# PART II: DIAGNOSTIC METHODS AND TESTING

**Part II** covers diagnostic evaluations of historical facility data for checking key assumptions implicit in the recommended statistical tests and for making appropriate adjustments to the data (e.g., consideration of outliers, seasonal autocorrelation, or non-detects). Also included is a discussion of groundwater sampling and how hydrologic factors such as flow and gradient can impact the sampling program.

**Chapter 9** provides a number of exploratory data tools and examples, which can generally be used in data evaluations. Approaches for fitting data sets to normal and other parametric distributions follows in **Chapter 10**. The importance of the normal distribution and its potential uses is also discussed. **Chapter 11** provides methods for assessing the equality of variance necessary for some formal testing. The subject of outliers and means of testing for them is covered in **Chapter 12**. **Chapter 13** addresses spatial variability, with particular emphasis on ANOVA means testing. In **Chapter 14**, a number of topics concerning temporal variation are provided. In addition to providing tests for identifying the presence of temporal variation, specific adjustments for certain types of temporal dependence are covered. The final **Chapter 15** of **Part II** discusses non-detect data and offers several methods for estimating missing data. In particular, methods are provided to deal with data containing multiple non-detection limits.

*This page intentionally left blank*

# CHAPTER 9. COMMON EXPLORATORY TOOLS

9.1 TIME SERIES PLOTS ..... 9-1
9.2 BOX PLOTS ..... 9-5
9.3 HISTOGRAMS ..... 9-8
9.4 SCATTER PLOTS ..... 9-13
9.5 PROBABILITY PLOTS ..... 9-16

Graphs are an important tool for exploring and understanding patterns in any data set. Plotting the data visually depicts the structure and helps unmask possible relationships between variables affecting the data set. Data plots which accompany quantitative statistical tests can better demonstrate the reasons for the results of a formal test. For example, a Shapiro-Wilk test may conclude that data are not normally distributed. A probability plot or histogram of the data can confirm this conclusion graphically to show why the data are not normally distributed (e.g., heavy skewness, bimodality, a single outlier, etc.).

Several common exploratory tools are presented in **Chapter 9**. These graphical techniques are discussed in statistical texts, but are presented here in detail for easy reference for the data analyst. An example data set is used to demonstrate how each of the following plots is created.

- Time series plots (**Section 9.1**)
- Box plots (**Section 9.2**)
- Histograms (**Section 9.3**)
- Scatter plots (**Section 9.4**)
- Probability plots (**Section 9.5**)

## 9.1 TIME SERIES PLOTS

Data collected over specific time intervals (e.g., monthly, biweekly, or hourly) have a temporal component. For example, air monitoring measurements of a pollutant may be collected once a minute or once a day. Water quality monitoring measurements may be collected weekly or monthly. Typically, groundwater sample data are collected quarterly from the same monitoring wells, either for detection monitoring testing or demonstrating compliance to a GWPS. An analyst examining temporal data may be interested in the trends over time, correlation among time periods, or cyclical patterns. Some graphical techniques specific to temporal data are the time plot, lag plot, correlogram, and variogram. The degree to which some of these techniques can be used will depend in part on the frequency and number of data collected over time.

A data sequence collected at regular time intervals is called a time series. More sophisticated time series data analyses are beyond the scope of this guidance. If needed, the interested user should consult with a statistician or appropriate statistical texts. The graphical representations presented in this section are recommended for any data set that includes a temporal component. Techniques described below will help identify temporal patterns that need to be accounted for in any analysis of the data. The analyst examining temporal environmental data may be interested in seasonal trends, directional trends, serial correlation, or stationarity. *Seasonal trends* are patterns in the data that repeat over time, i.e., the data

rise and fall regularly over one or more time periods. Seasonal trends may occur over long periods of time (large scale), such as a yearly cycle where the data show the same pattern of rising and falling from year to year, or the trends may be over a relatively short period of time (small scale), such as a daily cycle. Examples of seasonal trends are quarterly seasons (winter, spring, summer and fall), monthly seasons, or even hourly (e.g., air temperature rising and falling over the course of a day). *Directional trends* are increasing or decreasing patterns over time in monitored constituent data, which may be of importance in assessing the levels of contaminants. *Serial correlation* is a measure of the strength in the linear relationship of successive observations. If successive observations are related, statistical quantities calculated without accounting for the serial correlation may be biased. A time series is *stationary* if there is no systematic change in the mean (i.e., no trend) and variance across time. Stationary data look the same over all time periods except for random behavior. Directional trends or a change in the variability in the data imply non-stationarity.

