2

Properties of Distributions

Before delving into the genetics of quantitative variation, it is essential to have a basic understanding of statistics. The statistical concepts and techniques most frequently encountered in quantitative genetics are presented in this and the following chapter. For the reader with advanced training in statistical theory, much of what follows will probably be review, and some things may appear to be presented in a nonrigorous manner. Even so, it may still be profitable to skim the following pages to become familiar with the notation that will be used throughout the book. As an additional reward, a number of examples will provide some immediate contact with the field of quantitative genetics.

PARAMETERS OF THE UNIVARIATE DISTRIBUTIONS

Characters that are studied by biologists are of three types. Traits that are distributed into a range of discrete classes, such as scale counts in fish or leaf number in plants, are called **meristic characters**. Those that are measured on a continuous scale are known as **metric characters**. Length, weight, and growth rate attributes are examples of the latter. Attributes such as survival to a fixed age are known as **all-or-none** or **binary characters**. Of course, due to technical limitations, even measures of truly continuously distributed traits must always be artificially placed into discrete categories. Meter sticks, for example, are unable to distinguish between individuals that are 25.2 and 25.3 mm in length. Both would typically be placed in the 25–26 mm category, although the biological reality is that every conceivable length in the 25–26 mm range is possible.

Suppose that one performs a series of measurements on a collection of individuals. Compilation of the data provides some information on the relative incidence of different trait measures. A **univariate distribution** describes the relative frequencies of phenotypes for a single trait, whereas a **bivariate distribution** describes the mutual distribution of two characters. The joint distribution of more than two traits is referred to as a **multivariate distribution**. An example of a bivariate distribution for maternal weight and number of offspring is given for a population of rats in Table 2.1. The data are condensed into the univariate **marginal distributions** of the two traits in the last row and column.

Table 2.1 The bivariate distribution of mother's weight and number of offspring produced for a population of rats.

Maternal Weight (grams)	Number of Offspring *												
	1	2	3	4	5	6	7	8	9	10	11	12	Totals
50 -	–	–	–	1	3	1	–	–	–	–	–	–	5
60 -	–	–	–	1	6	2	–	–	–	–	–	–	9
70 -	–	–	2	10	17	12	4	–	1	–	–	–	46
80 -	1	1	11	8	18	10	9	3	2	–	–	–	63
90 -	2	5	7	18	30	28	12	5	1	–	–	–	108
100 -	3	5	10	25	37	35	21	7	2	1	–	–	146
110 -	1	4	12	19	38	37	29	6	2	–	–	–	148
120 -	2	6	9	21	36	26	30	14	6	–	1	–	151
130 -	4	4	9	12	35	29	17	17	6	1	1	1	136
140 -	1	4	6	9	12	27	15	6	2	1	–	–	83
150 -	–	3	–	2	13	11	6	6	2	–	–	–	43
160 -	–	2	–	1	11	11	9	3	4	–	–	–	41
170 -	1	–	1	1	2	4	2	2	1	–	1	–	15
180 -	–	–	1	1	–	2	2	2	–	–	–	–	8
190 -	–	–	–	–	–	–	–	–	–	1	–	–	1
Totals	15	34	68	129	258	235	156	71	29	4	3	1	1003

* Each number in the main body of the table refers to the number of observations in a particular bivariate class. For example, 38 animals weighed between 100 and 110 grams and produced 5 offspring. The final row and column are the marginal univariate distributions for the two traits. (From Pearson 1910.)

One of the goals of statistics is to fit fairly simple mathematical functions, known as **probability distributions,** to data. If a variable z takes on only discrete values (as with offspring number), the distribution of z is completely described by giving $P(z = z_i)$ for each possible outcome z_i, where P stands for probability. For example, for offspring number, letting $z_1 = 1$, then $P(z = z_1)$ is the proportion of mothers that produce a single offspring, which for the example in Table 2.1 is $15/1003$. Summing over all possible outcomes, $\sum_i P(z = z_i) = 1$, since the total probability of all possible events is one.

If, on the other hand, z is a continuously distributed variable (as with maternal weight), $P(z = z_i)$ makes no sense since the probability that z takes on any specific value is infinitesimally small. It is more meaningful to consider the probability that z lies within a specific range of values, say z_1 and z_2. This quantity is described

by the **probability density function** $p(z)$, which satisfies the integral

$$P(z_1 \leq z \leq z_2) = \int_{z_1}^{z_2} p(z)\, dz \tag{2.1}$$

If z_{min} and z_{max} are the upper and lower bounds to z, then $p(z) = 0$ outside of this range, and over the entire range $\int_{z_{min}}^{z_{max}} p(z)\, dz = 1$. Both of these properties are in accord with common sense — a probability is never negative, and the total probability of all possible outcomes is one. A large number of functions fulfill these properties, and they have been studied in considerable detail (Johnson and Kotz 1970a,b, 1972; Kendall and Stuart 1977).

