CHAPTER 4

Statistical Principles

Beyond the basic ideas of probability theory discussed in Chapter 3, the measurement and the analysis of random data involve uncertainties and estimation errors that must be evaluated by statistical techniques. This chapter reviews and illustrates various statistical ideas that have wide applications to commonly occurring data evaluation problems. The intent is to provide the reader with a minimum background in terminology and certain techniques of engineering statistics that are relevant to discussions in later chapters. More detailed treatments of applied statistics with engineering applications are available from Refs 1–3.

4.1 SAMPLE VALUES AND PARAMETER ESTIMATION

Consider a random variable x, as defined in Section 3.1, where the index k of the sample space is omitted for simplicity in notation. Further consider the two basic parameters of x that specify its central tendency and dispersion, namely the mean value and variance, respectively. From Equations (3.8) and (3.11), the mean value and variance are given by

$$\mu_x = E[x] = \int_{-\infty}^{\infty} xp(x)dx \tag{4.1}$$

$$\sigma_x^2 = E[(x-\mu_x)^2] = \int_{-\infty}^{\infty} (x-\mu_x)^2 p(x)dx \tag{4.2}$$

where $p(x)$ is the probability density function of the variable x. These two parameters of x cannot, of course, be precisely determined in practice because an exact knowledge of the probability density function will not generally be available. Hence, one must be content with estimates of the mean value and variance based on a finite number of observed values.

Random Data: Analysis and Measurement Procedures, Fourth Edition. By Julius S. Bendat and Allan G. Piersol

One possible method (there are others) for estimating the mean value and variance of x based on N independent observations would be as follows:

$$\bar{x} = \hat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.3}$$

$$s_b^2 = \hat{\sigma}_x^2 = \frac{1}{N}\sum_{i=1}^{N} (x-\bar{x})^2 \tag{4.4}$$

Here, $\bar{x}$ and s_b^2 are the *sample mean* and *sample variance*, respectively. The hats ($\hat{}$) over $\hat{\mu}_x$ and $\hat{\sigma}_x^2$ indicate that these sample values are being used as *estimators* for the mean value and variance of x. The subscript on s_b^2 means that this is a *biased* variance estimate (to be discussed later). The number of observations used to compute the estimates (sample values) is called the *sample size*.

The specific sample values in Equations (4.3) and (4.4) are not the only quantities that might be used to estimate the mean value and variance of a random variable x. For example, reasonable estimates of the mean value and the variance would also be obtained by dividing the summations in Equations (4.3) and (4.4) by $N-1$ instead of N. Estimators are never clearly right or wrong since they are defined somewhat arbitrarily. Nevertheless, certain estimators can be judged as being "good" estimators or "better" estimators than others.

Three principal factors can be used to establish the quality or "goodness" of an estimator. First, it is desirable that the expected value of the estimator be equal to the parameter being established. That is,

$$E[\hat{\phi}] = \phi \tag{4.5}$$

where $\hat{\phi}$ is an estimator for the parameter ϕ, If this is true, the estimator is said to be *unbiased*. Second, it is desirable that the mean square error of the estimator be smaller than for other possible estimators. That is,

$$E[(\hat{\phi}_1-\phi)^2] \leq E[(\hat{\phi}_i-\phi)^2] \tag{4.6}$$

where $\hat{\phi}_1$ is the estimator of interest and $\hat{\phi}_i$ is any other possible estimator. If this is true, the estimator is said to be more *efficient* than other possible estimators. Third, it is desirable that the estimator approach the parameter being estimated with a probability approaching unity as the sample size becomes large. That is, for any $\varepsilon > 0$,

$$\lim_{N\to\infty} \text{Prob}[|\hat{\phi}-\phi| \geq \varepsilon] = 0 \tag{4.7a}$$

If this is true, the estimator is said to be *consistent*. It follows from the Chebyshev inequality of Equation (3.23) that a sufficient (but not necessary) condition to meet the requirements of Equation (4.7a) is given by

$$\lim_{N\to\infty} E[(\hat{\phi}-\phi)^2] = 0 \tag{4.7b}$$

Note that the requirements stated in Equation (4.7) are simply convergence requirements in (a) the probability and (b) the mean square sense, as defined later in Section 5.3.4.

Consider the example of the mean value estimator given by Equation (4.3). The expected value of the sample mean $\bar{x}$ is

$$E[\bar{x}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}E\left[\sum_{i=1}^{N} x_i\right] = \frac{1}{N}(N\mu_x) = \mu_x \tag{4.8}$$

Hence, from Equation (4.5), the estimator $\hat{\mu}_x = \bar{x}$ is unbiased. The mean square error of the sample mean $\bar{x}$ is given by

$$E\left[(\bar{x}-\mu_x)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i-\mu_x\right)^2\right] = \frac{1}{N^2}E\left[\left(\sum_{i=1}^{N}(x_i-\mu_x)\right)^2\right]$$

From Section 3.2.1, since the observations x_i are independent, the cross product terms in the last expression will have an expected value of zero. It then follows that

$$E\left[(\bar{x}-\mu_x)^2\right] = \frac{1}{N^2}E\left[\sum_{i=1}^{N}(x_i-\mu_x)^2\right] = \frac{1}{N^2}(N\sigma_x^2) = \frac{\sigma_x^2}{N} \tag{4.9}$$

Hence, from Equation (4.7b), the estimator $\hat{\mu}_x = \bar{x}$ is consistent. It can be shown that the estimator is also efficient.

Now consider the example of the variance estimator given by Equation (4.4). The expected value of the sample variance s_b^2 is

$$E[s_b^2] = E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right] = \frac{1}{N}E\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]$$

However,

$$\begin{aligned}
\sum_{i=1}^{N}(x_i-\bar{x})^2 &= \sum_{i=1}^{N}(x_i-\mu_x+\mu_x-\bar{x})^2 \\
&= \sum_{i=1}^{N}(x_i-\mu_x)^2-2(\bar{x}-\mu_x)\sum_{i=1}^{N}(x_i-\mu_x)+\sum_{i=1}^{N}(\bar{x}-\mu_x)^2 \\
&= \sum_{i=1}^{N}(x_i-\mu_x)^2-2(\bar{x}-\mu_x)N(\bar{x}-\mu_x)+N(\bar{x}-\mu_x)^2 \\
&= \sum_{i=1}^{N}(x_i-\mu_x)^2-N(\bar{x}-\mu_x)^2
\end{aligned} \tag{4.10}$$

Because $E[(x_i-\mu_x)^2] = \sigma_x^2$ and $E[(\bar{x}-\mu_x)^2] = \sigma_x^2/N$, it follows that

$$E[s_b^2] = \frac{1}{N}(N\sigma_x^2-\sigma_x^2) = \frac{(N-1)}{N}\sigma_x^2 \tag{4.11}$$

Hence, the estimator $\hat{\sigma}_x^2 = s_b^2$ is *biased*. Although the sample variance s_b^2 is a biased estimator for σ_x^2, it is a consistent and an efficient estimator.

From the results in Equation (4.11), it is clear that an unbiased estimator for σ_x^2 may be obtained by computing a slightly different sample variance as follows:

$$s^2 = \hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{4.12}$$

The quantity defined in Equation (4.12) is an *unbiased* estimator for σ_x^2. For this reason, the sample variance defined in Equation (4.12) is often considered a "better" estimator than the sample variance given by Equation (4.4). The sample variance defined in Equation (4.12) will be used henceforth as an estimator for the variance of a random variable.

4.2 IMPORTANT PROBABILITY DISTRIBUTION FUNCTIONS

Examples of several theoretical probability distribution functions are given in Chapter 3. The most important of these distribution functions from the viewpoint of applied statistics is the Gaussian (*normal*) distribution. There are three other distribution functions associated with normally distributed random variables that have wide applications as statistical tools. These are the χ^2 distribution, the t distribution, and the F distribution. Each of these three, along with the normal distribution, will now be defined and discussed. Applications for each as an analysis tool will be covered in later sections.

