion of the 70 values of $\bar{x}$ that are between $E(\bar{x}) - \sigma_{\bar{x}}$ and $E(\bar{x}) + \sigma_{\bar{x}}$. How does this compare with the expected proportion assuming the sampling distribution of $\bar{x}$ is normal? The normal approximation is not expected to be close, owing to the small size of the population and of the sample. Also compute $E(\bar{x}) - 2\sigma_{\bar{x}}$ and $E(\bar{x}) + 2\sigma_{\bar{x}}$ and find the proportion of the 70 values of $\bar{x}$ that are between these two limits.

4.3 SAMPLE DESIGN

There are many methods of designing and selecting samples and of making estimates from samples. Each sampling method and estimator has a sampling distribution. Since the sampling distribution is assumed to be normal, alternative methods are compared in terms of $E(x')$ and $\sigma_{x'}$ (or $\sigma^2_{x'}$).

For simple random sampling, we have seen, for a sample of n, that every possible combination of n elements has an equal chance of being the sample selected. Some of these possible combinations (samples) are much better than others. It is possible to introduce restrictions in sampling so some of the combinations cannot occur or so some combinations have a

higher probability of occurrence than others. This can be done without introducing bias in the extimate x' and without losing a basis for estimating $\sigma_{x'}$. Discussion of particular sample designs is not a primary purpose of this chapter. However, a few simple illustrations will be used to introduce the subject of design and to help develop concepts of sampling variation.

Illustration 4.2. Suppose the population of 8 elements used in Table 4.1 is arranged so it consists of four sampling units as follows:

Sampling Unit	Elements	Values of X	Sample Unit Total
1	1,2	$X_1 = 2$, $X_2 = 1$	3
2	3,4	$X_3 = 6$, $X_4 = 4$	10
3	5,6	$X_5 = 7$, $X_6 = 8$	15
4	7,8	$X_7 = 11$, $X_8 = 9$	20

For sampling purposes the population now consists of four sampling units rather than eight elements. If we select a simple random sample of two sampling units from the population of four sampling units, it is clear that the sampling theory for simple random sampling applies. This illustration points out the importance of making a clear distinction between a sampling unit and an element that a measurement pertains to. A sampling unit corresponds to a random selection and it is the variation among sampling units (random selections) that determines the sampling error of an estimate. When the sampling units are composed of more than one element, the sampling is commonly referred to as cluster sampling because the elements in a sampling unit are usually close together geographically.

For a simple random sample of 2 sampling units, the variance of $\bar{x}_c$, where $\bar{x}_c$ is the sample average per sampling unit, is

$$S^2_{\bar{x}_c} = \frac{N-n}{N} \frac{S_c^2}{n} = 13.17$$

where

$$N = 4,\ n = 2,\ \text{and}\ S_c^2 = \frac{(3-12)^2 + (10-12)^2 + (15-12)^2 + (20-12)^2}{3} = \frac{158}{3}$$

Instead of the average per sampling unit one will probably be interested in the average per element, which is $\bar{x} = \frac{\bar{x}_c}{2}$, since there are two elements in each sampling unit. The variance of $\bar{x}$ is one-fourth of the variance of $\bar{x}_c$. Hence, the variance of $\bar{x}$ is $\frac{13.17}{4} = 3.29$.

There are only six possible random samples as follows:

Sample	Sampling Units	Sample average per sampling unit, $\bar{x}_c$	s_c^2
1	1,2	6.5	24.5
2	1,3	9.0	72.0
3	1,4	11.5	144.5
4	2,3	12.5	12.5
5	2,4	15.0	50.0
6	3,4	17.5	12.5

where $s_c^2 = \frac{\sum_i^n (x_i - \bar{x}_c)^2}{n-1}$ and x_i is a sampling unit total. Be sure to notice that s_c^2 (which is the sample estimate of S_c^2) is the variance among sampling units in the sample, not the variance among individual elements in the sample. From the list of six samples, it is easy to verify that $\bar{x}_c$ is an unbiased estimate of the population average per sampling unit and that s_c^2 is an unbiased estimate of $\frac{158}{3}$, the variance among the four sampling

units in the population. Also, the variance among the six values of $\bar{x}$ is 13.17 which agrees with the formula.

