Chapter 4
Simple random samples and their properties

4.1 INTRODUCTION

A sample is a part drawn from a larger whole. Rarely is there any interest in the sample *per se*; a sample is taken in order to learn something about the whole (the "population") from which it is drawn.

In an opinion poll, for example, a number of persons are interviewed and their opinions on an issue or issues are solicited in order to discover the attitude of the community as a whole, of which the polled persons are usually a small part. The viewing and listening habits of a relatively small number of persons are regularly monitored by ratings services, and, from these observations, projections are made about the preferences of the entire population for available television and radio programs. Large lots of manufactured products are accepted or rejected by purchasing departments in business or government following inspection of a relatively small number of items drawn from these lots. At border stations, customs officers enforce the laws by checking the effects of only a small number of travellers crossing the border. Auditors judge the extent to which the proper accounting procedures have been followed by examining a relatively small number of transactions, selected from a larger number taking place within a period of time. Industrial engineers often check the quality of manufacturing processes not by inspecting every single item produced but through samples selected from these processes. Countless surveys are carried out, regularly or otherwise, by marketing and advertising agencies to determine consumers' expectations, buying intentions, or shopping patterns.

Some of the best known measurements of the economy rely on samples, not on complete enumerations. The weights used in consumer price indexes, for example, are based on the purchases of a sample of urban families as ascertained by family expenditure surveys; the prices of the individual items are averages established through national samples of retail outlets. Unemployment statistics are based on monthly national samples of households. Similar samples regularly survey retail trade, personal incomes, inventories, shipments and outstanding orders of firms, exports, and imports.

In every case, a sample is selected because it is impossible, inconvenient, slow, or uneconomical to enumerate the entire population. Sampling is a method of collecting information which, if properly carried out, can be convenient, fast, economical, and reliable.

4.2 POPULATIONS, RANDOM AND NON-RANDOM SAMPLES

A *population* is the aggregate from which a sample is selected. A population consists of *elements*. For example, the population of interest may be a certain lot of manufactured items stored in a warehouse, all eligible voters in a county, all housewives in a given city, or all the accounts receivable of a certain firm.

A population is examined with respect to one or more attributes or variables. In a particular study, for example, the population of interest may consist of households residing in a metropolitan area. The objective of the study may be to obtain information on the age, income, and level of education of the head of the household, the brands and quantity of each brand of cereal regularly consumed, and the magazines to which the household subscribes.

A sample may be drawn in a number of ways. We shall be primarily concerned with *random samples*, that is, samples in which the selected items are drawn "at random" from the population. In random sampling, the sample elements are selected in much the same way that the winning ticket is drawn in some lotteries, or a hand of cards is dealt: before each draw, the population elements are thoroughly mixed so as to give each element the same chance of being selected. (There are more practical methods for selecting random samples, but more on this later.)

In practice, not all samples selected are random. A sample may be selected only from among those population elements that are easily accessible or conveniently located. For example, a sample from a ship's wheat cargo may be taken from the top layer only; a television reporter may interview the first persons that happen to pass by; or a sample of the city's households may be selected from the telephone directory (thereby ignoring households without a telephone and giving a greater probability of selection to households with more than one listed number). In other cases, a sample may be formed so that, in the judgment of its designer, it is "representative" of the entire population. For example, an interviewer may be instructed to select a "good cross-section" of shoppers, or to ensure that shoppers are selected according to certain "quotas"—such as 50% male and 50% female, or 40% teen and 60% adult.

Since some of these samples may be easier or cheaper to select than random samples, it is natural to ask why the preference is for random samples. Briefly, the principal reason for our interest is that *random samples*

have known desirable properties. We discuss these properties in detail below. Non-random samples, on the other hand, select the population elements with probabilities that are not known in advance. Although, properly interpreted, some of these samples can still provide useful information, the quality of their estimates is simply not known. For example, one intuitively expects that the larger the sample, the more likely it is that the sample estimate is close to the population characteristic of interest. And indeed it can be shown that random samples have this property. There is no guarantee, however, that samples selected by non-random methods will have this or other desirable properties.

The purpose of taking a sample is to learn something about the population from which it is selected. It goes without saying that there is no point in taking a sample if the population and its characteristics are known, or in making estimates when the true population characteristics are available. This appears obvious, yet it is surprising how often this basic principle is overlooked, as a result of a tendency to use elaborate sampling techniques without realizing that the available information describes an entire population and not part of one.

4.3 ESTIMATING POPULATION CHARACTERISTICS

Suppose that a population has N elements of which a proportion π belong to a certain category C, formed according to a certain attribute or variable. There could, of course, be many categories in which we may be interested, but whatever we say about one applies to all. The number of population elements belonging to this category is, obviously, $N\pi$. Since π and $N\pi$ are unknown, a sample is taken for the purpose of obtaining estimates of these characteristics. Suppose, then, that a random sample of n elements is selected, and R is the proportion of elements in the sample that belong to category C. It is reasonable, we suggest, to take R as an *estimate* of π and NR as an estimate of $N\pi$.

Think, for example, of a population of $N = 500,000$ subscribers to a mass-circulation magazine. The magazine, on behalf of its advertisers, would like to know what proportion of subscribers own their home, what proportion rent, and what proportion have other types of accommodation (e.g., living rent-free at parents' home, etc.). Suppose that a sample of $n = 200$ subscribers is selected at random from the list of subscribers. Interviews with the selected subscribers show that 31% own, 58% rent, and 11% have other types of accommodation. It is reasonable to use these numbers as *estimates* of the unknown proportions of *all* subscribers who own, rent, or have other accommodation. A reasonable estimate of the number of subscribers who rent is (500,000)(0.58) or 290,000, the estimate of the number owning is 155,000, and that of the number with other accommodation is 55,000.

Suppose now that with each of the N population elements there is

associated a numerical value of a certain variable X. For example, X could represent the number of cars owned by a subscriber to the magazine. If we knew $X_1, X_2, \ldots, X_N$—the values of X associated with each of the N population elements—the population average value (mean) of X could be calculated as $\mu = (\sum_{i=1}^{N} X_i)/N$. The total value of X (the sum of all X values in the population) could be calculated as $\sum_1^N X_i = N\mu$. We are usually interested in the population means or totals of many variables. As with proportions, however, whatever we say about one variable applies to all.

Invariably, μ and $N\mu$ are unknown. If a random sample is taken, it would be reasonable to estimate the population average by the sample average $\bar{X} = (\sum_{i=1}^{n} X_i)/n$, where $X_1, X_2, \ldots, X_n$ are the X values of the n elements in the sample, and the population total by $N\bar{X}$. For example, suppose that the average number of cars owned by 200 randomly selected subscribers is 1.2. It is reasonable to use this figure as an estimate of the unknown population average. The estimate of the number of cars owned by all subscribers is $(500,000)(1.2)$ or 600,000.

Indeed, it would be reasonable to estimate the population mode of a variable by the sample mode, the population variance by the sample variance, or the population median by the sample median. If the population elements are described by two variables, X and Y, the population correlation coefficient of X and Y can be estimated by the sample correlation coefficient of X and Y. All these population and sample characteristics are calculated in exactly the same manner, but the population characteristics are based on *all* the elements in the population, while the sample characteristics utilize the values of the elements selected *in the sample.*

As noted earlier, of all these population characteristics, the proportion of elements falling into a given category and the mean value of a variable are the most important in practice *and on these we shall concentrate in this and the following chapters.* Numerous estimates of proportions and means are usually made on the basis of a sample. Whatever we say about the estimate of one proportion or mean, however, applies to estimates of all proportions and means. Estimates of totals can easily be formed from those of means or proportions, as was illustrated above.

4.4 NON-RESPONSE, MEASUREMENT ERROR, ILL-TARGETED SAMPLES

Before examining the properties of these estimates, we must note some important restrictions to the results that follow. Throughout this and the next two chapters, *we shall assume that the population of interest is the one from which the sample is actually selected, that the selected population elements can be measured, and that measurement can be made without error.*

By "measuring," we understand determining the true category or value of a variable associated with a population element.

These assumptions are frequently violated in applications. Let us illustrate briefly.

Suppose that a market research survey requires the selection of a sample of households. As is often the case, there is no list of households from which to select the sample. The telephone book provides a tempting and convenient list. Clearly, though, the telephone-book population and the household population are not identical (there are unlisted numbers, households without telephone or with several telephones, non-residential telephone numbers, etc.).

Individuals often refuse to be interviewed or to complete questionnaires. The sample may have been carefully selected, but not all selected elements can be measured. If it can be assumed that the two subpopulations—those who respond and those who do not—have identical characteristics, the problem is solved. But if this is not the case, treating those that respond as a random sample from the entire population may result in misleading estimates.

Measurement error is usually not serious when it is objects that are being measured (although measuring instruments are sometimes inaccurate), but it could be so when dealing with people. For example, we may wish to believe that individuals reveal or report their income accurately, but, often, reported income is at variance with true income, even when participants are assured that their responses are confidential.

There are no simple solutions to these problems, and we shall not discuss them further, so that we can concentrate on other, equally important problems arising even when the assumptions are satisfied. Interested readers will find additional information in texts of marketing research and survey research methods.

4.5 ESTIMATES BASED ON RANDOM SAMPLES WITHOUT REPLACEMENT

The purpose of this section is to establish some properties of the sample proportion and the sample mean as estimators of the population proportion and mean respectively. These properties, summarized in the box which follows, form the basis for a number of useful results: they allow us, for example, to compare different sampling methods, and to determine the size of sample necessary to produce estimates with a desired degree of accuracy.

We shall not attempt to prove these properties, as this is a little difficult. Confirming them, however, by means of simple examples is straightforward. This is our first task and it will occupy us throughout this section. A discussion of the implications of these properties will follow.

We have in mind a population consisting of N elements. We suppose

For random samples of size n drawn *without replacement* from a population of size N, the expected value and variance of the probability distribution of the sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, are:

$$E(\bar{X}) = \mu, \qquad Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}. \tag{4.1}$$

The expected value and variance of the probability distribution of the sample proportion, R, are:

$$E(R) = \pi, \qquad Var(R) = \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1}. \tag{4.2}$$

The probability distribution of the sample frequency, $F = nR$, is hypergeometric with parameters N, n, and $k = N\pi$.

that these elements can be classified into a number of categories according to the variable or attribute of interest. Let C be one of these categories, and let π be the *population proportion of* C—the proportion of elements in the population that belong to C. We suppose further that with each of the population elements there is associated a value of a variable X of interest. The *population mean* of X (denoted by μ) is the average of all the N X values. The *population variance* of X (σ^2) is the average squared deviation of the N values of X about the population mean.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 \tag{4.3}$$

Because the selected elements will vary from one sample to another, so will the sample estimates (R and $\bar{X}$) of π and μ. R and $\bar{X}$ are random variables. Statistical theory has established the characteristics of their probability distributions summarized by Equations (4.1) and (4.2). A simple example will be used to explain and confirm these theoretical results.

Example 4.1 Suppose that every item in a lot of $N = 10$ small manufactured items is inspected for defects in workmanship and manufacture. The inspection results are as follows:

Item:	A	B	C	D	E	F	G	H	I	J
No. of defects:	1	2	0	1	0	1	1	0	1	2

The lot forms the population of this example: 3 items have 0 defects, 5 have 1 defect, and 2 items have 2 defects. The number of defects is the variable X of interest. The population mean of X (the average number of defects per item) is $\mu = 9/10 = 0.9$. The population variance of X (the variance of the number of defects) is

$$\sigma^2 = \frac{1}{10}[(1-0.9)^2 + (2-0.9)^2 + \cdots + (2-0.9)^2] = 0.49.$$

The very same items can also be viewed in a different way, according to whether or not they are "good" (have no defects) or "defective" (have one or more defects). From this point of view, the items in the lot can be classified into two categories: Good and Defective. We shall focus our attention on the second category, which becomes the "typical" category C of this illustration, but whatever we say about this one category applies to any category. Since the lot contains 3 good and 7 defective items, the population proportion of C (the proportion of defective items in the lot) is $\pi = 0.7$.

