CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-1

UNIVARIATE DESCRIPTION

A guide for report writers, reviewers, data analysts and interpreters on exploratory data analysis for one variable

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 application of statistics to contaminated site studies requires a clear and coherent understanding of the available data. For those directly involved in statistical analysis and interpretation, a clear and coherent understanding of the data will help them to select appropriate statistical tools and to make critical assumptions about statistical populations. For those who prepare statistical reports, it is important that their reports convey a clear and coherent understanding of the data to their audience; the readers of a report will not be able to form an opinion about the validity of the study's conclusions without a good understanding of the data on which it is based.

This guidance document discusses tools for exploratory data analysis, a statistical study's first step in which we investigate the available data, form tentative opinions and modify these opinions as our understanding of the data improves and evolves. The same tools that help us explore and interpret the available data are also ideal for presenting and summarizing our understanding of the data to those not directly involved in the study. This guidance document should therefore be of assistance not only to those who actually do the statistical analysis and interpretation, but also to those who are responsible for writing reports. This document is not intended to provide a rigid prescription for how to perform and present an exploratory data analysis; indeed, as noted in the final section of this document, such a rigid prescription would not permit us to exercise the curiosity that is one of the cornerstones of thorough exploratory data analysis. This document does intend, however, to encourage some much needed consistency in the performance and presentation of statistical studies by providing a simple and straightforward approach to exploratory data analysis.

This guidance document focuses on the exploratory data analysis of a single variable, such as the concentration of a single contaminant. Two other documents in this series focus on other aspects of exploratory data analysis. *BIVARIATE DESCRIPTION* focuses on tools for analyzing the relationship between pairs of variables; *SPATIAL DESCRIPTION* focuses on tools for analyzing the data in their spatial context.

PROVIDING DETAIL & CONVEYING INFORMATION

With all statistical presentations there is a tradeoff between the level of detail in the presentation and the amount of information that it conveys. Table 1 and Figure 1 demonstrate this tradeoff using data from a site contaminated with mercury. Table 1 provides the most detailed and complete information about the available data values and yet it does not immediately convey much information. By sacrificing some of the detail, the histogram shown in Figure 1 more immediately conveys useful information about the available data by giving us a quick appreciation of the fact that there are many low values around 1 ug/g and only a few erratic high ones above 10 ug/g. Though this fact can also be extracted from Table 1, the histogram makes it more readily apparent.

Table 1 Hg measurements (in ug/g) from a contaminated site.

1.08	1.10	7.27	0.30	5.01	1.97	1.58	0.67	0.06	0.22
0.28	3.22	0.46	1.13	0.78	7.44	0.08	0.45	14.91	0.26
0.32	11.47	8.93	0.93	1.81	9.40	0.52	12.89	5.40	1.52
1.25	15.87	2.31	0.70	11.31	4.64	0.69	0.06	0.45	3.59
0.54	0.76	4.80	7.92	8.46	0.90	0.71	8.94	2.01	9.83

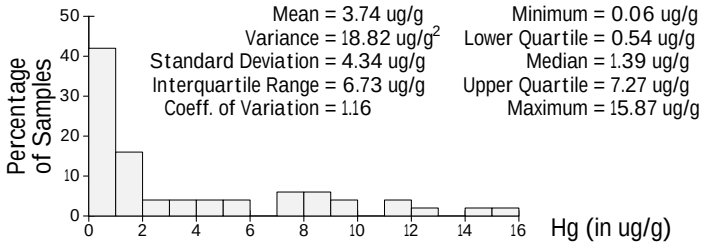

Figure 1 A histogram of the mercury data in Table 1.

As we compile a statistical description, we should keep this tradeoff in mind and should select graphical and numerical summaries that convey useful and salient information about the data. The various components of our statistical description will often be somewhat redundant; some of the statistics in Figure 1, for example, convey similar information as others. Such redundancy is not a flaw, however, as long as each component successfully conveys useful additional information. The guiding principle should be clarity of understanding — a good statistical presentation is one that enables others who are unfamiliar with the site to share our understanding of the data.

SUMMARY STATISTICS

Measures of center

The statistic most commonly used to summarize where the center of a distribution lies is the mean, which is simply the arithmetic average of the data values:

$$\text{Mean} = m = \frac{1}{n}\sum_{i=1}^{n} v_i$$

Though the mean is the traditional measure of the center of a distribution, it is strongly influenced by erratic high values

and may not correspond to our intuitive sense of the middle of the distribution. For the mercury data shown in Figure 1, for example, more than two-thirds of the values are smaller than the mean value of 3.74 ug/g, so it is not clear why this qualifies as a "central" value. For contaminated site data, which often span several orders of magnitude, it is common to find that the vast majority of the data values fall below the mean. Had the largest value in Table 1 been an order of magnitude higher, at 158.7 ug/g rather than 15.87 ug/g, the mean would nearly double to 6.60 ug/g higher than 75% of the data.