A time series plot of concentration data versus time makes it easy to identify lack of randomness, changes in location, change in scale, small scale trends, or large-scale trends over time. Small-scale trends are displayed as fluctuations over smaller time periods. For example, ozone levels over the course of one day typically rise until the afternoon, then decrease, and this process is repeated every day. Larger scale trends such as seasonal fluctuations appear as regular rises and drops in the graph. Ozone levels tend to be higher in the summer than in the winter, so ozone data tend to show both a daily trend and a seasonal trend. A time plot can also show directional trends or changing variability over time.

A time plot is constructed by plotting the measurements on the vertical axis versus the actual time of observation or the order of observation on the horizontal axis. The points plotted may be connected by lines, but this may create an unfounded sense of continuity. It is important to use the actual date, time or number at which the observation was made. This can create discontinuities in the plot but are needed as the data that should have been collected now appear as “missing values” but do not disturb the integrity of the plot. Plotting the data at equally spaced intervals when in reality there were different time periods between observations is not advised.

For environmental data, it is also important to use a different symbol or color to distinguish non-detects from detected data. Non-detects are often reported by the analytical laboratory with a “U” or “<” analytical qualifier associated with the reporting limit [RL]. In statistical terminology, they are left-censored data, meaning the actual concentration of the chemical is known only to be below the RL. Non-detects contrast with detected data, where the laboratory reports the result as a known concentration that is statistically higher than the analytical limit of detection. For example, the laboratory may report a trichloroethene concentration in groundwater of “5 U” or “< 5” µg/L, meaning the actual trichloroethene concentration is unknown, but is bounded between zero and 5 µg/L. This result is different than a detected concentration of 5 µg/L which is unqualified by the laboratory or data validator. Non-detects are handled differently than detected data when calculating summary statistics. A statistician should be consulted on the proper use of non-detects in statistical analysis. For radionuclides negative and zero concentrations should be plotted as reported by the laboratory, showing the detection status.

The scaling of the vertical axis of a time plot is of some importance. A wider scale tends to emphasize large-scale trends, whereas a narrower scale tends to emphasize small-scale trends. A wide scale would emphasize the seasonal component of the data, whereas a smaller scale would tend to

emphasize the daily fluctuations. The scale needs to contain the full range of the data. Directions for constructing a time plot are contained in **Example 9-1** and **Figure 9-1**.

►EXAMPLE 9-1

Construct a time series plot using trichloroethene groundwater data in **Table 9-1** for each well. Examine the time series for seasonality, directional trends and stationarity.

Table 9-1. Trichloroethene (TCE) Groundwater Concentrations

| | **Well 1** | | **Well 2** | |
|---|---|---|---|---|
| **Date Collected** | **TCE (mg/L)** | **Data Qualifier** | **TCE (mg/L)** | **Data Qualifier** |
| 1/2/2005 | 0.005 | U | 0.10 | U |
| 4/7/2005 | 0.005 | U | 0.12 | |
| 7/13/2005 | 0.004 | J | 0.125 | |
| 10/24/2005 | 0.006 | | 0.107 | |
| 1/7/2006 | 0.004 | U | 0.099 | U |
| 3/30/2006 | 0.009 | | 0.11 | |
| 6/28/2006 | 0.017 | | 0.13 | |
| 10/2/2006 | 0.045 | | 0.109 | |
| 10/17/2006 | 0.05 | | NA | |
| 1/15/2007 | 0.07 | | 0.10 | U |
| 4/10/2007 | 0.12 | | 0.115 | |
| 7/9/2007 | 0.10 | | 0.14 | |
| 10/5/2007 | NA | | 0.17 | |
| 10/29/2007 | 0.20 | | NA | |
| 12/30/2007 | 0.25 | | 0.11 | |

NA = Not available (missing data).
U denotes a non-detect.
J denotes an estimated detected concentration.