Let us now consider what will happen if we draw from this lot a random sample of three items in the following manner. First, the items will be thoroughly mixed, one item will be selected at random, and the number of its defects will be noted. Then, the remaining items in the lot will again be thoroughly mixed, a second item will be randomly selected, and the number of its defects noted. The same procedure will be repeated one more time for the third item. If, then, we were to select a sample of three items in this fashion (to select, in other words, a random sample of three items *without replacement*), what are the possible values of the estimators R and $\bar{X}$, and what are the probabilities of their occurrence?

Let X_1 denote the number of defects on the first item selected; also, let X_2 and X_3 represent the number of defects on the second and third selected items respectively. Table 4.1 shows all possible sample outcomes, that is, all possible sets of values of X_1, X_2, and X_3, and the corresponding probabilities. (Columns (6) and (7) will be explained shortly.)

Consider Outcome No. 11 as an example. The probability that the first item drawn will have one defect is $5/10$, since the first item is one of 10 items, 5 of which have one defect. The probability that the second item will have no defects given that the first item had one defect is $3/9$, since 9 items are left after the first draw and 3 of them have no defects. The probability that the third item will have one defect given that the first item had one defect and the second had no defects is $4/8$, since at this stage 8 items are left in the lot, of which 4—one less than the original number—have one defect. Thus, the probability that $X_1 = 1$, $X_2 = 0$, and $X_3 = 1$, is equal to the product $(5/10)(3/9)(4/8)$, or $60/720$. In general, the probability that $X_1 = x_1$, $X_2 = x_2$, and $X_3 = x_3$ in a random sample of size 3 without replacement is

$$p(x_1, x_2, x_3) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2),$$

Table 4.1
Random sampling without replacement;
an illustration

Outcome No.: (1)	X_1 (2)	X_2 (3)	X_3 (4)	$p(X_1, X_2, X_3)$ (5)	$\bar{X}$ (6)	R (7)
1	0	0	0	(3/10) (2/9) (1/8) = 6/720	0/3	0/3
2	0	0	1	(3/10) (2/9) (5/8) = 30/720	1/3	1/3
3	0	0	2	(3/10) (2/9) (2/8) = 12/720	2/3	1/3
4	0	1	0	(3/10) (5/9) (2/8) = 30/720	1/3	1/3
5	0	1	1	(3/10) (5/9) (4/8) = 60/720	2/3	2/3
6	0	1	2	(3/10) (5/9) (2/8) = 30/720	3/3	2/3
7	0	2	0	(3/10) (2/9) (2/8) = 12/720	2/3	1/3
8	0	2	1	(3/10) (2/9) (5/8) = 30/720	3/3	2/3
9	0	2	2	(3/10) (2/9) (1/8) = 6/720	4/3	2/3
10	1	0	0	(5/10) (3/9) (2/8) = 30/720	1/3	1/3
11	1	0	1	(5/10) (3/9) (4/8) = 60/720	2/3	2/3
12	1	0	2	(5/10) (3/9) (2/8) = 30/720	3/3	2/3
13	1	1	0	(5/10) (4/9) (3/8) = 60/720	2/3	2/3
14	1	1	1	(5/10) (4/9) (3/8) = 60/720	3/3	3/3
15	1	1	2	(5/10) (4/9) (2/8) = 40/720	4/3	3/3
16	1	2	0	(5/10) (2/9) (3/8) = 30/720	3/3	2/3
17	1	2	1	(5/10) (2/9) (4/8) = 40/720	4/3	3/3
18	1	2	2	(5/10) (2/9) (1/8) = 10/720	5/3	3/3
19	2	0	0	(2/10) (3/9) (2/8) = 12/720	2/3	1/3
20	2	0	1	(2/10) (3/9) (5/8) = 30/720	3/3	2/3
21	2	0	2	(2/10) (3/9) (1/8) = 6/720	4/3	2/3
22	2	1	0	(2/10) (5/9) (3/8) = 30/720	3/3	2/3
23	2	1	1	(2/10) (5/9) (4/8) = 40/720	4/3	3/3
24	2	1	2	(2/10) (5/9) (1/8) = 10/720	5/3	3/3
25	2	2	0	(2/10) (1/9) (3/8) = 6/720	4/3	2/3
26	2	2	1	(2/10) (1/9) (5/8) = 10/720	5/3	3/3
27	2	2	2	(2/10) (1/9) (0/8) = 0/720	6/3	3/3

where $p(x_2|x_1)$ denotes the probability that $X_2 = x_2$ given that $X_1 = x_1$, and $p(x_3|x_1, x_2)$ denotes the probability that $X_3 = x_3$ given that $X_1 = x_1$ and $X_2 = x_2$. Figure 4.3 shows part of the probability tree (you may wish to draw the complete tree to check the calculation of these probabilities).

Outcome No. 27 is impossible in sampling three items without replacement, since there are only 2 items in the lot having two defects. It can be omitted, or—as done here—included in the list of possible outcomes but

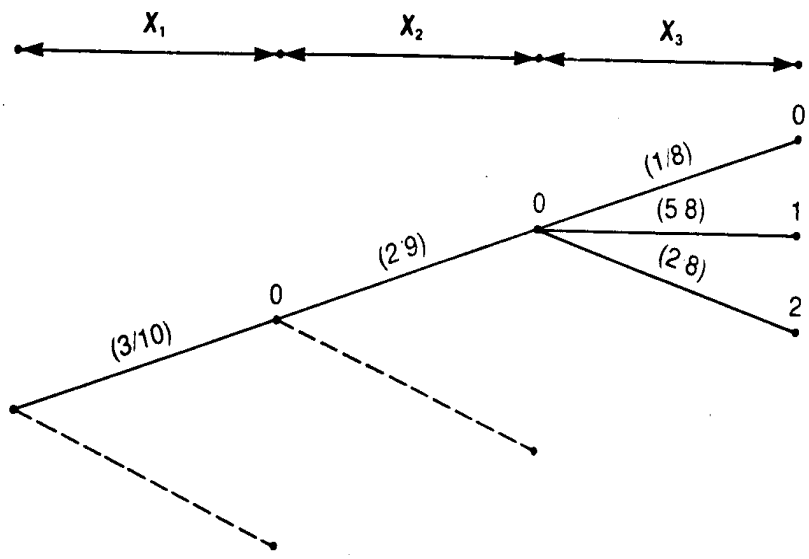

Figure 4.1
Probability tree, Example 4.2

with probability zero.

Let us now construct the probability distribution of the sample estimates of μ and π respectively: $\bar{X}$, the sample mean of X (in this case, the average number of defects per item in the sample); and R, the sample proportion of C (in this case, the proportion of defective items in the sample).

For each sample outcome shown in Table 4.1, there corresponds a value of $\bar{X}$ and one of R. These values are shown in columns (6) and (7) respectively.

Consider Outcome No. 8 as an example. If $X_1 = 0$, $X_2 = 2$, and $X_3 = 1$ (and the probability of this sample outcome is 30/720), the sample mean is $\bar{X} = (0+2+1)/3 = 1$ (column (6)), and the proportion of defective items in the sample is $R = 2/3$ (column (7)).

The probability distributions of the sample estimates can be constructed by grouping identical values and the associated probabilities. For example, the probability that the sample mean equals 1 is the sum of the probabilities of the sample outcomes 6, 8, 12, 14, 16, 20, and 22, that is, 240/720. The probabilities of all other possible values of the sample mean are determined similarly. The probability distribution of the sample mean is shown in Table 4.2.

In exactly the same manner we construct the probability distribution of R, as shown in columns (1) and (3) of Table 4.3. Column (2) will be

Table 4.2
Probability distribution of $\bar{X}$

$\bar{X}$	Probability
0	6/720
1/3	90/720
2/3	216/720
1	240/720
4/3	138/720
5/3	30/720
Total	720/720

Table 4.3
Probability distributions of R and F

R (1)	$F = nR$ (2)	Probability (3)
0	0	$6/720 = 0.008$
1/3	1	$126/720 = 0.175$
2/3	2	$378/720 = 0.525$
1	3	$210/720 = 0.292$
	Total	$720/720 = 1.000$

explained shortly.

We are now ready to confirm the properties summarized by Equations (4.1) and (4.2). Recall that for this example $N = 10$, $n = 3$, $\mu = 0.9$, $\sigma^2 = 0.49$, and $\pi = 0.7$.

Referring to Table 4.2, the mean of the probability distribution of $\bar{X}$ is

$$\begin{aligned} E(\bar{X}) &= (0)(6/720) + (1/3)(90/720) + \cdots + (5/3)(30/720) \\ &= 1944/2160 \\ &= 0.9. \end{aligned}$$

The variance of the distribution of $\bar{X}$ is

$$\begin{aligned} Var(\bar{X}) &= (0)^2(6/720) + (1/3)^2(90/720) + \cdots + (5/3)^2(30/720) - (0.9)^2 \\ &= (6072/6480) - 0.81 \\ &= 0.127037. \end{aligned}$$

According to (4.1), $E(\bar{X})$ should equal $\mu = 0.9$, and

$$Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{0.49}{3}\frac{10-3}{10-1} = 0.127037.$$

Our calculations, therefore, confirm the theoretical predictions.

The mean and variance of R in this example are

$$\begin{aligned} E(R) &= (0)(6/720) + (1/3)(126/720) + \cdots + (1)(210/720) \\ &= 0.7, \\ Var(R) &= (0)^2(6/720) + (1/3)^2(126/720) + \cdots + (1)^2(210/720) - (0.7)^2 \\ &= (3528/6480) - (0.49) \\ &= 0.054444. \end{aligned}$$

According to (4.2), $E(R)$ should equal $\pi = 0.7$, and

$$Var(R) = \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1} = \frac{(0.7)(1-0.7)}{3}\frac{10-3}{10-1} = 0.054444.$$

Again the calculations confirm the theory.

For every value of R there corresponds a value of *the sample frequency (F)* of C, the number of elements in the sample that belong to C—in this example, the number of defective items in the sample. Obviously, $F = nR$, and the probability distribution of F can be obtained easily from that of R; it is shown in columns (2) and (3) of Table 4.3.

In Section 2.8 of Chapter 2, we established that *the probability distribution of F is hypergeometric with parameters N, n, and $k = N\pi$.* (The sample frequency was denoted by W rather than F, but the meaning is the same.) This result can be confirmed again with the present example. The possible values of F are indeed 0, 1, 2, and 3, and the probabilities of these values shown in Table 4.3 match those listed in Appendix 4J for $N = 10$, $n = 3$, and $k = 7$ (check the notes in Appendix 4J concerning the interchange of k and n).

4.6 SAMPLING FROM LARGE POPULATIONS OR WITH REPLACEMENT

When the population size is large, $N \approx N - 1$, and the last term of the variances of $\bar{X}$ and R given in Equations (4.1) and (4.2) can be written as

$$\frac{N-n}{N-1} \approx \frac{N-n}{N} = 1 - \frac{n}{N}.$$

Therefore, *for large populations and small sample to population size ratios (n/N), the term $(N-n)/(N-1)$ is approximately equal to 1 and can be ignored.* In such cases, the variances of $\bar{X}$ and R do not depend on N.

Sampling with replacement differs from sampling without replacement essentially in that the population remains unchanged throughout the selection of the sample. Sampling without replacement from very large populations has approximately the same feature. To illustrate, suppose that the proportions of items with 0, 1, and 2 defects in a lot are 0.3, 0.5, and 0.2 as in the previous illustration, but that the lot size, N, is 100,000 instead of 10. The probability that in a random sample of size 3 with replacement $X_1 = 1$, $X_2 = 2$, and $X_3 = 0$ is (0.5)(0.2)(0.3) or 0.03. The probability of this outcome in a random sample of the same size without replacement from a lot of 100,000 is (50,000/100,000)(20,000/99,999)(30,000/99,998), which is obviously approximately 0.03. The same approximate equality holds for the probabilities of all other outcomes and consequently for the probability distributions of all estimators. Therefore, *the mean and variance of $\bar{X}$ and R for samples with replacement are given by Equations (4.1) and (4.2) with $(N - n)/(N - 1)$ replaced by 1.* That is, the variance of the probability distribution of $\bar{X}$ in samples with replacement is σ^2/n, while that of R is $\pi(1 - \pi)/n$.

It follows that *for large N and small n/N the hypergeometric distribution of F may be approximated by the binomial with parameters n and π.* (This can also be confirmed by an independent argument; see Section 4.9 of this chapter.)