For data that span several orders of magnitude, the median is less sensitive to extreme values and provides a stable statistic that corresponds more closely to our intuitive sense of the center of the distribution. The median is the number that appears halfway down the list of values when they are sorted from smallest to largest; for an even number of data, the median is the average of the middle two values. Since it depends only on the ordering of the data, the median would not be changed if the largest value was an order of magnitude higher.

Taken together, the mean and the median provide an indication of the influence of extreme values in a data set. If the two measures of the center are close to each other, then extreme values do not play much of a role. This is not typically the case with contaminated site data. It is not unusual to find that there are some influential extreme values that cause the mean of the data to be more than twice the median.

Measures of location

The statistics that are used to describe the location of other parts of the distribution can all be calculated easily from a sorted list of the data values. The minimum is the first value on the sorted list and the maximum is the last value on the sorted list. The "quartiles" provide two other useful measures of location. In the same way that the median splits the data set into halves, the quartiles split it into quarters. 25% of the data values are below the lower (or first) quartile and 25% of them are above the upper (or third) quartile.

In most contaminated site studies, the lowest values are below the detection limit. Rather than reporting these values as exactly half the detection limit, as is often done, it is more useful in a statistical summary to state the detection limit and to report how many values fall below it.

Measures of spread

In addition to describing where the center of the distribution lies, a complete statistical description should also report how the data values are spread around the center — are they all tightly grouped close to the center or are they scattered far away from the center? The statistics that are commonly used to describe the spread of the distribution are the variance, s^2, and the standard deviation, s. The sample variance is the average squared difference of the data values from their mean:

$$\text{Variance} = s^2 = \frac{1}{n}\sum_{i=1}^{n}(v_i - m)^2$$

The standard deviation is the square root of the variance.

Like the mean, and all other statistics that involve an averaging of the data, the variance and standard deviation are both sensitive to extreme values. Had the largest value in Table 1 been an order of magnitude higher, at 158.7 ug/g rather than 15.87 ug/g, the variance would soar from less than 20 to nearly 500 ug/g^2! With a single extreme value having such a profound influence, the variance and standard deviation are often difficult to interpret. For most contaminated site data, the interquartile range (IQR) is a more stable and interpretable alternative. The IQR is the difference between the upper quartile and lower quartile and provides a direct measurement of the spread of the middle 50% of the data values. Since it depends only on the quartiles, the IQR is insensitive to the exact values of the most extreme data.

Several of the guidance documents in this series warn that certain procedures should not be used if the data are not from a homogeneous population. Though it is difficult to give exact specifications for "homogeneity", an appropriate starting point is the measure of spread called the coefficient of variation, which is the ratio of the standard deviation to the mean: CV = s ÷ m. This measure of relative variation is often expressed in percent, rather than as a ratio. For data whose CV is 1 (or 100%), their standard deviation is as big as their mean. If the data are to be considered "homogeneous", their CV should be smaller than 1. By itself, this is not a guarantee that the data do come from a common population; qualitative information, such as the site history and the provenance of the data, also needs to be taken into account. If the CV is larger than 1, however, it is unlikely that the samples are from a single population; the large spread in the data values is likely a warning that different samples have been affected by different physical and chemical processes.

Measures of shape

The final summary statistic that is often included in a statistical description is a measure of the shape or symmetry of the distribution. The symmetry of a distribution can be described by comparing the mean to the median, and can also be captured in a statistic called the skewness. Though the skewness does have a specific formula, we rarely need to know its precise value and usually report only its sign, either positive or negative.

Positively skewed distributions have a lot of low values and a decreasing proportion of high values; the histogram of positively skewed data is asymmetric with a tail to the right, like the one shown in Figure 1. Negatively skewed distributions, which are rare in contaminated site studies, have a lot of high values and a decreasing proportion of low values; their histogram has a tail to the left. Occasionally we encounter contaminated site data whose skewness is very minor and whose histogram appears symmetric. The guidance document *CHOOSING A DISTRIBUTION* offers a rule of thumb that can be used to decide if the distribution can be deemed symmetric.

GRAPHICAL TOOLS

By themselves, summary statistics do not always convey all of the important information about a data set. Graphical presentations of the data provide valuable visual support to readers

who are trying to follow the details of a statistical study. A combination of graphical displays and numerical summaries is the most effective vehicle for conveying our understanding of the data to readers who are not familiar with the project.

Histograms

The most common graphical presentation of data is a histogram that shows how many samples fall in different categories. Using the mercury data presented on Table 1 on the previous page, Table 2 records how many samples fall within each of sixteen classes, from 0–1 ug/g to 15–16 ug/g. Figure 1 presents this information as a histogram on which the height of a bar is equal to the percentage of samples in that class.