4.2.1 Gaussian (Normal) Distribution

The probability density and distribution functions of a Gaussian distributed random variable x are defined by Equations (3.47) and (3.48) in Section 3.3. As noted in that section, a more convenient form of the Gaussian distribution is obtained by using the standardized variable z given by

$$z = \frac{x - \mu_x}{\sigma_x} \tag{4.13}$$

When Equation (4.13) is substituted into Equations (3.47) and (3.48), standardized Gaussian density and distribution functions with zero mean and unit variance ($\mu_z = 0$; $\sigma_x^2 = 1$) are obtained as given by

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{4.14a}$$

$$P(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\xi^2/2} d\xi \tag{4.14b}$$

The standardized Gaussian (normal) probability density and distribution functions in Equation (4.14) are plotted in Figure 3.5.

It is desirable for later applications to denote the value of z that corresponds to a specific probability distribution function value of $P(z) = 1 - \alpha$ by z_α. That is,

$$P(z_\alpha) = \int_{-\infty}^{z_\alpha} p(z)dz = \text{Prob}[z \leq z_\alpha] = 1-\alpha \tag{4.15a}$$

or

$$1-P(z_\alpha) = \int_{z_\alpha}^{\infty} p(z)dz = \text{Prob}[z > z_\alpha] = \alpha \tag{4.15b}$$

The value of z_α that satisfies Equation (4.15) is called the *100α percentage point* of the normal distribution. A limited tabulation of percentage points for the normal distribution is presented in Table A.2.

4.2.2 Chi-Square Distribution

Let $z_1, z_2, z_3, \ldots, z_n$ be n independent random variables, each of which has a Gaussian distribution with zero mean and unit variance. Let a new random variable be defined as

$$\chi_n^2 = z_1^2 + z_2^2 + z_3^2 + \cdots + z_n^2 \tag{4.16}$$

The random variable χ_n^2 is the chi-square variable with n degrees of freedom. The number of *degrees of freedom n* represents the number of independent or "free" squares entering into the expression. From Ref. 3, the probability density function of χ_n^2 is given by

$$p(\chi^2) = [2^{n/2}\Gamma(n/2)]^{-1} e^{-\chi^2/2}(\chi^2)^{(n/2)-1} \quad x^2 \geq 0 \tag{4.17}$$

where $\Gamma(n/2)$ is the gamma function. The corresponding distribution function of χ_n^2, given by the integral of Equation (4.17) from $-\infty$ to a specific value of χ_n^2, is called the *chi-square distribution with n degrees of freedom*. The 100α percentage point of the χ^2 distribution will be denoted by $\chi_{n;\alpha}^2$. That is,

$$\int_{\chi_{n;\alpha}^2}^{\infty} p(\chi^2)d\chi^2 = \text{Prob}[\chi_n^2 > \chi_{n;\alpha}^2] = \alpha \tag{4.18}$$

The mean value and variance of the variable χ_n^2 are

$$E[\chi_n^2] = \mu_{\chi^2} = n \tag{4.19}$$

$$E[(\chi_n^2 - \mu_{\chi^2})^2] = \sigma_{\chi^2}^2 = 2n \tag{4.20}$$

A limited tabulation of percentage points for the chi-square distribution function is presented in Table A.3.

Several features of the chi-square distribution should be noted. First, the chi-square distribution is a special case of the more general gamma function [2]. Second, the square root of chi-square with two degrees of freedom ($\sqrt{\chi_2^2}$) constitutes an important

case called the *Rayleigh distribution function* [3]. The Rayleigh distribution has wide applications to two-dimensional target problems and is also the limiting distribution function of both the envelope (see Section 3.4) and the peak values (see Section 5.5) for narrow bandwidth Gaussian random data as the bandwidth approaches zero. Third, a chi-square distribution approaches a Gaussian distribution as the number of degrees of freedom becomes large. Specifically, for $n > 100$, the quantity $\sqrt{2\chi_n^2}$ is distributed approximately as a Gaussian variable with a mean of $\mu = \sqrt{2n-1}$ and a variance of $\sigma^2 = 1$ [Ref. 1].

4.2.3 The *t* Distribution

Let y and z be independent random variables such that y has a χ_n^2 distribution function and z has a Gaussian distribution function with zero mean and unit variance. Let a new random variable be defined as

$$t_n = \frac{z}{\sqrt{y/n}} \tag{4.21}$$

The random variable t_n is Student's t variable with n degrees of freedom. From Ref. 2, the probability density function of t_n is given by

$$p(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)}\left[1+\frac{t^2}{n}\right]^{-(n+1)/2} \tag{4.22}$$

The corresponding distribution function of t_n, given by the integral of Equation (4.22) from $-\infty$ to a specific value of t_n, is called the *t distribution with n degrees of freedom*. The 100α percentage point of the t distribution will be denoted by $t_{n;\alpha}$. That is,

$$\int_{t_{n;\alpha}}^{\infty} p(t)dt = \text{Prob}[t_n > t_{n;\alpha}] = \alpha \tag{4.23}$$

The mean value and variance of the variable t_n are

$$E[t_n] = \mu_t = 0 \quad \text{for } n > 1 \tag{4.24}$$

$$E\left[(t_n-\mu_t)^2\right] = \sigma_t^2 = \frac{n}{n-2} \quad \text{for } n > 2 \tag{4.25}$$

A limited tabulation of percentage points for the t distribution function is presented in Table A.4. It should be noted that the t distribution approaches a standardized Gaussian distribution as the number of degrees of freedom n becomes large.

4.2.4 The *F* Distribution

Let y_1 and y_2 be independent random variables such that y_1 has a χ^2 distribution function with n_1 degrees of freedom and y_2 has a χ^2 distribution function with n_2 degrees of freedom. Let a new random variable be defined as

$$F_{n_1,n_2} = \frac{y_1/n_1}{y_2/n_2} = \frac{y_1 n_2}{y_2 n_1} \tag{4.26}$$

The random variable F_{n_1,n_2} is the F variable with n_1 and n_2 degrees of freedom From Ref. 3, the probability density function of F_{n_1,n_2} is given by

$$p(F) = \frac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2}F^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1F/n_2)]^{(n_1+n_2)/2}} \quad F \geq 0 \tag{4.27}$$

The corresponding distribution function of F_{n_1,n_2}, given by the integral of Equation (4.27) from $-\infty$ to a specific value of F_{n_1,n_2}, is called the *F distribution with n_1 and n_2 degrees of freedom*. The 100α percentage point of the F distribution will be denoted by $F_{n_1,n_2;\alpha}$. That is,

$$\int_{F_{n_1,n_2;\alpha}}^{\infty} p(F)dF = \text{Prob}[F_{n_1,n_2} > F_{n_1,n_2;\alpha}] = \alpha \tag{4.28}$$

The mean value and variance of F_{n_1,n_2} are

$$E[F_{n_1,n_2}] = \mu_F = \frac{n_2}{n_2-2} \quad \text{for } n_2 > 2 \tag{4.29}$$

$$E[(F_{n_1,n_2}-\mu_F)^2] = \sigma_F^2 = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)} \quad \text{for } n_2 > 4 \tag{4.30}$$

A limited tabulation of percentage points for the F distribution function is presented in Tables A.5(a), A.5(b), and A.5(c). It should be noted that the statistic t_n^2, the square of the variable defined in Equation (4.21), has an F distribution with $n_1 = 1$ and $n_2 = n$ degrees of freedom.

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 that 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 4.4–4.8.