SOLUTION

Step 1. Import the data into data analysis software capable of producing graphics.

Step 2. Sort the data by date collected.

Step 3. Determine the range of the data by calculating the minimum and maximum concentrations for each well, shown in the table below:

| | Well 1 | | Well 2 | |
|---|---|---|---|---|
| | **TCE (mg/L)** | **Data Qualifier** | **TCE (mg/L)** | **Data Qualifier** |
| Min | 0.004 | U | 0.099 | U |
| Max | 0.25 | | 0.17 | |

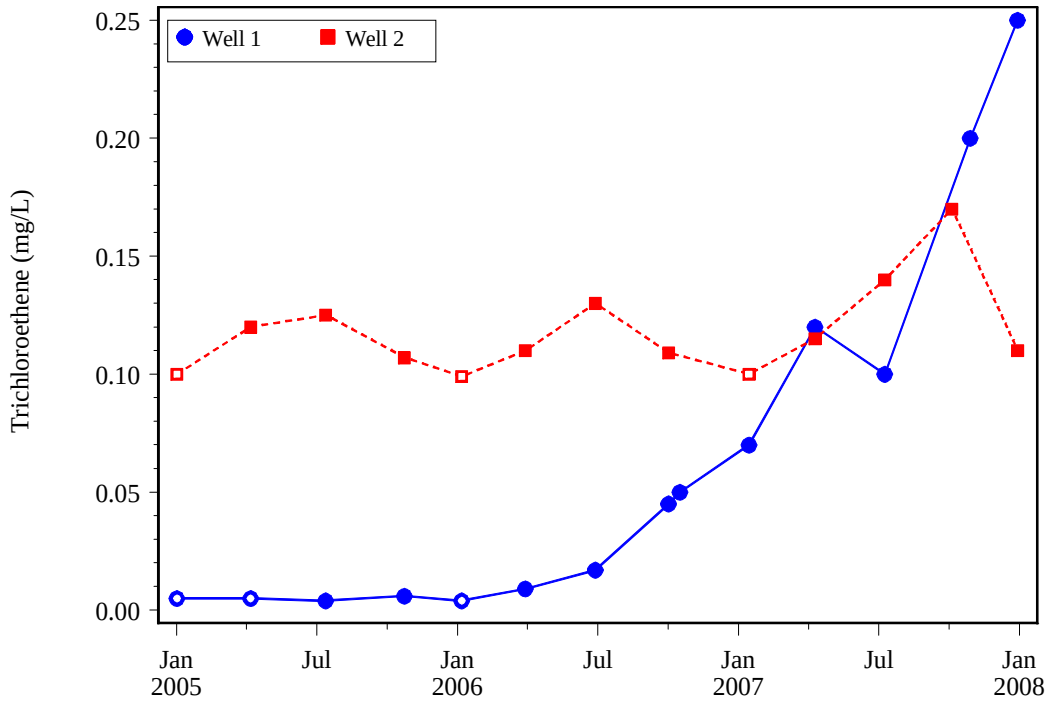
Step 4. Plot the data using a scale from 0 to 0.25 if data from both wells are plotted together on the same time series plot. Use separate symbols for non-detects and detected concentrations. One suggestion is to use "open" symbols (whose centers are white) for non-detects and "closed" symbols for detects.

Step 5. Examine each series for directional trends, seasonality and stationarity. Note that Well 1 demonstrates a positive directional trend across time, while Well 2 shows seasonality within each year. Neither well exhibits stationarity.

