CONTAMINATED SITES STATISTICAL APPLICATIONS GUIDANCE DOCUMENT NO. 12-5

NONPARAMETRIC METHODS

A guide for data analysts and interpreters on statistical methods that do not require a distribution model

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 validity of statistical statements can easily be challenged by questioning any distribution assumption. For example, we might be tempted to take data from a contaminated site, calculate that their mean is 50 ug/g and their standard deviation is 10 ug/g, and then use this information to predict that there is a less than a 1% chance that samples from the same population will exceed a threshold of 80 ug/g. This statement is defensible only if we can also defend the implicit assumption that the data values follow the classical bell-shaped normal distribution. The type of contaminant concentration data that we typically collect from contaminated sites very rarely follow a normal distribution, however, and any predictions that follow from this initial assumption are difficult to defend.

Though we try to make sure that our assumptions about underlying distributions are appropriate — choosing skewed distributions, for example, to model contaminant concentrations — we always run the risk that regardless of the distribution we choose, someone is going to challenge our predictions based on the fact that we *assumed* a particular distribution that we can never prove is correct. Fortunately, for many of the statistical problems that arise in contaminated site studies, there are methods that allow us to solve the problem without making any assumption about the underlying distribution. The predictions that we get from such *nonparametric* procedures will be defensible regardless of the assumption that anyone wants to make about the underlying distribution.

This guidance document presents some of the more common and practically useful nonparametric methods. In addition to demonstrating how they can be used in practice, this document also discusses the advantages and disadvantages of these nonparametric methods. There are two other documents in this series, *DISTRIBUTION MODELS* and *CHOOSING A DISTRIBUTION*, that discuss related issues.

ADVANTAGES OF NONPARAMETRIC METHODS

Inappropriateness of the normal distribution

The main advantage of nonparametric methods is that they do not require us to assume that data are normally distributed. Even though an assumption of normality underlies the vast majority of statistical procedures that are in common use, it is a very questionable assumption in contaminated site studies. Figure 1 shows a typical example of a histogram of sample values from a contaminated site along with some of the common summary statistics. These data have a mean that is much larger than their median; they show a large proportion of low values and a decreasing proportion of high ones. A normal distribution would show none of these characteristics; its mean and median would be very similar and its histogram would look symmetric, with similar proportions of low and high values.

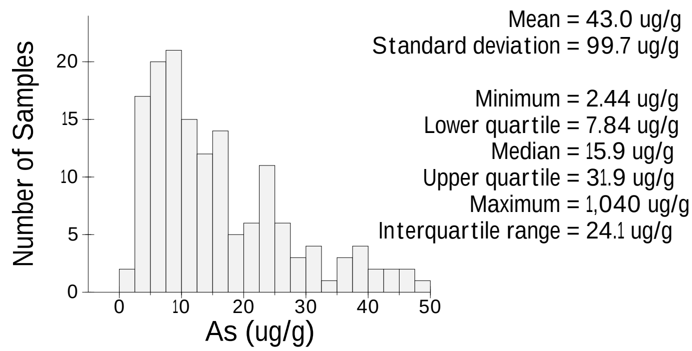


Figure 1 A histogram for measurements of the arsenic concentration in the soil from a contaminated landfill site.

Typical of the kinds of statistical predictions that depend on a prior assumption of normality is the use of the mean and standard deviation to build confidence intervals. Wherever we see m$\pm\sigma$ being used as a 68% confidence interval, or m$\pm 2\sigma$ as a 95% confidence interval, we are seeing a result that depends on an assumption of normality. If the unknown values that we are trying to predict do follow a normal distribution, then 68% of the values will fall within one standard deviation of the mean and 95% of them will fall within two standard deviations. If, however, the values do not follow a normal distribution (and this is more commonly the case in practice), then the traditional confidence intervals are meaningless.

In a nonparametric approach we make no assumption about the underlying distribution. This makes our predictions more robust in the sense that they do not depend on whether or not the underlying distribution is normal.

No need for any distribution model

Nonparametric methods are particularly useful in the early stages of a contaminated site study, where there are typically very few data yet available and, even if we intend ultimately to use a parametric technique that assumes some distribution model, we do not yet have enough data to allow us to choose an appropriate distribution model. Table 1 shows an example of a few measurements of the PCB concentration in the first ten samples collected from a contaminated site. Suppose that at this very early stage in the study we wanted to make some statement about whether the median for the entire population could be 10 ug/g. Any parametric technique would require us to first make an assumption about the underlying distribution from which these ten values come. With so few data at our disposal, it is very difficult to decide what kind of distribution might be an appropriate model for the PCB values. As discussed in greater detail later in this guidance document, this question about the median can be answered with a nonparametric technique that does not require us to assume anything about the underlying distribution.