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 3.3.1, 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 (see Section 3.1.1), 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 shown in Ref. 1 that

$$\sum_{i=1}^{N}(x_i-\bar{x})^2 = \sigma_n^2\chi_n^2 \quad n = N-1$$

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

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

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

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

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 consists of N_x independent observations of a random variable x, and the other consists of N_y independent 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 distributions 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:

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

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:

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

Whether the correct probability is zero or unity is usually not known. As the value of α becomes small (as the interval between $z_{1-\alpha/2}$ and $z_{\alpha/2}$ becomes wide), however, 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 on 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 on the sample values $\bar{x}$ and s by rearranging terms in Equation (4.39) as follows:

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

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 $100(1-\alpha)\%$. 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 on a sample variance s^2 from a sample of size N is

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

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 following 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% confidence interval for the mean value and variance of the random variable x.

From Equation (4.46b), 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

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

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 the 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% confidence intervals for the mean value and variance of the random variable x are as follows:

$$[56.85 \leq \mu_x < 60.37]$$

$$[22.91 \leq \sigma_x^2 < 54.22]$$

4.5 HYPOTHESIS TESTS

Consider the case where a given estimator $\hat{\phi}$ is computed from a sample of N independent observations of a random variable x. Assume there is reason to believe that the true parameter ϕ being estimated has a specific value ϕ_0. Now, even if $\phi = \phi_0$, the sample value $\hat{\phi}$ will probably not come out exactly equal to ϕ_0 because of the sampling variability associated with $\hat{\phi}$. Hence, the following question arises. If it is hypothesized that $\phi = \phi_0$, how much difference between $\hat{\phi}$ and ϕ_0 must occur before the hypothesis should be rejected as being invalid? This question can be answered in statistical terms by considering the probability of any noted difference between $\hat{\phi}$ and ϕ_0 based upon the sampling distribution of $\hat{\phi}$. If the probability of a given difference is small, the difference would be considered significant and the hypothesis that $\phi = \phi_0$ would be rejected. If the probability of a given difference is not small, the difference would be accepted as normal statistical variability and the hypothesis that $\phi = \phi_0$ would be accepted.

The preceding discussion outlines the simplest form of a statistical procedure called hypothesis testing. To clarify the general technique, assume that a sample value $\hat{\phi}$, which is an estimate of a parameter ϕ, has a probability density function of $p(\hat{\phi})$. Now, if a hypothesis that $\phi = \phi_0$ is true, then $p(\hat{\phi})$ would have a mean value of ϕ_0 as illustrated in Figure 4.1. The probability that $\hat{\phi}$ would fall below the lower level $\phi_{1-\alpha/2}$ is

$$\text{Prob}\left[\hat{\phi} \leq \phi_{1-\alpha/2}\right] = \int_{-\infty}^{\phi_{1-\alpha/2}} p(\hat{\phi})d\hat{\phi} = \frac{\alpha}{2} \tag{4.48a}$$

The probability that $\hat{\phi}$ would fall above the upper value $\phi_{\alpha/2}$ is

$$\text{Prob}\left[\hat{\phi} > \phi_{\alpha/2}\right] = \int_{\phi_{\alpha/2}}^{\infty} p(\hat{\phi})d\hat{\phi} = \frac{\alpha}{2} \tag{4.48b}$$

Hence, the probability that $\hat{\phi}$ would be outside the range between $\phi_{1-\alpha/2}$ and $\phi_{\alpha/2}$ is α. Now let α be small so that it is very unlikely that $\hat{\phi}$ would fall outside the range

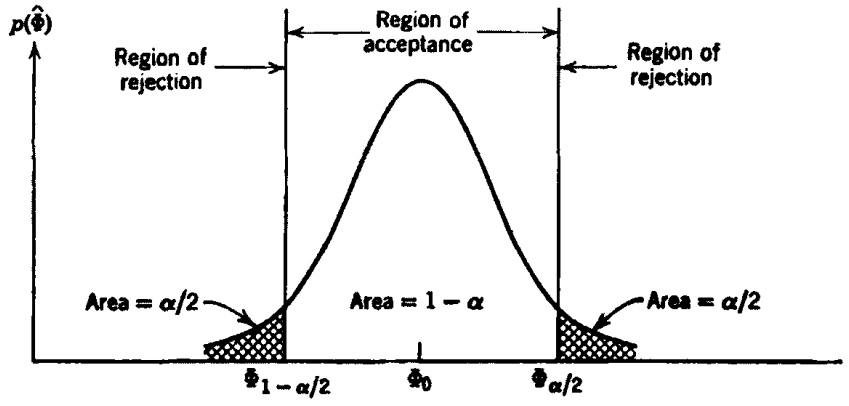


Figure 4.1 Acceptance and rejection regions for hypothesis tests.

between $\phi_{1-\alpha/2}$ and $\phi_{\alpha/2}$. If a sample were collected and a value of $\hat{\phi}$ were computed that in fact fell outside the range between $\phi_{1-\alpha/2}$ and $\phi_{\alpha/2}$, there would be a strong reason to question the original hypothesis that $\phi = \phi_0$ because such a value for $\hat{\phi}$ would be very unlikely if the hypothesis were true. Hence the hypothesis that $\phi = \phi_0$ would be rejected. On the other hand, if the value for $\hat{\phi}$ fell within the range between $\phi_{1-\alpha/2}$ and $\phi_{\alpha/2}$, there would be no strong reason to question the original hypothesis. Hence the hypothesis that $\phi = \phi_0$ would be accepted.

The small probability α used for the hypothesis test is called the *level of significance* of the test. The range of values of $\hat{\phi}$ for which the hypothesis will be rejected is called the *region of rejection* or *critical region*. The range of values of $\hat{\phi}$ for which the hypothesis will be accepted is called the *region of acceptance*. The simple hypothesis test outlined above is called a *two-sided test* because, if the hypothesis is not true, the value of ϕ could be either greater or less than ϕ_0. Hence, it is necessary to test for significant differences between ϕ and ϕ_0 in both directions. In other cases, a *one-sided test* might be sufficient. For example, let it be hypothesized that $\phi \geq \phi_0$. For this case, the hypothesis would be false only if ϕ were less than ϕ_0. Thus, the test would be performed using the lower side of the probability density function $p(\hat{\phi})$.

Two possible errors can occur when a hypothesis test is performed. First, the hypothesis might be rejected when in fact it is true. This possible error is called a *Type I error*. Second, the hypothesis might be accepted when in fact it is false. This possible error is called a *Type II error*. From Figure 4.1, a Type I error would occur if the hypothesis were true and $\hat{\phi}$ fell in the region of rejection. It follows that the probability of a Type I error is equal to α, the level of significance of the test.

In order to establish the probability of a Type II error, it is necessary to specify some deviation of the true parameter ϕ from the hypothesized parameter ϕ_0 that one desires to detect. For example, assume that the true parameter actually has a value of either $\phi = \phi_0 + d$ or $\phi = \phi_0 - d$, as illustrated in Figure 4.2. If it is hypothesized that $\phi = \phi_0$ when in fact $\phi = \phi_0 \pm d$, the probability that $\hat{\phi}$ would fall inside the acceptance region between $\phi_{1-\alpha/2}$ and $\phi_{\alpha/2}$ is β. Hence, the probability of a Type II error is β for detecting a difference of $\pm d$ from the hypothesized value ϕ_0.

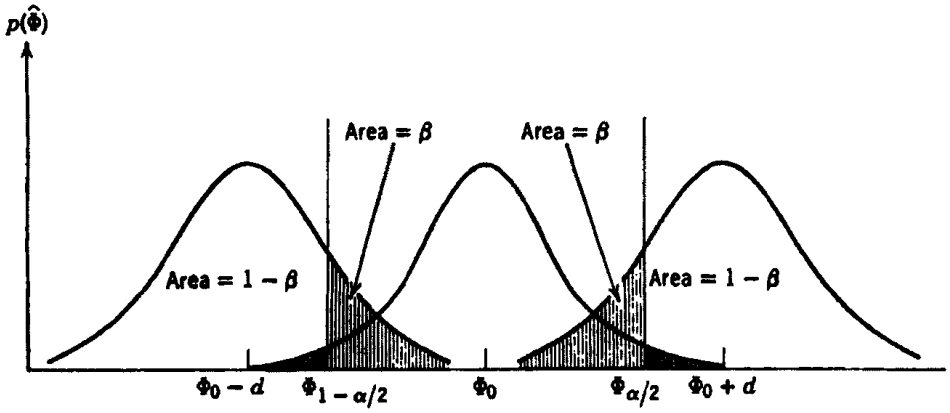


Figure 4.2 Type II error regions for hypothesis tests.