The six possible cluster samples are among the 70 samples listed in Table 4.1. Their sample numbers in Table 4.1 are 1, 9, 28, 43, 62, and 70. A "c" follows these sample numbers. The sampling distribution for the six samples is shown in Table 4.2 for comparison with simple random sampling. It is clear from inspection that random selection from these six is less desirable than random selection from the 70. For example, one of the two extreme averages, 3.25 or 8.75, has a probability of $\frac{1}{3}$ of occurring for the cluster sampling and a probability of only $\frac{1}{35}$ when selecting a simple random sample of four elements. In this illustration, the sampling restriction (clustering of elements) increased the sampling variance from 1.5 to 3.29.

It is of importance to note that the average variance among elements within the four clusters is only 1.25. (Students should compute the within cluster variances and verify 1.25). This is much less than 12.00, the variance among the 8 elements of the population. In reality, the variance among elements within clusters is usually less than the variance among all elements in the population, because clusters (sampling units) are usually composed of elements that are close together and elements that are close together usually show a tendency to be alike.

Exercise 4.2. In Illustration 4.2, if the average variance among elements within clusters had been greater than 12.00, the sampling variance for cluster sampling would have been less than the sampling variance for a simple random sample of elements. Repeat what was done in Illustration 4.2

using as sampling units elements 1 and 6, 2 and 5, 3 and 8, and 4 and 7. Study the results.

Illustration 4.3. Perhaps the most common method of sampling is to assign sampling units of a population to groups called strata. A simple random sample is then selected from each stratum. Suppose the population used in Illustration 4.1 is divided into two strata as follows:

Stratum 1 $\quad X_1 = 2,\ X_2 = 1,\ X_3 = 6,\ X_4 = 4$

Stratum 2 $\quad X_5 = 7,\ X_6 = 8,\ X_7 = 11,\ X_8 = 9$

The sampling plan is to select a simple random sample of two elements from each stratum. There are 36 possible samples of 4, two from each stratum. These 36 samples are identified in Table 4.1 by an s after the sample number so you may compare the 36 possible stratified random samples with the 70 simple random samples and with the six cluster samples. Also, see Table 4.2.

Consider the variance of $\bar{x}$. We can write

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2}{2}$$

where $\bar{x}_1$ is the sample average for stratum 1 and $\bar{x}_2$ is the average for stratum 2. According to Theorem 3.5

$$s^2_{\bar{x}} = \left(\frac{1}{4}\right)\left(s^2_{\bar{x}_1} + s^2_{\bar{x}_2} + 2s_{\bar{x}_1\bar{x}_2}\right)$$

We know the covariance, $s_{\bar{x}_1\bar{x}_2}$, is zero because the sampling from one stratum is independent of the sampling from the other stratum. And, since the sample within each stratum is a simple random sample,

$$s^2_{\bar{x}_1} = \frac{N_1 - n_1}{N_1}\,\frac{s_1^2}{n_1} \quad \text{where } s_1^2 = \frac{\sum_i^{N_1}(X_{1i} - \bar{X}_{1\cdot})^2}{N_1 - 1}$$

The subscript "1" refers to stratum 1. $s^2_{\bar{x}_2}$ is of the same form as $s^2_{\bar{x}_1}$. Therefore,

$$s^2_{\bar{x}} = \frac{1}{4}\left[\frac{N_1-n_1}{N_1}\ \frac{s_1^2}{n_1} + \frac{N_2-n_2}{N_2}\ \frac{s_2^2}{n_2}\right]$$

Since

$$\frac{N_1-n_1}{N_1} = \frac{N_2-n_2}{N_2} = \frac{1}{2}, \text{ and } n_1 = n_2 = 2,$$

$$s^2_{\bar{x}} = \frac{1}{8}\left[\frac{s_1^2+s_2^2}{2}\right] = \frac{1}{8}\left[\frac{4.92+2.92}{2}\right] = 0.49$$

The variance, 0.49, is comparable to 1.5 in Illustration 4.1 and to 3.29 in Illustration 4.2.