Example 1. Suppose that z is continuously distributed in the range of 0 to ∞ with probability density function

$$p(z) = \frac{1}{\lambda} e^{-z/\lambda}$$

This is the **negative exponential distribution** in which the density has a maximum at $z = 0$ and declines to zero as $z \to \infty$. Since the integral of $p(z)$ is $-e^{-z/\lambda}$,

$$\int_0^{\infty} p(z)\, dz = -e^{-z/\lambda}\Big|_0^{\infty} = 0 - (-1) = 1$$

showing that $p(z)$ fulfills the properties of a probability density.

What is the probability that a randomly drawn individual will have z in the range of $1/4$ to $1/2$?

$$P(1/4 \leq z \leq 1/2) = \int_{1/4}^{1/2} p(z)\, dz = -e^{-z/\lambda}\Big|_{1/4}^{1/2} = e^{-1/(4\lambda)} - e^{-1/(2\lambda)}$$

The numerical answer depends on the parameter λ. If, for example, $\lambda = 1/2$, then $P(1/4 \leq z \leq 1/2) = 0.239$.

Before moving on, we emphasize the importance of distinguishing between true **parameters** of distributions and **estimates** of those parameters obtained by sampling. True parameter values can only be obtained if every member of a population is measured with absolute accuracy. We must therefore almost always settle for approximations, the accuracy of which depends on the experimental setting, the measurement apparatus, and the sample size. Statisticians often denote parameters of a population with Greek symbols and to sample estimates with

Roman symbols. We will adhere to this protocol as much as possible, although there will be some instances where traditional quantitative-genetic notation prevents us from doing so.

The most useful probability density functions are defined completely by one or two parameters describing the central location and dispersion of the distribution. The most widely used measure of the location is the **arithmetic mean**, μ, also known as the **first moment about the origin**. If $p(z)$ is the probability density function of phenotype z, then weighting all values of z by their density leads to

$$\mu = \int_{-\infty}^{+\infty} z\, p(z)\, dz = E(z) \tag{2.2}$$

where $E(z)$ denotes the **expected value** or **expectation** of z. Here, we have arbitrarily put the limits $\pm\infty$ on the integral to ensure that the entire range of variation is covered. For discrete characters, $\mu = E(z) = \sum_i z_i P(z = z_i)$. For a character denoted by z, the sample estimate of the mean is generally denoted by $\bar{z}$, and estimated as the average of the n measures,

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$$

Example 2. What is the mean of the distribution discussed in Example 1? Since the integral of $(z/\lambda)\, e^{-z/\lambda}$ is $-(z+\lambda)\, e^{-z/\lambda}$,

$$\mu = \int_0^{\infty} z\, p(z)\, dz = -(z+\lambda)\, e^{-z/\lambda} \Big|_0^{\infty} = \lambda$$

Thus, the parameter λ is the mean of the distribution defined by the density function $p(x) = (1/\lambda)\, e^{-z/\lambda}$.

Higher-order moments provide measures of the dispersion of a frequency distribution. The most familiar and useful such measure is the population **variance** (a term introduced in Fisher's 1918 paper). Also known as the **second moment about the mean**, the variance is the expected squared deviation of an observation from its mean,

$$\sigma^2 = \int_{-\infty}^{+\infty} (z-\mu)^2\, p(z)\, dz = E\left[\, (z-\mu)^2 \,\right] \tag{2.3}$$

Because $\mu = E(z)$, this quantity can be expressed more simply by expanding $(z - \mu)^2$ to obtain

$$\sigma^2 = E(z^2 - 2z\mu + \mu^2) = E(z^2) - 2\mu E(z) + \mu^2 = E(z^2) - \mu^2 \qquad (2.4)$$

where we have used two useful properties of expectations,

$$E(x + y) = E(x) + E(y)$$
$$E(c\,x) = c\,E(x)$$

for a constant c. Several notations are used for the parametric variance of a distribution. When there is no ambiguity as to the variable being considered, σ^2 suffices. More generally, the variance of z is denoted by σ_z^2 or $\sigma^2(z)$.