Table 2 Frequency table of mercury data in Table 1.

Class (in ug/g)	No. of samples	% of total	Class (in ug/g)	No. of samples	% of total
0–1	21	42	8–9	3	6
1–2	8	16	9–10	2	4
2–3	2	4	10–11	0	0
3–4	2	4	11–12	2	4
4–5	2	4	12–13	1	2
5–6	2	4	13–14	0	0
6–7	0	0	14–15	1	2
7–8	3	6	15–16	1	2

It is often awkward to select a class width for histograms of contaminated site data since the data span several orders of magnitude. If a large class width is used in an attempt to display the entire range of the data values, then the first class on the histogram often gets the lion's share of the samples and the display does not provide much detail on the distribution of the lower values. If a small class width is used in an attempt to show more of the detail of the distribution of the lower values, then the number of classes needed to span the entire range becomes unmanageable. One solution to this common problem is to show two histograms, one that spans the entire range with wide classes and another that shows the details of the low end of the distribution with smaller classes.

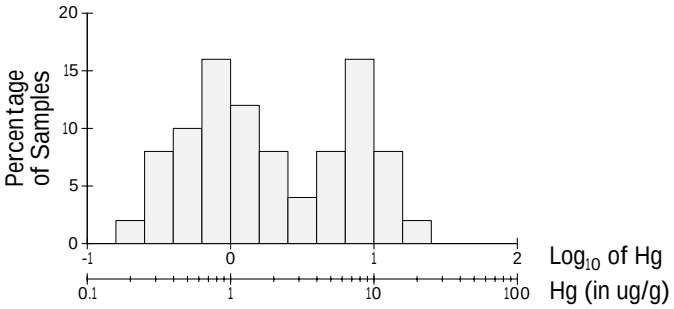

Figure 2 A logarithmic histogram of the data in Table 1.

Another way of dealing with data that span several orders of magnitude is to choose classes that have equal widths on a logarithmic scale. Figure 2 shows the mercury data from Table 1 plotted as a histogram on a logarithmic scale. One of the advantages of a logarithmic histogram is that it often makes different populations more apparent. On Figure 2, for example, the low background mercury values form one clear bump or "mode" around 1 ug/g, while the high contaminated values show another mode around 10 ug/g.

Cumulative plots and probablity plots

Using the mercury data presented on Table 1 on the previous page, Table 3 records the number and percentage of samples that fall below the sixteen thresholds from 1 to 16 ug/g; in the cumulative plot shown in Figure 3, the threshold values in the first column of Table 3 are used as the x-coordinates and the cumulative percentages in the last column are used as the y-coordinates. The cumulative plot in Figure 3 has a logarithmic x-axis to accommodate the skewness in the data; if the distribution had been more symmetric, a linear x-axis would have been more appropriate.

Table 3 Cumulative frequency table of mercury data in Table 1.

Threshold (in ug/g)	No. of samples below	% of total	Threshold (in ug/g)	No. of samples below	% of total
1	21	42	9	43	86
2	29	58	10	45	90
3	31	62	11	45	90
4	33	66	12	47	94
5	35	70	13	48	96
6	37	74	14	48	96
7	37	74	15	49	98
8	40	80	16	50	100

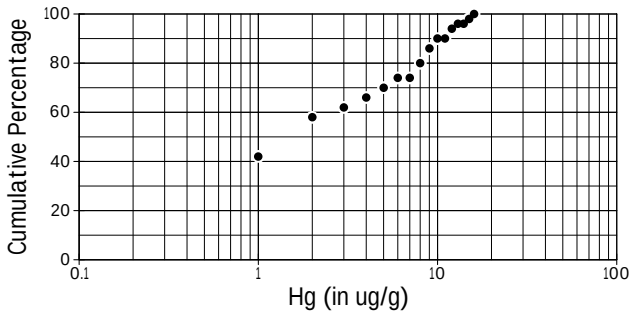

Figure 3 A cumulative plot of the mercury data in Table 1.

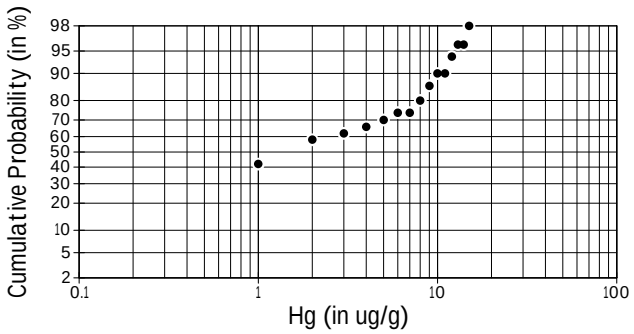

Figure 4 A probability plot of the mercury data in Table 1.

