we could select a sample of the beans from the jar and then estimate the proportion of the two types of beans. Obviously if we pick a sample of 10 beans from the jar, it is possible that all 10 may be black. This would lead us to the conclusion that 100% of the beans are black. The difference between what we conclude from a sample and reality or truth is called sampling error.

The hygienist is confronted with sampling error when only a sample of the possible exposures in a population is measured. The goal of any scientific measurement is to control all sources of error in a cost-effective manner. These include measurement error as well as statistical sampling error. However, sampling error constitutes the larger source of error associated with most industrial hygiene measurements.

As with measurement error, we can use statistics to help us estimate the magnitude of the sampling error. In the case of measurement error, the normal distribution was used. However, for statistical sampling error associated with measuring workplace chemical exposures, the normal distribution generally is not appropriate – the distribution of exposures over time or across workers is commonly not symmetrical (Esmen and Hammad, 1977; NIOSH, 1977; Rappaport, 1991). The hygienist may wish to measure things other than exposures to chemicals. For example, exposures to physical agents such as noise may be normally distributed (Behar and Plener, 1984).

Ways of using each of the various statistical distributions are discussed in detail in Chapters 9, 14, 16, and 17. However, it is important now to realize that in accounting for statistical sampling error associated with industrial hygiene exposure assessment, a distribution other than the normal is needed. The distribution that is commonly used is the lognormal distribution. Like the normal distribution, the lognormal distribution has many applications in industrial hygiene as well as other fields.

LOGNORMAL DISTRIBUTION

The good news is that the lognormal distribution is related to the normal distribution; a lognormally distributed population of observations can be made normally distributed by plotting the logarithms of the values. The bad news is that the similarities end here. The various tools that have been derived from the normal distribution for testing hypotheses, making inferences, and describing characteristics of populations are not so well developed for the lognormal distribution, and those that have been derived are rather complex. Nevertheless, failure to use the lognormal, when it is appropriate, can lead to important errors both in the testing of hypotheses and in describing the parameters (such as mean and variance) of a population.

As with the normal distribution, the lognormal is a mathematical model that is represented by an equation:

$$f(x) = \frac{1}{x\sigma_L\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_L^2}(\ln x - \mu_L)^2\right] \qquad \text{[Equation 7-7]}$$

If Equations 7-5 and 7-7 are compared, one can see several similarities. In the case of the normal probability density function (Equation 7-5), the two parameters are σ and μ. In the case of the lognormal, the two parameters are σ_L and μ_L. These two parameters are not the mean and the standard deviation of the members of the population, but rather the mean and the standard deviation of the logarithms of the members of the population.

Whereas the normal distribution is characteristically bell shaped, the shape of the lognormal distribution cannot be characterized so easily. Figure 7-7 shows that the shape of lognormal distributions can vary considerably. In fact, some lognormal distributions can be so slightly skewed (asymmetrical) that the normal model is adequate for characterizing them. As shown in Figure 7-7, as the geometric standard deviation and geometric mean change, the shape of the distribution changes even if the value for the mean remains constant. As the geometric mean increases, the geometric standard deviation *decreases* for a constant mean.

In order to develop equations for the lognormal, some terms must be defined. It is important that the reader memorize the definitions of these terms in order not to be confused.

1. x_i is defined as any observation from lognormal population.

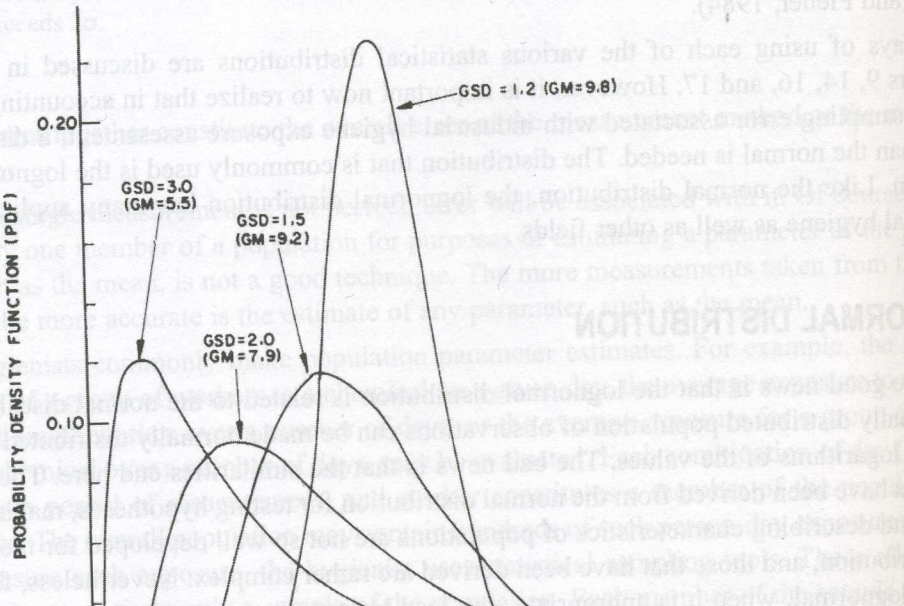

Figure 7-7. Four lognormal distributions with varying GSD and GM, but a constant mean value of 10 ppm (from NIOSH, 1977).

