CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-7

IDENTIFYING POPULATIONS

A guide for data analysts and interpreters on the identification of statistical populations

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

The essence of statistical inference is the borrowing of information from a group of data to make predictions about how a particular population behaves. The grouping of data and the definition of the population(s) of interest are fundamental and recurring problems in the application of statistical methods to contaminated site studies. On one hand, the uniqueness of each and every sample encourages us to split the data into smaller and smaller groups and to recognize multiple populations in our data. On the other hand, the need for a sufficient number of data to support statistical calculations encourages us to group data together into larger groups and to work with fewer populations, each of which contains more data.

As an example of this problem, consider the twenty sample values listed in Table 1 and shown as a histogram in Figure 1. Are there two populations here, a "background" of low values in the 0 to 10 ug/g range and another population of higher "contaminated" values? Or is there just a single population that happens to be highly skewed with a lot of low values and a decreasing proportion of higher ones?

Table 1

44	140	6.3	76	6.5
2.7	89	86	2.6	14
6.3	67	49	97	5.9
4.3	52	4.0	32	72
4.8	8.6	39	61	89
1.8	120	4.4	63	9.7
7.8	100	20	3.2	8.5
14	5.4	7.2	80	6.2
1.1	5.4	2.3	16	3.8
81	94	3.1	6.4	110

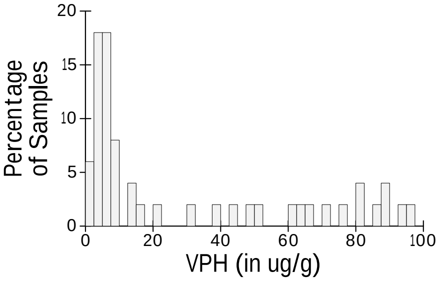


Figure 1 VPH concentrations in soil samples from a contaminated site.

If these data are viewed as a single population, then an appropriate distribution model would need to be asymmetric with a long tail; a lognormal distribution, for example, could do the job. On the other hand, if these data are viewed as a mixture of two populations, then it might not be necessary to use skewed distribution models; a combination of two normal distributions might be more appropriate. These two different approaches to data analysis and interpretation will lead to quite different predictions, particularly when trying to estimate the probability of extreme events.

Unfortunately, there is no statistical test that unambiguously proves that data belong in a single population or that they need to be split into separate populations. Trying to test for whether the data should be grouped or split is a chicken-and-egg problem. Until we assume some underlying population, we have no point of reference against which we can compare our actual data. Developing such a point of reference requires some data (or some bold assumptions) and, depending on which data we choose (or which assumptions we choose to make), we will either conclude that the data should be grouped or we will conclude that they should be split.

Despite the awkwardness of documenting that a particular grouping or splitting decision is appropriate, the issue of identifying the population(s) in a data set is critical. Without a clear definition of the population(s), other important issues, such as the evaluation of outlier data, can not be resolved.

This document presents guidance on identifying statistical populations. It begins with a discussion of graphical tools and then addresses statistical tests that can help with the decision of whether to treat the data as a single population or as several separate populations. There are other guidance documents in this series that contain related material. The one entitled *OUTLIERS* provides additional insight into methods for evaluating whether or not a particular sample should be treated as part of the population; *NONPARAMETRIC METHODS* provides alternatives to the statistical tests outlined here.

QUALITATIVE INFORMATION

The decision to group data into a single population or into several separate populations should, wherever possible, take into account qualitative information. An understanding of the historical use(s) of a site is invaluable in developing an appropriate statistical treatment of the available data. Field notes that describe local conditions in the immediate vicinity of each sample location are also very useful since these often provide critical clues to the physical, chemical and geological conditions that influence contaminant concentrations. A clear understanding of the goal of the study is also necessary in making appropriate decisions about the statistical treatment of the data; though it may be appropriate to group all of the data into a single population for assessing the total volume of soil that requires remediation at a contaminated site, it may be necessary to split the data into several populations if the goal of the study is detailed local mapping of contaminant concentrations.

There are several graphical tools and statistical tests that can be used to support decisions about grouping data together or splitting them into several separate populations. These should not be used by themselves, however, to justify a decision regarding statistical populations. If a probability plot, for example, suggests that there may be a mixture of two populations at the site, then this observation, which is based purely on a quantitative consideration of the data, should be reconciled with the qualitative information about the site. Do the two populations reflect natural background and industrial contamination? Are the magnitudes of the values in the population being designated as "background" consistent with other information about what the natural background levels should be? Could the two populations both be due to industrial contamination but from different sources? If so, are both sources part of the focus of the study? All of these questions, and dozens of similar ones, can not be answered with the data values themselves and need supporting historical and geological information.

