7-2 Sampling Distributions and the Central Limit Theorem

Statistical inference is concerned with making **decisions** about a population based on the information contained in a random sample from that population. For instance, we may be interested in the mean fill volume of a container of soft drink. The mean fill volume in the population is required to be 300 milliliters. An engineer takes a random sample of 25 containers and computes the sample average fill volume to be $\bar{x} = 298.8$ milliliters. The engineer will probably decide that the population mean is $\mu = 300$ milliliters even though the sample mean was 298.8 milliliters because he or she knows that the sample mean is a reasonable estimate of μ and that a sample mean of 298.8 milliliters is very likely to occur even if the true population mean is $\mu = 300$ milliliters. In fact, if the true mean is 300 milliliters, tests of 25 containers made repeatedly, perhaps every five minutes, would produce values of $\bar{x}$ that vary both above and below $\mu = 300$ milliliters.

The link between the probability models in the earlier chapters and the data is made as follows. Each numerical value in the data is the observed value of a random variable. Furthermore, the random variables are usually assumed to be independent and identically distributed. These random variables are known as a *random sample*.

Random Sample

The random variables $X_1, X_2, \dots, X_n$ are a random sample of size n if (a) the X_i's are independent random variables and (b) every X_i has the same probability distribution.

The observed data are also referred to as a *random sample*, but the use of the same phrase should not cause any confusion.

The assumption of a random sample is extremely important. If the sample is not random and is based on judgment or is flawed in some other way, statistical methods will not work properly and will lead to incorrect decisions.

The primary purpose in taking a random sample is to obtain information about the unknown population parameters. Suppose, for example, that we wish to reach a conclusion about the proportion of people in the United States who prefer a particular brand of soft drink. Let p represent the unknown value of this proportion. It is impractical to question every individual in the population to determine the true value of p. To make an inference regarding the true proportion p, a more reasonable procedure would be to select a random sample (of an appropriate size) and use the observed proportion $\hat{p}$ of people in this sample favoring the brand of soft drink.

The sample proportion, $\hat{p}$, is computed by dividing the number of individuals in the sample who prefer the brand of soft drink by the total sample size n. Thus, $\hat{p}$ is a function of the observed values in the random sample. Because many random samples are possible from a population, the value of $\hat{p}$ will vary from sample to sample. That is, $\hat{p}$ is a random variable. Such a random variable is called a **statistic.**

Statistic

A statistic is any function of the observations in a random sample.

We have encountered statistics before. For example, if $X, X_2, \dots, X_n$ is a random sample of size n, the **sample mean** $\bar{X}$, the **sample variance** S^2, and the **sample standard deviation** S are statistics. Because a statistic is a random variable, it has a probability distribution.

Sampling Distribution

The probability distribution of a statistic is called a sampling distribution.

For example, the probability distribution of $\bar{X}$ is called the **sampling distribution of the mean**. The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection. We now present perhaps the most important sampling distribution. Other sampling distributions and their applications will be illustrated extensively in the following two chapters.

Consider determining the sampling distribution of the sample mean $\bar{X}$. Suppose that a random sample of size n is taken from a normal population with mean μ and variance σ^2. Now each observation in this sample, say, $X_1, X_2, \dots, X_n$, is a normally and independently distributed random variable with mean μ and variance σ^2. Then because linear functions of independent, normally distributed random variables are also normally distributed (Chapter 5), we conclude that the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

has a normal distribution with mean

$$\mu_{\bar{X}} = \frac{\mu + \mu + \dots + \mu}{n} = \mu$$

and variance

$$\sigma^2{}_{\bar{X}} = \frac{\sigma^2 + \sigma^2 + \dots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

If we are sampling from a population that has an unknown probability distribution, the sampling distribution of the sample mean will still be approximately normal with mean μ and variance σ^2/n if the sample size n is large. This is one of the most useful theorems in statistics, called the **central limit theorem**. The statement is as follows:

Central Limit Theorem

> **If $X_1, X_2, \dots, X_n$ is a random sample of size n taken** from a population (either finite or infinite) **with mean μ and** finite variance σ^2 and if $\bar{X}$ is the sample mean, the limiting form of **the distribution** of
>
> $$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \qquad (7\text{-}1)$$
>
> as $n \to \infty$, is the standard normal distribution.