4.7 IMPLICATIONS

Let us now examine some implications of the properties illustrated and confirmed in the preceding sections. Once again, recall that the purpose of taking a sample is to obtain estimates of population characteristics; we agreed that it is reasonable to use $\bar{X}$ and R as estimators of μ and π respectively.*

The first implication appears rather obvious and innocuous, but is often overlooked in practice. Because the sample estimates depend on the selected elements, *there can be no guarantee that an estimate based on the sample actually selected will be equal to the population characteristic.* Unless the sample is without replacement and its size equals that of the population, the sample estimate may or may not be close to the population characteristic—and we can never tell with certainty. Look at Table 4.2 by way of illustration. The true population mean is 0.9. The estimates range from 0 to 1.667. Now, in this illustration the population is known, and the probability distributions of the estimators can be determined. In reality, of course, the population is unknown. All we observe are the sample elements,

* A word on terminology: an *estimator* of a population characteristic is a sample characteristic used to estimate the population characteristic; an *estimate* is the numerical value of an estimator. We have not always made that distinction in the past, but will try to be consistent from now on.

on the basis of which the estimates are calculated. We have no means of telling *with certainty* just how close the estimates are to the population characteristics.

Equations (4.1) and (4.2) show that the *expected* values of $\bar{X}$ and R equal μ and π respectively. The average, in other words, of a large number of estimates equals the population characteristic being estimated. While this provides little comfort in the case of a single sample, it is valuable to know that a particular estimator in the long run and on the average will neither underestimate nor overestimate, but will equal the population characteristic of interest. We shall say that an estimator is an *unbiased estimator* of a given population characteristic *if its expected value equals that population characteristic*. Thus, $\bar{X}$ is an unbiased estimator of μ, and R is an unbiased estimator of π. Other things being equal, an unbiased estimator is preferable to a biased one.

Perhaps an analogy will help clarify the concept of unbiasedness. Imagine someone tossing a ring at a peg located some distance away, trying to get as close to the peg as possible (for simplicity, imagine that the tosses land on a straight line to the peg and do not deviate to the right or left of this line). The patterns (distributions) of tosses of two persons, A and B, are shown in Figure 4.4.

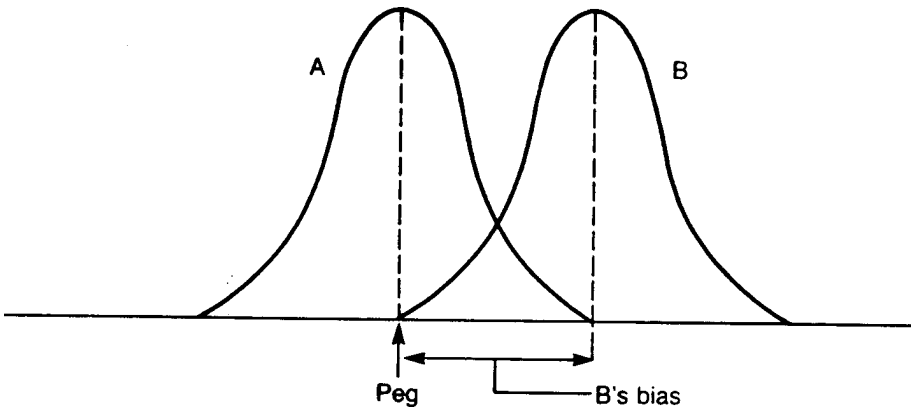

Figure 4.2
Patterns of landings, case A

The two patterns are assumed to be identical except for their mean location. A's mean location is the peg—the target. A may be called an unbiased marksman, B a biased one. B's bias is the distance between the target and the mean location of the tosses. Of the two, A may be considered to have better aim.

Consider now the variances of $\bar{X}$ and R given by Equations (4.1) and (4.2). Let us begin with that of $\bar{X}$, which can be written as

$$Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{\sigma^2}{N-1}\frac{N-n}{n} = (\frac{\sigma^2}{N-1})(\frac{N}{n} - 1).$$

The first term does not depend on the sample size; the second term becomes smaller and smaller as n approaches N, and equals 0 when $n = N$. In other words, the distribution of $\bar{X}$ tends to become more and more concentrated around μ as the sample size increases. The probability, therefore, that $\bar{X}$ will be within a certain range around μ (say, from $\mu - c$ to $\mu + c$, where c is some number) tends to 1 as n approaches N. Figure 4.5 illustrates this.

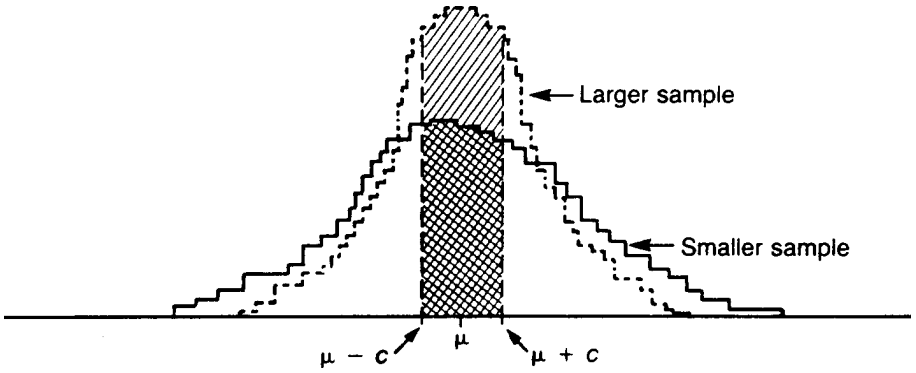

Figure 4.3
$Pr(\mu - c \leq \bar{X} \leq \mu + c)$ tends to increase as $n \to N$

The same conclusion holds true for $Var(R)$; it, too, decreases as n approaches N, and the probability that R will be within a specified range around π—no matter how narrow the range—tends to 1 as n approaches N.

An estimator is said to be a *consistent estimator* of a population characteristic if the probability of its being within a given range of the population characteristic approaches 1 as n approaches N. It can be shown that an unbiased estimator is consistent if its variance tends to 0 as n approaches N. Consistency is clearly a desirable property: it is useful to know that the variability of an estimator decreases and that the probability of being "close" to the population characteristic increases as the sample size increases.

Let us return to the example of the two ring tossers, and imagine that the distributions of their tosses are as shown in Figure 4.6. We assume that both A and B are "unbiased," that is, the mean location of their tosses is the

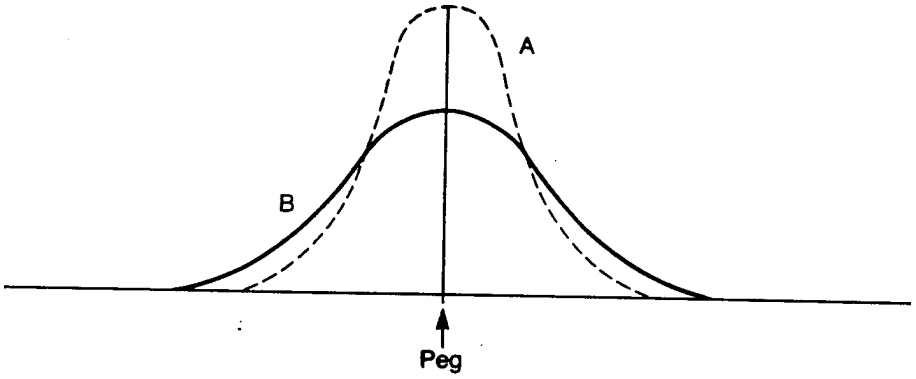

Figure 4.4
Patterns of landings, case B

target. However, A's tosses tend to cluster more tightly around the target than B's. In this case, A may be considered to have better aim.

The moral of the story is that *among unbiased estimators the one having the smallest variance should be preferred.* This principle will be used more than once. Its first application will be to show that, other things being equal, *sampling without replacement is preferable to sampling with replacement.*

Estimators based on samples with replacement are unbiased and consistent (as $n \to \infty$), as are those based on samples without replacement. It was noted earlier, however, that the variances of $\bar{X}$ and R based on samples without replacement differ from those with replacement by the term $(N-n)/(N-1)$. For large N,

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

which is less than 1. Therefore, the variance of $\bar{X}$ in a random sample of size n without replacement is smaller than that in a random sample of the same size selected with replacement; the same applies to $Var(R)$. In other words, for the same sample size, estimators based on sampling without replacement tend to vary less around the population characteristic than those based on sampling with replacement. Clearly, then, sampling without replacement should be preferred over sampling with replacement.*

* Although it is clear that among unbiased estimators the one with the lowest variance is preferable, it is not clear how to choose between an unbiased estimator with a large variance and one that has lower variance but is biased.

This result should be intuitively appealing. In sampling with replacement, every selected element is replaced in the population before the next element is drawn. However, it would appear that once a population element is drawn and inspected, all the information it can provide about the population is revealed. Replacing it, and thus risking that it might be drawn again, seems unnecessary and somewhat wasteful. (In an extreme case, a sample with replacement of size n could consist of the same element that happened to be drawn n times.) Statistical theory confirms this preference for sampling without replacement.*

These, then, are some of the desirable properties of random samples referred to earlier. Remember, random sampling is not the only possible method of selecting a sample. For example, we could always select the first n listed population elements; or we could select the first n listed elements meeting certain criteria. Unbiased and consistent estimators based on random samples can be formed. Other sampling methods may not (and frequently cannot) produce estimators with these properties.

Random sampling has other advantages, too: for instance, it is possible to determine how large a sample should be taken so as to obtain estimates with a desired degree of accuracy. The determination of appropriate sample sizes is discussed later in this chapter.

4.8 SELECTING A RANDOM SAMPLE IN PRACTICE

In Example 4.2, the population consisted of ten small manufactured items. In order to select a random sample, the items were first thoroughly mixed in their container before the first item was selected. Next, the remaining items were again thoroughly mixed and the second item was selected. The procedure was repeated one more time in order to select the third item.

---

* We note in passing that the sample variance,

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

is not an unbiased estimator of the population variance, σ^2. It can be shown that $E(S^2) = \frac{n-1}{n}\sigma^2$, so that S^2 tends to underestimate σ^2. An unbiased estimator is

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

This is often calculated in place of S^2 by computer programs. For the large samples used in business, however, the difference between S^2 and $\hat{S}^2$ as well as the bias of the former may be overlooked.

In general, random sampling requires that, in every draw, each eligible population element be given equal probability of selection. In sampling without replacement, eligible are the population elements that were not selected in earlier draws; in sampling with replacement, of course, all population elements are eligible in every draw. Clearly, the physical randomization of the population elements illustrated above satisfies this requirement.

Instead of mixing the population elements themselves—an impractical procedure in many cases—we could mix *substitutes* of the elements. For example, to select a random sample of people participating at a conference, we could prepare an equal number of identical tags, each marked with one participant's name or other identification, and then select a random sample of tags.

Physically mixing the population elements or their substitutes may not be the most efficient selection method when the population is large. If a list of the population elements is available, the selection of a random sample may be more easily accomplished with the use of a random device or of random numbers.

To illustrate, suppose that a population consists of 345 elements. These elements are listed in sequence and identified by the numbers 001, 002, ... , 345. Now suppose that we mark ten identical chips with the numbers 0 to 9, and put them into a hat. To select an element from the population, we draw three chips at random from the hat, *replacing them after each draw.* The first chip drawn will establish the first digit, the second and third draws will establish the second and third digits of the identification number. For example, if the chips marked 2, 1, and 8 are drawn, population Element No. 218 is selected. If the three-digit identification number formed is 000 or greater than 345, it is ignored. If the sample is without replacement, previously formed identification numbers are also ignored. A little thought will convince the reader that this procedure gives each eligible population element the same probability of selection.

Random numbers achieve the same objective but eliminate the need to mark and mix chips. The numbers shown in Table 4.4 can be thought of *as if* generated by repeated draws of chips marked 0, 1, 2, ... , 9 from a hat. They can be read in any orientation (horizontally, vertically, diagonally, etc.), in any direction (forward, backward, etc.), beginning with any number and proceeding one by one, or using every second, third, etc. number encountered. (There is no need, of course, to be overly elaborate: an ordinary reading is easiest and perfectly adequate.) The random numbers can be combined to form identification numbers of any number of digits.