In Illustration 4.2, the sampling units were groups of two elements and the variance among these groups (sampling units) appeared in the formula for the variance of $\bar{x}$. In Illustration 4.3, each element was a sampling unit but the selection process (randomization) was restricted to taking one stratum (subset) at a time, so the sampling variance was determined by variability within strata. As you study sampling plans, form mental pictures of the variation which the sampling error depends on. With experience and accumulated knowledge of what the patterns of variation in various populations are like, one can become expert in judging the efficiency of alternative sampling plans in relation to specific objectives of a survey.

If the population and the samples in the above illustrations had been larger, the distributions in Table 4.2 would have been approximately normal. Thus, since the form of the distribution of an estimate from a probability sample survey is accepted as being normal, only two attributes of an estimate need to be evaluated, namely its expected value and its variance.

In the above illustrations ideal conditions were implicitly assumed. Such conditions do not exist in the real world so the theory must be extended to fit, more exactly, actual conditions. There are numerous sources of error or variation to be evaluated. The nature of the relationship between theory and practice is a major governing factor determining the rate of progress toward improvement of the accuracy of survey results.

We will now extend error concepts toward more practical settings.

4.4 RESPONSE ERROR

So far, we have discussed sampling under implicit assumptions that measurements are obtained from all n elements in a sample and that the measurement for each element is without error. Neither assumption fits, exactly, the real world. In addition, there are "coverage" errors of various kinds. For example, for a farm survey a farm is defined but application of the definition involves some degree of ambiguity about whether particular enterprises satisfy the definition. Also, two persons might have an interest in the same farm tract giving rise to the possibility that the tract might be counted twice (included as a part of two farms) or omitted entirely.

Partly to emphasize that error in an estimate is more than a matter of sampling, statisticians often classify the numerous sources of error into one of two general classes: (1) Sampling errors which are errors associated with the fact that one has measurements for a sample of elements rather than measurements for all elements in the population, and (2) non-sampling errors--errors that occur whether sampling is involved or not. Mathematical error models can be very complex when they include a term for

each of many sources of error and attempt to represent exactly the real world. However, complicated error models are not always necessary, depending upon the purposes.

For purposes of discussion, two oversimplified response-error models will be used. This will introduce the subject of response error and give some clues regarding the nature of the impact of response error on the distribution of an estimate. For simplicity, we will assume that a measurement is obtained for each element in a random sample and that no ambiguity exists regarding the identity or definition of an element. Thus, we will be considering sampling error and response error simultaneously.

Illustration 4.4. Let $T_1,\ldots,T_N$ be the "true values" of some variable for the N elements of a population. The mention of true values raises numerous questions about what is a true value. For example, what is your true weight? How would you define the true weight of an individual? We will refrain from discussing the problem of defining true values and simply assume that true values do exist according to some practical definition. When an attempt is made to ascertain T_i, some value other than T_i might be obtained. Call the actual value obtained X_i. The difference, $e_i = X_i - T_i$, is the response error for the i^{th} element. If the characteristic, for example, is a person's weight, the observed weight, X_i, for the i^{th} individual depends upon when and how the measurement is taken. However, for simplicity, assume that X_i is always the value obtained regardless of the conditions under which the measurement is taken. In other words, assume that the response error, e_i, is constant for the i^{th} element. In this hypothetical case, we are actually sampling a population set of values $X_1,\ldots,X_N$ instead of a set of true values $T_1,\ldots,T_N$.

Under the conditions as stated, the sampling theory applies exactly to the set of population values $X_1,\ldots,X_N$. If a simple random sample of elements is selected and measurements for all elements in the sample are obtained, then $E(\bar{x}) = \bar{X}$. That is, if the purpose is to estimate $\bar{T} = \frac{\sum_{i}^{N} T_i}{N}$, the estimate is biased unless $\bar{T}$ happens to be equal to $\bar{X}$. The bias is $\bar{X} - \bar{T}$ which is appropriately called "response bias."

Rewrite $e_i = X_i - T_i$ as follows:

$$X_i = T_i + e_i \tag{4.2}$$

Then, the mean of a simple random sample may be expressed as

$$\bar{x} = \frac{\sum_{n} x_i}{n} = \frac{\sum_{n}(t_i+e_i)}{n},$$

or, as $\bar{x} = \bar{t} + \bar{e}$.

