Chapter 4

Statistical Inference: Estimation

4.1 SAMPLING

Statistical inference is one of the most important and crucial aspects of the decision-making process in economics, business, and science. Statistical inference refers to estimation and hypothesis testing (Chap. 5). *Estimation* is the process of inferring or estimating a population *parameter* (such as its mean or standard deviation) from the corresponding *statistic* of a sample drawn from the population.

To be valid, estimation (and hypothesis testing) must be based on a *representative* sample. This can be obtained by *random sampling*, whereby each member of the population has an equal chance of being included in the sample.

EXAMPLE 1. A random sample of 5 out of the 80 employees of a plant can be obtained by recording the name of each employee on a separate slip of paper, mixing the slips of paper thoroughly, and then picking 5 at random. A less cumbersome method is to use a table of random numbers (App. 4). To do this, we first assign each employee a number from 1 to 80. Then starting at random (say, from the third column and eleventh row) in App. 4, we can read 5 numbers (as pairs) either horizontally or vertically (eliminating all numbers exceeding 80). For example, reading vertically we get 13, 54, 19, 59, and 71.

4.2 SAMPLING DISTRIBUTION OF THE MEAN

If we take repeated random samples from a population and measure the mean of each sample, we find that most of these sample means, $\overline{X}$s, differ from each other. The probability distribution of these sample means is called the *sampling distribution of the mean.* However, the sampling distribution of the mean itself has a mean, given by the symbol $\mu_{\overline{X}}$, and a standard deviation or *standard error*, $\sigma_{\overline{X}}$.

Two important theorems relate the sampling distribution of the mean to the parent population.

Theorem 1. If we take repeated random samples of size n from a population:

$$\mu_{\overline{X}} = \mu \tag{4.1}$$

and

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{4.2a,b}$$

where Eq. (4.2*b*) is used for finite populations of size N when $n \geq 0.05N$ [see Prob. 4.5(*b*)].

Theorem 2. As the samples' size is increased (that is, as $n \to \infty$), the sampling distribution of the mean approaches the normal distribution regardless of the shape of the parent population. The approximation is sufficiently good for $n \geq 30$. This is the *central limit theorem.*

We can find the probability that a random sample has a mean $\overline{X}$ in a given interval by first calculating the z values for the interval, where

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \tag{4.3}$$

and then looking up these values in App. 3, as explained in Sec. 3.5.

EXAMPLE 2. In Fig. 4-1, the mean of the sampling distribution of the mean, $\mu_{\overline{X}}$, is equal to the mean of the parent population, μ, regardless of the samples' size, n. However, the greater is n, the smaller is the spread or standard error of the mean, $\sigma_{\overline{X}}$. If the parent population is normal, the sampling distributions of the mean are also normally distributed, even in small samples. According to the central limit theorem, even if the parent population is not normally distributed, the sampling distributions of the mean are approximately normal for $n > 30$.

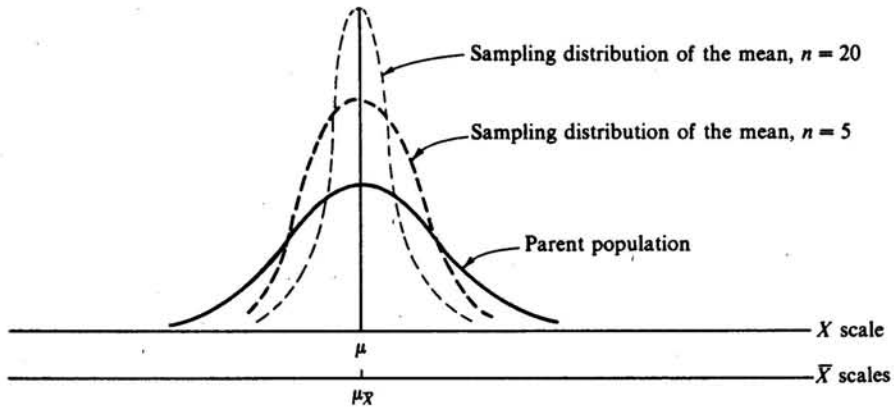

Fig. 4-1

EXAMPLE 3. Assume that a population is composed of 900 elements with a mean of 20 units and a standard deviation of 12. The mean and standard error of the sampling distribution of the mean for a sample size of 36 is

$$\mu_{\bar{X}} = \mu = 20 \text{ units}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{36}} = 2$$