For example, to select a random sample of five three-digit numbers between 001 and 345 without replacement, we may begin with the second row of Table 4.4 and form the following consecutive three-digit numbers: 420, 065, 289, 435, 537, 817, 382, 284, 214, 455, 121. Ignoring numbers

Table 4.4
Random numbers

0	4	0	4	6	8	4	5	2	0	3	0	2	1	9	1	5	9	6	7	8	0	2	4	3	5	7	7	0	6
4	2	0	0	6	5	2	8	9	4	3	5	5	3	7	8	1	7	3	8	2	2	8	4	2	1	4	4	5	5
1	2	1	3	7	7	0	1	6	5	6	7	9	2	2	4	0	1	3	2	3	6	9	0	3	6	9	7	6	4
4	9	4	2	6	9	1	1	1	5	7	4	2	2	5	2	6	2	2	2	6	4	2	1	7	1	4	1	8	5
7	8	0	6	4	8	7	5	6	6	4	3	9	5	9	5	1	2	0	2	5	8	6	4	3	4	1	3	2	7
7	2	1	4	9	6	8	4	4	5	1	0	4	4	7	3	5	3	6	4	3	8	6	7	0	6	2	6	2	1
6	5	4	9	1	1	9	6	2	6	0	0	0	8	1	2	5	1	5	2	0	8	1	5	6	0	1	3	6	5
7	1	0	0	8	1	6	7	4	6	5	4	4	7	9	8	8	2	6	4	8	1	6	7	6	2	6	0	5	3
6	0	4	0	6	3	3	4	9	4	6	3	7	7	8	2	5	0	9	4	7	6	1	9	4	4	8	0	2	3
9	6	6	7	9	7	3	7	7	1	0	6	1	9	5	8	3	1	0	0	5	9	8	5	9	4	4	7	2	6
4	8	9	3	5	8	5	0	8	4	9	0	8	4	9	2	3	9	1	5	4	5	6	7	8	7	6	2	4	8
3	9	6	0	4	2	5	1	5	3	6	4	8	7	6	0	9	5	1	1	9	9	8	9	9	4	9	3	1	5
3	6	3	3	8	7	2	1	3	8	6	0	6	9	1	9	5	4	0	5	5	8	1	7	7	0	8	2	5	0
8	2	5	4	6	8	8	1	3	6	4	8	8	7	4	2	1	7	9	7	8	2	9	6	6	1	2	0	8	8
2	4	3	4	3	7	9	3	6	9	3	9	3	9	2	9	9	9	4	0	7	2	0	0	9	2	0	2	9	6
1	6	7	3	9	7	2	7	0	9	4	4	7	8	2	6	0	4	3	3	1	7	5	2	0	8	7	9	2	6
6	4	1	1	0	2	7	1	8	7	3	7	3	2	1	4	5	7	1	9	4	9	3	7	7	9	1	9	5	4
2	5	5	1	5	1	5	7	5	1	0	2	1	0	2	5	0	1	0	3	4	0	7	9	1	0	9	9	1	0
3	3	3	6	7	1	0	6	3	1	2	6	9	1	7	7	6	3	1	1	1	0	5	0	1	4	5	3	4	5
4	3	8	4	5	8	2	2	3	5	6	6	1	0	3	2	8	1	7	7	3	4	8	9	6	2	1	2	2	5
1	3	0	3	8	9	1	3	2	3	9	5	4	8	9	2	9	2	6	7	6	4	8	4	0	8	8	5	7	3
8	3	6	5	2	5	7	5	0	6	4	0	7	5	7	6	7	2	0	9	2	4	5	3	1	5	9	5	2	2
7	7	5	4	1	0	8	7	0	9	1	4	3	0	0	4	1	9	5	4	9	4	9	8	6	1	1	2	0	8
5	1	6	0	8	0	6	5	0	7	3	3	3	8	8	7	1	5	9	1	3	0	1	0	5	8	2	5	9	1
9	9	1	7	7	7	9	4	8	0	9	5	2	5	2	8	0	4	6	3	6	2	9	5	9	4	7	9	0	3
3	4	5	4	2	1	4	5	8	6	0	9	3	3	7	9	8	1	5	6	9	4	2	3	8	7	9	4	0	4
5	2	3	4	1	5	2	3	2	5	5	4	1	5	1	3	5	9	3	8	6	3	8	1	6	1	5	7	4	2
3	5	7	6	0	5	9	3	2	5	1	8	3	2	0	5	1	0	2	6	2	7	4	6	9	1	6	7	7	6
8	4	6	5	2	0	7	8	9	2	2	4	4	0	1	6	9	2	0	1	4	2	2	4	7	1	8	2	8	4
5	9	3	5	5	6	5	1	1	3	1	5	8	5	8	7	7	3	3	2	8	2	7	0	5	0	0	8	9	2
8	6	2	0	6	5	6	6	5	0	7	9	9	9	1	8	7	3	4	1	2	5	3	0	4	5	5	6	8	7
8	9	5	7	4	1	7	9	7	9	0	5	4	5	8	8	0	3	2	6	9	0	9	2	5	1	7	5	6	2
1	3	4	1	5	2	4	4	0	3	5	4	2	7	1	6	6	5	6	4	1	5	0	9	9	3	4	6	6	6
4	7	7	1	1	8	2	3	4	4	6	2	4	8	2	0	3	1	3	2	3	4	0	1	2	1	7	0	0	4
3	5	3	0	7	2	2	7	2	5	2	5	5	1	5	5	1	5	0	7	6	6	5	0	6	4	9	2	5	1
8	0	1	3	2	4	6	6	7	3	5	1	6	1	6	7	1	9	6	3	4	2	7	3	5	6	4	1	4	7
8	1	1	2	2	1	5	0	6	5	9	0	4	7	3	0	7	5	3	8	9	3	1	8	1	0	3	4	5	7
1	1	0	1	1	2	9	4	6	6	7	2	7	9	1	0	0	6	0	1	8	5	6	2	7	4	5	5	6	0
3	5	7	1	3	8	2	2	1	7	7	9	0	3	9	3	6	8	7	0	0	8	0	5	8	5	3	9	3	2

greater than 345, the sample consists of the elements numbered 065, 289, 284, 214, 121.

Needless to say, the numbers in Table 4.4 were not generated by drawing chips from a hat. Most computer languages provide a *random number generator*—a routine for generating any number of random numbers. These routines are usually carefully tested to ensure that the digits 0, 1, . . . , 9 they generate behave like the draws of chips from a hat, i.e., that they appear with equal relative frequency (1/10) in the long run and are independent of one another. One such routine was used for the random numbers shown in Table 4.4.

Selection of a random sample, it will be noted, *requires a list* identifying the elements of the population (unless, of course, the elements are such that physical randomization is possible). What, then, is one to do if such a list is not available? How, for instance, is a marketing research firm to select a random sample of adults from the population of all adults in the country, when, obviously, no comprehensive list of adults is maintained anywhere?

Actually, lists of the type needed in marketing research are not scarce. For example, lists of households sorted by neighborhood, street, and street number are available for many metropolitan areas. They are produced and maintained by firms specializing in just this service.

In some cases, it may be reasonable to *assume* that a certain method of selecting a sample is, for all practical purposes, random. For example, for a study of air travellers, a research firm had its interviewers approach every 20th person passing through the security gates of the airport during the week in which the study was conducted. It can be argued that because the order and arrival time of air travellers at the airport is determined by the interaction of innumerable factors, the above selection is as random as any. Unless, then, there are reasons for suspecting that the selection of population elements results in biased estimators, such "essentially random" methods provide a reasonable substitute for strictly random ones.

When a list is unavailable, an "essentially random" method impossible, or alternative random sampling impractical, then—sadly—one must proceed without a sample, using—cautiously—whatever information is available. This is far wiser and safer than the lamentable but all-too-frequent practice of treating any sample as random, and, without questioning its origin, producing from it conclusions which only a random sample justifies.

4.9 SAMPLING FROM AN INDEPENDENT RANDOM PROCESS

A sample, once more, is a part of a whole, and is taken in order to obtain estimates of certain features of the whole. In the situations described so far, the whole is a finite population of elements, from which some are drawn without replacement. There are situations, however, notably in manufac-

turing, in which different definitions of "part" and "whole" are found to be useful.

Imagine a manufacturing process producing a certain product one item at a time—for example, a metal stamping machine continuously producing metal discs of a given size. Each item must meet certain specifications—for example, the diameter of each disc must not exceed 1.211 cm or be less than 1.209 cm. An item is called Good if it meets the specifications, and Defective if it does not.

Now, it may be reasonable to *assume* that, at a given time, the probabilities that an item will be defective and good do not vary from one item to another. It may also be reasonable to *further assume* that the quality of an item (that is, whether it is good or defective) does not depend on the quality of any other item. A manufacturing process satisfying these assumptions is an *independent process* with two outcomes.

Not all manufacturing processes satisfy these assumptions. The first assumption is violated, for example, when the equipment gradually wears out, so that the probability that, say, the 10,000th item will be defective is greater than that of the 10th. The second assumption is not satisfied when the outcomes tend to follow a pattern. For example, suppose that defectives—whenever they occur—occur in clusters of two:

$$\ldots GGG\underbrace{DD}GG\underbrace{DD}GGGGGG\underbrace{DD}GG\ldots$$

The probability of a defective following a defective is different from that of a defective following a good item; the quality of the "next" item depends on that of "this" item, and the assumption of independence is violated.

Suppose, however, that a manufacturing process does satisfy the above assumptions. Let π be the probability that an item will be defective, and $1-\pi$ that it will be good. π can be interpreted as the expected proportion of defectives in a very large run of items that *could be* produced under the current conditions.

Consider now selecting a sample of *any* n items from this process. These items could be selected one after the other as they are produced, or at a constant interval (say, every 10th item, or every 5 minutes), or at varying intervals—it does not matter how.

It may be intuitively clear that *a sample from this process has the same properties as a random sample with replacement from a lot of N items of which a proportion π are defective.*

If this is not clear, consider two simple cases: (a) a random sample of $n = 2$ items with replacement from a lot containing a proportion π of defectives; and (b) a sample of $n = 2$ items from an independent process producing in the long run a proportion π of defectives. As illustrated by the probability tree of Figure 4.7, in either case, the possible sample outcomes

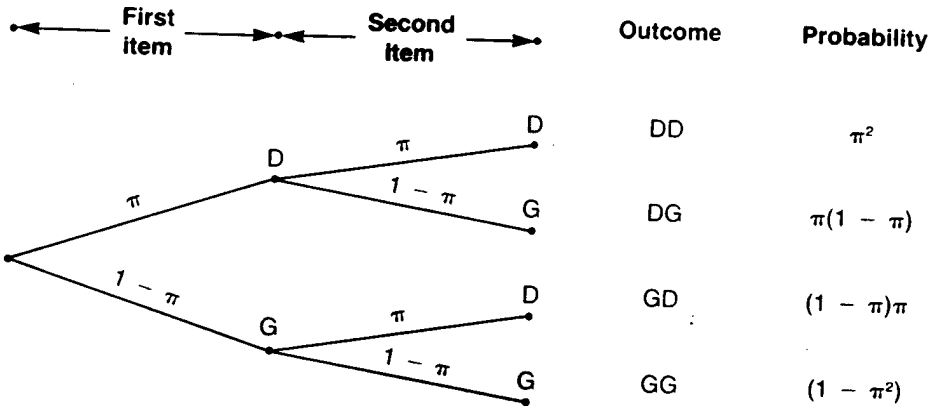

Figure 4.5
Sampling with replacement or from independent process

are DD, DG, GD, and GG, and their probabilities π^2, $\pi(1-\pi)$, $\pi(1-\pi)$, and $(1-\pi)^2$ respectively.

It follows that the proportion of defectives (R) in a sample of n items from an independent process has the same mean and variance as that of the sample proportion in a random sample with replacement, that is, $E(R) = \pi$ and $Var(R) = \pi(1-\pi)/n$. R, therefore, is an unbiased estimator of the proportion of defectives in the long run. In addition, the number of defectives ($F = nR$) in a sample of size n from an independent process has a *binomial* distribution with parameters n and $p = \pi$.

All these can be verified in the simple example of Figure 4.7. The probability distribution of the number of defectives in a sample of 2 items is shown in columns (1) and (2) of Table 4.5.

