4.3 SAMPLING DISTRIBUTIONS AND ILLUSTRATIONS

Consider a random variable x with a probability distribution function $P(x)$. Let $x_1, x_2, \ldots, x_N$ be a sample of N observed values of x. Any quantity computed from these sample values will also be a random variable. For example, consider the mean value $\bar{x}$ of the sample. If a series of different samples of size N were selected from the same random variable x, the value of $\bar{x}$ computed from each sample would generally be different. Hence, $\bar{x}$ is also a random variable with a probability distribution function $P(\bar{x})$. This probability distribution function is called the *sampling distribution* of $\bar{x}$.

Some of the more common sampling distributions which often arise in practice will now be considered. These involve the probability distribution functions defined and discussed in Section 4.2. The use of these sampling distributions to establish confidence intervals and perform hypothesis tests is illustrated in Sections 4.4 through 4.8.

4.3.1 *Distribution of Sample Mean with Known Variance*

Consider the mean value of a sample of N independent observations from a random variable x as follows.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.31}$$

First consider the case where the random variable x is normally distributed with a mean value of μ_x and a known variance of σ_x^2. From Section 3.5.1, the sampling distribution of the sample mean $\bar{x}$ will also be normally distributed. From Equation (4.8), the mean value of the sampling distribution of $\bar{x}$ is

$$\mu_{\bar{x}} = \mu_x \tag{4.32}$$

and from Equation (4.9), the variance of the sampling distribution of $\bar{x}$ is

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{N} \tag{4.33}$$

Hence, from Equation (4.13), the following sampling distribution applies for the sample mean $\bar{x}$:

$$\frac{(\bar{x} - \mu_x)\sqrt{N}}{\sigma_x} = z \tag{4.34}$$

where z has a standardized normal distribution, as defined in Section 4.2.1.

It follows that a probability statement concerning future values of the sample mean may be made as follows.

$$\text{Prob}\left[\bar{x} > \left(\frac{\sigma_x z_\alpha}{\sqrt{N}} + \mu_x\right)\right] = \alpha \tag{4.35}$$

Now, consider the case where the random variable x is not normally distributed. From the practical implications of the Central Limit Theorem, the following result occurs. As the sample size N becomes large, the *sampling distribution of the sample mean $\bar{x}$ approaches a normal distribution regardless of the distribution of the original variable x.* In practical terms, a normality assumption for the sampling distribution of $\bar{x}$ becomes reasonable in many cases for $N > 4$ and quite accurate in most cases for $N > 10$. Hence for reasonably large sample sizes, Equation (4.34) applies to the sampling distribution of $\bar{x}$ computed for any random variable x, regardless of its probability distribution function.

4.3.2 *Distribution of Sample Variance*

Consider the variance of a sample of N independent observations from a random variable x as follows.

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{4.36}$$

If the variable x is normally distributed with a mean of μ_x and a variance of σ_x^2, it is readily shown using Equations (4.10), (4.16), and (4.34) that

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N}(x_i - \mu_x)^2 - N(\bar{x} - \mu_x)^2$$

$$= \sigma_x^2 \sum_{i=1}^{N} z_i^2 - \frac{N\sigma_x^2}{N} z^2 = \sigma_x^2 \sum_{i=1}^{N-1} z_i^2$$

$$= \sigma_x^2 \chi_n^2 \qquad n = N - 1$$

where χ_n^2 has a chi-square distribution with $n = N - 1$ degrees of freedom, as defined in Section 4.2.2. Hence the sampling distribution of the sample variance s^2 is given by

$$\frac{ns^2}{\sigma_x^2} = \chi_n^2 \qquad n = N - 1 \tag{4.37}$$

It follows that a probability statement concerning future values of the sample variance s^2 may be made as follows.

$$\text{Prob}\left[s^2 > \frac{\sigma_x^2 \chi_{n;\alpha}^2}{n}\right] = \alpha \tag{4.38}$$

4.3.3 *Distribution of Sample Mean with Unknown Variance*

Consider the mean value of a sample of N independent observations from a random variable x, as given by Equation (4.31). If the variable x is normally distributed with a mean value of μ_x and an unknown variance, it is seen from Equations (4.21) and (4.37) that

$$\frac{(\bar{x}-\mu_x)}{s/\sqrt{N}} = \frac{\sigma_x z/\sqrt{N}}{\sqrt{\sigma_x^2\chi_n^2/n}/\sqrt{N}} = \frac{z}{\sqrt{\chi_n^2/n}} = t_n$$

where t_n has a Student t distribution with $n = N - 1$ degrees of freedom, as defined in Secton 4.2.3. Hence the sampling distribution of the sample mean $\bar{x}$ when σ_x^2 is unknown is given by

$$\frac{(\bar{x}-\mu_x)\sqrt{N}}{s} = t_n \qquad n = N-1 \tag{4.39}$$

It follows that a probability statement concerning future values of the sample mean $\bar{x}$ may be made as follows.