Step 6. Examine each series for missing values. Inquire from the project laboratory why data are missing or collected at unequal time intervals. A response from the laboratory for this data set noted that on 10/5/2007 the sample was accidentally broken in the laboratory from Well 1, so Well 1 was resampled on 10/29/2007. Well 1 was resampled on 10/17/2006 to confirm the historically high concentration collected on 10/2/2006. Well 2 was not sampled on 10/17/2006 because the data collected on 10/2/2006 from Well 2 did not merit a resample, as did Well 1.

Step 7. Examine each series for elevated detection limits. Inquire why the detection limits for Well 2 are much larger than detection limits for Well 1. A reason may be that different laboratories analyzed the samples from the two wells. The laboratory analyzing samples from Well 1 used lower detection limits than did the laboratory analyzing samples from Well 2.◀

Figure 9-1. Time Series Plot of Trichloroethene Groundwater for Wells 1 and 2 from 2005-2007.




Open symbols denote non-detects. Closed symbols denote detected concentrations.

## 9.2 BOX PLOTS

Box plots (also known as Box and Whisker plots) are useful in situations where a picture of the distribution is desired, but it is not necessary or feasible to portray all the details of the data. A box plot displays several percentiles of the data set. It is a simple plot, yet provides insight into the location, shape, and spread of the data and underlying distribution. A simple box plot contains only the $0^{th}$ (minimum data value), $25^{th}$, $50^{th}$, $75^{th}$ and $100^{th}$ (maximum data value) percentiles. A box-plot divides the data into 4 sections, each containing 25% of the data. Whiskers are the lines drawn to the minimum and maximum data values from the $25^{th}$ and $75^{th}$ percentiles. The box shows the interquartile range (IQR) which is defined as the difference between the $75^{th}$ and the $25^{th}$ percentiles. The length of the central box indicates the spread of the data (the central 50%), while the length of the whiskers shows the breadth of the tails of the distribution. The $50^{th}$ percentile (median) is the line within the box. In addition, the mean and the 95% confidence limits around the mean are shown. Potential outliers are categorized into two groups:

- data points between 1.5 and 3 times the IQR above the $75^{th}$ percentile or between 1.5 and 3 times the IQR below the $25^{th}$ percentile, and
- data points that exceed 3 times the IQR above the $75^{th}$ percentile or exceed 3 times the IQR below the $25^{th}$ percentile.

The mean is shown as a star, while the lower and upper 95% confidence limits around the mean are shown as bars. Individual data points between 1.5 and 3 times the IQR above the 75$^{th}$ percentile or below the 25$^{th}$ percentile are shown as circles. Individual data points at least 3 times the IQR above the 75$^{th}$ percentile or below the 25$^{th}$ percentile are shown as squares.

Information from box plots can assist in identifying potential data distributions. If the upper box and whisker are approximately the same length as the lower box and whisker, with the mean and median approximately equal, then the data are distributed symmetrically. The normal distribution is one of a number that is symmetric. If the upper box and whisker are longer than the lower box and whisker, with the mean greater than the median, then the data are right-skewed (such as lognormal or square root normal distributions in original units). Conversely, if the upper box and whisker are shorter than the lower box and whisker with the mean less than the median, then the data are left-skewed.

A box plot showing a normal distribution will have the following characteristics: the mean and median will be in the center of the box, whiskers to the minimum and maximum values are the same length, and there would be no potential outliers. A box plot showing a lognormal distribution (in original units) typical of environmental applications will have the following characteristics: the mean will be larger than the median, the whisker above the 75$^{th}$ percentile will be longer than the whisker below the 25$^{th}$ percentile, and extreme upper values may be indicated as potential outliers. Once the data have been logarithmically transformed, the pattern should follow that described for a normal distribution. Other right-skewed distributions transformable to normality would indicate similar patterns.

It is often helpful to show box plots of different sets of data side by side to show differences between monitoring stations (see **Figure 9-2**). This allows a simple method to compare the locations, spreads and shapes of several data sets or different groups within a single data set. In this situation, the width of the box can be proportional to the sample size of each data set. If the data will be compared to a standard, such as a preliminary remediation goal (PRG) or maximum contaminant level (MCL), a line on the graph can be drawn to show if any results exceed the criteria.