If n had been 64 instead of 36 (so that $n > 0.05N$), then

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{12}{\sqrt{64}}\sqrt{\frac{900-64}{900-1}} = \frac{12}{8}\sqrt{\frac{836}{899}} = (1.5)(0.96) = 1.44$$

instead of $\sigma_{\bar{X}} = 1.5$, without the *finite correction factor*.

EXAMPLE 4. The probability that the mean of a random sample, $\bar{X}$, of 36 elements from the population in Example 3 falls between 18 and 24 units is computed as follows:

$$z_1 = \frac{\bar{X}_1 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{18-20}{2} = -1 \quad \text{and} \quad z_2 = \frac{\bar{X}_2 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{24-20}{2} = 2$$

Looking up z_1 and z_2 in App. 3, we get

$$P(18 < \bar{X} < 24) = 0.3413 + 0.4772 = 0.8185, \text{ or } 81.85\%$$

See Fig. 4-2.

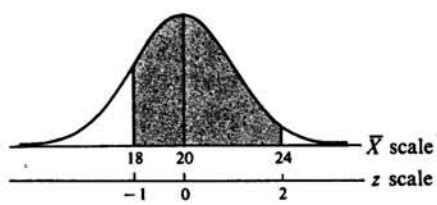

Fig. 4-2

4.3 ESTIMATION USING THE NORMAL DISTRIBUTION

We can get a point or an interval estimate of a population parameter. A *point estimate* is a single number. Such a point estimate is *unbiased* if in repeated random samplings from the population, the expected or mean value of the corresponding statistic is equal to the population parameter. For

example, $\bar{X}$ is an unbiased (point) estimate of μ because $\mu_{\bar{X}} = \mu$, where $\mu_{\bar{X}}$ is the expected value of $\bar{X}$. The sample standard deviation s [as defined in Eqs. (2.10*b*) and (2.11*b*)] is an unbiased estimate of σ [see Prob. 4.13(*b*)], and the sample proportion $\bar{p}$ is an unbiased estimate of p (the proportion of the population with a given characteristic).

An *interval estimate* refers to a range of values together with the probability, or *confidence level*, that the interval includes the unknown population parameter. Given the population standard deviation or its estimate, and given that the population is normal or that a random sample is equal to or larger than 30, we can find the 95% confidence interval for the unknown population mean as

$$P\left(\bar{X} - 1.96\sigma_{\bar{X}} < \mu < \bar{X} + 1.96\sigma_{\bar{X}}\right) = 0.95 \tag{4.4}$$

This states that in repeated random sampling, we expect that 95 out of 100 intervals like Eq. (4.4) include the unknown population mean and that our confidence interval (based on a single random sample) is one of these.

A confidence interval can be constructed similarly for the population *proportion* (see Example 7) where

$$\mu_{\bar{p}} = \frac{\mu}{n} = p \qquad \text{(the proportion of successes in the population)} \tag{4.5}$$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \text{(the standard error of the proportion)} \tag{4.6}$$

EXAMPLE 5. A random sample of 144 with a mean of 100 and a standard deviation of 60 is taken from a population of 1,000. The 95% confidence interval for the unknown population mean is

$$\mu = \bar{X} \pm 1.96\sigma_{\bar{X}} \qquad \text{since } n > 30$$

$$= \bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \text{since } n > 0.05N$$

$$= 100 \pm 1.96\frac{60}{\sqrt{144}}\sqrt{\frac{1{,}000-144}{1{,}000-1}} \qquad \text{using } s \text{ as an estimate of } \sigma$$

$$= 100 \pm 1.96(5)(0.93)$$

$$= 100 \pm 9.11$$

Thus μ is between 90.89 and 109.11 with a 95% degree of confidence. Other frequently used confidence intervals are the 90 and 99% levels, corresponding to the z values of 1.64 and 2.58, respectively (see App. 3).