Table 1 PCB values (in ug/g).

<1	51.2	17.9	34.6	<1	22.4	11.5	48.2	7.8	31.4

Ease of calculation and interpretation

The common nonparametric methods are very simple to apply. They usually work with the ranks of the data or with simple counts of values above and below the median and are therefore easy to calculate manually or to implement on a computer.

Another advantage of nonparametric statistics is that they are often easier for non-statisticians to understand and interpret. As discussed later, some of the graphical displays that are based on nonparametric statistics, such as the percentiles of the distribution, are more straightforward than other more traditional displays and still convey as much useful information.

Ability to work with no-detects

One of the other advantages of many nonparametric techniques is that they can accommodate values below detection limit without assigning such samples some arbitrary value (such as half the detection limit). As we will see later, we can make statistical tests with data such as those shown in Table 1 even though we do not know the exact value of every sample.

DISADVANTAGES OF NONPARAMETRIC METHODS

Not as efficient

The principal limitation of nonparametric methods is that they are not as efficient or powerful as parametric methods that are based on a known underlying distribution. For example, if we are trying to use statistics to document that two groups of data should be treated as separate populations, and if we already know that it is reasonable to assume that the data values in both groups are normally distributed, then a parametric test, such as the t-test, will be able to discriminate more effectively between the means of the two groups than would the corresponding nonparametric test described below.

Unable to extrapolate beyond data

The assumption of a specific distribution model is very powerful and buys us a lot of predictive power. Once we claim to know the distribution that the data represent, and we have chosen the parameters of our assumed distribution (such as the mean and the standard deviation for the normal distribution), we are then able to leverage our assumption and predict the behaviour of the entire distribution. For example, a parametric approach gives us the ability to predict the 99th percentile even if we haven't actually got a sample value that high yet. With its fundamental philosophy of avoiding unnecessary distribution models and letting the data speak for themselves, a nonparametric approach has no additional information to leverage beyond the data themselves; if we have only ten sample values, it will not be possible to predict the 99th percentile with a nonparametric approach.

Need more data

By letting data speak for themselves rather than letting a distribution model do the speaking for them, nonparametric methods provide statistical predictions that are not compromised by unnecessary distribution assumptions. The price for this strict adherence to data, however, is that nonparametric methods cannot make strong statistical statements with few data.

NONPARAMETRIC DATA ANALYSIS

Percentile-based statistics

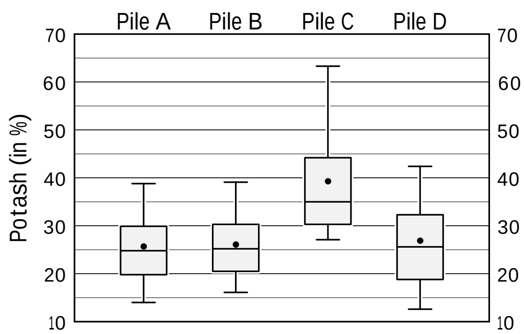
The two statistics that are most commonly used to describe a distribution are the mean and standard deviation. The first of these is a measure of the center of the distribution, the second is a measure of the spread of the distribution. The popularity of these two particular statistics is due, in large part, to the fact that they are the common parameters for the normal distribution. Though they are commonly used, these two statistics are often of little value for exploratory data analysis since they are both strongly influenced by extreme values. With the arsenic data shown in Figure 1, for example, it is questionable whether the mean of 43.0 ug/g is really describing the center of the distribution, or whether the standard deviation of 99.7 ug/g is telling us anything useful about the spread of the values. In this particular example, as in many other actual data sets from contaminated site studies, a few extremely high values have a profound influence on these two statistics.

Nonparametric methods rely on "rank" or "order" statistics that are simply the percentiles of a distribution. Rather than use the mean to describe the center of the distribution, nonparametric approaches more commonly use the median or 50th percentile. The difference between the upper quartile (75th percentile) and the lower quartile (25th percentile) is called the "interquartile range" and is the nonparametric alternative to the standard deviation for describing the spread.

For most people, the median corresponds more closely to their visual sense of where the center of the histogram lies than does the mean. Similarly, their visual sense for the spread of the distribution is closer to the interquartile range than to the standard deviation. Most of us, statisticians and non-statisticians alike, have a stronger intuitive feel for what the interquartile range is measuring — the spread of the middle half of the data — than we have for whatever it is that the standard deviation is measuring — the square root of the average squared deviations from the mean?! For the purposes of communicating statistical information to a non-technical audience, nonparametric statistics are therefore an excellent supplement to the more conventional mean and standard deviation.