Table 4.5
Sample of $n = 2$ items
from independent process

F	Probability	R
(1)	(2)	(3)
0	$(1-\pi)^2$	0.0
1	$2\pi(1-\pi)$	0.5
2	π^2	1.0
	1.0	

The binomial distribution with parameters $n = 2$ and π is generated

by:

$$p(x) = \frac{2!}{x!(2-x)!}\pi^x(1-\pi)^{2-x},$$

for x =0, 1, and 2. Calculation of $p(0)$, $p(1)$, and $p(2)$ produces the entries in column (2) of Table 4.5. The probability distribution of the proportion of defectives in the sample, R, is given in columns (3) and (2), from which we calculate:

$$E(R) = (0)(1-\pi)^2 + (0.5)[2\pi(1-\pi)] + (1)(\pi^2) = \pi,$$

$$\begin{aligned} Var(R) &= (0)^2(1-\pi)^2 + (0.5)^2[2\pi(1-\pi)] + (1)^2(\pi^2) - \pi^2 \\ &= \pi(1-\pi)/2, \end{aligned}$$

exactly as claimed.

Clearly, the two outcomes of an independent process could be any two mutually exclusive and collectively exhaustive categories—not just Good and Defective. If C is the category of interest, π the probability that an item will fall into C, and R the proportion of the n items in the sample that belong to C, then $E(R) = \pi$, $Var(R) = \pi(1-\pi)/n$, and the probability distribution of $F = nR$ is binomial with parameters n and π.

Not all manufacturing processes, evidently, are of the two-outcome variety. Imagine, for example, a machine that fills bottles with a liquid. Empty bottles are moved on a conveyor belt to a nozzle, one bottle at a time. The nozzle descends, and the liquid begins to flow into the bottle. A sensing device detects when the liquid reaches the nozzle, trips a circuit, and closes the nozzle to stop the flow. The nozzle retreats, the filled bottle is moved away, and its place under the nozzle is taken by the next bottle. The cycle begins anew, in a smooth and continuous process. Each bottle is supposed to contain a certain volume of liquid, but the actual content varies somewhat due to imperfections in the machine and the liquid material. It is the varying volume that characterizes this filling process.

In general, consider a manufacturing process producing items with measurement X (representing length, width, volume, hardness, temperature, number of defects, etc.). *Assume* that, at a given time, the process satisfies two conditions: (a) the probability distribution of an item's measurement is the same for all items; and (b) an item's measurement is independent of that of any other item. A process satisfying these conditions will be called an *independent process* (since the two-outcome process is a special case, there is no need to add the qualifier "with multiple outcomes").

These assumptions, it must be emphasized, do not characterize every process. As noted in the case of a two-outcome process, equipment wear may cause a gradual change in the measurement of a unit, and the presence of any pattern would violate the assumption of independence (for example,

if, whenever one bottle contains more than the nominal volume of liquid, the next one tends to have less).

Let $p(x)$ be the common probability distribution of any item's measurement, $\mu = E(X)$ the mean, and $\sigma^2 = Var(X)$ the variance of this distribution. Let $X_1, X_2, \ldots, X_n$ be the measurements of n items selected from an independent process *in any manner whatever.*

It should be clear, by analogy with the simpler two-outcome special case, that this sample *has the same properties as a random sample of size n selected with replacement from a finite population in which the variable X is distributed according to $p(x)$.* In particular, the expected value of the sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, equals μ, and its variance equals σ^2/n.

These results are summarized in the box that follows.

For samples of size n *drawn in any manner from an independent process* or *at random and with replacement from a finite population*, the expected value and variance of the probability distribution of the sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, are:

$$E(\bar{X}) = \mu, \qquad Var(\bar{X}) = \frac{\sigma^2}{n}. \tag{4.4}$$

The expected value and variance of the probability distribution of the sample proportion, R, are:

$$E(R) = \pi, \qquad Var(R) = \frac{\pi(1-\pi)}{n}. \tag{4.5}$$

The probability distribution of the sample frequency, $F = nR$, is binomial with parameters n and π.

The concept of a random process is useful also in areas other than manufacturing. For example, a driver's record with an insurance company in a period of time (say, one year) may be considered a random process with two outcomes: driver makes a claim, driver does not make a claim. If the process is assumed to be an independent one, and π is the probability of a claim in any one period, then the probability distribution of the number of claims in n periods is binomial with parameters n and π.

The daily closing price of a stock on the exchange could be viewed as the measurement of a random process, and successive daily closing prices as a sample from the process. But does such a process satisfy the two

conditions of an independent process? This question has received a great deal of attention in the literature of finance. The assumption that closing prices have the same probability distribution is generally accepted, provided that the period of time spanned by these prices is reasonably short. It is the assumption of independence that has generated the greatest controversy. For it is an essential tenet of some "technical analysts" (or "chartists," as some dedicated observers of stock prices are known) that stock prices tend to *follow a pattern*, and that any pattern can be exploited.

To illustrate, suppose that an increase in the daily closing price of a stock tends to be followed by a decline on the next day, and vice versa. If one were to buy at the beginning of the day following a price decline, and sell at the end of the day, one would tend to make a profit. This is a simple—and unrealistic—pattern, and a simple trading strategy exploiting it. More complex patterns require commensurately complicated strategies.

It would take too long to describe the complicated patterns often perceived by technical analysts; in fact, the patterns are sometimes so complex that the assistance of a technical analyst is neccessary to detect them. If a study of the history of a stock's price shows that a pattern exists, that pattern can be exploited. The manner in which a given pattern may be exploited could also be complicated, requiring the further assistance of a technical analyst. The empirical evidence of the last 30 years, however, suggests stock prices behave, for all practical purposes, like an independent process. The price of a stock at a given point in time, in other words, is unrelated to its price at any other time. The study of the price history of the stock, therefore, ought to be of no value in attempting to predict any future price.

4.10 LARGE-SAMPLE DISTRIBUTIONS, CENTRAL LIMIT THEOREM

The observant reader must have noticed the absence of any statement in the results presented so far concerning the form of the probability distribution of the sample mean, $\bar{X}$. Whereas the distribution of F is binomial or hypergeometric, depending on whether sampling is with or without replacement, and that of R can be obtained from the distribution of F, nothing was said about the form of the distribution of $\bar{X}$.

The reason is that nothing precise can be said in general about this distribution. If the sample is with replacement or from an independent process *and* the distribution of X has one of a few mathematically tractable forms, then the distribution of the sample mean is predictable. For example, if the distribution of X is normal with mean μ and variance σ^2, it can be shown that the distribution of $\bar{X}$ is also normal with the same mean $E(\bar{X}) = \mu$ and variance $Var(\bar{X}) = \sigma^2/n$. If the distribution of X is Poisson, with parameter m, then it can be shown that the distribution of $n\bar{X} =$

$X_1 + X_2 + \cdots + X_n$ is also Poisson with parameter nm. But such elegant and simple results do not extend to other cases.

In the literature of statistics, however, there is an old and venerable result which practitioners often invoke to fill the gap. Roughly speaking, this result—one of several versions of the so-called *central limit theorem*—states that *when the sample size is large, the probability distribution of the sample mean $\bar{X}$ of a random sample with replacement or of a sample from an independent process is approximately normal with parameters μ and σ^2/n.* Other versions of the theorem establish that, *under the same conditions, the probability distributions of F and R are approximately normal with parameters equal to the mean and variance of the variable.*

In the case of F and R, the theorem provides a convenient approximation to their known distributions. (Recall that the probability distribution of F is binomial, while the distribution of R can be simply obtained from that of F.) In the case of $\bar{X}$, the theorem often serves as the only basis for assessing its unknown distribution.

The central limit theorem, it will be noted, does not indicate at which sample size the approximation becomes satisfactory. Indeed, the theorem is proven only for *infinitely large* samples. Experiments show that sometimes the approximation is remarkably good for samples as small as 5 or 10, while at other times even samples of size 1,000 are insufficient for a satisfactory approximation.

As a general rule, the more the distribution of the variable X itself in the population or process resembles the normal (i.e., bell-shaped and symmetric), or the closer the population or process proportion (π) is to 0.5, the better the normal approximations for a sample of a given size.

Figure 4.8 shows the probability distribution of R for random samples of size $n = 10$ and $n = 20$ with replacement when $\pi = 0.5$. Superimposed is the normal approximation. The approximation is clearly very good even for a sample as small as 10.

Figure 4.9 shows the probability distribution of the sample mean for samples of size $n = 100$, $n = 500$, and $n = 1,000$ drawn with replacement from a very skew population distribution. The sample size must be greater than 1,000 for the normal approximation to be as good as in Figure 4.8.

Sampling without replacement from a finite population adds another element of complexity, yet it can be shown that the same results hold as in sampling with replacement, *provided that both the sample size n and the population size N are large, and $n < N$.* Unfortunately, it is not possible to state simply and in advance precisely just how large n and N must be for the approximations to be satisfactory. Certainly, if the population runs into the hundreds of thousands or millions, and the sample size into the hundreds or thousands, the conditions are well satisfied. National opinion polls and television ratings typically rely on samples of about 1,000 to 3,000

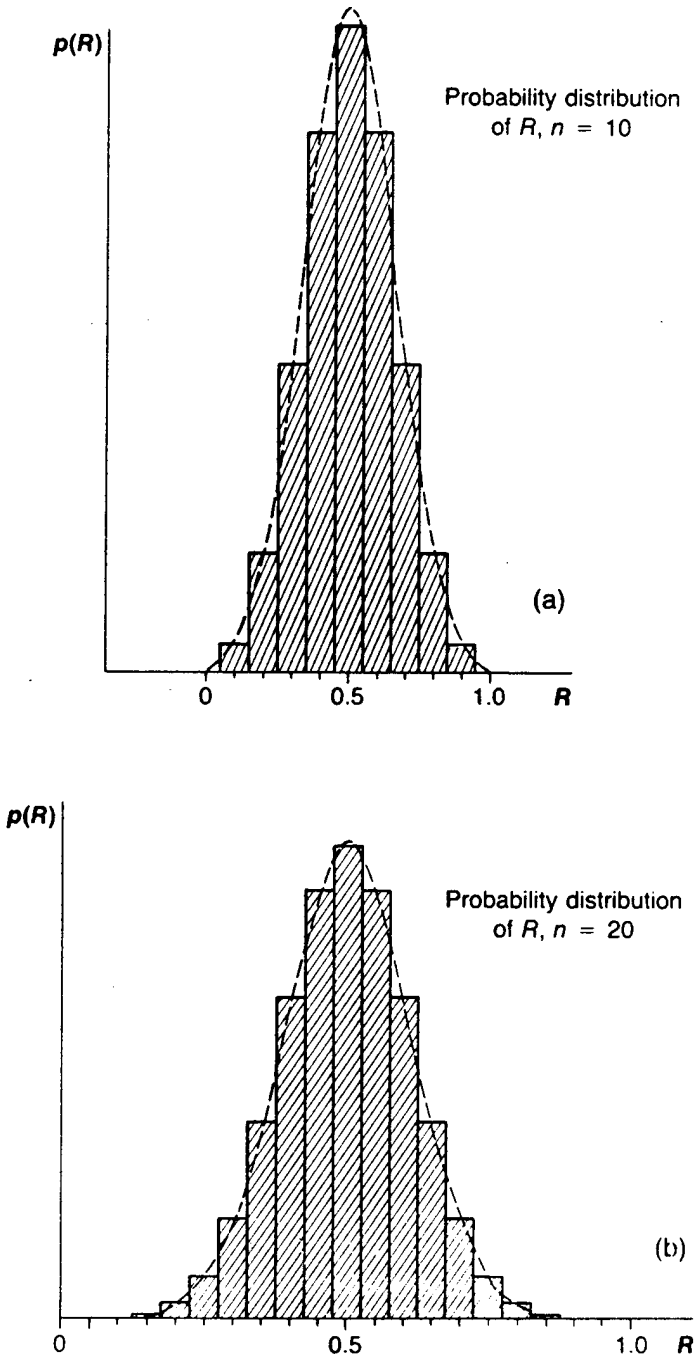

Figure 4.6
Probability distributions of R

from the nation's population of voters and households (these samples, however, are not the *simple* random samples of this chapter, but are selected by methods to be described later). Marketing research studies, though more modest in size, usually involve large samples from large populations. Assuming symmetry and proportions close to 0.5 (the conditions ensuring good approximation for a sample with replacement), the normal approximation is probably satisfactory in a sample as small as 100, drawn without replacement from a population of 200 or more.