A slight complication arises when one wishes to estimate the parameter σ^2 from a random sample of the population. As noted above, the true parameters μ and $E(z^2)$ cannot be known with certainty unless the entire population is sampled. Because the estimated mean $(\bar{z})$ is a function of the data, individual measures tend to be closer to the observed mean than to the true mean, and as a consequence, observed values of $\overline{z^2} - \bar{z}^2$ tend to be slightly less than the parametric value $[\,E(z^2) - \mu\,]$. Thus, the estimator $(\,\overline{z^2} - \bar{z}^2\,)$ is biased in the sense that it tends to underestimate the parameter $\sigma^2(z)$ to a degree that decreases with increasing sample size (n). A major goal of applied statistics is to obtain unbiased estimators that account for these kinds of small sample size limitations. In the case of the variance, the solution is simple (Example 2, Appendix 1), with

$$\text{Var}(z) = \frac{n\,(\,\overline{z^2} - \bar{z}^2\,)}{n - 1} \qquad (2.5)$$

providing an unbiased estimate of $\sigma^2(z)$ (for the derivation of this expression, see Example 2, Appendix 1.) This equation should be used whenever the true population variance, $\sigma^2(z)$, is being estimated from actual sample data.

The variance is measured in units that are the square of those of the mean, but it is often desirable to describe the dispersion of a frequency distribution on the same scale as the mean. The square root of the variance of z is called the **standard deviation** of z. The parametric value is denoted by $\sigma(z)$, σ_z, or just σ, and the statistic by $\text{SD}(z) = \sqrt{\text{Var}(z)}$. The ratio of the standard deviation to the mean, the **coefficient of variation,** is frequently used as a relative measure of dispersion. It is known that the statistic $\text{CV}(z) = \text{SD}(z)/\bar{z}$ is a downwardly biased estimator of the parametric index (σ/μ), but the bias is expected to be negligible in most cases (Haldane 1955).

Statisticians generally rely on the variance as a measure of the dispersion of a distribution. However, additional moments can be informative. For example, the third moment about the mean (μ_3) is a useful measure of the asymmetry of

a distribution. Also known as the **skewness**, μ_3 is the expected cubic deviation from the mean. As in the case of the variance, it can be expressed in terms of the moments about the origin,

$$\begin{aligned}\mu_3 &= \int_{-\infty}^{+\infty} (z-\mu)^3\, p(z)\, dz = E\left[(z-\mu)^3\right] \\ &= E(z^3) - 3\mu E(z^2) + 3\mu[E(z)]^2 - [E(z)]^3 \\ &= E(z^3) - 3\mu E(z^2) + 2\mu^3\end{aligned} \tag{2.6}$$

Thiele (1889) found that an unbiased sample estimator for μ_3 is

$$\text{Skw}(z) = \frac{n^2\,(\,\overline{z^3} - 3\,\overline{z^2}\,\bar{z} + 2\,\bar{z}^3\,)}{(n-1)(n-2)} \tag{2.7}$$

where $\overline{z^3}$ denotes the observed mean cubed value of z. The degree of asymmetry can also be described with a dimensionless index, the **coefficient of skewness**, which is estimated by the ratio

$$k_3 = \frac{\text{Skw}(z)}{\text{Var}(z)^{3/2}} \tag{2.8}$$

k_3 is positive when the longer tail of a distribution is to the right, negative when the tail is to the left, and zero for a perfectly symmetrical distribution.

From the above, it follows that

$$\mu_r = \int_{-\infty}^{+\infty} (z-\mu)^r p(z)\, dz \tag{2.9}$$

is a general expression for the rth moment about the mean. It also follows that μ_r can always be expressed in terms of moments about the origin $[\,E(z),\, E(z^2),\, \ldots,\, E(z^r)\,]$. As was shown for the variance and the skewness, these terms are obtainable from the binomial expansion of $(z-\mu)^r$.

Finally, we note that when moments are calculated from data that are grouped into classes, as in Table 2.1, a certain amount of bias is introduced because the true measures are assumed to be concentrated at the midpoints of the classes. Provided the total distribution is continuous and tails off smoothly at its extremities, this bias can often be eliminated by application of Sheppard's (1898) corrections. In the case of the variance, the corrected estimate is obtained by subtracting from $\text{Var}(z)$ the quantity $\omega^2/12$, where ω is the width of the interval. No correction is required for the third moment about the mean. For details on higher-order moments, see Kendall and Stuart (1977, p. 77).