The probability $1-\beta$ is called the *power of the test*. Clearly, for any given sample size N, the probability of a Type I error can be reduced by reducing the level of significance α. However, this will increase the probability β of a Type II error (reduce the power of the test). The only way to reduce both α and β is to increase the sample size N for the estimate $\hat{\phi}$. These ideas form the basis for selecting the necessary sample sizes for statistical experiments.

Example 4.2. **Illustration of Hypothesis Test Design.** Assume there is reason to believe that the mean value of a random variable x is $\mu_x = 10$. Further assume that the variance of x is known to be $\sigma_x^2 = 4$. Determine the proper sample size to test the hypothesis that $\mu_x = 10$ at the 5% level of significance, where the probability of a Type II error is to be 5% for detecting a difference of 10% from the hypothesized value. Determine the region of acceptance to be used for the test.

An unbiased estimate for μ_x is given by the sample mean value $\bar{x}$ as defined in Equation (4.3). The appropriate sampling distribution of $\bar{x}$ is given by Equation (4.34) as

$$\bar{x} = \frac{\sigma_x}{\sqrt{N}} z + \mu_x$$

where z is normally distributed with zero mean and unit variance. Note that this sampling distribution of $\bar{x}$ is precise if x is normally distributed and is still a good approximation if x is not normally distributed.

The upper and lower limits of the acceptance region for the hypothesis test are as follows:

$$\text{Upper limit} = \frac{\sigma_x}{\sqrt{N}} z_{\alpha/2} + \mu_x$$

$$\text{Lower limit} = \frac{\sigma_x}{\sqrt{N}} z_{1-\alpha/2} + \mu_x$$

Now if the true mean value were in fact $\mu'_x = \mu_x \pm d$, a Type II error would occur with probability β if the sample value $\bar{x}$ fell below the upper limit or above the lower limit. In terms of the sampling distributions of $\bar{x}$ with a mean value $\mu'_x = \mu_x + d$ or $\mu'_x = \mu_x - d$,

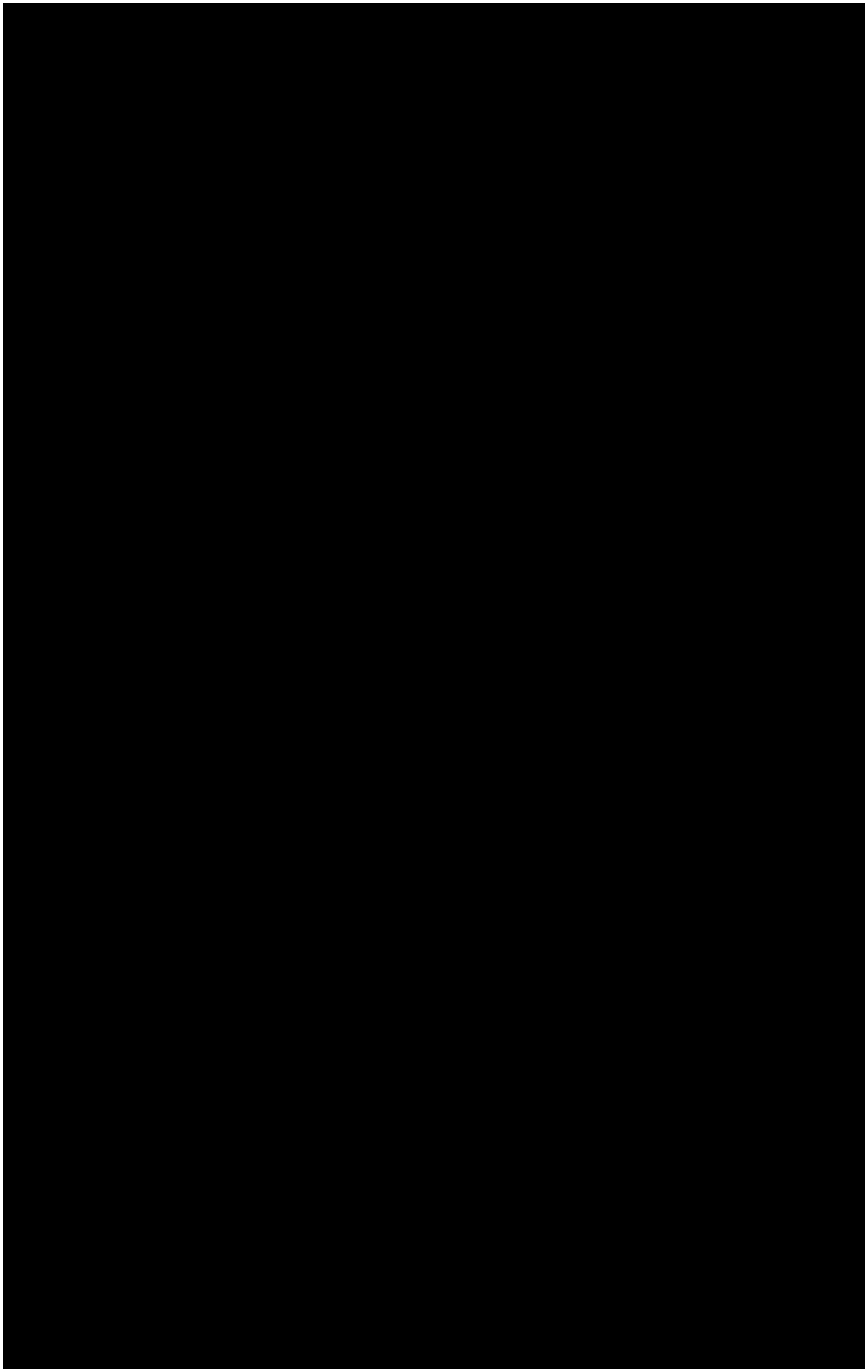
$$\text{Upper limit} = \frac{\sigma_x}{\sqrt{N}} z_{1-\beta} + \mu_x + d$$

$$\text{Lower limit} = \frac{\sigma_x}{\sqrt{N}} z_{\beta} + \mu_x - d$$

Hence the following equalities apply:

$$\frac{\sigma_x}{\sqrt{N}} z_{\alpha/2} + \mu_x = \frac{\sigma_x}{\sqrt{N}} z_{1-\beta} + \mu_x + d$$

$$\frac{\sigma_x}{\sqrt{N}} z_{1-\alpha/2} + \mu_x = \frac{\sigma_x}{\sqrt{N}} z_{\beta} + \mu_x - d$$



$$X^2 = \sum_{i=1}^{K} \frac{(f_i - F_i)^2}{F_i} \tag{4.49}$$

It is shown in Ref. 2 that the distribution of X^2 in Equation (4.49) is approximately the same as for χ_n^2 discussed in Section 4.2.2. The number of degrees of freedom n in this case is equal to K minus the number of different independent linear restrictions imposed on the observations. There is one such restriction due to the fact that the frequency in the last class interval is determined once the frequencies in the first $K-1$ class intervals are known. If the comparison is made by fitting the expected theoretical density function to the frequency histogram for the observed data, then one additional constraint results from each independent parameter of the theoretical density function that must be computed to make the fit. For example, if the expected theoretical density function is a normal density function with unknown mean and variance, then two additional constraints are involved, because two parameters (a mean and a variance) must be computed to fit a normal density function. Hence, for the common case where the chi-square goodness-of-fit test is used as a test for normality, the number of degrees of freedom for X^2 in Equation (4.49) is $n = K - 3$.