From now on, and purely as a rule of thumb, whenever we speak of "large" N and n, we shall understand $n < N$, $N \geq 200$ and $n \geq 100$.

4.11 HOW LARGE SHOULD A SAMPLE BE?

Let us note at the very beginning that in many cases in business, the size of the sample is dictated by the budget. Since—as intuition correctly tells us—the accuracy of a random sample increases with its size, the practical solution to the sample size problem is often simply to select as many observations as can be afforded.

In most situations, however, the size of the sample can be controlled. We may decide to take a small, relatively inexpensive sample, or a larger, more expensive, and more accurate sample. In theory at least, the optimum sample size is that at which the best balance occurs between the conflicting goals of accuracy and economy.

In what follows, we present essentially two methods for determining the required sample size. The first method results in a formula that is quite easy to apply, but requires certain conditions to be observed. The second is not subject to any conditions, but requires calculations best left to a computer.

We begin by assuming that the purpose of the sample is to estimate the proportion (relative frequency, π) of elements in the population that belong to a given category.

From Equations (4.2) we know that its estimator, the relative frequency of the given category in the planned sample (R), can be regarded as a random variable with mean equal to π, and variance $Var(R)$ which decreases as the sample size increases. The variance, as should be well known by now, measures the variability of a random variable around the mean of its distribution. The mean of the distribution of R is π. Therefore, the larger the sample, the smaller is the variability of R around π, and the greater tends to be the probability that R will be within any given interval around π (say, $\pi \pm c$, that is, from $\pi - c$ to $\pi + c$, where c is some given number; see Figure 4.5). This probability approaches 1 as the sample size approaches the population size.

A measure of the *accuracy* of the sample is the probability that R will be within $\pm c$ of π (that is, in the interval from $\pi - c$ to $\pi + c$). The question therefore is: How large must the sample size be so that the probability

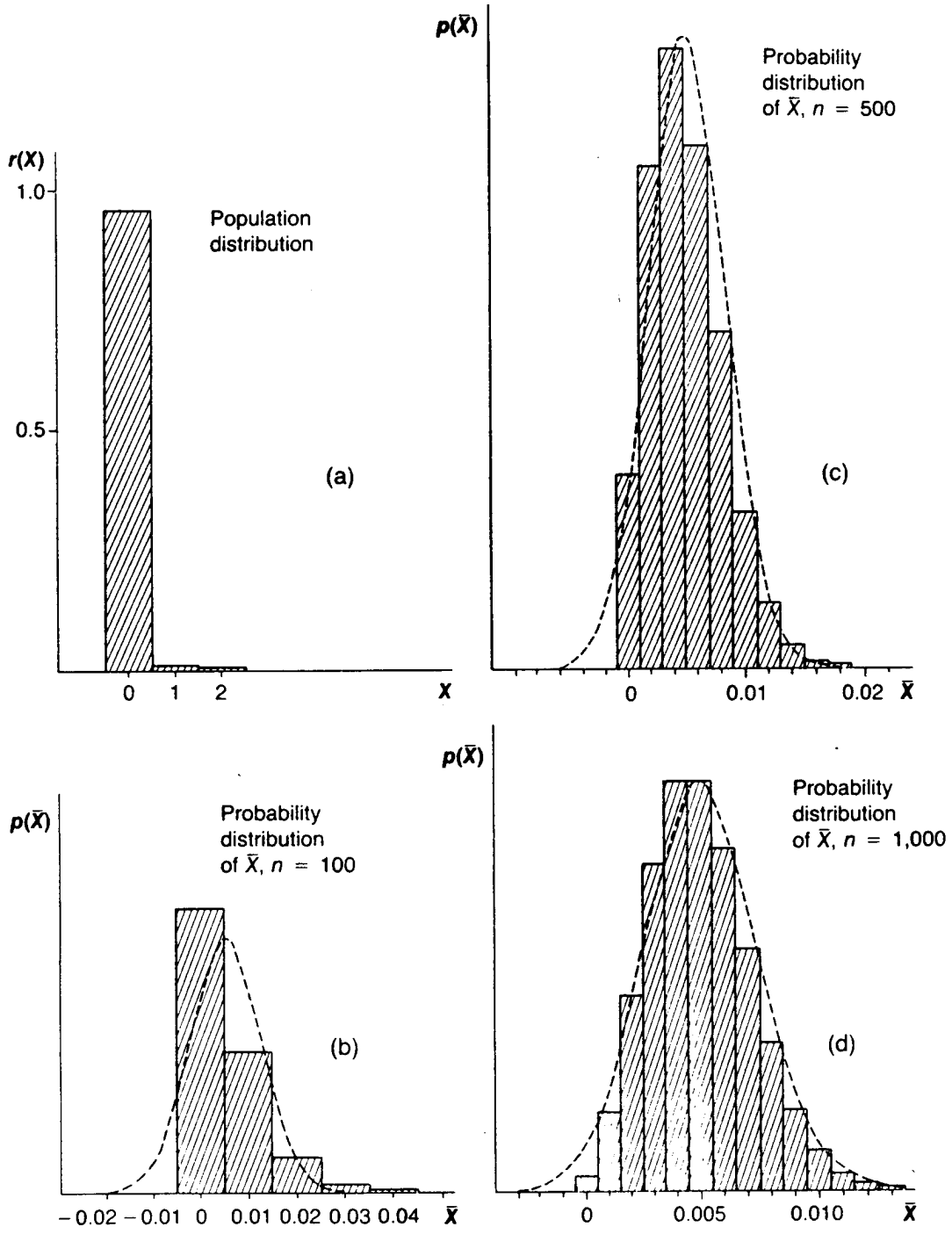

Figure 4.7
Probability distributions of $\bar{X}$

that R will be within $\pm c$ of π is at least $1-\alpha$? Both c and $1-\alpha$ are given numbers determining the desired accuracy. For a specific problem, therefore, the question may be: How large must n be so that the probability is at least 90% that R will be within ± 0.01 of π, whatever the true value of π happens to be? In this case, $c = 0.01$ and $1-\alpha = 0.90$.

Provided that the accuracy requirements are suitably demanding (this too will be explained soon), the required sample size can be calculated by Equations (4.6) and (4.7) in the box that follows.

The size of a sample without replacement required to estimate a population relative frequency (π) within $\pm c$ with probability $(1-\alpha)$ is approximately:

$$n_2 = \frac{n_1}{1 + (n_1/N)}, \tag{4.6}$$

where

$$n_1 = \left(\frac{U_{\alpha/2}}{c}\right)^2 \pi(1-\pi). \tag{4.7}$$

n_1 gives approximately the required size of a sample with replacement or from an independent process.

Equations (4.6) and (4.7) require that N and n_2 (in samples without replacement), or n_1 (in samples with replacement or from an independent process) be large. π is the unknown population proportion and must be replaced by an estimate or by $\pi = 0.5$ (we explain this below). $U_{\alpha/2}$ is a number such that the probability of the standard normal variable U exceeding that value equals $\alpha/2$, i.e., $Pr(U > U_{\alpha/2}) = \alpha/2$. Values of $U_{\alpha/2}$ corresponding to selected values of $1-\alpha$ are given in Table 4.6.

Table 4.6
$U_{\alpha/2}$ for selected $1-\alpha$

$1-\alpha$	$U_{\alpha/2}$	$1-\alpha$	$U_{\alpha/2}$
0.99	2.576	0.80	1.282
0.95	1.960	0.60	0.842
0.90	1.645	0.50	0.674

Before explaining the derivation of these formulae, let us illustrate their application.

Example 4.2 How large a random sample without replacement should be taken of a district's $N = 50,000$ households so that the estimate (R) of the proportion of households buying a given product is within ± 0.01 of the population proportion (π) with probability 95%?

In this case, $1 - \alpha = 0.95$, $U_{\alpha/2} = 1.96$, and $c = 0.01$. A survey taken more than two years ago indicated the product was used by 40% of the households at that time. Using 0.40 as an estimate of π, the required size of a sample with replacement is

$$n_1 = \left(\frac{1.96}{0.01}\right)^2 (0.4)(1 - 0.4) = 9220.$$

If the sample is without replacement, the sample should be of size

$$n_2 = \frac{9220}{1 + (9220/50000)} = 7784.$$

Note that n_1, n_2, and N are all large.

The same accuracy requirements are met by a sample of size 9,220 with replacement or one of size 7,784 without replacement—another demonstration (if one was needed) of the superiority of samples without replacement over samples with replacement.

We would now like to explain the derivation of Equations (4.6) and (4.7).

¶ According to the central limit theorem, for large N and n, the probability distribution of R is approximately normal with mean and variance given by Equations (4.2). To keep the notation simple, let σ_R denote the standard deviation of R. Therefore, the distribution of the ratio $(R - \pi)/\sigma_R$ is standard normal. Let $U_{\alpha/2}$ represent a number such that the probability that a standard normal random variable (U) will exceed this number is $\alpha/2$. (See Figure 4.10.) By the symmetry of the normal distribution, we have $Pr(-U_{\alpha/2} \leq U \leq U_{\alpha/2}) = 1 - \alpha$.

Substituting $(R - \pi)/\sigma_R$ for U, we get

$$Pr(-U_{\alpha/2} \leq \frac{R - \pi}{\sigma_R} \leq U_{\alpha/2}) = 1 - \alpha.$$

Manipulating the inequality in brackets, we get successively:

$$\begin{aligned} 1 - \alpha &= Pr(-\sigma_R U_{\alpha/2} \leq R - \pi \leq \sigma_R U_{\alpha/2}) \\ &= Pr(\pi - \sigma_R U_{\alpha/2} \leq R \leq \pi + \sigma_R U_{\alpha/2}). \end{aligned}$$

Therefore, if we set $c = \sigma_R U_{\alpha/2}$, the probability that R will be within $\pm c$ of π is indeed $1 - \alpha$. σ_R is a function of n. The problem is then to find that value of n

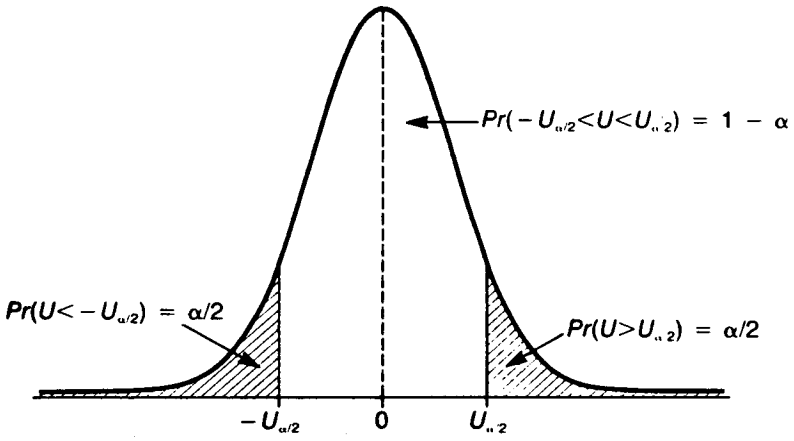

Figure 4.8
Standard normal distribution

which makes c equal to $\sigma_R U_{\alpha/2}$. Replacing σ_R by its equal from Equation (4.2), we have

$$c = U_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}\frac{N-n}{N-1}}.$$

Squaring both sides and rearranging terms, we get

$$\frac{N-n}{n(N-1)} = \frac{c^2}{U^2_{\alpha/2}\pi(1-\pi)}.$$

For large N, $N-1 \approx N$. Therefore, the above expression can be written as

$$\frac{nN}{N-n} = \left(\frac{U_{\alpha/2}}{c}\right)^2 \pi(1-\pi) = n_1.$$

It follows that $nN = n_1(N-n)$, or $n(N+n_1) = n_1 N$, or, finally,

$$n = \frac{n_1 N}{N+n_1} = \frac{n_1}{1+(n_1/N)} = n_2.$$

In other words, if a random sample of size n_2 is taken, the probability is indeed $1-\alpha$ that R will be within $\pm c$ of π.