From the theory of expected values, we have

$$E(\bar{x}) = E(\bar{t}) + E(\bar{e})$$

Since $E(\bar{x}) = \bar{X}$ and $E(\bar{t}) = \bar{T}$ it follows that

$$\bar{X} = \bar{T} + E(\bar{e})$$

Thus, $\bar{x}$ is a biased estimate of $\bar{T}$ unless $E(\bar{e}) = 0$, where $E(\bar{e}) = \frac{\sum_{i}^{N} e_i}{N}$. That is, $E(\bar{e})$ is the average of the response errors, e_i, for the whole population.

For simple random sampling the variance of $\bar{x}$ is

$$S^2_{\bar{x}} = \frac{N-n}{N}\,\frac{S^2_X}{n} \quad \text{where} \quad S^2_X = \frac{\sum_{i}^{N}(X_i-\bar{X})^2}{N-1}$$

How does the response error affect the variance of X and of $\bar{x}$? We have already written the observed value for the i^{th} element as being equal to

its true value plus a response error, that is, $X_i = T_i + e_i$. Assuming random sampling, T_i and e_i are random variables. We can use Theorem 3.5 from Chapter III and write

$$S_X^2 = S_T^2 + S_e^2 + 2S_{T,e} \tag{4.3}$$

where S_X^2 is the variance of X, S_T^2 is the variance of T, S_e^2 is the response variance (that is, the variance of e), and $S_{T,e}$ is the covariance of T and e. The terms on the right-hand side of Equation (4.3) cannot be evaluated unless data on X_i and T_i are available; however, the equation does show how the response error influences the variance of X and hence of $\bar{x}$.

As a numerical example, assume a population of five elements and the following values for T and X:

	T_i	X_i	e_i
	23	26	3
	13	12	-1
	17	23	6
	25	25	0
	7	9	2
Average	17	19	2

Students may wish to verify the following results, especially the variance of e and the covariance of T and e:

$$S_X^2 = 62.5 \qquad S_T^2 = 54.0 \qquad S_e^2 = 7.5 \qquad S_{T,e} = 0.5$$

As a verification of Equation (4.3) we have

$$62.5 = 54.0 + 7.5 + (2)(0.5)$$

From data in a simple random sample one would compute $s_{\bar{x}}^2 = \frac{\sum_{i}^{n}(x_i-\bar{x})^2}{n-1}$ and use $\frac{N-n}{N}\frac{s_x^2}{n}$ as an estimate of the variance of $\bar{x}$. Is it clear that s_x^2 is an unbiased estimate of S_X^2 rather than of S_T^2 and that the impact of variation in e_i is included in s_x^2 ?

To summarize, response error caused a bias in $\bar{x}$ as an estimate of $\bar{T}$ that was equal to $\bar{X} - \bar{T}$. In addition, it was a source of variation included in the standard error of $\bar{x}$. To evaluate bias and variance attributable to response error, information on X_i and T_i must be available.

Illustration 4.5. In this case, we assume that the response error for a given element is not constant. That is, if an element were measured on several occasions, the observed values for the i^{th} element could vary even though the true value, T_i, remained unchanged. Let the error model be

$$X_{ij} = T_i + e'_{ij}$$

where X_{ij} is the observed value of X for the i^{th} element when the observation is taken on a particular occasion, j,

T_i is the true value of X for the i^{th} element,

and e'_{ij} is the response error for the i^{th} element on a particular occasion, j.

Assume, for any given element, that the response error, e'_{ij}, is a random variable. We can let $e'_{ij} = \bar{e}_i + e_{ij}$, where $\bar{e}_i$ is the average value of e_{ij} for a fixed i, that is, $\bar{e}_i = E(e'_{ij}|i)$. This divides the response error for the i^{th} element into two components: a constant component, $\bar{e}_i$, and a variable component, e_{ij}. By definition, the expected value of e_{ij} is zero for any given element. That is, $E(e_{ij}|i) = 0$.

