CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-6

CHOOSING A DISTRIBUTION

A guide for data analysts and interpreters on how to select an appropriate distribution model and document the choice

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

Many statistical procedures used in contaminated site studies involve assumptions about the underlying distribution of data values. If these assumptions are poorly founded, our statistical interpretations may be very misleading; we must be clear about our assumptions in order not to waste time on meaningless calculations. Even if our assumptions are well founded, a failure to state them clearly or to justify them in a report may leave doubt in a reviewer's mind about the validity of our conclusions; we owe it to those who eventually review our work to provide a clear statement of what has been assumed and why.

This document discusses the choice of a distribution model and recommends procedures for providing supporting documentation. It begins with an example that demonstrates how different assumptions about the underlying distribution can lead to very different remediation decisions. It then presents three common distribution models and describes statistical tools and procedures that can help us make an appropriate selection. It closes with a section that discusses whether or not a distribution model is necessary; many of the issues that we confront in statistical applications on contaminated site studies can be addressed adequately without assuming any distribution model.

There are other documents in this series that the reader should also examine. *DISTRIBUTION MODELS* provides an introduction to the distribution models discussed here and provides more detail on their statistical characteristics than is covered in this document. *NONPARAMETRIC METHODS* discusses statistical methods that do not require any distribution assumption. *IDENTIFYING POPULATIONS* addresses the problem of separating data into subpopulations, a common concern when the distribution of the available data appears multimodal and may reflect a mixture of two or more distributions.

INTRODUCTORY EXAMPLE

Even with exactly the same sample data, different choices of distribution models can lead to different remediation decisions. To take a simple example, consider the data shown in Table 1; these are measurements of the total volatile petroleum hydrocarbon (VPH) concentration in ten discrete samples taken at an early stage from a site containing roughly 50,000 cubic metres of soil, some of which may be contaminated.

Table 1 VPH values (in ug/g).

11	41	3	196	52	7	107	81	22	16

As a preliminary step, we might need to get a ballpark estimate of how much soil is considered industrial quality according to BC Environment regulations and how much has to be regarded as waste and removed from the site. We therefore decide to use statistics to help us estimate the proportion of material that exceeds the BC Environment industrial soil quality threshold of 200 ug/g VPH. Three of many possible distribution models we could adopt are:

- The data are from a normal distribution.
- The data are from an exponential distribution.
- The distribution of total VPH values over the entire area is exactly the same as that currently shown by the ten available samples (i.e. the highest value is exactly 196 ug/g).

If we assume that the sample values given in Table 1 are independent, we can use them to estimate a mean of 53.6 ug/g and a standard deviation of 60.6 ug/g for the underlying population. Using methods described in *DISTRIBUTION MODELS*, we can calculate that if the data are normally distributed, then 0.8% of the underlying distribution would exceed 200 ug/g This corresponds to about 400 cubic metres of soil that could not be considered industrial quality and would have to be removed from the site. Under the second assumption, we can use the same information to calculate that if the data are exponentially distributed, then 2.4% of the underlying distribution would exceed 20 ug/g. This corresponds to about 1,200 cubic metres of soil that could not be considered industrial quality. Finally, if we adopt the third assumption, then all of the soil could be considered industrial quality and none would have to be removed from the site.

The three assumptions lead to very different predictions about how much material will need to be remediated, with the difference between their corresponding costs amounting to several hundreds of thousands of dollars.

As this example shows, the choice of an appropriate distribution model may be critical, especially when it is based on few samples and is used to predict the probability of extreme events.

SOME COMMON DISTRIBUTION MODELS

This guidance document discusses the three distribution models shown in Figure 1: the normal, lognormal and exponential distributions; these are the distribution models whose properties and statistical characteristics are described in *DISTRIBUTION MODELS*. While these are three of the more common and useful distribution models for statistical studies of contaminated sites, there are many other distribution models that may also be useful for particular problems at specific sites; Johnson and Kotz (1970) provide details on a wide variety of alternatives.

HOW TO CHOOSE A DISTRIBUTION MODEL

When confronted with the need to document why we have chosen a particular distribution model, there are several different types of arguments that we can present. Some of these are specific quantitative calculations, others are more qualitative. Ideally, we should be able to use both kinds of arguments.

Symmetric or not?

One calculation that can help us decide whether to choose a symmetric distribution, such as the normal distribution, or an asymmetric one, such as the lognormal or exponential distributions, is to compare the mean to the median. With symmetric distributions the two should be just about the same; with the asymmetric distributions commonly encountered in contaminated site studies, the mean is often much larger than the median. A boxplot provides a quick graphical check of whether the mean is close enough to the median to warrant an assumption that the underlying distribution is symmetric; if the mean plots outside the box (i.e. above the 75th percentile) then a symmetric distribution is not an appropriate model.