For a sample with replacement, we take N to be very large, in which case $n_1/N \approx 0$, and $n_2 = n_1$. The explanation is complete. ¶

The required sample size depends on $\pi(1-\pi)$. The following table

shows the value of $\pi(1-\pi)$ for selected values of π:

π	$\pi(1-\pi)$
0.1 or 0.9	0.09
0.2 or 0.8	0.16
0.3 or 0.7	0.21
0.4 or 0.6	0.24
0.5	0.25

Notice that the term $\pi(1-\pi)$ is maximized when $\pi = 0.5$. Other things being equal, therefore, *the required sample size is greatest when* $\pi = 0.5$. π, of course, is unknown. If there is no information at all concerning π, it could be assumed conservatively that $\pi = 0.5$ and the required sample size would be calculated under this assumption; with a sample of that size, the probability is *at least* $1-\alpha$ that R will be within $\pm c$ of π no matter what the true value of π happens to be.

If available information suggests that π is no closer—in either direction—to 0.5 than π_0, then π_0 should be used in place of 0.5 in the calculations. For example, if it is felt that the population proportion cannot possibly exceed 20%, $\pi = 0.20$ should be used; if it is felt that this proportion cannot possibly be less than 60%, likewise use $\pi = 0.60$.

If the sample is planned in order to estimate several π's—a very frequent case in practice—and the sample size is calculated using $\pi = 0.5$ (under the assumption that some π is likely to be 0.5 or close to that number), it is clear that the probability is *at least* $1-\alpha$ that *any* R will be within $\pm c$ of the corresponding π. This procedure also aplies to the special case in which we wish to estimate the distribution of a variable or attribute in the population.

A final comment. Equations (4.6) and (4.7) were developed using the normal approximation to the probability distribution of R—an approximation that requires both n and N to be large. This seems like a circular argument, for it is n that we wish to determine using this approximation. What is actually assumed is that the accuracy requirements (as specified by c and $1-\alpha$) are suitably high so that the n_1 and n_2 calculated using (4.6) and (4.7) are large enough to justify the normal approximation. If the sample size calculated for given c and $1-\alpha$ is small (less than 100 by our rule of thumb), this indicates not only low accuracy requirements, but also an inappropriate application of the formulae.

If the primary objective of the sample is to estimate the mean or total of a variable, Equations (4.8) and (4.9), shown in the box, may be applied.

These expressions also require that n_1 *or* n_2 *and* N *be large.* σ^2 is the population variance of X and must be replaced by an estimate.

The sample size required for the estimate ($\bar{X}$) of the population mean (μ) of a variable X to be within $\pm c$ of μ (that is, in the interval from $\mu - c$ to $\mu + c$) with probability $1 - \alpha$ in a sample from an independent process or a random sample with replacement is approximately

$$n_1 = \left(\frac{U_{\alpha/2}}{c}\right)^2 \sigma^2. \tag{4.8}$$

For a sample without replacement it is approximately

$$n_2 = \frac{n_1}{1 + (n_1/N)}, \tag{4.9}$$

where n_1 is given by Equation (4.8).

The proof relies on the central limit theorem and is similar to that presented earlier in this section.

Just as an estimate of π is necessary to calculate the required sample size for proportions, so an estimate of the population variance (σ^2) of the variable is needed for the application of the last formulae.

Example 4.3 (Suggested by a case study in W. E. Deming, *Some Theory of Sampling*, Wiley, 1950.) A telephone company wants to ascertain the condition of telephone poles and the cost of any required repairs in the region it services. There are altogether 10,000 poles, a list of which is maintained by the company. From this list, a random sample of 100 poles was selected without replacement. Crews were sent out to examine the condition of the poles selected, and to calculate the cost of needed repairs. The results of this inspection were as follows: sample average cost of needed repairs, $\bar{X} = \$83$; sample variance of repair costs, $S^2 = 121$. How many additional poles must be sampled if the estimate of the total cost of repairing all telephone poles (formed by pooling the observations in the pilot and the planned sample) is to be within $\pm\$5{,}000$ of the true total cost with probability 90%?

Here, $N = 10,000$, $1 - \alpha = 0.90$, $U_{\alpha/2} = 1.645$. For the estimate of the total cost to be within $\pm\$5{,}000$ of the true value, the estimate of the average cost must be within $\pm(5{,}000/10{,}000)$ or 0.50 of the population average. Using the variance of the pilot sample as an estimate of the population variance and $c = 0.50$, the required sample size is found by first calculating

n_1,

$$n_1 = \left(\frac{1.645}{0.50}\right)^2 (121) = 1310,$$

and then substituting in n_2,

$$n_2 = \frac{1310}{1 + (1310/10000)} = 1158.$$

Therefore, 1,158−100 = 1,058, or, in round figures, 1,060 additional poles must be inspected.

One should not be under the impression that sample size calculations need be extremely precise. Although the computations of this section were precise, this was mainly to avoid ambiguity. We noted in several places that approximations and estimates were involved. The very last calculations, for example, could very well be rounded to n_1 =1,300, and n_2 =1,200 or 1,100 for practical purposes. Excessive accuracy is neither required nor desirable.

The above formulae are based on the normal approximation and require large n_1, or large N and n_2. In fact, if the purpose of the sample is to estimate a population proportion, it is possible to determine the required sample size for *any* c and $1 - \alpha$, and not only those resulting in large n_1 or n_2. The method described below is conceptually quite simple, but its calculations require the use of the computer.

The key to this method is the hypergeometric distribution of the sample frequency $F = nR$, from which we can determine the probability that R will be within $\pm c$ of π for given N, n, and π.

Suppose, for example, that $c = 0.05$ and $N = 100$. To begin with, assume that $\pi = 0.30$, and that a sample of size $n = 40$ without replacement is considered. The probability that R (the sample proportion) will be between ± 0.05 of 0.30 (that is, in the interval 0.25 to 0.35) in a sample of size 40 without replacement is the probability that F (the sample frequency—the number of observations in the sample that belong to the category of interest) will be between 10 (10/40 = 0.25) and 14 (14/40 = 0.35). Since the distribution of F is hypergeometric with parameters $N = 100$, $n = 40$, and $k = N\pi = 30$, $Pr(10 \leq F \leq 14)$ can be calculated easily with the help of a computer program; it is 0.7347. Because of a symmetric feature of the hypergeometric distribution, 0.7347 is also the probability that R will be within ± 0.05 of π when $\pi = 0.7$. Table 4.7 shows this probability, and others for selected values of π and n.

The example illustrates two features which hold for other parameter values as well. First, for given π, the larger the sample, the greater tends to be the probability that R will be within $\pm c$ of π. Second, for given n, the probability that R will be within $\pm c$ of π tends to be least when $\pi = 0.5$.

Table 4.7
Probability that R will be within ± 0.05 of π:
random sampling without replacement, $N = 100$

π	$k = N\pi$	$n = 20$	$n = 40$	$n = 80$
0.1 or 0.9	10 (90)	0.7953	0.9158	0.9996
0.2 or 0.8	20 (80)	0.6515	0.7988	0.9961
0.3 or 0.7	30 (70)	0.5862	0.7347	0.9877
0.4 or 0.6	40 (60)	0.5553	0.7024	0.9800
0.5	50	0.5461	0.6925	0.9772

Thus, to determine how large n must be for this probability to be *at least* $1 - \alpha$ regardless of the true value of π, we need examine in Table 4.7 (or one similarly constructed) only the line corresponding to $\pi = 0.5$, and find the smallest n for which the probability is greater than or equal to $1 - \alpha$.

Table 4.7 shows that if a random sample without replacement is desired from a population of size $N = 100$, the sample size to be such that the probability is at least 95% that any sample proportion will be within ± 0.05 of the corresponding population proportion, then the sample must consist of about 80 observations.

The words "tends to" in the description of the two features above should be noted. For it is not always true that for given π the probability increases with n, or that for given n the probability decreases as π approaches 0.5. In part, this is due to the discreteness of F and to the need to interpret strictly the requirement that R be within $\pm c$ of π. If the sample size must be determined precisely, the computer program can easily generate a table similar to Table 4.7 for every possible value of π and sample size, from an inspection of which the smallest sample meeting the accuracy requirement can be established.

Exactly the same approach may be used to determine the required size of a sample from an independent process or of a random sample from a finite population with replacement. The only difference is that the *binomial* distribution takes the place of the hypergeometric in the calculations.

4.12 IN SUMMARY

A sample is taken in order to estimate some unknown population characteristics. In each draw of a random sample, each eligible population element has the same probability of being selected. This is accomplished by thoroughly mixing the population elements or their surrogates prior to each draw, or with the use of random numbers or of a random device and a list of the population elements.

Unbiased and consistent estimators of the two population characteristics of greatest practical interest, the population mean and proportion, are

the sample mean ($\bar{X}$) and the sample proportion (R) respectively.

These estimators are unbiased, whether the sample is with or without replacement. However, $\bar{X}$ and R in random samples without replacement are more precise (that is, have lower variance) than the same estimators based on samples with replacement of the same size. Sampling without replacement, therefore, should be the preferred sampling method.

Manufacturing processes often behave like independent processes: items are produced one at a time, the probability distribution of an item's measurement (e.g., length, weight, volume) does not vary from item to item, and any item's measurement does not depend on and does not influence that of any other item. A sample from an independent process, selected in any manner whatever, has the same properties as a random sample with replacement from a finite population.

The larger the sample, the more accurate it tends to be, that is, the greater tends to be the probability that its estimators will be within a specified range around the population characteristics. The methods described in this chapter allow the calculation of the sample size needed to obtain estimators with the desired degree of accuracy.

PROBLEMS

4.1 Consider a situation similar to that of Example 4.2. Every item in a lot of $N = 10$ manufactured items has been inspected for defects in manufacture with the following results:

Number of defects	Frequency
0	5
1	3
2	2
Total	10

The lot forms the population of this exercise.

(a) Calculate the population mean (μ), variance (σ^2), and standard deviation (σ) of the number of defects. Calculate the proportion of defective items in the population (π).

(b) Suppose that a random sample of $n = 2$ items *without* replacement is considered. In the manner of Tables 4.2 and 4.3, determine the probability distributions of $\bar{X}$, F (the number), and R (the proportion of defectives in the sample).

(c) Verify that the distribution of F is hypergeometric with appropriate parameters.

(d) Calculate *by two different methods* the mean and variance of $\bar{X}$, F, and R. (Make sure that the results are identical; otherwise, find the error, fix it, and repeat the calculations.)

4.2 Consider the same situation as in Problem 4.1. Suppose that a random sample of $n = 2$ items *with* replacement is considered. Answer questions (a), (b), (d) of Problem 4.1. In place of 4.1(c), show that the probability distribution of F is binomial with appropriate parameters.

4.3 A lot consists of 10 ball bearings. Specifications call for these bearings to be one-quarter of an inch in diameter. The inspection supervisor had all ball bearings measured, and the results are shown below.

Diameter	Number of ball bearings
−1	3
0	5
+1	2
Total	10

The diameter is measured as the difference from the specification (0.250 inches), in thousandths of an inch.

(a) Construct the probability distribution of the average diameter ($\bar{X}$) of two ball bearings drawn from the lot at random *with* replacement. Calculate *by two different methods* the mean and variance of this probability distribution.

(b) Same as (a) except that the sample is *without* replacement.

(c) Construct the probability distribution of the number of ball bearings in a random sample of size $n = 2$ *with* replacement which do *not* meet the specification (i.e., whose diameter is not equal to 0). Calculate *by two different methods* the mean and variance of this distribution.

(d) Same as (c), except that the sample is without replacement.

4.4 A lot of 10 manufactured items contains 3 defective and 7 good items.

(a) With the help of a probability tree, determine the probability distribution of the number of defective items in a random sample with replacement of 3 items from this lot.

(b) Same as (a), except the sample is without replacement.

(c) Verify that the probability distribution in (a) is binomial with appropriate parameters.

(d) Verify that the probability distribution in (b) is hypergeometric with appropriate parameters.

4.5 (a) Determine the expected value and variance of $N\bar{X}$ and NR. *Hint:* These are linear functions of $\bar{X}$ and R. N is assumed to be a known constant.

(b) Explain why $N\bar{X}$ is an unbiased estimator of $N\mu$ and NR one of $N\pi$.

(c) Suppose Y is a "minimum variance unbiased (MVU)" estimator of μ (or π), that is, Y has the smallest variance among all unbiased estimators of μ (or π). Explain why NY is a MVU estimator of $N\mu$ (or $N\pi$).

