CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-4

DISTRIBUTION MODELS

A guide for reviewers, data analysts and interpreters on the statistical properties of common distribution models

This guidance document is one of a series that outlines important basic statistical concepts and procedures that are useful in contaminated sites studies. BC Environment recommends that these suggestions be followed where applicable, but is open to other techniques provided that these alternatives are technically sound. Before a different methodology is adopted it should be discussed with BC Environment.

April 2001

THE GENERAL IDEA

Statistical applications for contaminated site studies commonly make use of theoretical distributions such as the normal and lognormal distribution. This document presents some of the common choices for distribution models and discusses their properties; it is intended to serve two audiences:

- reviewers who need an overview of the common distribution models and their characteristics, and
- data analysts and interpreters who need information on how to calculate percentiles and other summary statistics for some of the common distributions.

This guidance document discusses the normal, lognormal and exponential distributions, three of the more common and useful distribution models for statistical studies of data from contaminated sites. Before discussing the characteristics of these particular distributions, there are two things that should be made clear. First, it may not be necessary to choose a distribution model at all; for many of the statistical problems commonly encountered in contaminated site studies, a distribution model is not necessary. Second, in addition to the three distribution models discussed here, there are many others that might also be useful for particular problems at specific sites; Johnson and Kotz (1970) provide details on a wide variety of alternatives.

Other guidance documents in this series discuss these two closely related topics. The document entitled *NONPARAMETRIC METHODS* discusses whether or not a distribution model is necessary and presents some statistical methods that do not require any distribution assumption. If it is necessary to choose a distribution model, the document entitled *CHOOSING A DISTRIBUTION* provides advice on how to select an appropriate model and how to document the reasons for this choice. Readers are strongly encouraged to read these other two documents so that they have a more complete appreciation of the various issues surrounding the selection of a distribution model.

The distribution models discussed in this document have only one mode; histograms of actual data from contaminated sites sometimes show two or more modes. Such multimodal behaviour is usually due to a mixture of two or more populations. It is common to find with heavy metals, for example, that the data represent a mixture of two distributions, one that reflects the naturally occurring background concentrations and another that reflects the concentrations of material affected by industrial and other anthropogenic contamination. The document entitled *IDENTIFYING POPULATIONS* discusses the issue of separating data into different subpopulations.

THE NORMAL DISTRIBUTION

Overview

The most commonly used (and chronically misused) distribution model in statistics is the normal distribution. Data values from a normally distributed population have a histogram that looks like the one in Figure 1: fairly symmetric with the most common values representing the middle of the distribution and with extremely low or high values being equally uncommon.

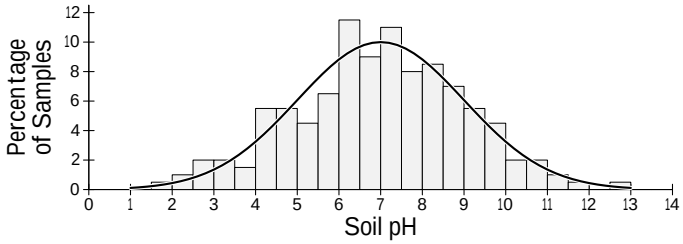

Figure 1 A histogram of normally distributed data.

The use of the normal distribution is often defended by invoking the Central Limit Theorem, which states that any distribution will tend to look more and more like a normal distribution as we average values together. While this is an interesting statistical fact, it has some important restrictions that limit its practical relevance for contaminated site studies. The most important of the assumptions that underlie the Central Limit Theorem is the assumption that the values being averaged together are independent. Violation of the independence assumption is, apart from systematic errors, the most critical violation of the common statistical assumptions. Soil or water contamination is not the result of the averaging of several independent events. Sample values from contaminated site studies rarely have the pleasing symmetry of the normal distribution; it is much more "normal" to see a lot of low values and a decreasing proportion of erratic high ones.

Despite the fact that the normal distribution usually does a poor job of modelling the distribution of contaminant concentrations, it does have a useful role to play in some specific applications. Certain variables, such as porosity in many groundwater studies or the pH of soil or water, have distributions that are fairly symmetric and that rarely have the kind of extreme values that are characteristic of contaminant concentrations.