It is easy to demonstrate the central limit theorem with a **computer simulation experiment**. Consider the lognormal distribution in Fig. 7-1. This distribution has parameters $\theta = 2$ (called the **location** parameter) and $\omega = 0.75$ (called the **scale** parameter), resulting in mean $\mu = 9.79$ and standard deviation $\sigma = 8.51$. Notice that this lognormal distribution does not look very much like the normal distribution; it is defined only for positive values of the random variable X and is skewed considerably to the right. We used computer software to draw 20 samples at random from this distribution, each of size $n = 10$. The data from this sampling experiment are shown in Table 7-1. The last row in this table is the average of each sample $\bar{x}$.

The first thing that we notice in looking at the values of $\bar{x}$ is that they are not all the same. This is a clear demonstration of the point made previously that any statistic is a random

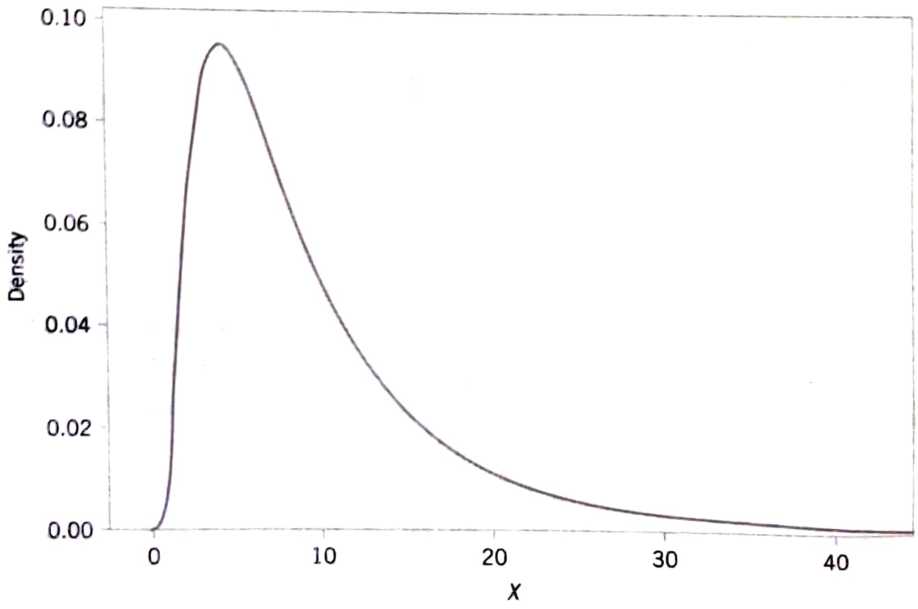


FIGURE 7-1 A lognormal distribution with $\theta = 2$ and $\omega = 0.75$.

variable. If we had calculated any sample statistic (s, the sample median, the upper or lower quartile, or a percentile), they would also have varied from sample to sample because they are random variables. Try it and see for yourself.

According to the central limit theorem, the distribution of the sample average x is normal. Figure 7-2 is a normal probability plot of the 20 sample averages $\bar{x}$ from Table 7-1. The observations scatter generally along a straight line, providing evidence that the distribution of the sample mean is normal even though the distribution of the population is very non-normal. This type of sampling experiment can be used to investigate the sampling distribution of any statistic.

The normal approximation for $\bar{X}$ depends on the sample size n. Figure 7-3(a) is the distribution obtained for throws of a single, six-sided true die. The probabilities are equal ($1/6$) for all the values obtained: 1, 2, 3, 4, 5, or 6. Figure 7-3(b) is the distribution of the average score obtained when tossing two dice, and Fig. 7-3(c), 7-3(d), and 7-3(e) show the distributions of average scores obtained when tossing 3, 5, and 10 dice, respectively. Notice that, although the population (one die) is relatively far from normal, the distribution of averages is approximated reasonably well by the normal distribution for sample sizes as small as five. (The dice throw distributions are discrete, but the normal is continuous.)

The central limit theorem is the underlying reason why many of the random variables encountered in engineering and science are normally distributed. The observed variable of the results from a series of underlying disturbances that act together to create a central limit effect.

TABLE • 7-1 Twenty samples of size $n = 10$ from the lognormal distribution in Figure 7-1.