Substituting $\bar{e}_i + e_{ij}$ for e'_{ij}, the model becomes

$$X_{ij} = T_i + \bar{e}_i + e_{ij} \tag{4.4}$$

The model, Equation (4.4), is now in a good form for comparison with the model in Illustration 4.4. In Equation (4.4), $\bar{e}_i$, like e_i in Equation (4.2) is constant for a given element. Thus, the two models are alike except for the added term, e_{ij}, in Equation (4.4) which allows for the possibility that the response error for the i^{th} element might not be constant.

Assume a simple random sample of n elements and one observation for each element. According to the model, Equation (4.4), we may now write the sample mean as follows:

$$\bar{x} = \frac{\Sigma t_i}{n} + \frac{\Sigma \bar{e}_i}{n} + \frac{\Sigma e_{ij}}{n}$$

Summation with respect to j is not needed as there is only one observation for each element in the sample. Under the conditions specified the expected value of $\bar{x}$ may be expressed as follows:

$$E(\bar{x}) = \bar{T} + \bar{e}$$

where $\bar{T} = \frac{\sum_{i}^{N} T_i}{N}$ and $\bar{e} = \frac{\sum_{i}^{N} \bar{e}_i}{N}$

The variance of $\bar{x}$ is complicated unless some further assumptions are made. Assume that all covariance terms are zero. Also, assume that the conditional variance of e_{ij} is constant for all values of i; that is, let $V(e_{ij} | i) = S_e^2$. Then, the variance of $\bar{x}$ is

$$S_{\bar{x}}^2 = \frac{N-n}{N} \frac{S_T^2}{n} + \frac{N-n}{N} \frac{S_{\bar{e}}^2}{n} + \frac{S_e^2}{n}$$

where $S_T^2 = \frac{\sum_i^N (T_i - \bar{T})^2}{N-1}$, $S_e^2 = \frac{\sum_i^N (\bar{e}_i - \bar{e})^2}{N-1}$,

and S_e^2 is the conditional variance of e_{ij}, that is, $V(e_{ij}|i)$. For this model the variance of $\bar{x}$ does not diminish to zero as $n \to N$. However, assuming N is large, the variance of $\bar{x}$, which becomes $\frac{S_e^2}{N}$, is probably negligible.

Definition 4.2. Mean-Square Error. In terms of the theory of expected values the mean-square error of an estimate, x', is $E(x'-T)^2$ where T is the target value, that is, the value being estimated. From the theory it is easy to show that

$$E(x'-T)^2 = [E(x')-T]^2 + E[x'-E(x')]^2$$

Thus, the mean-square error, mse, can be expressed as follows:

$$mse = B^2 + \sigma_{x'}^2 \tag{4.5}$$

where $$B = E(x') - T \tag{4.6}$$

and $$\sigma_{x'}^2 = E[x'-E(x')]^2 \tag{4.7}$$

Definition 4.3. Bias. In Equation (4.5), B is the bias in x' as an estimate of T.

Definition 4.4. Precision. The precision of an estimate is the standard error of the estimate, namely, $\sigma_{x'}$ in Equation (4.7).

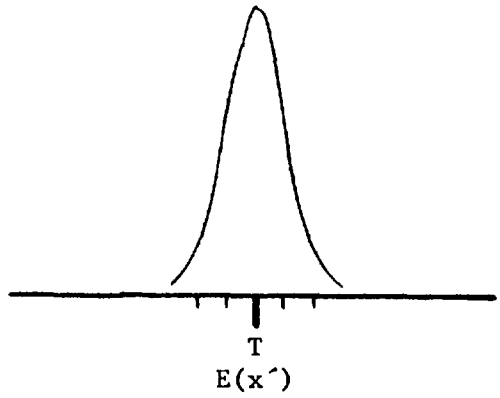
Precision is a measure of repeatability. Conceptually, it is a measure of the dispersion of estimates that would be generated by repetition of the same sampling and estimation procedures many times under the same conditions. With reference to the sampling distribution, it is a measure of the dispersion of the estimates from the center of the distribution and

does not include any indication of where the center of the distribution is in relation to a target.