Having established the proper degrees of freedom for X^2, a hypothesis test may be performed as follows. Let it be hypothesized that the variable x has a probability density function of $p(x) = p_0(x)$. After grouping the sampled observations into K class intervals and computing the expected frequency for each interval assuming $p(x) = p_0(x)$, compute X^2 as indicated in Equation (4.49). Because any deviation of $p(x)$ from $p_0(x)$ will cause X^2 to increase, an one-sided (upper tail) test is used. The region of acceptance is

$$X^2 \leq \chi_{n;\alpha}^2 \tag{4.50}$$

where the value of $\chi_{n;\alpha}^2$ is available from Table A.3. If the sample value X^2 is greater than $\chi_{n;\alpha}^2$, the hypothesis that $p(x) = p_0(x)$ is rejected at the α level of significance. If X^2 is less than or equal to $\chi_{n;\alpha}^2$, the hypothesis is accepted at the α level of significance.

There are two basic ways to apply the chi-square goodness-of-fit test. The first way is to select class intervals in a manner that will provide equal expected frequencies within each interval. Excluding a uniform distribution hypothesis, this procedure will result in different interval widths from one class interval to another. The second way is to select class intervals of equal width. Again, except for the uniform distribution hypothesis, this procedure will result in different expected frequencies from one class interval to another. Chi-square tests for normality are usually performed using the constant interval width approach. Given sample data with a standard deviation of s, a class interval width of $\Delta x \simeq 0.4s$ is often used. A more fundamental requirement is that the expected frequencies in all class intervals must be sufficiently large to make Equation (4.49) an acceptable approximation to χ_n^2. A common recommendation is that $F_i > 3$ in all intervals. In a normality test where the expected frequencies diminish on the tails of the distribution, this requirement is complied with by letting the first and last intervals extend to $-\infty$ and $+\infty$, respectively, such that $F_1, F_K > 3$.

Table 4.1 Sample Observations Arranged in Increasing Order

−7.6	−3.8	−2.5	−1.6	−0.7	0.2	1.1	2.0	3.4	4.6
−6.9	−3.8	−2.5	−1.6	−0.7	0.2	1.1	2.1	3.5	4.8
−6.6	−3.7	−2.4	−1.6	−0.6	0.2	1.2	2.3	3.5	4.8
−6.4	−3.6	−2.3	−1.5	−0.6	0.3	1.2	2.3	3.6	4.9
−6.2	−3.5	−2.3	−1.5	−0.5	0.3	1.3	2.3	3.6	5.0
−6.1	−3.4	−2.3	−1.4	−0.5	0.3	1.3	2.4	3.6	5.2
−6.0	−3.4	−2.2	−1.4	−0.4	0.4	1.3	2.4	3.7	5.3
−5.7	−3.4	−2.2	−1.2	−0.4	0.4	1.4	2.5	3.7	5.4
−5.6	−3.3	−2.1	−1.2	−0.4	0.5	1.5	2.5	3.7	5.6
−5.5	−3.2	−2.1	−1.2	−0.3	0.5	1.5	2.6	3.7	5.9
−5.4	−3.2	−2.0	−1.1	−0.3	0.6	1.6	2.6	3.8	6.1
−5.2	−3.1	−2.0	−1.1	−0.2	0.6	1.6	2.6	3.8	6.3
−4.8	−3.0	−1.9	−1.0	−0.2	0.7	1.6	2.7	3.9	6.3
−4.6	−3.0	−1.9	−1.0	−0.2	0.8	1.7	2.8	4.0	6.5
−4.4	−2.9	−1.8	−1.0	−0.1	0.9	1.8	2.8	4.2	6.9
−4.4	−2.9	−1.8	−0.9	−0.0	0.9	1.8	2.9	4.2	7.1
−4.3	−2.9	−1.8	−0.9	0.0	1.0	1.8	3.1	4.3	7.2
−4.1	−2.7	−1.7	−0.8	0.1	1.0	1.9	3.2	4.3	7.4
−4.0	−2.6	−1.7	−0.8	0.1	1.1	1.9	3.2	4.4	7.9
−3.8	−2.6	−1.6	−0.7	0.2	1.1	2.0	3.3	4.4	9.0

Example 4.3. **Illustration of Test for Normality.** A sample of $N = 200$ independent observations of the digitized output of a thermal noise generator are presented in Table 4.1. The sample values have been rank ordered from the smallest to largest value for convenience. Test the noise generator output for normality by performing a chi-square goodness-of-fit test at the $\alpha = 0.05$ level of significance.

The calculations required to perform the test are summarized in Table 4.2. For an interval width of $\Delta x = 0.4s$, the standardized values of the normal distribution that define the class interval boundaries are as shown under z_α in the table. These interval boundaries are converted to volts in the next column. From Table A.2, the probability P that a sample value will fall in each class interval is determined using the z_α values. The product of P and the sample size N yields the expected frequency in each interval as listed under F in Table 4.2. Note that the first and last class intervals are selected so that $F > 3$. A total of 12 class intervals result. The observed frequencies are now counted using the interval boundaries in volts as indicated in Table 4.1. The normalized squared discrepancies between the expected and observed frequencies are then calculated and summed to obtain $X^2 = 2.43$. Note that the appropriate degrees of freedom is $n = K - 3 = 9$. The acceptance region for the test is found in Table A.3 to be $X^2 \leq \chi^2_{9;0.05} = 16.92$. Hence, the hypothesis of normality is accepted at the $\alpha = 0.05$ level of significance.

4.5.2 Nonparametric Trend Test

Situations often arise in data analysis where it is desired to establish if a sequence of observations or parameter estimates include an underlying trend. This is partiularly

Table 4.2 Calculations for Goodness-of-Fit Test

Interval Number	Upper Limit of Interval						
	z_α	$x = sz + \bar{x}$	P	$F = NP$	f	$\lvert F-f \rvert$	$\frac{(F-f)^2}{F}$
1	−2.0	−6.36	0.0228	4.5	4	0.5	0.06
2	−1.6	−5.04	0.0320	6.4	8	1.6	0.40
3	−1.2	−3.72	0.0603	12.1	10	2.1	0.36
4	−0.8	−2.40	0.0968	19.4	21	1.6	0.13
5	−0.4	−1.08	0.1327	26.5	29	2.5	0.24
6	0	0.24	0.1554	31.1	31	0.1	0.00
7	0.4	1.56	0.1554	31.1	27	4.1	0.54
8	0.8	2.88	0.1327	26.5	25	1.5	0.08
9	1.2	4.20	0.0968	19.4	20	0.6	0.02
10	1.6	5.52	0.0603	12.1	13	0.9	0.07
11	2.0	6.84	0.0320	6.4	6	0.4	0.03
12	∞	∞	0.0228	4.5	6	1.5	0.50
			1.000	200	200		2.43
$N = 200$	$\bar{x} = 0.24$		$s = 3.30$	$n = K - 3 = 9$			$X^2 = 2.43$

true in the analysis of nonstationary data discussed later in Chapter 12. Because the observations or parameter estimates of interest may have a wide range of probability distribution functions, it is convenient to perform such evaluations with *distribution-free* or *nonparametric* procedures, where no assumption is made concerning the probability distribution of the data being evaluated. One such procedure that is easy to apply and useful for detecting underlying trends in random data records is the *reverse arrangement test.*

Consider a sequence of N observations of a random variable x, where the observations are denoted by x_i, $i = 1, 2, 3, \ldots, N$. Now, count the number of times that $x_i > x_j$ for $i < j$. Each such inequality is called a reverse arrangement. The total number of reverse arrangements is denoted by A.