GRAPHICAL TOOLS

Probability plots

One of the most useful graphical displays for exploratory data analysis is a probability plot, an example of which is shown in Figure 2. On such a graph, data values are plotted on the x-axis against the cumulative probability on the y-axis. Using Figure 2 as an example, about 35% of the data values are below 100 ug/g, and slightly more than 90% are below 1000 ug/g.

The scaling on the axes of a probability plot is often confusing to non-statisticians. The y-axis is squashed in the middle and stretched at the ends; the distance between 50% and 60%, for example, is smaller than the distance between 80% and 90%. This kind of y-axis scale is used on probability plots because it makes it easier to tell whether the data are close to being normally distributed. If both the x and y axes have a conventional linear scale, then cumulative curves of normally distributed data will plot as an S-curve. It is difficult to tell if an S-shaped curve is close to the kind of S-shape that normal data would produce; it is much easier to tell if the data are plotting on a straight line. By stretching out the y-axis for the very low and very high values, and squashing it for the middle ones, we end up with distorted graph paper on which cumulative curves of normally distributed data will plot as a straight line.

If the distribution of data is skewed, with many low values and a decreasing proportion of high ones, it is common to use a logarithmic scale on the x-axis; this style of log-probability plot is the one that has been used in Figure 2. With logarithmic scaling on the x-axis and the distorted probability scale on the y-axis, cumulative data will plot as a straight line if the data are lognormally distributed.

Whether the data or their logarithms are normally distributed or whether they follow some other distribution, a probability plot is useful for exploring possible sub-populations. If the probability plot has kinks, with some of the values not following the trend of the others, this is often taken as evidence that the data should be separated into different groups. With the example in Figure 3, the mercury data values show a consistent trend up to about the 90th percentile; the highest 10% of the data, however, do not follow the same trend as the lowest 90%. This behaviour indicates that the highest 10% of the values could be treated as a separate population. As discussed earlier, however, the decision to treat the highest 10% as a different population should be supported by qualitative information regarding the history of the site and the source of contamination.

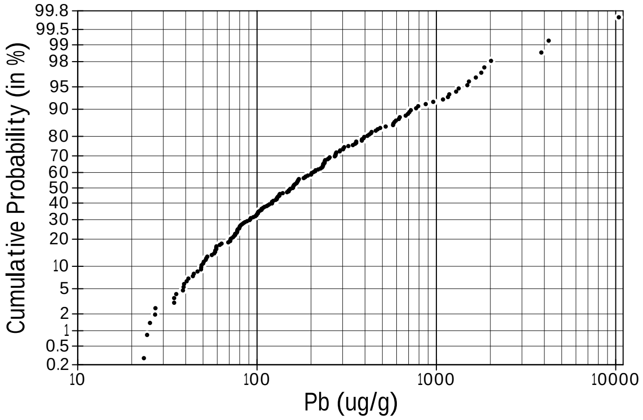


Figure 2 A log-probability plot for lead concentration measurements from the soil in the vicinity of a smelter.

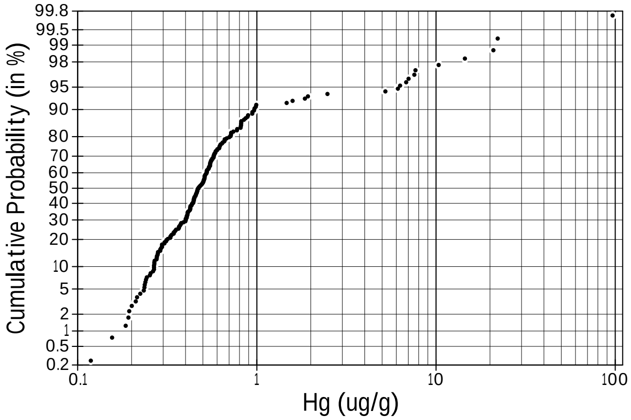


Figure 3 A log-probability plot showing multiple populations.

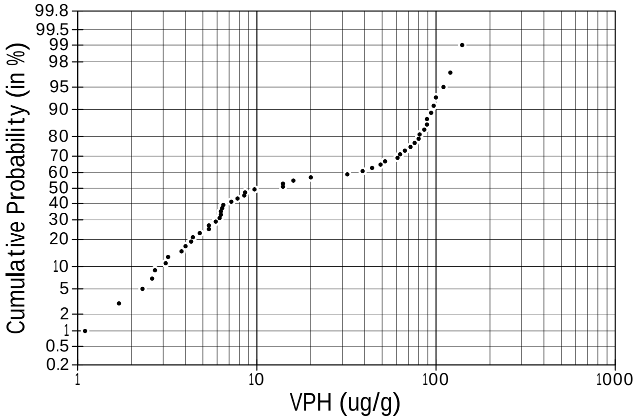
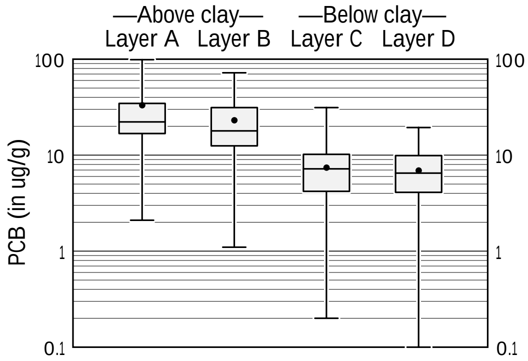


Figure 4 A log-probability plot that shows a mixture of "background" and "contaminated" populations.