	Sample									
Obs	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	3.9950	8.2220	4.1893	15.0907	12.8233	15.2285	5.6319	7.5504	2.1503	3.1390
2	7.8452	13.8194	2.6186	4.5107	3.1392	16.3821	3.3469	1.4393	46.3631	1.8314
3	1.8858	4.0513	8.7829	7.1955	7.1819	12.0456	8.1139	6.0995	2.4787	3.7612
4	16.3041	7.5223	2.5766	18.9189	4.2923	13.4837	13.6444	8.0837	19.7610	15.7647
5	9.7061	6.7623	4.4940	11.1338	3.1460	13.7345	9.3532	2.1988	3.8142	3.6519
6	7.6146	5.3355	10.8979	3.6718	21.1501	1.6469	4.9919	13.6334	2.8456	14.5579
7	6.2978	6.7051	6.0570	8.5411	3.9089	11.0555	6.2107	7.9361	11.4422	9.7823
8	19.3613	15.6610	10.9201	5.9469	8.5416	19.7158	11.3562	3.9083	12.8958	2.2788
9	7.2275	3.7706	38.3312	6.0463	10.1081	2.2129	11.2097	3.7184	28.2844	26.0186
10	16.2093	3.4991	6.6584	4.2594	6.1328	9.2619	4.1761	5.2093	10.0632	17.9411
$\bar{x}$	9.6447	7.5348	9.5526	8.5315	8.0424	11.4767	7.8035	5.9777	14.0098	9.8727
Obs	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	7.5528	8.4998	2.5299	2.3115	6.1115	3.9102	2.3593	9.6420	5.0707	6.8075
2	4.9644	3.9780	11.0097	18.8265	3.1343	11.0269	7.3140	37.4338	5.5860	8.7372
3	16.7181	6.2696	21.9326	7.9053	2.3187	12.0887	5.1996	3.6109	3.6879	19.2486
4	8.2167	8.1599	15.5126	7.4145	6.7088	8.3312	11.9890	11.0013	5.6657	5.3550
5	9.0399	15.9189	7.9941	22.9887	8.0867	2.7181	5.7980	4.4095	12.1895	16.9185
6	4.0417	2.8099	7.1098	1.4794	14.5747	8.6157	7.8752	7.5667	32.7319	8.2588
7	4.9550	40.1865	5.1538	8.1568	4.8331	14.4199	4.3802	33.0634	11.9011	4.8917
8	7.5029	10.1408	2.6880	1.5977	7.2705	5.8623	2.0234	6.4656	12.8903	3.3929
9	8.4102	6.4106	7.6495	7.2551	3.9539	16.4997	1.8237	8.1360	7.4377	15.2643
10	7.2316	11.5961	4.4851	23.0760	10.3469	9.9330	8.6515	1.6852	3.6678	2.9765
$\bar{x}$	7.8633	11.3970	8.6065	10.1011	6.7339	9.3406	5.7415	12.3014	10.0828	9.1851

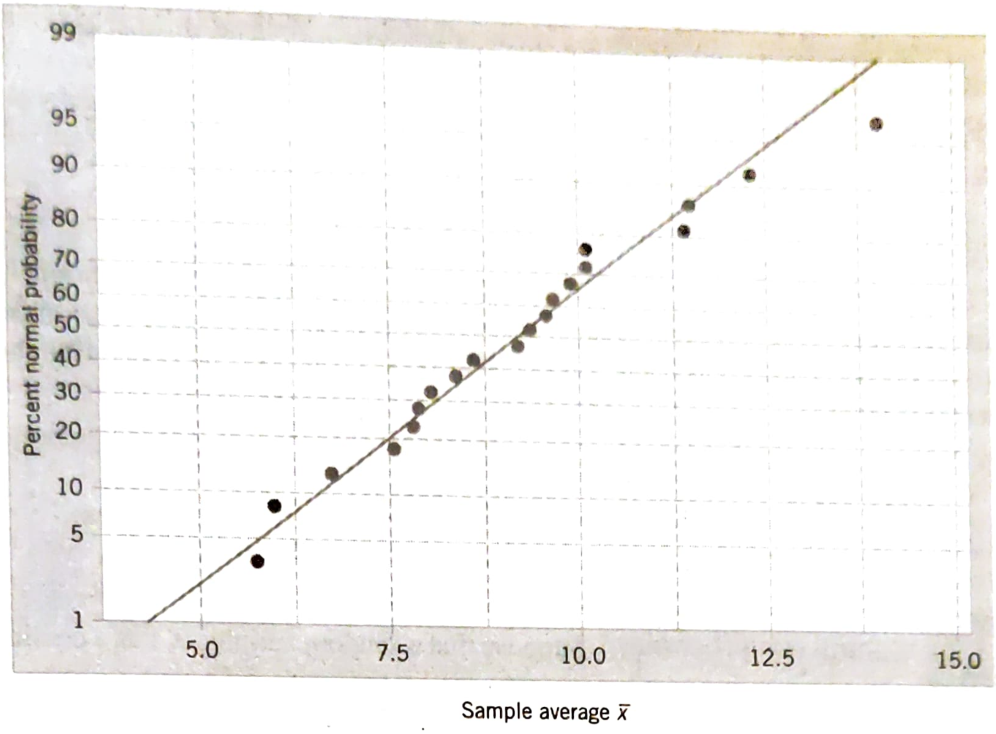
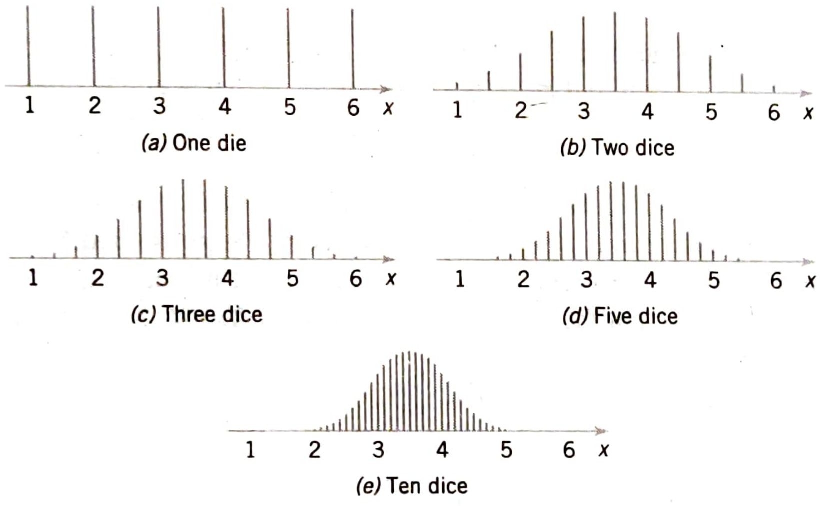