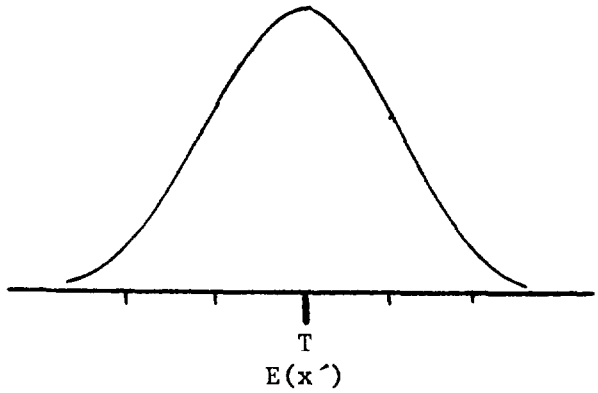
In Illustrations 4.1, 4.2, and 4.3, the target value was implicitly assumed to be $\bar{X}$; that is, T was equal to $\bar{X}$. Therefore, B was zero and the mean-square error of x' was the same as the variance of x'. In Illustrations 4.4 and 4.5 the picture was broadened somewhat by introducing response error and examining, theoretically, the impact of response error on $E(x')$ and $\sigma_{x'}$. In practice many factors have potential for influencing the sampling distribution of x'. That is, the data in a sample are subject to error that might be attributed to several sources.

From sample data an estimate, x', is computed and an estimate of the variance of x' is also computed. How does one interpret the results? In Illustrations 4.4 and 4.5 we found that response error could be divided into bias and variance. The error from any source can, at least conceptually, be divided into bias and variance. An estimate from a sample is subject to the combined influence of bias and variance corresponding to each of the several sources of error. When an estimate of the variance of x' is computed from sample data, the estimate is a combination of variances that might be identified with various sources. Likewise the difference between $E(x')$ and T is a combination of biases that might be identified with various sources.

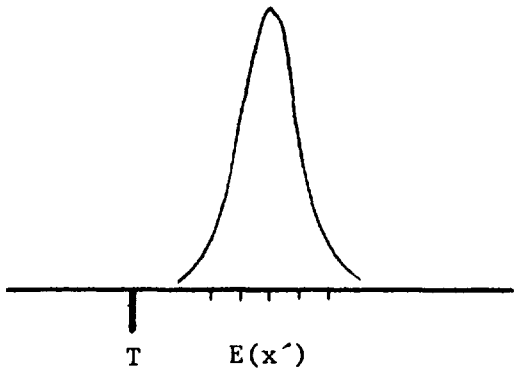
Figure 4.2 illustrates the sampling distribution of x' for four different cases: A, no bias and low standard error; B, no bias and large standard error; C, large bias and low standard error; and D, large bias and large standard error. The accuracy of an estimator is sometimes defined as the square root of the mean-square error of the estimator. According