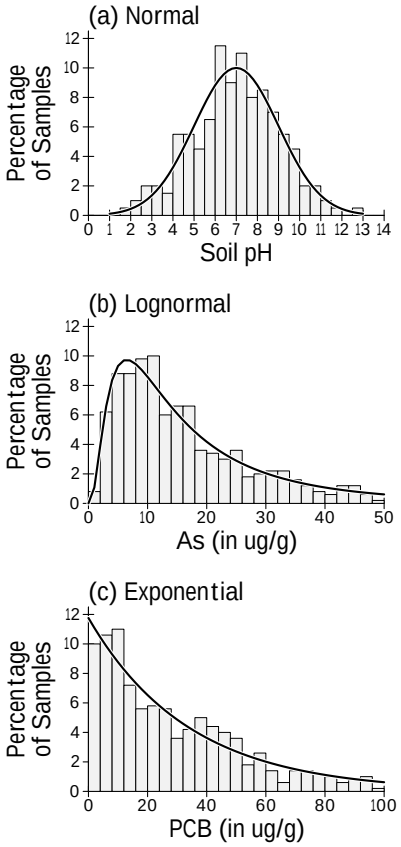

Figure 1 Examples of (a) normally distributed data, (b) lognormally distributed data and (c) exponentially distributed data.

Even if the mean is below the 75th percentile, this does not mean that a symmetric distribution is appropriate; with large data sets, we should expect much closer agreement between the mean and the median if we are going to assume a symmetric distribution. For a data set containing N values, a symmetric distribution is not an appropriate model if the difference between the mean and the median is larger than the standard deviation divided by $\sqrt{N}$.

Histograms, cumulative plots and probability plots

Figure 1 shows a graphical presentation that helps document the rationale for a distribution model: a plot of the relative frequencies predicted by the model along with the histogram of the data. Though this style of presentation may help to sort out hopelessly inappropriate models, it can also be somewhat deceptive. Figure 2 shows a histogram of arsenic measurements from a contaminated site. Superimposed on this histogram is the relative frequency curve predicted by a lognormal distribution with a mean of 14.1 ug/g and a standard deviation of 10.0 ug/g Though this looks like a good fit, it is very misleading to claim that the data come from a distribution whose mean is 14.1 ug/g because the actual mean of the data is 43.0 ug/g, more than three times that of our theoretical model!

Figure 3 shows a cumulative plot of the arsenic data along with the cumulative curve predicted by the same theoretical lognormal model we considered earlier. From this plot it is clear that although the lognormal does a reasonably good job with the lower values, it does a very poor job with the high values. The plot shown in Figure 2 is deceptive because it doesn't show how badly we do at the very high end. If we are trying to give convincing graphical support for our distribution model, we should show how the cumulative plot of the actual data compares to the corresponding theoretical curve predicted by our distribution model.

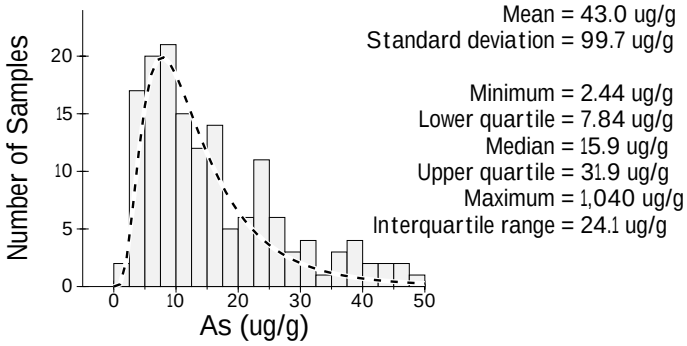

Figure 2 A histogram of arsenic concentrations along with a lognormal distribution model.

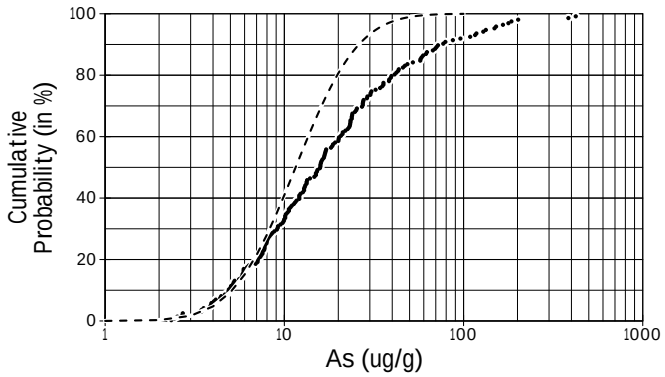

Figure 3 Cumulative plot of arsenic concentrations from Figure 2 along with the same lognormal distribution model.