Classifying stockpiled material based on the average concentration of the stockpile is the other common situation in which the normal distribution may be an appropriate model. The distribution of the average concentration of a contaminant over

large homogeneous volumes of material will definitely be more symmetric than the distribution of the contaminant concentration from discrete samples. The average concentrations of several stockpiles may, in some situations, be viewed as averages of many independent values from a common population. They can be viewed as averages since the true average grade of any individual stockpile is the average of the vast numbers of discrete samples that make up that stockpile; they can be viewed as coming from a common population as long as all of the stockpiled material is drawn from a homogeneous area; and they can often be viewed as independent because the stockpiling process usually removes the spatial correlation that might have existed in the *in situ* material. For these reasons, we may be justified in assuming that average concentrations of stockpiled material will follow a normal distribution.

Calculation of percentiles

The smooth symmetric curve in Figure 1 shows the relative frequencies that the normal distribution model predicts. One of the attractions of the normal distribution is that this theoretical curve of relative proportions is fully determined by the mean and standard deviation of the distribution.

The "standard" normal distribution is one with a mean of 0 and a standard deviation of 1; many statistical textbooks contain tables of the percentiles of the standard normal distribution. The percentiles of any other normal distribution can be calculated by first calculating the corresponding percentile of the standard normal distribution, and then multiplying the result by the standard deviation and adding the mean. For example, suppose we need to calculate the 90th percentile of a normal distribution whose mean is 50 ug/g and whose standard deviation is 10 ug/g. From a table that gives the percentiles of the standard normal distribution, such as Table 26.1 in Abramowitz and Stegun (1970), we know that 1.28 is the 90th percentile of the standard normal distribution. The 90th percentile of our normal distribution is therefore

$$\begin{aligned} \text{90th percentile} &= 1.28 \times \text{Standard deviation} + \text{Mean} \\ &= 1.28 \times 10 + 50 = 62.8 \text{ ug/g} \end{aligned}$$

In addition to the tables provided in many books, there are also some approximations that can be implemented on a programmable calculator or a computer; Kennedy and Gentle (1980) provide a good discussion on the numerical approximations for the percentiles of the standard normal distribution.

68% and 95% confidence intervals

Many statistical procedures make use of the fact that a value drawn randomly from a normal distribution has a 68% chance of falling within one standard deviation of the mean, a 95% chance of falling within two standard deviations of the mean and a 99% chance of falling within three standard deviations from the mean (see Figure 2). Wherever we see a statistical statement involving 68% or 95% "confidence intervals", we can be fairly sure that an assumption of normality has been made. Not all statistical statements involving 68% and 95% depend on an assumption of normality, but the vast majority do.

An example of a remediation decision for which the normal distribution is commonly assumed is the classification of stockpiled material. We don't know the true average contaminant concentration of the stockpile, but we may choose to view this unknown average concentration as a value drawn randomly from a normal distribution. Having decided to model the unknown average concentration in this way, we now need to choose the parameters for this distribution. Using samples from the stockpile, and the methods outlined in the guidance document entitled *ESTIMATING A GLOBAL MEAN*, we may decide that our normal distribution has a mean of 80 ug/g and a standard deviation of 10 ug/g. Having selected the parameters for our normal distribution, we are now able to make predictions about the chance that the unknown average concentration will exceed various thresholds. With a mean of 80 ug/g and a standard deviation of 10 ug/g, there is a 95% chance that the unknown average concentration will be between 60 ug/g and 100 ug/g (two standard deviations from the mean). If we are concerned only with the chance that the unknown average will exceed a regulatory limit of 100 ug/g, then the symmetry of the normal distribution entails that there is only a 2.5% chance that the unknown average concentration will exceed 100 ug/g.