FIGURE 7-2 Normal probability plot of the sample averages from Table 7-1.



(a) One die

(b) Two dice

(c) Three dice

(d) Five dice

(e) Ten dice

FIGURE 7-3 Distributions of average scores from throwing dice. *Source*: [Adapted with permission from Box, Hunter, and Hunter (1978).]

When is the sample size large enough so that the central limit theorem can be assumed to apply? The answer depends on how close the underlying distribution is to the normal. If the underlying distribution is symmetric and unimodal (not too far from normal), the central limit theorem will apply for small values of n, say 4 or 5. If the sampled population is very non-normal, larger samples will be required. As a general guideline, if $n > 30$, the central limit theorem will almost always apply. There are exceptions to this guideline are relatively rare. In most cases encountered in practice, this guideline is very conservative, and the central limit theorem will apply for sample sizes much smaller than 30. For example, consider the dice example in Fig. 7-3.

Example 7-1

Resistors An electronics company manufactures resistors that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. Find the probability that a random sample of $n = 25$ resistors will have an average resistance of fewer than 95 ohms.

Note that the sampling distribution of $\bar{X}$ is normal with mean $\mu_{\bar{X}} = 100$ ohms and a standard deviation of

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

Therefore, the desired probability corresponds to the shaded area in Fig. 7-4. Standardizing the point $\bar{X} = 95$ in Fig. 7-4, we find that

$$z = \frac{95-100}{2} = -2.5$$

and therefore,

$$P(\bar{X} < 95) = P(Z < -2.5) = 0.0062$$

Practical Conclusion: This example shows that if the distribution of resistance is normal with mean 100 ohms and standard deviation of 10 ohms, finding a random sample of resistors with a sample mean less than 95 ohms is a **rare event**. If this actually happens, it casts doubt as to whether the true mean is really 100 ohms or if the true standard deviation is really 10 ohms.

The following example makes use of the central limit theorem.

Example 7-2 Central Limit Theorem Suppose that a random variable X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 40$.

The mean and variance of X are $\mu = 5$ and $\sigma^2 = (6-4)^2/12 = 1/3$. The central limit theorem indicates that the distribution of $\bar{X}$ is approximately normal with mean $\mu_{\bar{X}} = 5$ and variance $\sigma_{\bar{X}}^2 = \sigma^2/n = 1/[3(40)] = 1/120$. See the distributions of X and $\bar{X}$ in Fig. 7-5.

Now consider the case in which we have two independent populations. Let the first population have mean μ_1 and variance σ_1^2 and the second population have mean μ_2 and variance σ_2^2. Suppose that both populations are normally distributed. Then, using the fact that linear combinations of independent normal random variables follow a normal distribution (see Chapter 5), we can say that the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is normal with mean

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2 \qquad (7\text{-}2)$$

and variance

$$\sigma_{\bar{X}_1-\bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \qquad (7\text{-}3)$$

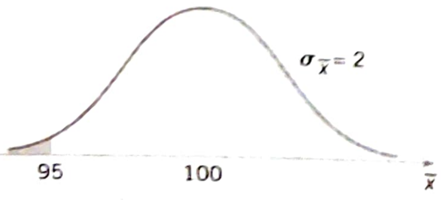


FIGURE 7-4 Probability for Example 7-1.