Example 3. Utilizing the data for maternal weight from Table 2.1, we now summarize the procedures for obtaining estimates of the first three moments.

grams *	z	$n(z)$	$z\,n(z)$	$z^2 n(z)$	$z^3 n(z)$
50 -	55	5	275	15,125	831,875
60 -	65	9	585	38,025	2,471,625
70 -	75	46	3,450	258,750	19,406,250
80 -	85	63	5,355	455,175	38,689,875
90 -	95	108	10,260	974,700	92,596,500
100 -	105	146	15,330	1,609,650	169,013,250
110 -	115	148	17,020	1,957,300	225,089,500
120 -	125	151	18,875	2,359,375	294,921,875
130 -	135	136	18,360	2,478,600	334,611,000
140 -	145	83	12,035	1,745,075	253,035,875
150 -	155	43	6,665	1,033,075	160,126,625
160 -	165	41	6,765	1,116,225	184,177,125
170 -	175	15	2,625	459,375	80,390,625
180 -	185	8	1,480	273,800	50,653,000
190 -	195	1	195	38,025	7,414,875
Totals		$n =$ 1,003	$\sum z\,n(z) =$ 119,255	$\sum z^2\,n(z) =$ 14,812,275	$\sum z^3\,n(z) =$ 1,913,429,875

* For each weight category, z is taken arbitrarily to be the midpoint of the measurement interval, so that for the interval 50-60, we take $z = 55$. The frequency of observations in each category, $f(z)$, is equal to $n(z)/n$, where $n(z)$ is the number of observations with phenotype z, and $n = \sum n(z)$ is the total sample size.

The moments about the origin are obtained by dividing the weighted sums in the table by n,

$$\bar{z} = \sum z f(z) = \sum z\,n(z)/n = \frac{119,255}{1,003} = 118.90$$

$$\overline{z^2} = \sum z^2 f(z) = \sum z^2 n(z)/n = \frac{14,812,275}{1,003} = 14,767.97$$

$$\overline{z^3} = \sum z^3 f(z) = \sum z^3 n(z)/n = \frac{1,913,429,875}{1,003} = 1,907,706.75$$

The variance estimated from the pooled data is

$$\text{Var}(z) = \frac{n\,(\,\overline{z^2} - \bar{z}^2\,)}{n-1} = 631.39$$

and application of Sheppard's correction, with $\omega = 10$, reduces this to

$$\text{Var}(z) = 631.39 - \frac{\omega^2}{12} = 623.06$$

The coefficient of variation is then

$$\text{CV}(z) = \frac{[\,\text{Var}(z)\,]^{1/2}}{\bar{z}} = 0.21$$

Finally, the skewness and coefficient of skewness are

$$\text{Skw}(z) = \frac{n^2\,(\,\overline{z^3} - 3\,\overline{z^2}\,\bar{z} + 2\,\bar{z}^3\,)}{(n-1)(n-2)} = 1,805.40$$

$$k_3 = \frac{\text{Skw}(z)}{[\,\text{Var}(z)\,]^{3/2}} = 0.12$$

THE NORMAL DISTRIBUTION

When large data sets of the type compiled in Table 2.1 are displayed in the form of frequency histograms (Figure 2.1), they often approximate a bell-shaped distribution. Three famous mathematicians, DeMoivre (1738), LaPlace (1778), and Gauss (1809), worked out the properties of a very useful description of this form — the **normal distribution**, also referred to as the **Gaussian distribution**. If z is a normally distributed variable, its density function is given by

$$p(z) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right] \tag{2.10}$$

where $\exp \simeq 2.7183$ is the base of natural logarithms, and $\pi \simeq 3.1416$. The normal distribution is a function of only two parameters, the population mean (μ) and variance (σ^2). The normal density attains a maximum when $z = \mu$ and declines continuously and symmetrically in both directions as z deviates from μ (Figure 2.1). A normally distributed variable with mean μ and variance σ^2 is often denoted by $z \sim \text{N}(\mu, \sigma^2)$, where $\sim$ means "is distributed as." In discussions in future chapters, we often use the notation $\varphi(z, \mu, \sigma^2)$ to denote the probability density of a normal, to remind the reader that it is also a function of the mean and variance.