Boxplots

A boxplot provides a concise graphical format for displaying the key nonparametric statistics. Figure 2 shows an example of a set of boxplots for the K_2O (potash) concentrations from discrete samples taken from four different stockpiles of cement kiln dust. The box in the middle of a boxplot extends from the lower quartile to the upper quartile; the bar in the middle of the box shows where the median lies. There are a couple of different conventions for how to draw the arms that stick out of the box; the one used in Figure 2 shows them extending all the way to the minimum on the low side and to the maximum on the high side. The other common convention is to draw the arms only part way to the extremes and to plot a star at each of the very extreme values. Boxplots commonly also pay homage to the fact that the mean is by far the most common summary statistic of all and, even though it is not a percentile-based statistic, it is usually shown with some special symbol — a black dot in the examples shown in Figure 2.



	Pile A	Pile B	Pile C	Pile D	
Number of data	39	31	63	85	Number of data
Mean	25.7	26.1	39.3	26.9	Mean
Maximum	38.8	39.1	63.3	42.4	Maximum
Upper quartile	29.9	30.3	44.2	32.3	Upper quartile
Median	24.8	25.2	35.0	25.6	Median
Lower quartile	19.8	20.5	30.3	18.8	Lower quartile
Minimum	14.0	16.1	27.1	12.6	Minimum

Figure 2 Side-by-side boxplots.

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 or not it is symmetric. The boxplot therefore offers most of the useful information that a histogram contains, but in a more compact form that is more amenable to side-by-side comparisons between different groups of data.

NONPARAMETRIC TESTS

Chebyshev's inequality for confidence intervals

Earlier, we pointed out that the use of $m \pm \sigma$ for calculating 68% confidence intervals is fine for the normal distribution but does not work for other distributions. There is a century-old non-parametric result known as "Chebyshev's inequality" that allows us to build confidence intervals using the mean and standard deviation even if we don't know the underlying distribution. Chebyshev's inequality says that for any constant k, the proportion of data that are within k standard deviations from the mean cannot be less than $1 - (1 \div k)^2$. If we take k=2, for example, this inequality tells us that at least 75% of the distribution must be within two standard deviations of the mean; for k=10, at least 99% of the distribution must be within ten standard deviations of the mean.

Compared to confidence intervals predicted from any distribution model, those predicted using Chebyshev's inequality are broader. For example, the opening example on the first page of this document involved a distribution with a mean of 50 ug/g and a standard deviation of 10 ug/g; with these statistical parameters, an assumption of normality leads to the conclusion that less than 1% of the data should exceed 80 ug/g. For this same threshold, which happens to be three standard deviations above the mean, Chebyshev's inequality states that any possible distribution must have at least 89% of the data within three standard deviations of the mean; no more than 11% could possibly be more than three standard deviations from the mean. This gives us a pessimistic upper bound on how much of the distribution *might* exceed 80 ug/g if our assumption of normality is inappropriate: for any distribution whatsoever, it is not possible to get more than 11% of the values to be greater than three standard deviations above the mean.

The sign test for the median

Earlier in Table 1 we showed ten PCB values and asked if the median could be as low as 10 ug/g. The "sign test" is a non-parametric procedure in which all data values above the proposed median are given + signs and all others are given − signs. We can test whether the median could be as low as some specified threshold, T, by noting that if T is, indeed, the median, then regardless of the shape of the distribution, each data value has the same probability of getting a + sign as a − sign:

$$p_+ = p_- = \frac{1}{2}$$

In a sample of size N, the number of observations with a + sign, N_+, will follow a binomial distribution. The probability of getting more than n + signs is:

$$\text{Prob}[N_+ \geq n] = \left[\frac{1}{2}\right]^N \times \sum_{i=n}^{N} \frac{N!}{(N\text{-}i)! \times i!}$$

These binomial probabilities are tabulated in most reference and textbooks on probability and statistics. For large values of N, most introductory probability books, such as Blake (1979), discuss good approximations to these binomial probabilities.

Using the data from Table 1 and a proposed median of 10 ppm, seven of the values would get + signs. The no-detect samples do not create any difficulty; even though we do not know exactly the PCB concentration of these samples, we can still assign them − signs since they are definitely below 10 ppm. The probability of getting seven or more + signs out of a total of ten tries is:

$$\left[\frac{1}{2}\right]^{10} \times \left[\frac{10!}{3! \times 7!} + \frac{10!}{2! \times 8!} + \frac{10!}{1! \times 9!} + \frac{10!}{0! \times 10!}\right] = 0.172$$

Regardless of the underlying distribution, the chance that its median is 10 ug/g or lower given the ten observed values shown in Table 1 is about 17%.