2. x_{L_i} is defined as the logarithmic transformation of x_i.
3. The true distribution mean and the standard deviation of the x_i are defined as μ and σ.
4. The true distribution mean and the standard deviation of the x_{L_i} are μ_L and σ_L.
5. The true geometric mean of the x_i is defined as

$$\mu_g = \exp(\mu_L) \qquad \text{[Equation 7-8]}$$

6. The true geometric standard deviation of the x_i (GSD) is defined as

$$\sigma_g = \exp(\sigma_L) \qquad \text{[Equation 7-9]}$$

7. $\mu \neq \mu_L \neq \mu_g$ and $\sigma \neq \sigma_L \neq \sigma_g$.

There are six parameters that are important to an understanding of the lognormal distribution, whereas for the normal only two parameters are defined. However, any two of the six lognormal parameters may be used to define a lognormal distribution, as the interconvertibility of the equations in Table 7-1 shows.

Figure 7-8 shows the mean, geometric mean, and mode of a lognormal distribution. They are not equal as they are for the normal distribution. The mode is the value that occurs most frequently in a distribution. The median of a lognormal is equal to μ_g, the geometric mean. This is derived as follows. We define $\ln x_i = x_{L_i}$. Because the x_{L_i} are normally distributed, the median x_{L_i} = mean x_{L_i} or μ_L. Now $\exp(\mu_L) = \mu_g$, but this exponential transformation does not change the relative placement of the values. In other words, μ_L is the median or middle value of the x_{L_i}, and so $\exp(\mu_L)$ is still a median or middle value. The geometric mean, μ_g, is the median of a lognormal, not its mean. The geometric mean is always larger than the mode, and the true mean is always larger than the geometric mean, as Figure 7-8 shows.

Among the terms for variation or spread of a lognormally distributed population, the geometric standard deviation is more widely used than σ or σ_L. For a lognormal distribution, σ or σ_L gives us information about the spread of the distribution; however, as shown previously, without normalizing the standard deviation (e.g., calculating the coefficient of variation), we do not have any information concerning the relationship of the spread (variance) of the distribution to the mean. In other words, a standard deviation of 1 will have different importance if the mean is 1,000 than if the mean is 10. Standard deviations are positive values with no theoretical bound.

On the other hand, the geometric standard deviation has considerably different characteristics from the standard deviation. It is a measure of relative variability similar to the CV. (It can be related to the CV as in Table 7-1). For lognormal distributions of importance to hygienists, the geometric standard deviation ranges from approximately 1.2 up to values of 10, and most are in the range of 1.5 to 3.5 (NIOSH, 1977; Rappaport, 1991; Buringh and Lanting, 1991). *This is true regardless of the mean or its units.* To further examine this characteristic of the lognormal distribution and the geometric standard deviation, see Table 7-2. For the first

Table 7-1
Equations for parameters of the lognormal distribution

Given	To Obtain	Use
μL	μ_g =	$\exp(\mu_L)$
μ_c, σ_c	μ_g =	$\mu_c^2 / (\mu_c^2 + \sigma_c^2)^{0.5}$
σ_L	σ_g =	$\exp(\sigma_L)$
μ_c, σ_c	σ_g =	$\exp[(\ln[1 + \sigma_c^2/\mu_c^2])^{0.5}]$
μ_L, σ_L	μ_c =	$\exp[\mu_L + 0.5\sigma_L^2]$
μ_g, σ_L	μ_c =	$\mu_g \exp[0.5\sigma_L^2]$
μ_L, σ_L	σ_c =	$([\exp(2\mu_L+\sigma_L^2)][\exp(\sigma_L^2)- 1])^{0.5}$
μ_g, σ_L	σ_c =	$(\mu_g^2[\exp(\sigma_L^2)][\exp(\sigma_L^2) - 1])^{0.5}$
μ_g	μ_L =	$\ln \mu_g$
μ_c, μ_L	μ_L =	$\ln \mu_c - 0.5(\sigma_L^2)$
σ_g	σ_L =	$\ln \sigma_g$
μ_c, σ_c	σ_L =	$(\ln[1 + \sigma_c^2/\mu_c^2])^{0.5}$

μ_c, σ_c: mean and standard deviation of concentration
μ_g, σ_g: geometric mean and geometric std. dev. of concentration
μ_L, σ_L: mean and standard deviation of the natural logs of concentration

two columns of concentrations shown, one contains observations that are 1,000 times the size of the other. Consequently, the mean and the standard deviation of the second sample are 1,000 times larger than the first. In the third and fourth columns, the logarithms are given for the two sets of concentrations, and column three reflects the natural logs of column one. Note that the standard deviations of the logarithms for the two samples are equal, and consequently the geometric standard deviations (the exponentials of the standard deviations of the logs) are also equal. Hence, even though the magnitudes of the two data sets are considerably different, the geometric standard deviations are the same; so the geometric standard deviation is independent of the order of magnitude of the values in a population.

THE RELATIONSHIP OF SAMPLING AND MEASUREMENT ERRORS

Although the concept will be revisited later, it is important for the purposes of completing the ideas presented in this chapter to consider the magnitude of statistical sampling error associated with a geometric standard deviation in the range of 1.5 to 3.5 compared to typical measurement error values. Disregarding for the moment the units of what we are measuring, consider a measurement method that has a precision or CV of ± 10% (bias corrected or more often near 0). This measurement error is typical for industrial hygiene airborne concentration