Figure 4, which uses the VPH data shown in Table 1, shows another type of behaviour that probability plots often exhibit: the data values seem to plot on two different trends with a gradual transition between the two. Probability plots of this type are usually due to overlapping mixtures of several populations. Such mixtures are common in contaminated sites where the lower end of industrial contamination overlaps with the higher end of natural background contamination. Sinclair (1974) discusses how the information from such probability plots can be used to develop distribution models for each of the mixed populations as well as to calculate the proportion of each population in the mixture.

Side-by-side boxplots

In many contaminated sites there is qualitative information about the site or the soil conditions that allows us, if we deem it necessary, to subdivide the data into different groups. When considering whether such a splitting is appropriate or not, it is often useful to be able to compare the distributions in the different groups under consideration. For example, we might expect the level of PCB contamination at depth to indicate whether or not the surface contamination has passed through a clay layer. This situation differs somewhat from those that we considered in the previous section. With all of the probability plot examples shown earlier we started with the data in a single group and used the probability plot as a tool to study whether it might be more appropriate to subdivide them. With the PCB example given in this section, we already have a grouping in mind — depth from surface and position relative to the clay layer — and we are trying to decide whether this distinction is important or whether no subdivision is necessary.



No. of data	69	58	39	31	No. of data
Mean	33.2	23.1	7.4	6.9	Mean
Stand. Dev.	65.2	47.1	20.4	19.2	Stand. Dev.
Maximum	98.8	72.1	31.3	19.4	Maximum
Upper quartile	34.6	31.3	10.2	9.9	Upper quartile
Median	22.2	17.9	7.2	6.5	Median
Lower quartile	16.8	12.5	4.2	4.1	Lower quartile
Minimum	2.1	1.1	0.2	0.1	Minimum

Figure 5 Side-by-side boxplots of PCB contamination.

One way of comparing distributions from several groups of data is to plot their histograms and tabulate some key summary statistics. Side-by-side boxplots, such as those shown in Figure 5, provide a more useful graphical comparison. The box in the middle of each individual boxplot extends from the lower quartile to the upper quartile of the distribution; the bar in the middle of the box is the median and the "arms" define the range (minimum to maximum). The black dot shows the mean of the distribution.

A boxplot presents most of the relevant univariate information that we need from an exploratory data analysis. It gives us a sense for where the middle of the distribution lies, how spread out it is and whether it is symmetric or not. The boxplot therefore offers most of the useful information that a histogram contains, but in a more compact format that is more amenable to side-by-side comparisons between different groups of data. Where side-by-side boxplots for two groups of data show that their boxes do not overlap — the central 50% of one group does not overlap with the central 50% of the other group — this is evidence that supports treating the two groups separately. As with the other graphical and statistical tools for examining populations, a decision to split data into separate populations should not be based on boxplots alone but should also be supported with qualitative information that explains why the distributions are different.

Scatterplots

Where probability plots suggest a mixture of two overlapping populations, it is often difficult to identify the population to which each sample belongs since intermediate values could be high values of one population or low values of the other. Most contaminated site studies involve a suite of possible contaminants and scatterplots can often be the key to sorting out which samples belong to which populations.

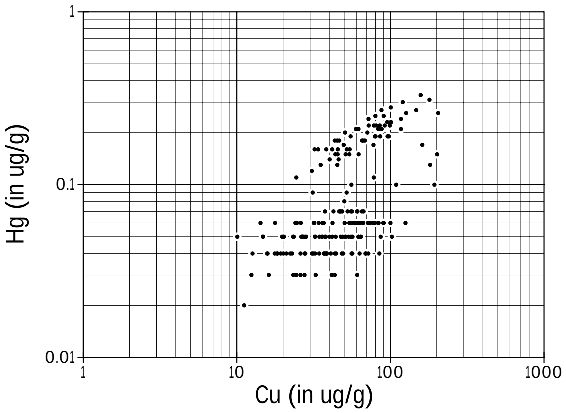


Figure 6 Separate populations on a scatterplot.

Figure 6 shows an example in which the copper contamination consists of two populations that overlap. Using only the copper concentration it is not possible to assign each sample to the background or contaminated population since the low end of the contaminated distribution overlaps with the high end of the background distribution. At this site, mercury concentrations have also been measured. When the mercury and copper concentrations are plotted together on a scatterplot, the separate populations become clearer. With the use of both variables the separation of the data into two separate populations is much more straightforward than when the copper or mercury concentrations are used separately.