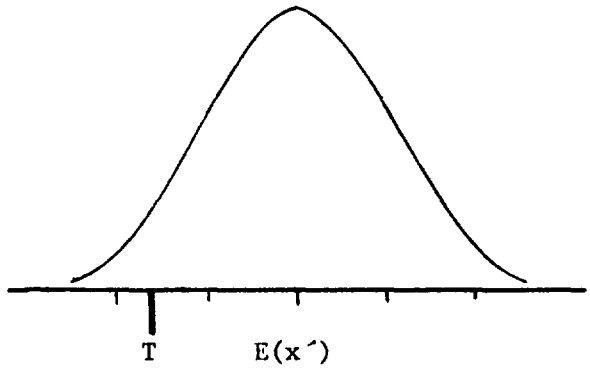
A: No bias--low standard error



B: No bias--large standard error



C: Large bias--low standard error



D: Large bias--large standard error

Figure 4.2--Examples of four sampling distributions

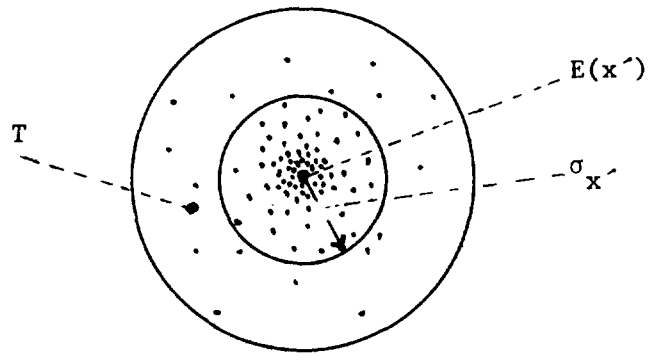


Figure 4.3--Sampling distribution--
Each small dot corresponds to an estimate

to that definition, we could describe estimators having the four sampling distributions in Figure 4.2 as follows: In case A the estimator is precise and accurate; in B the estimator lacks precision and is therefore inaccurate; in C the estimator is precise but inaccurate because of bias, and in D the estimator is inaccurate because of bias and low precision.

Unfortunately, it is generally not possible to determine, exactly, the magnitude of bias in an estimate, or of a particular component of bias. However, evidence of the magnitude of bias is often available from general experience, from knowledge of how well the survey processes were performed, and from special investigations. The author accepts a point of view that the mean-square error is an appropriate concept of accuracy to follow. In that context, the concern becomes a matter of the magnitude of the mse and the size of B relative to $\sigma_{\bar{x}}$. That viewpoint is important because it is not possible to be certain that B is zero. Our goal should be to prepare survey specifications and to conduct survey operations so B is small in relation to $\sigma_{\bar{x}}$. Or, one might say we want the mse to be minimum for a given cost of doing the survey. Ways of getting evidence on the magnitude of bias is a major subject and is outside the scope of this publication.

As indicated in the previous paragraph, it is important to know something about the magnitude of the bias, B, relative to the standard error, $\sigma_{\bar{x}}$. The standard error is controlled primarily by the design of a sample and its size. For many survey populations, as the size of the sample increases, the standard error becomes small relative to the bias. In fact, the bias might be larger than the standard error even for samples of moderate size, for example a few hundred cases, depending upon the circumstances. The point is that if the mean-square error is to be small, both

B and $\sigma_{\bar{x}}$ must be small. The approaches for reducing B are very different from the approaches for reducing $\sigma_{\bar{x}}$. The greater concern about non-sampling error is bias rather than impact on variance. In the design and selection of samples and in the processes of doing the survey an effort is made to prevent biases that are "sampling" in origin. However, in survey work one must be constantly aware of potential biases and on the alert to minimize biases as well as random error (that is, $\sigma_{\bar{x}}$).

The above discussion puts a census in the same light as a sample. Results from both have a mean-square error. Both are surveys with reference to use of results. Uncertain inferences are involved in the use of results from a census as well as from a sample. The only difference is that in a census one attempts to get a measurement for all N elements, but making n = N does not reduce the mse to zero. Indeed, as the sample size increases, there is no positive assurance that the mse will always decrease; because, as the variance component of the mse decreases, the bias component might increase. This can occur especially when the population is large and items on the questionnaire are such that simple, accurate answers are difficult to obtain. For a large sample or a census, compared to a small sample, it might be more difficult to control factors that cause bias. Thus, it is possible for a census to be less accurate (have a larger mse) than a sample wherein the sources of error are more adequately controlled. Much depends upon the kind of information being collected.

4.5 BIAS AND STANDARD ERROR

The words "bias," "biased," and "unbiased" have a wide variety of meaning among various individuals. As a result, much confusion exists,

especially since the terms are often used loosely. Technically, it seems logical to define the bias in an estimate as being equal to B in Equation (4.6), which is the difference between the expected value of an estimate and the target value. But, except for hypothetical cases, numerical values do not exist for either $E(x')$ or the target T. Hence, defining an unbiased estimate as one where $B = E(x') - T = 0$ is of little, if any, practical value unless one is willing to accept the target as being equal to $E(x')$. From a sampling point of view there are conditions that give a rational basis for accepting $E(x')$ as the target. However, regardless of how the target is defined, a good practical interpretation of $E(x')$ is needed.