It is important to plot the data as reported by the laboratory for non-detects or negative radionuclide data. Proxy values for non-detects should not be plotted since we want to see the distribution of the original data. Different symbols can be used to display non-detects, such as the open symbols described in **Section 9.1**. The mean will be biased high if using the RL of non-detects in the calculation, but the purpose of the box plot is to assess the distribution of the data, not quantifying a precise estimate of an unbiased mean. Displaying the frequency of detection (number of detected values / number of total samples) under the station name is also useful. Unlike time series plots, box plots cannot use missing data, so missing data should be removed before producing a box plot.

Directions for generating a box plot are contained in **Example 9-2**, and an example is shown in **Figure 9-2.** It is important to remove lab and field duplicates from the data before calculating summary statistics such as the mean and UCL since these statistics assume independent data. The box plot assumes the data are statistically independent.

►EXAMPLE 9-2

Construct a box plot using the trichloroethene groundwater data in **Table 9-1** for each well. Examine the box plot to assess how each well is distributed (normal, lognormal, skewed, symmetric, etc.). Identify possible outliers.

SOLUTION

Step 1. Import the data into data analysis software capable of producing box plots.

Step 2. Sort the data from smallest to largest results by well.

Step 3. Compute the $0^{th}$ (minimum value), $25^{th}$, $50^{th}$ (median), $75^{th}$ and $100^{th}$ (maximum value) percentiles by well.

Step 4. Plot these points vertically. Draw a box around the $25^{th}$ and $75^{th}$ percentiles and add a line through the box at the $50^{th}$ percentile. Optionally, make the width of the box proportional to the sample size. Narrow boxes reflect smaller sample sizes, while wider boxes reflect larger sample sizes.

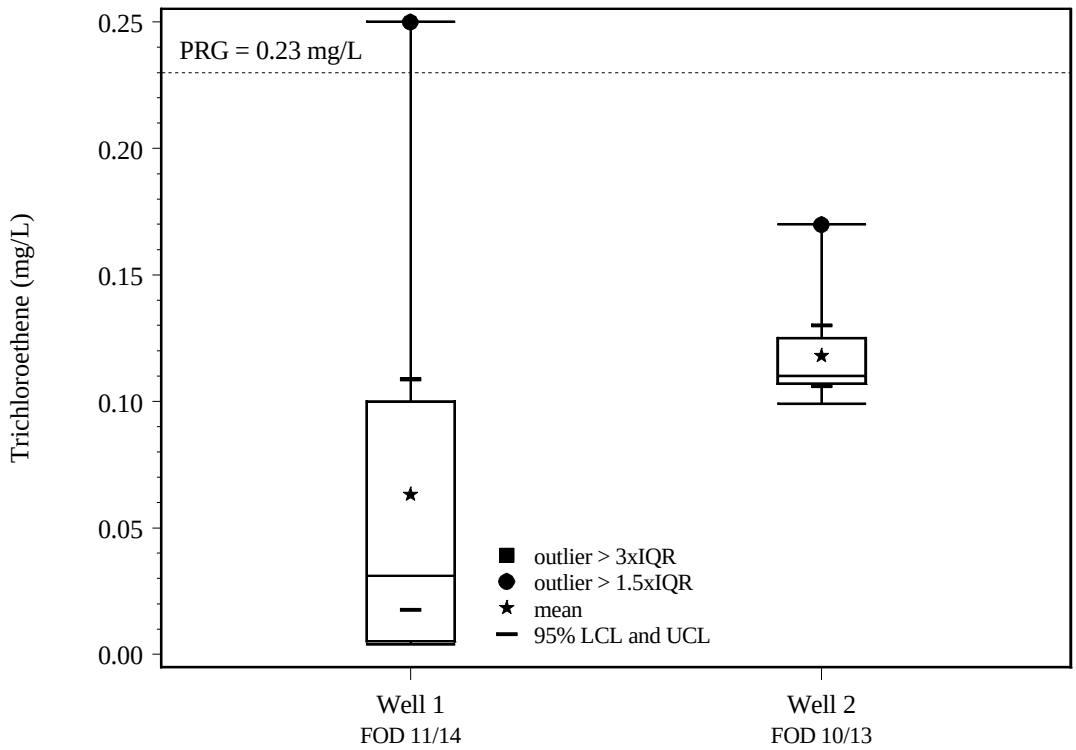
Step 5. Compute the mean and the lower and upper 95% confidence limits. Denote the mean with a star and the confidence limits as bars. Also, identify potential outliers between 1.5×IQR and 3×IQR beyond the box with a circle. Identify potential outliers exceeding 3×IQR beyond the box with a ` square.