$$\text{Prob}\left[\bar{x} > \left(\frac{st_{n;\alpha}}{\sqrt{N}} + \mu_x\right)\right] = \alpha \tag{4.40}$$

4.3.4 *Distribution of Ratio of Two Sample Variances*

Consider the variances of two samples, one consisting of N_x independent observations of a random variable x, and the other consisting of N_y observations of a random variable y, as given by Equation (4.36). If the variable x is normally distributed with a mean value of μ_x and a variance of σ_x^2, and the variable y is normally distributed with a mean value of μ_y and a variance σ_y^2, it is seen from Equations (4.26) and (4.37) that

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} = \frac{\sigma_x^2\chi_n^2/n_x\sigma_x^2}{\sigma_y^2\chi_n^2/n_y\sigma_y^2} = F_{n_x,n_y}$$

where F_{n_x,n_y} has an F distribution with $n_x = N_x - 1$ and $n_y = N_y - 1$ degrees of freedom, as defined in Section 4.2.4. Hence the sampling distribution of the ratio of the sample variances s_x^2 and s_y^2 is given by

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} = F_{n_x,n_y} \qquad \begin{matrix} n_x = N_x - 1 \\ n_y = N_y - 1 \end{matrix} \tag{4.41}$$

It follows that a probability statement concerning future values of the ratio of the sample variances s_x^2 and s_y^2 may be made as follows.

$$\text{Prob}\left[\frac{s_x^2}{s_y^2} > \frac{\sigma_x^2}{\sigma_y^2}F_{n_x,n_y;\alpha}\right] = \alpha \tag{4.42}$$

Note that if the two samples are obtained from the same random variable, $x = y$, then Equation (4.41) reduces to

$$\frac{s_1^2}{s_2^2} = F_{n_1,n_2} \qquad \begin{matrix} n_1 = N_1 - 1 \\ n_2 = N_2 - 1 \end{matrix} \tag{4.43}$$

4.4 CONFIDENCE INTERVALS

The use of sample values as estimators for parameters of random variables is discussed in Section 4.1. However, those procedures result only in point estimates for a parameter of interest; no indication is provided as to how closely a sample value estimates the parameter. A more meaningful procedure for estimating parameters of random variables involves the estimation of an interval, as opposed to a single point value, which will include the parameter being estimated with a known degree of uncertainty. For example, consider the case where the sample mean $\bar{x}$ computed from N independent observations of a random variable x is being used as an estimator for the mean value μ_x. It is usually more desirable to estimate μ_x in terms of some interval, such as $\bar{x} \pm d$, where there is a specified uncertainty that μ_x falls within that interval. Such intervals can be established if the sampling distribution of the estimator in question is known.

Continuing with the example of a mean value estimate, it is shown in Section 4.3 that probability statements can be made concerning the value of a sample mean $\bar{x}$ as follows.

$$\text{Prob}\left[z_{1-\alpha/2} < \frac{(\bar{x}-\mu_x)\sqrt{N}}{\sigma_x} \le z_{\alpha/2}\right] = 1 - \alpha \tag{4.44}$$

The above probability statement is technically correct *before* the sample has been collected and $\bar{x}$ has been computed. After the sample has been collected, however, the value of $\bar{x}$ is a fixed number rather than a random variable. Hence it can be argued that the probability statement in Equation (4.44) no longer applies since the quantity $(\bar{x}-\mu_x)\sqrt{N}/\sigma_x$ either *does* or *does not* fall within the indicated limits. In other words, after a sample has been collected, a technically correct probability statement would be as follows.

$$\text{Prob}\left[z_{1-\alpha/2} < \frac{(\bar{x}-\mu_x)\sqrt{N}}{\sigma_x} \le z_{\alpha/2}\right] = \begin{cases} 0 \\ 1 \end{cases} \tag{4.45}$$

Whether the correct probability is zero or unity is usually not known. However, as the value of α becomes small (as the interval between $z_{1-\alpha/2}$ and $z_{\alpha/2}$ becomes wide), one would tend to guess that the probability is more likely to be unity than zero. In slightly different terms, if many different

samples were repeatedly collected and a value of $\bar{x}$ were computed for each sample, one would tend to expect the quantity in Equation (4.45) to fall within the noted interval for about $1 - \alpha$ of the samples. In this context, a statement can be made about an interval within which one would expect to find the quantity $(\bar{x} - \mu_x)\sqrt{N}/\sigma_x$ with a small degree of uncertainty. Such statements are called *confidence statements*. The interval associated with a confidence statement is called a *confidence interval*. The degree of trust associated with the confidence statement is called the *confidence coefficient*.