It has become common practice among survey statisticians to call an estimate unbiased when it is based on methods of sampling and estimation that are "unbiased." For example, in Illustration 4.4, $\bar{x}$ would be referred to as an unbiased estimate--unbiased because the method of sampling and estimation was unbiased. In other words, since $\bar{x}$ was an unbiased estimate of $\bar{X}$, $\bar{x}$ could be interpreted as an unbiased estimate of the result that would have been obtained if all elements in the population had been measured.

In Illustration 4.5 the expected value of $\bar{x}$ is more difficult to describe. Nevertheless, with reference to the method of sampling and estimation, $\bar{x}$ was "unbiased" and could be called an unbiased estimate even though $E(\bar{x})$ is not equal to $\bar{T}$.

The point is that a simple statement which says, "the estimate is unbiased" is incomplete and can be very misleading, especially if one is not familiar with the context and concepts of bias. Calling an estimate unbiased is equivalent to saying the estimate is an unbiased estimate of

its expected value. Regardless of how "bias" is defined or used, $E(x')$ is the mean of the sampling distribution of x; and this concept of $E(x')$ is very important because $E(x')$ appears in the standard error, $\sigma_{x'}$, of x' as well as in B. See Equations (4.6) and (4.7).

As a simple concept or picture of the error of an estimate from a survey, the writer likes the analogy between an estimate and a shot at a target with a gun or an arrow. Think of a survey being replicated many times using the same sampling plan, but a different sample for each replication. Each replication would provide an estimate that corresponds to a shot at a target.

In Figure 4.3, each dot corresponds to an estimate from one of the replicated samples. The center of the cluster of dots is labeled $E(x')$ because it corresponds to the expected value of an estimate. Around the point $E(x')$ a circle is drawn which contains two-thirds of the points. The radius of this circle corresponds to $\sigma_{x'}$, the standard error of the estimate. The outer circle has a radius of two standard errors and contains 95 percent of the points. The target is labeled T. The distance between T and $E(x')$ is bias, which in the figure is greater than the standard error.

In practice, we usually have only one estimate, x', and an estimate, $s_{x'}$, of the standard error of x'. With reference to Figure 4.3, this means one point and an estimate of the radius of the circle around $E(x')$ that would contain two-thirds of the estimates in repeated samplings. We do not know the value of $E(x')$; that is, we do not know where the center of the circles is. However, when we make a statement about the standard error of x', we are expressing a degree of confidence about how close a

particular estimate prepared from a survey is to $E(x')$; that is, how close one of the points in Figure 4.3 probably is to the unknown point $E(x')$. A judgment as to how far $E(x')$ is from T is a matter of how T is defined and assessment of the magnitude of biases associated with various sources of error.

Unfortunately, it is not easy to make a short, rigorous, and complete interpretative statement about the standard error of x'. If the estimated standard error of x' is three percent, one could simply state that fact and not make an interpretation. It does not help much to say, for example, that the odds are about two out of three that the estimate is within three percent of its expected value, because a person familiar with the concepts already understands that and it probably does not help the person who is unfamiliar with the concepts. Suppose one states, "the standard error of x' means the odds are two out of three that the estimate is within three percent of the value that would have been obtained from a census taken under identically the same conditions." That is a good type of statement to make but, when one engages in considerations of the finer points, interpretation of "a census taken under identically the same conditions" is needed--especially since it is not possible to take a census under identically the same conditions.

In summary, think of a survey as a fully defined system or process including all details that could affect an estimate, including: the method of sampling; the method of estimation; the wording of questions; the order of the questions on the questionnaire; interviewing procedures; selection, training, and supervision of interviewers; and editing and processing of

data. Conceptually, the sampling is then replicated many times, holding all specifications and conditions constant. This would generate a sampling distribution as illustrated in Figures 4.2 or 4.3. We need to recognize that a change in any of the survey specifications or conditions, regardless of how trivial the change might seem, has a potential for changing the sampling distribution, especially the expected value of x'. Changes in survey plans, even though the definition of the parameters being estimated remains unchanged, often result in discrepancies that are larger than the random error that can be attributed to sampling.

The points discussed in the latter part of this chapter were included to emphasize that much more than a well designed sample is required to assure accurate results. Good survey planning and management calls for evaluation of errors from all sources and for trying to balance the effort to control error from various sources so the mean-square error will be within acceptable limits as economically as possible.