When cumulative plots are presented on special graph paper called "probability paper" they are usually called "probability plots". Figure 4 presents the data from Table 3 as a probability plot. The probability axis on this kind of plot is squashed in the middle and stretched at the top and bottom. The reason for plotting cumulative curves on this distorted grid is that it simplifies the checking of whether the distribution of the data values is close to that of a mathematical model called the "normal" or "gaussian" distribution. The histogram of normally distributed data will be shaped like a bell. Rather than checking how bell-like the histogram is, it is easier to check how straight

the probability plot is. The distorted grid of probability paper is designed in such a way that the cumulative curve of normally distributed data will plot as a straight line.

Boxplots

Figure 5 shows another graphical tool that is becoming popular in exploratory data analysis. The box goes from the lower quartile to the upper quartile and therefore spans the middle 50% of the data values. The bar in the middle of the box shows the median and the dot shows the mean. The arms that stick out of the box go to the minimum and maximum. As with other graphical presentations, logarithmic scaling often makes the boxplot more informative if the distribution of data values is skewed. The simple boxplot captures most of the critical information about a distribution — its center, its spread and its skewness — in a format that is more compact than a histogram.

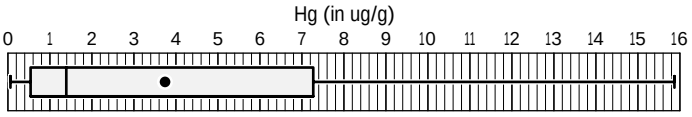

Figure 5 A boxplot of the mercury data in Table 1.

DATA BASE COMPILATION AND VERIFICATION

An exploratory data analysis is only as good as the data on which it is based; when there are errors in the data base, our exploratory data analysis is often useless and misleading. With the data from contaminated site studies often having to be transcribed, keypunched or electronically merged from some other source, there are ample opportunities for human error. Before attempting an exploratory data analysis, we need to know how the data base was created. If the data base is not accompanied by a clear audit trail that explains all of the steps involved in its creation, then it should be verified against original records wherever possible.

One of the best ways to verify a data base is by using teams of two people to proofread the data. One person reads out loud the data values from a hardcopy of the data base while the other checks each value against *original* records, such as laboratory reports or surveyor's notes. Though two person proofing is a tedious exercise, it is a very important one if the integrity of the data base is uncertain. Experience has shown that it provides a much more complete and exhaustive verification of a data base than automated or computer-based techniques, which can do no more than check one electronic version of the data base against another. If the original data were not recorded electronically, but were recorded manually and later transcribed, then verifying one electronic version against another cannot catch mistakes that crept into the data before or during the creation of the first electronic version.

Once the integrity of the data base is well documented, every effort should be made to maintain this integrity. In many contaminated site studies, where there are several phases of data collection, the integrity of the early data base is lost as various people merge new data and modify old data to suit their individual purposes. Concentrations may need to be converted from one unit of measurement to another, or the coordinate system used by one group may need to be transformed to the coordinate system used by another group. With every such modification of the data there are opportunities for error. Such opportunities increase with every new person who has access to the data base and is able to make modifications. Once the integrity of a data base is lost, restoring it will either require considerable effort or will be completely impossible.

On large projects, where there are more than 100 samples or where several groups have been collecting data, one person should have the responsibility for maintaining the authoritative and verified data base. Others may obtain copies for their own work but none of their individual changes should be accepted in the single authoritative version until the data base coordinator approves the change and prepares documentation that explains exactly what changes were made and why.

RECOMMENDED PRACTICE

It is not possible to give a rigid prescription for exploratory data analysis since a thorough understanding of the data requires both creativity and curiosity. The sequence of steps that worked on one project will not always work on another one. The following general guidelines, however, should improve any exploratory data analysis for contaminated site studies:

1. Before exploratory data analysis, the integrity of the data should be documented, either by reference to a report on procedures used to compile and verify the data or by a complete check of all data against original records.
2. Complete listings of all data used in statistical studies should be included as appendices to reports; these do not, however, constitute an appropriate statistical summary. Statistical summaries of univariate data should include:
   (a) Graphical presentations of the data, such as histograms, probability plots or boxplots. If the data are skewed, then logarithmic scaling will often make such graphical presentations more informative.
   (b) Summary statistics that describe the center, location, spread and shape of the distribution of data values. If the data are skewed, then the mean and standard deviation should not be used alone to summarize the data but should be accompanied by other measures, such as the median and interquartile range, that are not so sensitive to extreme values.

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.

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

Isaaks, E.H. and Srivastava, R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

Moore, D.S., *Statistics: Concepts and Controversies*, W.H. Freeman and Company, New York, 1985.