A general definition for A is as follows. From the set of observations $x_1, x_2, \ldots, x_N$, define

$$h_{ij} = \begin{cases} 1 & \text{if } x_i > x_j \\ 0 & \text{otherwise} \end{cases} \tag{4.51}$$

Then

$$A = \sum_{i=1}^{N-1} A_i \tag{4.52}$$

where

$$A_i = \sum_{j=i+1}^{N} h_{ij} \tag{4.53}$$

For example,

$$A_1 = \sum_{j=2}^{N} h_{1j} \qquad A_2 = \sum_{j=3}^{N} h_{2j} \qquad A_3 = \sum_{j=4}^{N} h_{3j} \quad \text{etc.}$$

To help clarify the meaning of reverse arrangements, consider the following sequence of $N=8$ observations:

$$x_1 = 5, \quad x_2 = 3, \quad x_3 = 8, \quad x_4 = 9, \quad x_5 = 4, \quad x_6 = 1, \quad x_7 = 7, \quad x_8 = 5$$

In the above sequence $x_1 > x_2$, $x_1 > x_5$, and $x_1 > x_6$, which gives $A_1 = 3$ reverse arrangements for x_1. Now, choosing x_2 and comparing it against subsequent observations (i.e., for $i=2$ and $i<j=3,4,\ldots,8$), one notes $x_2 > x_6$ only, so that the number of reverse arrangements for x_2 is $A_2 = 1$. Continuing on, it is seen that $A_3 = 4$, $A_4 = 4$, $A_5 = 1$, $A_6 = 0$, and $A_7 = 1$. The total number of reverse arrangements is, therefore,

$$A = A_1 + A_2 + \cdots + A_7 = 3+1+4+4+1+0+1 = 14$$

If the sequence of N observations is independent observations of the same random variable, then the number of reverse arrangements is a random variable A, with a mean variable and a variance as follows [Ref. 4]:

$$\mu_A = \frac{N(N-1)}{4} \tag{4.54}$$

$$\sigma_A^2 = \frac{2N^3 + 3N^2 - 5N}{72} = \frac{N(2N+5)(N-1)}{72} \tag{4.55}$$

A limited tabulation of 100α percentage points for the distribution function of A is presented in Table A.6.

Example 4.4. **Illustration of Reverse Arrangement Test.** Assume a sequence of $N=20$ observations of a random variable produces results as noted below:

(1) 5.2	(6) 4.0	(11) 5.9	(16) 5.6
(2) 6.2	(7) 3.9	(12) 6.5	(17) 5.2
(3) 3.7	(8) 5.3	(13) 4.3	(18) 3.9
(4) 6.4	(9) 4.0	(14) 5.7	(19) 6.2
(5) 3.9	(10) 4.6	(15) 3.1	(20) 5.0

Test the sequence of $N=20$ observations for a trend at the $\alpha = 0.05$ level of significance. The number of reverse arrangements in the observations is as follows:

$A_1 = 10$	$A_6 = 3$	$A_{11} = 7$	$A_{16} = 3$
$A_2 = 15$	$A_7 = 1$	$A_{12} = 8$	$A_{17} = 2$
$A_3 = 1$	$A_8 = 7$	$A_{13} = 2$	$A_{18} = 0$
$A_4 = 15$	$A_9 = 2$	$A_{14} = 5$	$A_{19} = 1$
$A_5 = 1$	$A_{10} = 3$	$A_{15} = 0$	

The total number of reverse arrangements is $A = 86$.

Let it be hypothesized that the observations are independent observations of a random variable x, where there is no trend. The acceptance region for this hypothesis is

$$[A_{20;1-\alpha/2} < A \leq A_{20;\alpha/2}]$$

From Table A.6, for $\alpha = 0.05$, $A_{20;1-\alpha/2} = A_{20;0.975} = 64$ and $A_{20;\alpha/2} = A_{20;0.025} = 125$. Hence, the hypothesis is accepted at the 5% level of significance because $A = 86$ falls within the range between 64 and 125.

4.6 CORRELATION AND REGRESSION PROCEDURES

Techniques of correlation and regression analysis are fundamental to much of the material developed in this book. The concept of correlation between two random variables has already been introduced in Chapter 3 and will be expanded on in Chapter 5. The concept of linear regression is basic to the techniques of frequency response function estimation from input/output data, as formulated in Chapters 6 and 7. The material in these chapters, however, is developed in a frequency domain context that may obscure associations with more familiar classical presentations. Hence, a brief review of correlation and regression concepts from the viewpoint of elementary statistics may be helpful as an introduction to this later material.

4.6.1 Linear Correlation Analysis

For a wide class of problems, a matter of primary interest is whether or not two or more random variables are interrelated. For example, is there a relationship between cigarette smoking and life expectancy, or between measured aptitude and academic success? In an engineering context, such problems often reduce to detecting a relationship between some assumed excitation and an observed response of a physical system of interest. The existence of such interrelationships and their relative strength can be measured in terms of a correlation coefficient ρ as defined in Section 3.2.1. For the simple case of two random variables x and y, the correlation coefficient is given by Equation (3.36) as

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} \tag{4.56}$$

where C_{xy} is the covariance of x and y as defined in Equation (3.34).

Now assume the random variables x and y are sampled to obtain N pairs of observed values. The correlation coefficient may be estimated from the sample data by

$$\begin{aligned} r_{xy} = \hat{\rho}_{xy} = \frac{s_{xy}}{s_x s_y} &= \frac{\sum_{j=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2\right]^{1/2}} \\ &= \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\,\bar{y}}{\left[\left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right)\left(\sum_{i=1}^{N} y_i^2 - N\bar{y}^2\right)\right]^{1/2}} \end{aligned} \tag{4.57}$$

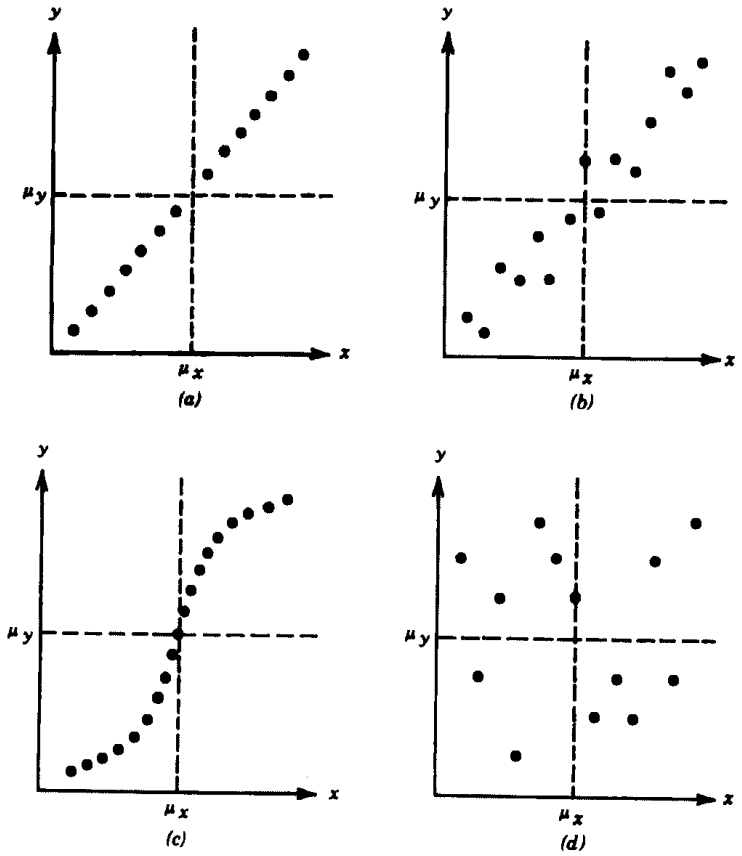


Figure 4.3 Illustration of varying degrees of correlation, (*a*) Perfect linear correlation, (*b*) Moderate linear correlation, (*c*) Nonlinear correlation, (*d*) No correlation.