4.6 A population consists of 5 individuals. It is known that 1 of these has an annual income (X) of \$30(000), while the other 4 have an income of \$20(000). You *plan* to take a random sample without replacement of 3 individuals from this population.

(a) Determine the probability distribution of the average income ($\bar{X}$) of the 3 individuals in the sample. (Express $\bar{X}$ in \$000. List results as fractions—do not convert to decimals.)

(b) Calculate the mean, $E(\bar{X})$, and the variance, $Var(\bar{X})$, of this distribution.

(c) Show that $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2(N-n)/n(N-1)$, where μ and σ^2 are the population mean and variance of X.

4.7 A manufacturing process produces items with measurement X having the following probability distribution:

x	$p(x)$
-1	0.1
0	0.3
$+1$	0.6
	1.0

(a) Assuming the conditions of an independent process are satisfied, verify Equations (4.4) for the mean measurement ($\bar{X}$) of a sample of two items selected in any manner whatever from this process.

(b) Verify Equations (4.5) for the proportion of items having measurement $X = 0$ in a sample of two items selected in any manner whatever from this process.

4.8 The expected value and variance of the average of n independent random variables is given by Equations 2.29 and 2.30 of Chapter 2. Apply these to derive Equations (4.4) for the sample mean of an independent process.

4.9 (a) A random variable, Y, takes the value 1 with probability p and the value 0 with probability $1 - p$. Show that $E(Y) = p$ and $Var(Y) = p(1 - p)$.

(b) A sample of n items from an independent manufacturing process will be taken in order to estimate the long-run proportion of defective items, π. Let Y_1, Y_2, ... , Y_n be *codes* such that $Y_i = 1$ if the ith item in the sample is Defective, and $Y_i = 0$ if it is Good. (i) Explain why $Pr(Y_i = 1) = \pi$, $Pr(Y_i = 0) = 1 - \pi$, and the Y's are independent. (ii) Show that $R = \bar{Y} = (Y_1 + Y_2 + \cdots + Y_n)/n$. (iii) Apply Equations 2.29 and 2.30 of Chapter 2 to derive Equations (4.5).

4.10 A random sample of 6 light bulbs was selected from a lot of 100 for the purpose of estimating its quality. The life of each selected light bulb was measured by letting the bulb burn until it burned out. The test results were as follows:

Bulb No.:	Life duration (hours)
1	950
2	1,210
3	1,070
4	840
5	1,420
6	980

(a) Calculate the sample mean and the sample variance of life duration.
(b) Estimate the average life and the total life of the bulbs in the lot.
(c) Estimate the proportion of bulbs in the lot with life exceeding 1,000 hours.
(d) Estimate the variance of life duration of the bulbs in the lot.
(e) Briefly defend your choice of these estimates.

4.11 A lot of 10 manufactured items is considered acceptable ("good") if it does not contain more than 1 defective, and unacceptable ("bad") if it contains more than 1 defective. (In industry, the number distinguishing a good from a bad lot is called the *acceptable quality level [AQL]*, and is usually expressed as a percentage

of the lot size—in this case, 1/10 or 10%.) The actual number of defectives in the lot is not known. A random sample of size 3 without replacement will be taken, and the lot as a whole will be accepted only if there are no defectives in the sample; the lot will be rejected if one or more defectives show up in the sample.

(a) Calculate the probabilities of accepting and rejecting a lot containing 0, 1, 2, . . . , 10 defectives.

(b) The probability of accepting a bad lot is referred to in industry as the "buyer's risk"; that of rejecting a good lot as the "supplier's risk." These "risks" depend on the unknown number of defectives in the lot. Determine the buyer's and seller's risk for each possible number of defective items in this lot.

(c) What is the maximum probability of accepting a lot containing more than 1 defective (that is, the maximum buyer's risk)? What is the maximum probability of rejecting a lot containing no more than 1 defective (that is, the maximum supplier's risk)? Does the current rule for deciding the fate of the lot treat the buyer and the seller equitably?

(d) Let us make the decision rule more general:

$$\begin{aligned} &\text{Accept the lot if } F \leq c, \\ &\text{Reject the lot if } F > c, \end{aligned}$$

where F is the number of defectives in a sample of size 3. Determine the value of c so that the probability of accepting a bad lot does not exceed 60%. Determine c so that the probability of rejecting a good lot does not exceed 20%.

(e) What would you do (describe, do not calculate) to ensure that the maximum seller's risk is the same as the buyer's risk?

4.12 A company purchases a certain type of electronic equipment in lots of 10,000 items. The decision to accept or reject each lot is based on a random sample of 100 items drawn without replacement. If less than 2% of the items in the sample are defective, the lot is accepted; otherwise it is rejected and returned to the supplier. What is the probability of rejecting a lot containing 1% defectives? 2%? What is the probability of accepting a lot containing 3% defectives? 5%? *Hint:* According to the central limit theorem, for large n and N the probability distribution of R is approximately normal with mean $E(R)$ and variance $Var(R)$ given by Equations (4.2).

4.13 Suppose that a random sample of size n will be drawn from a population consisting of N elements. Show that the probability that a given element will be included in the sample is equal to

$$\text{(a) } \frac{n}{N}, \qquad \text{or} \qquad \text{(b) } 1 - \frac{(N-1)^n}{N^n},$$

depending on whether the sample is (a) without or (b) with replacement. *Hint:* The probability that the element will be included is equal to one minus the probability that it will not be included. Also, the population can be thought of as consisting of two categories: the element itself and all other elements.

4.14 μ and σ^2 are the population mean and variance of a variable X. Consider the following two estimators of the population mean based on a random sample of two observations *with* replacement.

$$\bar{X} = \frac{1}{2}X_1 + \frac{1}{2}X_2$$

$$W = \frac{1}{3}X_1 + \frac{2}{3}X_2$$

X_1 and X_2 are the X values of the first and second sample observations respectively.

(a) Show that both $\bar{X}$ and W are unbiased estimators of μ.

(b) Which of these two estimators of μ is preferable? Why?

Hint: Both estimators are linear functions of the sample observations. The expected value and variance of linear functions of random variables are given by Equations (2.25) and (2.26) of Chapter 2.

4.15 The purpose of this exercise is to demonstrate that a recommended procedure for selecting a random sample does indeed satisfy the requirements.

A population has 6 elements, identified by the letters A through F and numbered serially from 1 to 6. We plan to select a random sample of size 2 without replacement. Suppose that the first sample element has been selected and that it is A. There are 5 elements eligible for the second draw, B through F. The next (one-digit) random number will be used. If it is 2, B will be the second sample element; if 3, C; and so on. If the random number is 1 (the number for A), 0, 7, 8, or 9, another random number will be used, and the same procedure will be repeated. Clearly, it is possible that many random numbers may be required before a selection is made.

Show that every eligible population element in the second draw is given the same probability of selection. *Hint:* $1 + a + a^2 + \cdots = 1/(1-a)$, if $0 < a < 1$.

4.16 A population has 8,943 elements, numbered serially from 0001 to 8943. Select a random sample of 10 elements without replacement, showing your method clearly.

4.17 (a) A population consists of $N = 5,000$ elements. How large a sample should be drawn from this population if it is desired to estimate all population relative frequencies within ± 0.075 with probability of 90%? Assume the sample is (i) with, and (ii) without replacement.

(b) A population consists of $N = 7,500$ elements. How large a sample should be drawn from this population if it is desired to estimate the population mean within ± 20 with probability 99%? A pilot survey suggests that the population variance is likely to be in the range 80 to 100. Assume the sample is (i) with, and (ii) without replacement.

4.18 Calculate the probability that R will be within ± 0.10 of π for each possible value of π and a random sample without replacement of size 5 from a population consisting of $N = 10$ elements. *Hint:* The probability that R will be between 0 and 0.2 is the probability that F will be 0 or 1; the probability that R will be between 0.1 and 0.3 is the probability that F will be equal to 1; and so on.

4.19 (a) Calculate the probability that R will be within ± 0.05 of π when $\pi = 0.5$ for random samples of size 100, 200, and 300 with replacement from a population of size $N = 1,000$. Approximate answers are satisfactory.

(b) Same as (a), except that sampling is without replacement.

4.20 A population consists of $N = 10$ elements. A random sample of size n without replacement is considered, for the purpose of estimating π, the proportion of elements in the population that belong to a given category C.

Determine if a sample of size $n = 5$ is sufficiently large for R (the proportion of elements in the sample belonging to C) to be within ± 0.10 of π (that is, for

$\pi - 0.10 \leq R \leq \pi + 0.10$) with probability at least 70%, whatever π happens to be.

4.21 In order to determine the general pattern of railway freight movements and the changing trends of railway traffic in the country, the Transport Commission annually selects a "one per cent continuous sample" of waybills. The participating railways are requested to forward copies of 1% of all line-haul carload waybills at their stations. This 1% sample is drawn by taking all waybills bearing the serial numbers ending in "01."

(a) Comment on this method of sample selection. Can this "one per cent sample" be considered a random sample without replacement from the population of all waybills?

(b) Given a population of 100,000 waybills, calculate the probability that in a "one per cent" *random* sample of waybills any sample relative frequency will be within ± 0.01 of the corresponding population relative frequency.

4.22 Many firms are able to maintain fairly accurate lists of customers by utilizing warranty registration cards. Manufacturers of watches, calculators, washers, dryers, toasters, refrigerators, blenders, and many other appliances usually guarantee the product against defects in manufacture for a certain number of years from the date of purchase. In order to register the warranty, buyers must fill out and return a card showing their name and address, the model purchased, the type of store from which it was purchased, the method of payment, etc. These cards provide useful information to the company when they are first received (for example, in determining the distribution of sales by type of store or location, or of the time elapsing between the date of sale and the date of shipment), and a list of customers which may be used in later surveys.

Yoshita is a manufacturer of television and other electronic equipment. On the basis of warranty registration cards, it maintains a file of 21,528 customers who purchased a TV set. From this file, a random sample of 500 customers was selected (without replacement, of course) for the purpose of estimating the proportion owning a VCR and the average age of these VCRs. Of the 500 sampled customers, 361 stated they owned a VCR. The average age of these VCRs was 14.3 months.

(a) Using a table of random numbers, show how the first 5 customers should be selected.

(b) Estimate the number of customers owning a VCR. Is this an unbiased estimator? Why?

(c) Estimate the average age of all VCRs owned by customers. Is this an unbiased estimator? Why? Can you write the variance of this estimator?

4.23 GreenTurf is a company making more than 100 garden and agricultural fertilizers, consisting of different combinations of three basic ingredients. One of these ingredients is nitrogen. The company estimates the total quantity of nitrogen used on the basis of a random sample of production orders. Each order shows the quantity of nitrogen and of the other ingredients used for a particular job. There were 4,000 production orders issued during the year, numbered from 0001 to 4000.

(a) Describe briefly but precisely how you would select a *random* sample of 100 production orders without replacement.

(b) The auditor prefers to draw a sample in a different way. He would select at random one production order among those numbered 0001 to 0040, and every 40th order in sequence thereafter. For example, if the first order selected is No. 0005, the remaining selected orders would have numbers 0045, 0085, ..., 3965.

This method (known as *systematic sampling*) will indeed produce a sample of 100. Is a systematic sample a random one? Under what conditions can such a sample be treated as random for all practical purposes?

(c) A random sample without replacement of 100 production orders was taken. The average quantity of nitrogen per order was 150 lb, and the standard deviation of the quantities of nitrogen in the sample was 40 lb. Estimate the total quantity of nitrogen used.

(d) How large should *next* year's sample be so that the estimated total quantity of nitrogen used will be within ±10,000 lb of the true total with probability 99%? Assume this year's number of production orders (4,000).

4.24 According to a recent survey of 1,545 households in the "Golden Horseshoe" area, conducted by Operational Management Inc., the market shares of the three major food chains were estimated as follows:

Loblaws	27%
Dominion Stores	52%
Miracle Food	8%
Other	13%

"Market share" is the proportion of households who buy regularly from a chain.

Assume that the households surveyed constitute a random sample of households from the population of about one million households in the Golden Horseshoe area.

How large should the *next* random sample without replacement be so that the estimated market shares will be within ±0.01 of the true (population) shares with probability at least 95%?