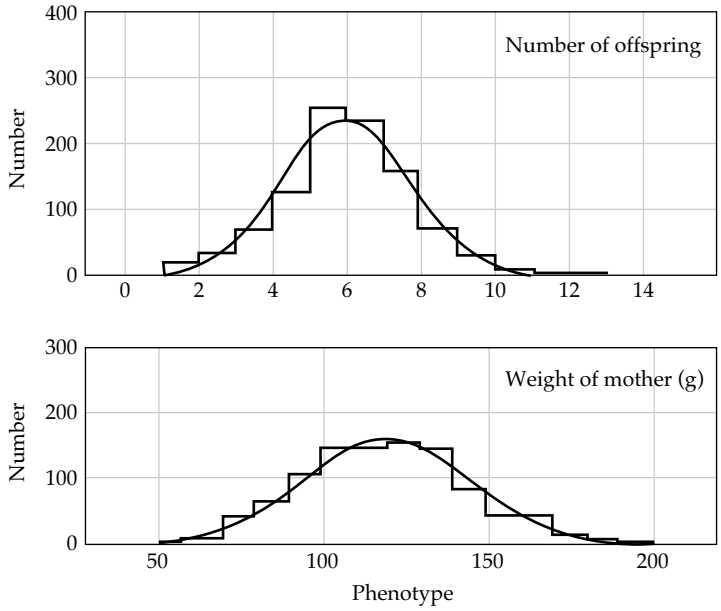

Figure 2.1 Frequency histograms for the two univariate distributions in Table 2.1 and their normal approximations based on the observed means and variances.

The normal distribution plays a central role in statistical theory for two reasons. First, the normal probability density function has many simple mathematical features that allow the derivation of practical statistical tests. Second, even when actual distributions of phenotypes are inconsistent with the normal density function, after an appropriate scale transformation (Chapter 11), many can be rendered approximately normal. A general reason why many traits are distributed normally or nearly so is provided by the **central limit theorem**, which states that the sum of a number of independent random variables approaches normality as the number of variables increases. This is expected to be the case, for example, for a metric character influenced by many environmental factors and a large number of unlinked genes, each with small additive effects. As a consequence, the normal distribution has been relied upon extensively in quantitative genetics. Whenever an assumption regarding the form of a phenotype distribution is necessary, the normal distribution is generally invoked as a first approximation. The normal density function is also often used to define a **Gaussian fitness function** in the theory of stabilizing selection, the "mean" serving as a measure of the optimum phenotype and the "variance" being inversely related to the intensity of selection (because the fitness function becomes flatter as the width increases).

There are, of course, limitations of the normal density function and of distribution functions in general. For instance, the normal distribution gives small

positive values, rather than zero, for negative z, an unrealistic situation for traits such as body size or bone length, which cannot take on negative values. Nevertheless, if the mean of a distribution is sufficiently greater than zero, the theoretical incidence of negative values is minuscule and not problematical. It should also be emphasized that the normal distribution is a continuous function, giving positive values for noninteger values of z. It is, therefore, not strictly applicable to meristic traits such as egg number or spine count, although it provides a close approximation when the number of classes is large.

It is often convenient to work with a standardized form of Equation 2.10. A **standard normal deviate,** $z' = (z - \mu)/\sigma$, is the deviation of a measure from the population mean in units of standard deviations. Applying a useful property of distribution theory in the following example, we show that if z is normally distributed with mean μ and variance σ^2, then z' is normal with zero mean and unit variance, i.e.,

$$p(z') = (2\pi)^{-1/2} \exp\left[-\frac{(z')^2}{2}\right] \tag{2.11}$$

Example 4. It is known that if y is a function of z, denoted by $f(z)$, then its probability density function is

$$p(y) = \left|\frac{df(z)}{dz}\right|^{-1} p(z)$$

where $|\cdots|$ denotes absolute value. This transformation is valid provided that $df(z)/dz$ exists and is nonzero for all z values for which $p(z) > 0$. This criterion is met by the standard normal deviate.