The sign test can be adapted to test for any percentile by changing the equation given above to accommodate the fact that p_+ and p_- are no longer the same:

$$\text{Prob}[N_+ \geq n] = \sum_{i=n}^{N} \frac{N!}{(N\text{-}i)! \times i!} \times {p_+}^{i} \times {p_-}^{(N\text{-}i)}$$

The Wilcoxon rank-sum test

Nonparametric methods for testing the difference between two groups of data usually deal with the ranks of the data. In a group of N data, the ranks are simply numbers from 1 to N that order the data from smallest to largest: the smallest data value has a rank of 1, the second smallest has a rank of 2 and so on up to the largest data value, which has a rank of N.

With two groups of data, the first containing N_1 samples and the second containing N_2 samples, the Wilcoxon rank-sum statistic, W, is created as follows:

1. Combine both groups of data, creating a large group with N samples.
2. Assign ranks to the data.
3. Let W be the sum of the ranks of all the data that came from the first group.

To test whether the two groups are significantly different, the Wilcoxon rank-sum test compares W against tabulated values of critical values. These tables are given for various values of N_1 and N_2. They show the range of values that W can have if the two groups of data actually come from the same population. If the observed value of W falls outside the range given in such tables, we accept this as evidence that the differences between the data values in the two groups are too large to be explained by chance alone; a more plausible explanation than mere chance is that the data values in each group were drawn from different populations.

If the values of N_1 and N_2 are larger than those that appear in reference tables, there is an another way to check if W is too extreme. We calculate the following test statistic

$$z = \frac{W - \frac{N_1 \cdot (N_1 + N_2 + 1)}{2}}{\sqrt{\frac{N_1 \cdot N_2 \cdot (N_1 + N_2 + 1)}{12}}}$$

and check to see if $|z|$ is greater than 3. If it is, then the chance that the differences between the two groups are due to chance alone is less than 1%, so values of z outside the range -3 to +3 are accepted as evidence that there are significant statistical differences between the two groups.

As an example of the application of the Wilcoxon rank-sum test, consider the problem of checking whether the following four PCB values might belong in the same group as the ten shown earlier in Table 1: <1, 5.2, 9.2 and 1.9 ug/g. These four values seem to be low compared to those seen earlier, but could this just be chance?

When the four new values are combined with the other ten to make a group of 14 samples, the three lowest values are all no-detects. Since we can't sort out the order of these three and don't know which should get the rank of 1, which should get the rank of 2 and which should get the rank of 3, we assign the average rank of 2 to each of these three tied values. The four new values therefore get ranks of 2, 4, 5 and 7; the sum of these ranks is 18. Tabulated values of the Wilcoxon rank-sum statistic (Finkelstein and Levin, p. 563 – 564) show that with a group of 4 samples being compared to a group of 10 samples, there is a 90% chance that W will be between 16 and 44. So although the new values tend to be on the low side, we cannot reject the possibility that they could actually be from the same population as the original ten values shown earlier.

RECOMMENDED PRACTICE

1. When presenting a statistical summary of data collected from a contaminated site, use percentile-based statistics, such as the quartiles and the median to supplement the more traditional mean and standard deviation.
2. Use boxplots as an alternative to histograms for graphical display purposes, especially when documenting a comparison between two or more groups of data.
3. Wherever a statistical prediction calls for a prior assumption about the underlying distribution, use a nonparametric alternative as a way of checking the sensitivity of the conclusion to the distribution assumption.

REFERENCES AND FURTHER READING

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

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

Conover, W., *Nonparametric Statistics*, John Wiley & Sons, 1980.

Finkelstein, M.O., and Levin, B., *Statistics for Lawyers*, Springer-Verlag, 1990.

Gibbons, J.D., *Nonparametric Methods for Quantitative Analysis*, 2nd ed., American Sciences Press, 1985.

Gibbons, R.D., "General statistical procedure for ground water detection monitoring at waste disposal facilities," *Ground Water*, v. 28, p. 235 – 248, 1990.

Kendall, M.G., *Rank Correlation Methods*, 4th ed., Griffin, 1970.

Lehmann, E.L., *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, 1975.

Millard, S.P. and Deverel, S.J., "Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits," *Water Resources Research*, v. 24, n. 12, p. 2087 – 2098, 1988.

Mosteller, F., and Rourke, R.E.K., *Sturdy Statistics: Nonparametric and Order Statistics*, Addison-Wesley, 1973.

United States Environmental Protection Agency, "40 CFR Part 264: Statistical methods for evaluating ground-water monitoring from hazardous waste facilities; final rule," *Federal Register*, v. 53, n. 196, p. 39720 – 39731, U.S. Government Printing Office, 1988.