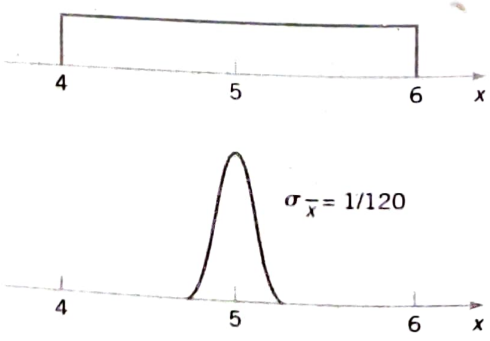


FIGURE 7-5 The distribution of X and $\bar{X}$ for Example 7-2.

If the two populations are not normally distributed and if both sample sizes n_1 and n_2 are more than 30, we may use the central limit theorem and assume that $\bar{X}_2$ and $\bar{X}_2$ follow approximately independent normal distributions. Therefore, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal with mean and variance given by Equations 7-2 and 7-3, respectively.

If either n_1 or n_2 is fewer than 30, the sampling distribution of $X_1 - X_2$ will still be approximately normal with mean and variance given by Equations 7-2 and 7-3 provided that the population from which the small sample is taken is not dramatically different from the normal. We may summarize this with the following definition.

Approximate Sampling Distribution of a Difference in Sample Means

If we have two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 and if $\overline{X}_1$ and $\overline{X}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \tag{7-4}$$

is approximately standard normal if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of Z is exactly standard normal.

Example 7-3 Aircraft Engine Life The effective life of a component used in a jet-turbine aircraft engine is a random variable with mean 5000 hours and standard deviation 40 hours. The distribution of effective life is fairly close to a normal distribution. The engine manufacturer introduces an improvement into the manufacturing process for this component that increases the mean life to 5050 hours and decreases the standard deviation to 30 hours. Suppose that a random sample of $n_1 = 16$ components is selected from the "old" process and a random sample of $n_2 = 25$ components is selected from the "improved" process. What is the probability that the difference in the two samples means $\overline{X}_2 - \overline{X}_1$ is at least 25 hours? Assume that the old and improved processes can be regarded as independent populations.

To solve this problem, we first note that the distribution of $\overline{X}_1$ is normal with mean $\mu_1 = 5000$ hours and standard deviation $\sigma_1 / \sqrt{n_1} = 40 / \sqrt{16} = 10$ hours, and the distribution of $\overline{X}_2$ is normal with mean $\mu_2 = 5050$ hours and standard deviation $\sigma_2 / \sqrt{n_2} = 30 / \sqrt{25} = 6$ hours. Now the distribution of $\overline{X}_2 - \overline{X}_1$ is normal with mean $\mu_2 - \mu_1 = 5050 - 5000 = 50$ hours and variance $\sigma_2^2 / n_2 + \sigma_1^2 / n_1 = (6)^2 + (10)^2 = 136$ hours^2. This sampling distribution is shown in Fig. 7-6. The probability that $\overline{X}_2 - \overline{X}_1 \geq 25$ is the shaded portion of the normal distribution in this figure.

Corresponding to the value $\overline{x}_2 - \overline{x}_1 = 25$ in Fig. 7-4, we find that

$$z = \frac{25 - 50}{\sqrt{136}} = -2.14$$

and consequently,

$$P(\overline{X}_2 - \overline{X}_1 \geq 25) = P(Z \geq -2.14) = 0.9838$$

Therefore, there is a high probability (0.9838) that the difference in sample means between the new and the old process will be at least 25 hours if the sample sizes are $n_1 = 16$ and $n_2 = 25$.

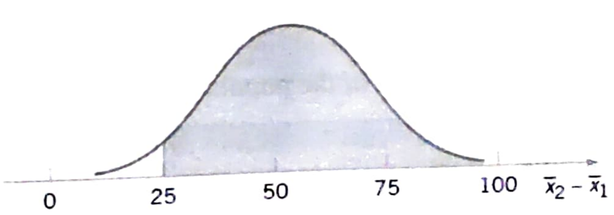


FIGURE 7-6 The sampling distribution of $\overline{X}_2 - \overline{X}_1$ in Example 7-3.