STATISTICAL TESTS

In addition to the graphical tools that can be used to support a decision to treat two groups of data separately, there are some statistical tests that can also provide support for such a decision. Statistical tests are often used when the decision to recognize separate populations is not immediately obvious. With the data shown in Figure 5, for example, it is clear that the PCB concentrations above the clay layer are different from those below the clay layer. For the A and B soil layers that are above the clay layer, however, it is not as obvious whether or not these should be treated as different populations. Though their means are different, this could either be due to chance or could also be due to the underlying populations being different in the two layers; even if the two sets of data were drawn from the same underlying population, we would still expect to see some differences in their means. There are statistical tests that help us to decide if a difference between the statistics of two groups of data is due to chance alone or if it is more likely that the two groups were drawn from different populations.

The t-test

The t-test is used to determine if the difference between the means of two populations is "statistically significant". This test begins by assuming that the two groups of data are from the same population and then tries to refute this assumption.

If sets of N data are drawn from a common population, the means of these different sets will fluctuate around the mean of the parent population. How much fluctuation we should expect depends on the value of N; if N is large then the mean of the actual data will be closer to the mean of the parent population than if N is small. If the N values in each set are independent, then σ_m, the standard deviation of the means of the different sets, is related to σ, the standard deviation of the parent population, by the following equation: $\sigma_m = \sigma \div \sqrt{N}$.

In addition to using this equation that describes how much the sample means can fluctuate from the mean of their underlying population, the t-test assumes that the means will be normally distributed. Under all of these assumptions — that the means are, in fact, the same, that the samples are all independent and that the means are normally distributed — the following statistic should follow a standard normal distribution:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

N_1, m_1 and σ_1 are the number of samples in the first group, their mean and their standard deviation; N_2, m_2 and σ_2 are the corresponding values for the second group. If the value of t calculated from the equation above is well within the range of values expected for a standard normal distribution, -3 to +3, then the difference between the means can well be explained by the random fluctuations that we expect between groups of samples drawn from the same distribution. If t is less than -3 or greater than +3 then it is very unlikely that random fluctuations alone are causing the difference. The conventional approach is to interpret extreme values of t as evidence that the two groups come from different populations.

As an example of the use of the t-test, let us determine whether the Layer A and Layer B samples from Figure 5 have significantly different means. The values that we need to substitute in the equation for the t-statistic are the following:

$$N_1 = 69 \quad m_1 = 33.2 \quad \sigma_1 = 65.2$$

$$N_2 = 58 \quad m_2 = 23.1 \quad \sigma_2 = 47.1$$

With these values, the calculated value of the t-statistic is 1.01, well within the -3 to +3 range that we expect for a standard normal distribution. This tells us that the difference between the mean value of the 69 Layer A samples and the 58 Layer B samples is not large enough to lead us to believe that the underlying populations are different.

The t-test given above is what is called a "two-sided" t-test since it takes into account that neither of the sample means is exactly the same as that of the underlying population. In situations where the mean of the underlying population is known (not a very common situation in contaminated site studies), the equation given above can be turned into a "one-sided" t-test by setting m_2 to the true mean and σ_2 to zero. An example of a situation in which we may prefer to do a one-sided t-test is the comparison of laboratory measurements of a reference standard to the accepted value of the standard. In this case, the true mean is the accepted value of the standard and we are interested in whether the mean of repeated measurments is significantly different from this accepted reference value.

The philosophy of the t-test is to make an assumption, namely that the data are from the same population, and then use extreme values of the t statistic to argue that this assumption is not very plausible. It should be noted, however, that in believing that the calculated t-statistic should come from a standard normal distribution, we are making several other assumptions. It is possible that the assumption of a common population is correct and that it is one of the other ones — independence of the samples or normality of the means — that is incorrect.

There are some statistical tests for differences between populations that do not make distribution assumptions; some of these are discussed in the guidance document entitled *NONPARAMETRIC METHODS*. If the samples are not independent (a common case in contaminated site studies) then the difference between the two means can be larger than the t-test assumes. If two groups of data fail the t-test under an assumption of independence — if their t-statistic is between -3 and +3 — then they would also fail a modified version of the test when correlation between the samples is taken into account.

REFERENCES AND FURTHER READING

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

Davis, J.C., *Statistics and Data Analysis in Geology*, 2nd edition, John Wiley & Sons, New York, 1986.

Sinclair, A.J., "Selection of threshold values in geochemical data using probability graphs," *Journal of Geochemical Exploration*, v. 3, p. 129 – 149, 1974.

Understanding Robust and Exploratory Data Analysis, (Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds.), John Wiley & Sons, New York, 1983.