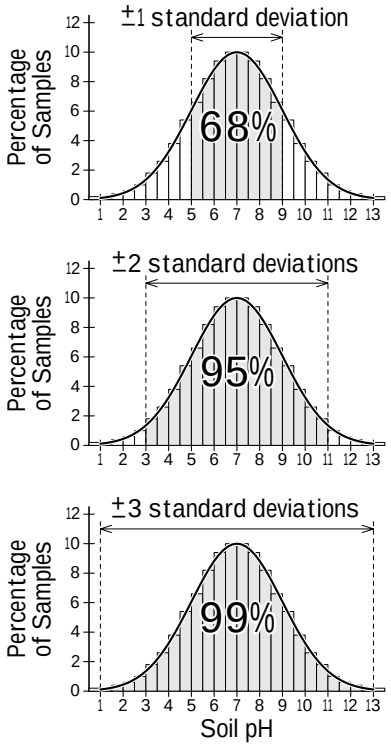

Figure 2 Probability of a value drawn at random from a normal distribution falling within one, two and three standard deviations of the mean.

THE LOGNORMAL DISTRIBUTION

Overview

Distributions of contaminant concentrations rarely have the kind of symmetry that makes the normal distribution an appropriate model; it is common to find that data from contaminated site studies contain many low values and a decreasing proportion of high values. A distribution model that captures this kind of asymmetry is the lognormal distribution. Figure 3 shows an example of the histogram of data drawn from a lognormal distribution; it has a lot of low values and a steadily decreasing proportion of high ones.

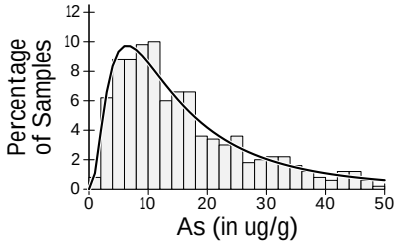

Figure 3 A histogram of lognormally distributed data.

The lognormal distribution, as the name implies, is one for which the logarithms of the data values are normally distributed. It is used for many earth science problems in which the data values span several orders of magnitude and have an

asymmetric distribution. In the same way that the use of the normal distribution is often defended by arguing that the data are the result of a large number of independent additive events, the lognormal distribution can be defended by arguing that the data are the result of a large number of independent multiplicative events. There are a few papers in the technical literature that use arguments based on reaction rates and chemical reactions to support the point of view that the genesis of certain kinds of contamination does, indeed, involve a series of independent multiplicative events. Despite such observations, it is fair to say that the success of the lognormal model is not really due to independent multiplicative events, but due to the fact that by offering us an asymmetric distribution, the lognormal model corrects the major practical deficiency of the normal distribution.

Calculation of percentiles

The lognormal distribution shares with the normal distribution the fact that it is completely determined by the mean and standard deviation; it does not, unfortunately, share its computational convenience. The calculation of a percentile is typically accomplished by first calculating the percentile in terms of the logarithm and then exponentiating the result. In order to calculate the percentile in terms of the logarithm, we first need to know the mean and standard deviation of the logarithms. The following equations describe how the mean, m, and the standard deviation, s, of lognormally distributed values are related to the mean, α, and the standard deviation, β, of their logarithms:

$$m = \exp\left(\alpha + \frac{\beta^2}{2}\right) \qquad s = m\sqrt{\exp\left(\beta^2\right) - 1}$$

$$\alpha = \log(m) - \frac{\beta^2}{2} \qquad \beta = \sqrt{\log\left[1 + \left(\frac{s}{m}\right)^2\right]}$$

where all of the logarithms are natural (base e) logarithms.

As an example of how to calculate a percentile for a lognormal distribution, we can take the lognormally distributed data shown in Figure 3 and find their 90th percentile. The mean of the arsenic values shown in Figure 3 is 18.9 ug/g and their standard deviation is 20.1 ug/g. Using the equation given above for β, we can calculate that the standard deviation of their logarithms should be 0.87; and, using the equation for α, their mean should be 2.62. As discussed earlier in the section on calculating percentiles for a normal distribution, tables from reference books tell us that the 90th percentile of a normal distribution is 1.28 standard deviations above the mean. So, in terms of the logarithms, the 90th percentile would be

$$\begin{aligned} \text{90th percentile of logs} &= 1.28 \times \text{Standard deviation} + \text{Mean} \\ &= 1.28 \times 0.87 + 2.62 = 3.73 \end{aligned}$$