Probability paper

With certain distribution models, such as the normal and lognormal ones, there is a convenient way to check if the cumulative plot of the actual data is close to what the theoretical model predicts. Rather than plot the actual and theoretical curves, as we did in Figure 3, we can use special probability paper whose axes have been scaled in such a way that the cumulative probabilities of the actual data will plot on a straight line if the data do, in fact, represent the distribution model we have chosen. With normal probability paper, the cumulative probability axis is squashed in the middle and stretched at the ends so that a cumulative normal distribution, which plots as an S-curve on an arithmetic scale, will plot as a straight line. If the cumulative probabilities of the data plot as a straight line

on normal probability paper, then we have some justification for choosing a normal distribution as a model.

In addition to distorting the cumulative probability axis, lognormal probability paper also uses a logarithmic scale on the data value axis. If the cumulative probabilities of the data plot as a straight line on lognormal probability paper, then we have some justification for choosing a lognormal distribution as a model.

Statistical tests

For any distribution model that we might choose, there will always be some differences between the proportion of the actual data values that fall within a particular class on our histogram and the proportion that should fall within that class according to our theoretical model. The chi-square test (Bratley et al., 1983) provides a way of testing whether these differences between the actual and theoretical proportions are significant enough that we should abandon our distribution model. Unfortunately, this test is very permissive; in those few cases where it rejects our distribution assumption, we would likely reach the same conclusion through a comparison of the actual and theoretical cumulative probability plots. Other disadvantages of the chi-square test are that it needs more data to work well than are commonly available in contaminated site studies, and that independence is a necessary assumption.

Is there a good default model?

It is tempting to tackle the problem of choosing a distribution model by taking the point of view that one particular model is the best choice barring any strong evidence to the contrary. Unfortunately, there is no distribution that can serve as a good default for the wide variety of statistical problems that arise in contaminated site studies since the shape of the distribution depends on the volume of material in question. A histogram of discrete sample values from a contaminated site will look much more skewed than the histogram of the average concentrations of large stockpiles. Since one of the main reasons for choosing one distribution model over another is its skewness, it is important when choosing a preliminary distribution model to be clear on what volume the data are based.

When dealing with values defined on a relatively small volume, such as concentrations of discrete or composite samples, we should expect the distribution of data values to be skewed. Until we have enough data to confirm or refute a specific distribution, data values based on small volumes should be assumed to follow an asymmetric distribution, such as the lognormal or exponential distribution; the normal distribution is *not* a good default choice for such data.

A normal distribution is a good default choice only if we know that we are dealing with values defined on a large and homogeneous volume, such as the average concentration of an entire stockpile. It is important that the volume be both large and homogeneous. A common yardstick for measuring homogeneity is the coefficient of variation, which is discussed in *UNIVARIATE DESCRIPTION*; if the coefficient of variation is greater than 1, then it is not appropriate to assume homogeneity. If the material is not homogeneous, the contaminant concentrations typically span several orders of magnitude, and the distribution of the average concentration of entire stockpiles, even large ones, may still be noticeably skewed.

As an example of the linkage between the choice of a distribution model and the volume of material under consideration, consider the problem of interpreting the results of a composite sample taken from a stockpile. If we are trying to address the issue of whether any single discrete sample value in the composite might have exceeded some threshold then we are concerned with data values that are based on a very small volume of material: a single discrete sample. For this issue, our preliminary assumption should be that the data values follow a skewed distribution. With the same sample information, however, we might be trying to address the issue of whether the average concentration of the entire stockpile is above some threshold; we are now interested in data defined on a much larger volume: an entire stockpile. For this issue, we could adopt a normal distribution as our preliminary model if we can assume that the stockpiled material is homogeneous; as discussed in the document entitled *STOCKPILING*, such an assumption of homogeneity is best supported by a careful and thorough *in situ* characterization study.

WHY CHOOSE A DISTRIBUTION?

Before choosing a distribution model, it is worth considering whether we really need one. How is our work made any easier or better by assuming a distribution model?

A historical perspective

The reason why statisticians seem to spend so much time worrying about an appropriate distribution model lies in the early part of this century, when data were scarce and computers nonexistent. In an era without computers or calculators, the initial focus of a statistical study was on finding a tractable, well understood mathematical model that described the distribution of the data values. Though the data set in a typical statistical study from the early part of this century would now be considered quite a small data set, it was still difficult to deal with the raw data. Even with as little as 20 or 30 data values, simple mathematical calculations, such as the mean or the standard deviation, are tedious when they have to be done manually.

The life of statisticians in the early part of this century was made much easier by the pioneering work of mathematical statisticians like Sir Robert Fisher, who added a great deal to the knowledge of how certain distributions behave. With a large and growing literature on the properties of various distribution models, statistical studies were considerably simplified if the actual data were replaced by a model. It is possible to make much quicker progress with a normal distribution model, for example, than to struggle through manual calculations with actual data. Once the parameters of the distribution model have been chosen, typically the mean and standard deviation, it is possible to make many different kinds of predictions about the behaviour of the entire population. We could calculate its percentiles, for example, its skewness, its peakedness, its mode, its median, and so on — all without having to grind the actual data through another set of calculations.