For the case of the mean value estimate, a confidence interval can be established for the mean value μ_x based upon the sample value $\bar{x}$ by rearranging terms in Equation (4.45) as follows.

$$\left[\bar{x} - \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}} \leq \mu_x < \bar{x} + \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}}\right] \tag{4.46a}$$

Furthermore, if σ_x is unknown, a confidence interval can still be established for the mean value μ_x based upon the sample values $\bar{x}$ and s by rearranging terms in Equation (4.39) as follows.

$$\left[\bar{x} - \frac{s t_{n;\alpha/2}}{\sqrt{N}} \leq \mu_x < \bar{x} + \frac{s t_{n;\alpha/2}}{\sqrt{N}}\right] \qquad n = N - 1 \tag{4.46b}$$

Equation (4.46) uses the fact that $z_{1-\alpha/2} = -z_{\alpha/2}$ and $t_{n;1-\alpha/2} = -t_{n;\alpha/2}$. The confidence coefficient associated with the intervals is $1 - \alpha$. Hence the confidence statement would be as follows: "The true mean value μ_x falls within the noted interval with a confidence coefficient of $1 - \alpha$," or, in more common terminology, "with a confidence of $100(1 - \alpha)$ percent." Similar confidence statements can be established for any parameter estimates where proper sampling distributions are known. For example, from Equation (4.37), a $1 - \alpha$ confidence interval for the variance σ_x^2 based upon a sample variance s^2 from a sample of size N is

$$\left[\frac{n s^2}{\chi^2_{n;\alpha/2}} \leq \sigma_x^2 < \frac{n s^2}{\chi^2_{n;1-\alpha/2}}\right] \qquad n = N - 1 \tag{4.47}$$

Example 4.1. **Illustration of Confidence Intervals.** Assume a sample of $N = 31$ independent observations are collected from a normally distributed random variable x with the folllowing results:

60	61	47	56	61	63
65	69	54	59	43	61
55	61	56	48	67	65
60	58	57	62	57	58
53	59	58	61	67	62
54					

Determine a 90 percent confidence interval for the mean value and variance of the random variable x.

From Equation (4.46), a $1 - \alpha$ confidence interval for the mean value μ_x based on the sample mean $\bar{x}$ and the sample variance s^2 for a sample size of $N = 31$ is given by

$$\left[\left(\bar{x} - \frac{s t_{30;\alpha/2}}{\sqrt{31}}\right) \leq \mu_x < \left(\bar{x} + \frac{s t_{30;\alpha/2}}{\sqrt{31}}\right)\right]$$

From Table A.4, for $\alpha = 0.10$, $t_{30;\alpha/2} = t_{30;0.05} = 1.697$, so the interval reduces to

$$[\bar{x} - 0.3048s) \leq \mu_x < (\bar{x} + 0.3048s)]$$

From Equation (4.47), a $1 - \alpha$ confidence interval for the variance σ_x^2 based on the sample variance s^2 for a sample size of $N = 31$ is given by

$$\left[\frac{30s^2}{\chi^2_{30;\alpha/2}} \leq \sigma_x^2 < \frac{30s^2}{\chi^2_{30;1-\alpha/2}}\right]$$

From Table A.3, for $\alpha = 0.10$, $\chi^2_{30;\alpha/2} = \chi^2_{30;0.05} = 43.77$ and $\chi^2_{30;1-\alpha/2} = \chi^2_{30;0.95} = 18.49$, so the interval reduces to

$$[0.6854s^2 \leq \sigma_x^2 < 1.622s^2]$$

It now remains to calculate the sample mean and variance, and substitute these values into the interval statements. From Equation (4.3), the sample mean is

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 58.61$$

From Equation (4.12), the sample variance is

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N-1}\left\{\sum_{i=1}^{N} x_i^2 - N(\bar{x})^2\right\} = 33.43$$

Hence the 90 percent confidence intervals for the mean value and variance of the random variable x are as follows.

$$[56.85 \leq \mu_x < 60.37]$$
$$[22.91 \leq \sigma_x^2 < 54.22]$$