This result needs to be exponentiated to get the 90th percentile of our original arsenic values:

$$\text{90th percentile of original values} = \exp(3.73) = 41.8 \text{ ug/g}$$

In addition to the exact calculations that can be done with the equations given above, there are some rules of thumb that may be useful to get a quick idea of where various high percentiles of a lognormal distribution lie. Most of these back-of-the-envelope calculations make use of the coefficient of variation (CV), which is the ratio of the standard deviation to the mean, and express the percentile as a multiple of the median (not the mean). For a lognormal distribution with a CV of 1 (the standard deviation is equal to the mean), the 90th percentile is roughly three times the median, the 95th percentile is nearly four times the median and the 99th percentile is nearly seven times the median. If the CV climbs to 2 (the standard deviation is twice the mean), then the 90th percentile is five times the median, the 95th percentile is eight times the median and the 99th percentile is almost twenty times the median.

The arsenic data shown in Figure 3 have a median of 13.7 ug/g, and their coefficient of variation is close to 1. We can use the rules of thumb given above to conclude that the 90th percentile will be fairly close to three times the median, or roughly 41 ug/g — a quick, but still very good, approximation to the exact value of 41.8 ug/g that we calculated earlier.

THE EXPONENTIAL DISTRIBUTION

Overview

The lognormal distribution is not the only distribution that allows us to capture the fact that low values are more common than high ones. One of the other common distributions that has the same kind of asymmetry as the lognormal distribution is the exponential distribution. Figure 4 shows an example of the histogram of data drawn from an exponential distribution. Like the lognormal distribution, it has a lot of low values and a steadily decreasing proportion of high ones. It differs from the lognormal distribution in the behaviour of the very lowest values. For the exponential distribution, lower values are always more common than higher ones; for the lognormal distribution, the very lowest values are actually not quite as common as some of the slightly higher values. In Figure 4, the tallest bar on the histogram is the first one; on Figure 3, however, the first bar is not the tallest one. The curve drawn with the heavier line in Figure 4 shows the relative frequencies that the exponential distribution model predicts.

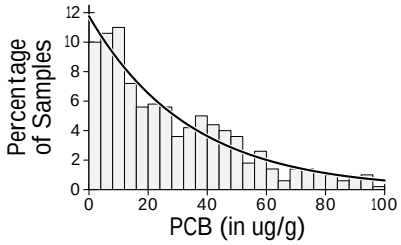

Figure 4 A histogram of exponentially distributed data.

While normal and lognormal distributions need two parameters, the mean and the standard deviation, the exponential distribution is completely determined by its mean. The standard deviation of an exponential distribution happens to be equal to the mean, so this distribution may be appropriate for data whose coefficient of variation is close to 1. The calculation of percentiles for an exponential distribution is more straightforward than for the lognormal distribution since the mean is the only parameter involved. Percentiles for an exponential distribution can be calculated using the following equation:

$$\text{p-th percentile} = -m \times \log\left[1 - \frac{p}{100}\right]$$

where m is the mean and the logarithm is the natural (base e) logarithm. Using the PCB data shown in Figure 4 as an

example, their mean is 34.2 ug/g, so their 90th percentile is calculated as follows:

$$\begin{aligned} \text{90th percentile} &= -34.2 \times \log\left[1 - \frac{90}{100}\right] \\ &= 78.7 \text{ ug/g} \end{aligned}$$

In addition to the exact calculation given above, there are some rules of thumb that can be used to get a quick approximation of some of the high percentiles. For an exponential distribution, the 95th percentile is three times the mean and the 99th percentile is almost five times the mean. The 90th percentile is not very close to being a simple multiple of the mean; it happens to be 2.3 times the mean.