Letting $z' = f(z) = (z - \mu)/\sigma$, then $df(z)/dz = \sigma^{-1}$. Substituting the normal probability density function for $p(z)$ recovers the **standard normal** or (**unit normal**) **distribution**,

$$p(z') = \left|\frac{1}{\sigma}\right|^{-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right] = (2\pi)^{-1/2} \exp\left[-\frac{(z')^2}{2}\right]$$

Because the normal distribution is symmetrical, the third moment (μ_3) is equal to zero. The fourth moment has an expected value equal to $3\sigma^4$. Thus, if we let $\mathrm{Kur}(z)$ be the sample estimate of μ_4, where Kur denotes **kurtosis**, the index

$$k_4 = \frac{\mathrm{Kur}(z) - 3\,[\,\mathrm{Var}(z)\,]^2}{[\,\mathrm{Var}(z)\,]^2} \tag{2.12a}$$

where

$$\text{Kur}(z) = \frac{n^2(n+1)(\overline{z^4} - 4\,\overline{z^3}\,\bar{z} + 6\,\overline{z^2}\,\bar{z}^2 - 3\,\bar{z}^4)}{(n-1)(n-2)(n-3)} \tag{2.12b}$$

provides a measure of the peakedness of a distribution. For a truly normal distribution, $k_4 = 0$. A distribution with a high narrow peak relative to the normal ($k_4 > 0$) is said to be **leptokurtic**. A broader peak than normal ($k_4 < 0$) is referred to as **platykurtic.**

CONFIDENCE INTERVALS

Estimates such as $\bar{z}$ and $\text{Var}(z)$ vary from one sample to the next because of sampling error, so it is useful to know how far an observed statistic is likely to deviate from the true parameter that is being estimated. Although the true values are unknown, if something is known about the sampling error of the estimate, it is possible to evaluate the probability that the observed value lies within a specific range of the true value. Generally, we do not estimate the sampling error of statistics by sampling populations over and over again, but by using known algebraic expressions that themselves depend on sample statistics.

As an example, consider an estimate $\bar{z}$ of the mean of a distribution. An important issue here is the probability α that the parameter μ is within a certain range $\bar{z} \pm \Delta$. By symmetry, this is the same as the probability that $\bar{z}$ lies within the range $\mu \pm \Delta$. Transforming to standardized variables by letting $z' = (\bar{z} - \mu)/\sigma(\bar{z})$, where $\sigma(\bar{z})$ is the sampling variance of the mean, then the probability of interest is defined to be

$$\alpha = P[\,(\bar{z} - \Delta) \leq \mu \leq (\bar{z} + \Delta)\,] = \int_{-\Delta/\sigma(\bar{z})}^{+\Delta/\sigma(\bar{z})} p(z')dz' \tag{2.16}$$

The range $\bar{z} \pm \Delta$ defines the **confidence limits** or **interval** for the mean associated with the α probability level. In applications of Equation 2.16, it is generally assumed that the statistic is unbiased (so that the expected value of the statistic equals the true parameter value) and normally distributed. In the case of the mean, this implies that replicate estimates of the mean ($\bar{z}$) should be normally distributed about the parametric value (μ) with sampling variance $\sigma^2(\bar{z})$. Equation 2.16 is then simply an integration over the standardized normal density.

Although Equation 2.16 cannot be integrated directly, tables relating the standardized limits (Δ/σ) to α are provided in most statistics texts. The quantity Δ/σ, usually denoted as t, defines the distance (in standard errors) that the deviation between observed statistic and parametric value will lie with probability α. (Whereas the standard deviation is a measure of the dispersion of individual measures, the term **standard error** is usually reserved as a measure of the dispersion of statistics.) For any particular probability level, t decreases with increasing

sample size, asymptotically approaching a constant. For sample sizes exceeding 50 or so, $t \simeq 1.96$ for $\alpha = 0.95$, and $t \simeq 2.58$ for $\alpha = 0.99$.

The remaining problem is to obtain an estimate of the sampling variance of the statistic (the square of the standard error). In the case of the mean, it is well known that an unbiased estimator of the sampling variance is $\mathrm{Var}(z)/n$, where $\mathrm{Var}(z)$ is the variance of individual measures, and n is the number of measures (Appendix 1). Thus, the 95% confidence interval for the mean is approximately $\bar{z} \pm 1.96\,[\,\mathrm{Var}(z)/n\,]^{1/2}$.

Unfortunately, expressions for the sampling variances of other statistics (such as the variance, higher-order moments, coefficients of variation, etc.) are usually much more complicated than those for the mean. Appendix 1 outlines procedures used to obtain expressions for sampling variances for such statistics. These expressions are usually referred to as **large-sample variance** estimators because they are functions of observed statistics whose reliability increases with increasing sample size. A common procedure in statistics is to use twice the square root of the large-sample variance as a crude estimate of the 95% confidence limit. We emphasize that this assumes that the statistic has a sampling distribution that is close to normal, that the estimator is unbiased, and that the sample size is large enough that the large-sample variance (itself an estimate) is reasonably reliable.