Like ρ_{xy}, the sample correlation coefficient r_{xy} will lie between -1 and $+1$, and will have a bounding value only when the observations display a perfect linear relationship. A nonlinear relationship and/or data scatter, whether it be due to measurement errors or imperfect correlation of the variables, will force the value of r_{xy} toward zero, as illustrated in Figure 4.3.

To evaluate the accuracy of the estimate r_{xy}, it is convenient to work with a particular function of r_{xy} given by

$$w = \frac{1}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right] \tag{4.58}$$

From Ref. 1, the random variable w has an approximately normal distribution with a mean and a variance of

$$\mu_w = \frac{1}{2}\ln\left[\frac{1+\rho_{xy}}{1-\rho_{xy}}\right] \tag{4.59}$$

$$\sigma_w^2 = \frac{1}{N-3} \tag{4.60}$$

Using the above relationships, confidence intervals for ρ_{xy} based on a sample estimate r_{xy} may be readily established as outlined in Section 4.4.

Because of the variability of correlation estimates, it is usually desirable to verify that a nonzero value of the sample correlation coefficient indeed reflects the existence of a statistically significant correlation between the variables of interest. This may be accomplished by testing the hypothesis that $\rho_{xy} = 0$, where a significant correlation is indicated if the hypothesis is rejected. From Equations (4.59) and (4.60), the sampling distribution of w given $\rho_{xy} = 0$ is normal with a mean of $\mu_w = 0$ and a variance of $\sigma_w^2 = 1/(N-3)$. Hence the acceptance region for the hypothesis of zero correlation is given by

$$\left[-z_{\alpha/2} \leq \frac{\sqrt{N-3}}{2} \ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right] < z_{\alpha/2}\right] \tag{4.61}$$

where z is the standardized normal variable. Values outside the above interval would constitute evidence of statistical correlation at the α level of significance.

Example 4.5. **Illustration of Linear Correlation Analysis.** The heights and weights of $N = 25$ male university students selected at random are presented in Table 4.3. Is there reason to believe that the height and weight of male students are correlated at the $\alpha = 0.05$ level of significance?

Let x be height and y be weight. From the data in Table 4.3, the following values needed in Equation (4.61) are calculated:

$$\sum_{i=1}^{N} x_i y_i = 299,056 \qquad \sum_{i=1}^{N} x_i^2 = 124,986 \qquad \sum_{i=1}^{N} y_i^2 = 723,604$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1766}{25} = 70.64 \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{4224}{25} = 168.96$$

Substituting the above values into Equation (4.57) yields the estimated correlation coefficient as follows:

$$r_{xy} = \frac{299,056 - (25)(70.64)(168.96)}{[(124,986 - 25(70.64)^2)(723,604 - 25(168.96)^2]^{1/2}} = 0.44$$

Table 4.3 Height and Weight Data for Male Students

	x = height in inches								y = weight in pounds				
x	70	74	70	65	69	73	72	69	72	76	74	72	
y	140	210	148	145	182	165	155	170	174	155	185	185	
x	68	70	71	68	73	65	73	74	64	72	72	67	73
y	165	220	185	180	170	135	175	180	150	170	165	145	170

From Equation (4.58), the quantity $w = 0.472$; thus $\sqrt{N-3}w = 2.21$. Now using Equation (4.61), the hypothesis that $\rho_{xy} = 0$ is rejected at the 5% level of significance since $\sqrt{N-3}w = 2.21$ falls outside the acceptance region bounded by $\pm z_{\alpha/2} = \pm 1.96$. Hence, there is reason to believe that significant correlation exists between the height and weight of male students.

4.6.2 Linear Regression Analysis

Correlation analysis can establish the degree to which two or more random variables are interrelated. Beyond this, however, a model for the relationship may be desired so that predictions can be made for one variable based on specific values of other variables. For instance, a significant linear relationship between the height and weight of male university students is indicated by the correlation analysis of data presented in Example 4.5. A logical second step would be to evaluate the relationship further so that the weight of students can be predicted for any given height. Procedures for dealing with problems of this type come under the heading of regression analysis.

Consider the simple case of two correlated random variables x and y. Referring again to Example 4.5, x might be student height and y student weight. A linear relationship between the two variables would suggest that for a given value of x, a value of y would be predicted by

$$\tilde{y} = A + Bx \tag{4.62}$$

where A and B are the intercept and slope, respectively, of a straight line. For the case of data that display perfect linear correlation ($r_{xy} = 1$), the predicted value $\tilde{y}_i$ would always equal the observed value y_i for any given x_i. In practice, however, data usually do not display a perfect linear relationship. There generally is some scatter due to extraneous random effects, and perhaps distortion due to nonlinearities, as illustrated in Figure 4.3. Nevertheless, if a linear relationship is assumed and unlimited data are available, appropriate values of A and B can be determined that will predict the expected value of y_i for any given x_i. That is, $\tilde{y}_i$ will not necessarily equal the observed value y_i associated with the corresponding x_i, but it will be an average for all such values that might have been observed.

The accepted procedure for determining the coefficients in Equation (4.62) is to select those values of A and B that minimize the sum of the squared deviations of the observed values from the predicted values of y. This procedure is called a *least squares fit*. Specifically, noting that the deviation of the observed values from the predicted values is

$$y_i - \tilde{y}_i = y_i - (A + Bx_i) \tag{4.63}$$

it follows that the sum of this squared deviations is given by

$$Q = \sum_{i=1}^{N} (y_i - A - Bx_i)^2 \tag{4.64}$$

Hence, a least squares fit is provided by those values of A and B that make

$$\frac{\partial Q}{\partial A} = \frac{\partial Q}{\partial B} = 0 \tag{4.65}$$

In practice, the available data will be limited to a sample of N pairs of observed values for x and y. This means that Equation (4.65) will yield only estimates of A and B, to be denoted by a and b, respectively. Substituting Equation (4.64) into Equation (4.65) and solving for the estimates of A and B yields

$$a = \bar{y} - b\bar{x} \tag{4.66a}$$

$$b = \frac{\sum_{i=1}^{N}(x_i - \bar{x})y_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\,\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2} \tag{4.66b}$$

These estimates can now be used to write a prediction model for y given x as follows:

$$\hat{y} = a + bx = (\bar{y} - b\bar{x}) + bx = \bar{y} + b(x - \bar{x}) \tag{4.67}$$

The straight line defined by Equation (4.67) is called the *linear regression line for y on x*. By switching the dependent and independent variables in Equation (4.66), a regression line for x on y could also be calculated. Specifically,

$$\hat{x} = \bar{x} + b'(y - \bar{y}) \tag{4.68}$$

where

$$b' = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\,\bar{y}}{\sum_{i=1}^{N} y_i^2 - N\bar{y}^2} \tag{4.69}$$

Comparing the product of Equations (4.66b) and (4.69) to Equation (4.57), it is seen that the slopes of the regression lines for y on x and x on y are related to the sample correlation coefficient of x and y by

$$r_{xy} = [bb']^{1/2} \tag{4.70}$$

Now consider the accuracy of the estimates a and b given by Equation (4.66). Assuming a normal distribution of y given x, it is shown in Ref. 1 that a and b are unbiased estimates of A and B, respectively, with sampling distributions related to the t distribution as follows:

$$\frac{a - A}{\left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)^{1/2}} = s_{y|x} t_{N-2} \tag{4.71}$$

$$\frac{b - B}{\left(\frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)^{1/2}} = s_{y|x} t_{N-2} \tag{4.72}$$

Of particular interest is the sampling distribution of $\hat{y}$ associated with a specific value of $x = x_0$. This is given by

$$\frac{\hat{y}-\tilde{y}}{\left(\frac{1}{N}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)^{1/2}} = s_{y|x}t_{N-2} \tag{4.73}$$

In Equations (4.71)–(4.73), the term $s_{y|x}$ is the sample standard deviation of the observed values of y_i about the prediction $\hat{y}_i = a + bx_i$ and is given by

$$s_{y|x} = \left[\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N-2}\right]^{1/2} = \left[\left(\frac{n-1}{n-2}\right)s_y^2\left(1-r_{xy}^2\right)\right]^{1/2} \tag{4.74}$$

The above relationships provide a basis for establishing confidence intervals for A, B, and $\tilde{y}$ based on the estimates a, b, and $\hat{y}$.