ASYMMETRIC CONFIDENCE INTERVALS

When discussing the normal distribution, we pointed out that 68% of the values fall within one standard deviation of the mean and 95% fall within two standard deviations. This property of the normal distribution is often used as the basis for making statements about the "confidence intervals" for an estimate. The typical assumption is that the quantity we are trying to estimate can be modelled by a normal distribution, that our estimate represents the mean of this distribution, and that we have somehow been able to express the uncertainty in our estimate as a standard deviation. The guidance document entitled *ESTIMATING A GLOBAL MEAN* gives an example of this type of procedure where, under an assumption of independence, the mean of a statistical population can be estimated by m, the mean of the available samples, and the standard deviation of this estimate is $\sigma_m = s \div \sqrt{N}$ where s is the standard deviation of the individual samples and N is the number of available samples. Though this approach is valid for quantifying the uncertainty on the mean of any distribution of values, whether normal or not, an assumption of normality is made as soon as we use this information to report $m \pm \sigma_m$ as our "68% confidence interval" or $m \pm 2\sigma_m$ as our "95% confidence interval".

Table 1 Lead concentrations (in ug/g).

12	191	872	13	52	92	43	17	5	59

With data that are clearly skewed (as are most data from contaminated site studies), the uncertainty about the mean is not likely to follow a normal distribution, especially if there are only a few samples available for estimation. Table 1 shows an example of 10 samples of lead concentrations in the soil from a contaminated site. Using these ten data, and assuming that they are independent, we can estimate that the mean of the population from which they came is 135.6 ug/g and that the standard deviation of this estimate is 83.7 ug/g. Up to this point, we have made no assumption about the underlying distribution, we have simply applied the equation given above. Given the very evident skewness of these data, it makes little sense to assume that the uncertainty on our estimate is going to follow a normal distribution. The normal 95% confidence interval, for example, would be 135.6±167.4 ug/g; a dose of common sense tells us that there's not a lot of meaning in a confidence interval that goes down to -31.8 ug/g on the low side.

When the quantity we are trying to estimate is better modelled by a skewed distribution, it is more useful to calculate confidence intervals directly from the percentiles than to use the classical $\pm\sigma$ and $\pm 2\sigma$ intervals. Regardless of the distribution, there is a 95% chance that a value will fall between the 2.5th percentile and the 97.5th percentile, so we can use these percentiles directly to report a 95% confidence interval. This approach works for any distribution, even a normal one, and is the only sensible way to report confidence intervals where the distribution is not normal. To continue with the example of the data in Table 1, it is more appropriate to assume that their unknown true mean follows a lognormal distribution with the mean and standard deviation reported above. Using the method outlined earlier, in which we calculated the percentile in terms of the logarithm and then exponentiated the result, the 2.5th percentile of our unknown mean is 37.9 ug/g and the 97.5th percentile is 351.4 ug/g. Using this information, we can report an asymmetric 95% confidence interval of 37.9 – 351.4 ug/g for our estimate of the mean.

RECOMMENDED PRACTICE

1. If a distribution model is necessary for some calculation, and if a normal, lognormal or exponential model has been chosen, the equations given in this guidance document can be used to calculate percentiles. It is recommended that the exact equations be used wherever possible and that the rules of thumb be used only for rough calculations. In all cases where a percentile or confidence interval is calculated from some distribution model, the type of model should be reported along with its parameters.
2. If a mean and standard deviation are being used to calculate confidence intervals, the classical $\pm\sigma$ 68% confidence interval and $\pm 2\sigma$ 95% confidence interval should not be used unless there is good reason to believe that the quantity being estimated follows a normal distribution. If a skewed distribution is more appropriate, the 95% confidence interval should be reported as the range from the 2.5th percentile to the 97.5th percentile.

REFERENCES AND FURTHER READING

This guidance document does not provide specific guidance on when to choose a distribution model or how to choose an appropriate distribution model. These issues are addressed in the guidance documents entitled *NONPARAMETRIC METHODS* and *CHOOSING A DISTRIBUTION*. In addition to the other guidance documents in this series, the following references provide useful supplementary material.

Abramowitz, M. and Stegun, I.A., (eds.), *Handbook of Mathematical Functions*, Dover, New York, 1970.

Blake, I.F., *An Introduction to Applied Probability*, John Wiley & Sons, New York, 1979.

Johnson, N.L. and Kotz, S., *Distributions in Statistics — Continuous Univariate Distributions, Volume 1*, Houghton Mifflin, Boston, 1970.

Kennedy, W.J. and Gentle, J.E., *Statistical Computing*, Marcel Dekker, New York, 1980.