EXAMPLE 6. A manager wishes to estimate the mean number of minutes that workers take to complete a particular manufacturing process within ± 3 min and with 90% confidence. From past experience, the manager knows that the standard deviation, σ, is 15 min. The minimum required sample size ($n > 30$) is found as follows:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

$$z\sigma_{\bar{X}} = \bar{X} - \mu$$

$$1.64\frac{\sigma}{\sqrt{n}} = \bar{X} - \mu \qquad \text{assuming } n < 0.05N$$

$$1.64\frac{15}{\sqrt{n}} = 3 \qquad \text{since the total confidence interval, } \bar{X} - \mu, \text{ is 3 min}$$

$$1.64\frac{15}{3} = \sqrt{n}$$

$$n = 67.24, \text{ or } 68 \text{ (rounded to the next higher integer)}$$

EXAMPLE 7. A state education department finds that in a random sample of 100 persons who attended college, 40 received a college degree. To find the 99% confidence interval for the proportion of college graduates out of all the persons who attended college, we proceed as follows. First, we note that this problem involves the binomial distribution (see Sec. 3.3). Since $n > 30$ and both $np > 5$ and $n(1-p) > 5$, the binomial distribution approaches the normal distribution (which is simpler to use: see Sec. 3.5). Then,

$$p = \bar{p} \pm z\sigma_{\bar{p}}$$

and

$$p = \bar{p} \pm z\sqrt{\frac{p(1-p)}{n}} \qquad \text{assuming } n < 0.05N$$

$$= 0.4 \pm 2.58\sqrt{\frac{(0.4)(0.6)}{100}} \qquad \text{using } \bar{p} \text{ as an estimate of } p$$

$$\simeq 0.4 \pm 2.58(0.05)$$

$$\simeq 0.4 \pm 0.13$$

Thus p is between 0.27 and 0.53 with a 99% level of confidence.

4.4 CONFIDENCE INTERVALS FOR THE MEAN USING THE t DISTRIBUTION

When the population is normally distributed but σ is not known and $n < 30$, we cannot use the normal distribution for determining confidence intervals for the unknown population mean, but we can use the t distribution. This is symmetrical about its zero mean but is flatter than the standard normal distribution, so that more of its area falls within the tails. While there is a single standard normal distribution, there is a different t distribution for each sample size, n. However, as n becomes larger, the t distribution approaches the standard normal distribution (see Fig. 4-3) until, when $n \geq 30$, they are approximately equal.

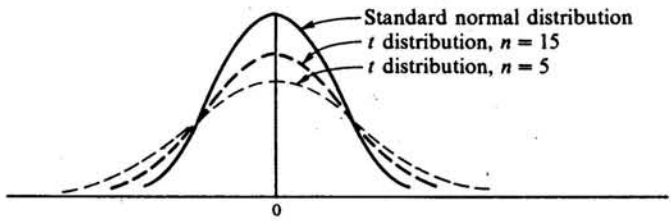

Fig. 4-3

Appendix 5 gives the values of t *to the right of which* we find 10, 5, 2.5, 1, and 0.5% of the total area under the curve for various *degrees of freedom*. Degrees of freedom, df, are defined in this case as $n - 1$ (or the sample size minus 1 for the single parameter, μ, we wish to estimate). The 95% confidence interval for the unknown population mean when the t distribution is used is given by

$$P\left(\bar{X} - t\frac{s}{\sqrt{n}} < \mu < \bar{X} + t\frac{s}{\sqrt{n}}\right) = 0.95 \tag{4.7}$$

where t refers to the t values such that 2.5% of the total area under the curve falls within each tail (for the degrees of freedom involved) and $s/\sqrt{n}$ is used instead of $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.