Example 4.6. **Illustration of Linear Regression Analysis.** Using the data presented in Table 4.3 for Example 4.5, determine a regression line that will provide a linear prediction for the average weight of male university students as a function of their height. Determining a 95% confidence interval for the average weight of male students who are 70 in. tall.

As in Example 4.5, let x be height and y be weight. The values needed to determine the slope and intercept of the regression line for y on x have already been calculated in Example 4.5. Substituting these values into Equation (4.66) yields the estimated slope and the intercept as follows:

$$b = \frac{299,056-(25)(70.64)(168.96)}{168.96-(25)(70.64)^2} = 2.85$$

$$a = 168.96-(2.85)(70.64) = -32.6$$

Hence, the regression line estimating the average weight of male university students given height is

$$\hat{y} = -32.6+2.85x$$

which yields an estimated weight of $\hat{y} = 167.1$ lb for a height of $x = 70$ in.

To establish a confidence interval for the average weight $\tilde{y}$ based on the estimate $\hat{y} = 167.1$ lb, it is necessary to calculate $s_{y|x}$ given by Equation (4.74). A more convenient equation for $s_{y|x}$ from the computational viewpoint is

$$s_{y|x} = \left[\frac{1}{N-2}\left(\sum_{i=1}^{N}(y_i-\bar{y})^2-\frac{\left[\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)\right]^{1/2}$$

where the terms in the above expression are further simplified for computational purposes by noting that

$$\sum_{i=1}^{N}(v_i-\bar{v})^2=\sum_{i=1}^{N}v_i^2-N\bar{v}^2 \qquad \sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{N}x_iy_i-N\bar{x}\,\bar{y}$$

Substitution of the appropriate values into these expressions yields

$$s_{y|x}=\left[\frac{1}{23}\left(9917-\frac{(673)^2}{236}\right)\right]^{1/2}=18.65$$

Then, from Equation (4.73), a 95% confidence interval for the average weight of male university students with a height of 70 in. is

$$\hat{y}\pm s_{y|x}t_{N-2;\alpha/2}\left[\frac{1}{N}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right]^{1/2}$$

$$=167.2\pm(18.65)t_{23;0.025}\left[\frac{1}{25}+\frac{(70-70.64)^2}{236}\right]^{1/2}$$

$$=167.2\pm7.9=159.3 \text{ to } 175.1 \text{ lb}$$

This concludes Example 4.6.

The techniques of correlation and regression analysis are readily extended for applications involving more than two random variables. As noted earlier, such extensions are fundamental to the analysis of multiple-input/output problems developed in Chapter 7. Hence, further discussions of this subject are deferred to that chapter.

PROBLEMS

4.1 Given the random variable $y=cx$ where c is a constant and x is a random variable with a mean value and a variance of μ_x and σ_x^2, respectively, prove that the following relationships are true.

(a) $\mu_y=c\mu_x$.

(b) $\sigma_y^2=c^2\sigma_x^2$.

4.2 Given a random variable x with a probability density function of

$$p(x)=\frac{1}{2\sqrt{2\pi}}e^{-(x-1)^2/8}$$

What are the mean value and variance of x?

4.3 Given two independent random variables, x and y, with mean values of μ_x and μ_y, and variances of σ_x^2 and σ_y^2, determine the

(a) mean value of the product xy.

(b) variance of the difference $x - y$.

4.4 The normalized random error (coefficient of variation) ε_r of an unbiased parameter estimate $\hat{\phi}$ is defined as the ratio of the standard deviation of the estimate to the expected value of the estimate, that is, $\varepsilon_r = \sigma_{\hat{\phi}}/\mu_{\hat{\phi}}$. Determine the normalized random error of a variance estimate s^2 computed from $N = 200$ sample observations using Equation (4.12).

4.5 Given four independent standardized normally distributed random variables, z_1, z_2, z_3, and z_4, define the distribution functions of the following combinations of these variables. For each case, specify the associated degrees of freedom or mean value and variance, as appropriate.

(a) $z_1^2 + z_2^2 + z_3^2 + z_4^2$.

(b) $z_1 + z_2 - z_3 - z_4$.

(c) $\dfrac{z_4}{\{[z_1^2 + z_2^2 + z_3^2]/3\}^{1/2}}$.

(d) $\dfrac{[z_1^2 + z_2^2 + z_3^2]/3}{z_4^2}$.

4.6 What distribution function would be used to establish confidence intervals for the following parameters of two independent normally distributed random variables, x and y?

(a) Interval for μ_x based on a sample mean $\bar{x}$ and known variance σ_x^2.

(b) Interval for σ_x^2/σ_y^2 based on a ratio of sample variances s_x^2/s_y^2.

(c) Interval for σ_x^2 based on a sample variance s_x^2.

(d) Interval for μ_x based on a sample mean $\bar{x}$ and sample variance s_x^2.

4.7 A correlation study is performed using a sample of $N = 7$ pairs of observations $(x_1y_1, x_2y_2, \ldots, x_7y_7)$. A sample correlation coefficient of $r_{xy} = 0.77$ is calculated. Test the hypothesis that ρ_{xy} is greater than zero at the $\alpha = 0.01$ level of significance.

4.8 Assume the sample mean values of two correlated random variables are $\bar{x} = 1$ and $\bar{y} = 2$. Further assume that the sample correlation coefficient is $r_{xy} = 0.5$. If the regression line for y on x is given by $\hat{y} = 1 + x$,

(a) what is the slope b' of the regression line for x on y?

(b) what is the equation for the regression line for x on $y(\hat{x} = a' + b'y)$?

4.9 Given a sample of N independent observations of a random variable x with a known mean value of zero, an *efficient* estimator for the variance of x is

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

(a) Prove the above estimator is unbiased.

(b) Write an expression relating the above estimator to a chi-square variable with the appropriate degrees of freedom specified.

(c) What is the variance of the above estimator? (*Hint:* The variance of χ_n^2 is $2n$.)

4.10 Assume a time sequence of $N = 20$ measurements are made of a normally distributed random variable x with the following results:

Time	Value	Time	Value	Time	Value	Time	Value
1	10.1	6	10.6	11	10.9	16	11.4
2	10.4	7	11.3	12	10.1	17	10.1
3	9.9	8	9.7	13	10.5	18	11.5
4	10.0	9	10.2	14	10.7	19	10.3
5	10.0	10	11.2	15	10.8	20	10.9

Test the time sequence of measurements for a trend at the 5% level of significance in two ways, namely,

(a) by computing the reverse arrangements and performing a nonparametric test.

(b) by comparing the slope b of the linear regression line and testing the hypothesis that $B = 0$.

REFERENCES

1. Guttman, I., Wilks, S. S., and Hunter, J. S., *Introductory Engineering Statistics*, 3rd ed., Wiley, New York, 1982.
2. Ross, S. M., *Introduction to Probability and Statistics for Engineers and Scientists*, Wiley, New York, 1987.
3. Hines, W. H., and Montgomery, D. C., *Probability and Statistics in Engineering and Management Sciences*, 3rd ed., Wiley, New York, 1990.
4. Kendall, M. G., and Stuart, A., *The Advanced Theory of Statistics*, Vol. 2, *Inference and Relationships*, Hafner, New York, 1961.