With the advent of modern computers, however, the need for a

distribution model becomes questionable. With computers able to rapidly sort data, even if there are thousands of values, and able to calculate even the most complicated statistics in a few seconds, why should a tractable and well-studied mathematical model be of much interest?

Advantages of distribution models

Even though their computational convenience is now largely a matter of historical curiosity, distribution models possess other advantages.

Some kind of model is necessary if we are trying to make predictions about events that are so rare that they are never (or hardly ever) observed. Those weird and wonderful statistics about how much more likely it is that we'll get hit by a meteorite than suffer a fatal accident related to nuclear reactors, are all based on distribution models for low probability events. Statistical predictions of this type are very sensitive to the way that the distribution model behaves for extreme values. As we saw in the introductory example with VPH concentrations, there can be considerable variability in predictions about the chance of exceeding a threshold that no data value has yet exceeded. If our distribution model predicts a rapid decrease in the occurrence of extreme values (like the normal distribution does), then we're not going to calculate a very high chance of exceeding the threshold; if, on the other hand, the model predicts a slower decrease in the occurrence of extreme values (like the exponential distribution does), then we're going to calculate a higher chance of exceeding the threshold.

Another advantage of choosing a distribution model is that it gives us a very compact way of describing a data set. Where a need exists to communicate the essential features of a data set to other people, it is often easier to say something like "the VPH concentrations follow an exponential distribution with a mean of 54 ug/g" than to list all of the available data. When used in this way, the distribution model is useful only if our audience is already familiar with its shape and statistical parameters.

The use of distribution models as a kind of shorthand notation for describing data takes on a sharper focus when, in certain fields of study, the use of specific distributions is so common that workers in the same discipline can use the parameters of the distribution as diagnostic features. For example, although the parameters of the Weibull distribution, commonly called λ and α, are not likely to be familiar to most people, they are so commonly understood by many of the researchers who study the failure rates of communication systems that experimental data sets from this area of application are often summarized with these two parameters alone.

The final advantage of some distribution models is that they simplify certain inferences and predictions. The most notably convenient and computationally simple distribution model is the normal distribution. In order to quantify the uncertainty on an estimate, our job is made much easier if we assume that the errors we might make with our estimate are normally distributed. Having made this assumption, all that is needed to develop confidence intervals is an estimate of the standard deviation of the estimation errors. Once this is available, the 95% confidence interval goes from two standard deviations below to two standard deviations above the estimate. When used in this way, the distribution model is not something that we choose after thoughtful consideration of our data, it is something that we *hope* is appropriate because it makes our calculations easier. When this hope has no justificati on,eagerness for a simple and tractable calculation usually leads to misapplication of a distribution model.

Disadvantages of distribution models

The main disadvantage of using a distribution model is that it may not be appropriate for a particular set of data. While one of the commonly used distributions can usually do a good job of fitting most of the data, few of them do a good job for all of the data values. Typically, there are departures between what a distribution model predicts for the occurrence of extreme values and what the data actually show. With our interest in contaminated site studies often focused on the high values, the good fit of a model over the lower 90% of the data may be useless if it does a poor job of fitting the critical upper 10%.

RECOMMENDED PRACTICE

1. All distribution assumptions should be made explicit in reports.
2. If the difference between the mean and the median of a data set containing N samples is greater than the standard deviation divided by $\sqrt{N}$, then it cannot be assumed that the data come from a symmetric distribution, such as the normal distribution.
3. The appropriateness of a distribution model should be documented graphically by comparing the cumulative probabilities of actual data to the cumulative probabilities predicted by the theoretical model. Such a comparison should be done on probability paper, if available.
4. If there are too few data to adequately support or refute a distribution model, then discrete samples should be assumed to follow an asymmetric distribution. Average values over large volumes, such as stockpiles, may be assumed to follow a symmetric distribution, such as the normal distribution, if *in situ* characterization has demonstrated the material to be homogeneous.

REFERENCES AND FURTHER READING

In addition to the other guidance documents in this series, the following references provide useful supplementary material:

Bratley, P., Fox, B.L. and Schrage, L.E., *A Guide to Simulation*, Springer-Verlag, 1983.

Jaeger, R.M., *Statistics — A Spectator Sport*, Sage Publications, 1990.

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

Moore, D.S., *Statistics — Concepts and Controversies* W.H. Freeman and Company, 1985.

Size, W. B., (ed.), *Use and Abuse of Statistical Methods in the Earth Sciences*, IAMG Studies in Mathematical Geology, Volume 1, Oxford University Press, 1987.