Step 6. Draw the whiskers from each end of the box to the furthest data point to show the full range of the data.

INTERPRETATION

The box plots in **Figure 9-2** show the similarities and differences in the distributions of trichloroethene in Wells 1 and 2. The mean of trichloroethene in Well 1 is significantly lower than the mean in Well 2. The variance of the data from Well 1 is significantly larger than the variance from Well 2. A parametric t-test or nonparametric Wilcoxon Rank Sum test can quantitatively confirm these conclusions. Since the mean exceeds the median for both wells and the whiskers at the top of each box are much longer than the whiskers at the bottom of each box, we can conclude both distributions are skewed to the right, resembling a lognormal distribution. In fact, the Shapiro-Wilk test quantitatively confirms that both distributions are lognormally distributed. Both wells have their largest concentrations between 1.5 and 3 times the IQR, as denoted by a black circle. No data point lies outside 3 times the IQR. Since the data for both wells are lognormally distributed, the maximum concentrations in each well should not be removed just because they exceed 1.5 times the IQR. Long tails are expected for the lognormal distribution. The width of the 95% confidence limits confirms the large variability in Well 1 compared to the width of the confidence limits in Well 2. Well 1 has one concentration exceeding the PRG of 0.23 mg/L, while Well 2 has all concentrations below the PRG. The width of each box is similar since the sample size as shown in the frequency of detection (FOD) are nearly the same (11 detects out of 14 samples for Well 1 and 10 detects out of 13 samples for Well 2). ◄

Figure 9-2. Box Plots of Trichloroethene Data for Wells 1 & 2




## 9.3 HISTOGRAMS

A histogram is a visual representation of the data collected into groups. This graphical technique provides a visual method of identifying the underlying distribution of the data. The data range is divided into several bins or classes and the data is sorted into the bins. A histogram is a bar graph conveying the bins and the frequency of data points in each bin. Other forms of the histogram use a normalization of the bin frequencies for the heights of the bars. The two most common normalizations are relative frequencies (frequencies divided by sample size) and densities (relative frequency divided by the bin width). **Figure 9-3** is an example of a histogram using frequencies and **Figure 9-4** is a histogram of densities. Histograms provide a visual method of accessing location, shape and spread of the data. Also, extreme values and multiple modes can be identified. The details of the data are lost, but an overall picture of the data is obtained. A stem and leaf plot offers the same insights into the data as a histogram, but the data values are retained.

The visual impression of a histogram is sensitive to the number of bins selected. A large number of bins will increase data detail, while fewer bins will increase the smoothness of the histogram. A good starting point when choosing the number of bins is the square root of the sample size $n$. The minimum number of bins for any histogram should be at least 4. Another factor in choosing bins is the choice of endpoints. When feasible, using simple bin endpoints can improve the readability of the histogram. Simple bin endpoints include multiples of $5k$ units for some integer $k > 0$ (e.g., 0 to <5, 5 to <10, etc. or 1 to <1.5, 1.5 to <2, etc.). Finally, when plotting a histogram for a continuous variable (e.g.,

concentration), it is necessary to decide on an endpoint convention; that is, what to do with data points that fall on the boundary of a bin. Also, use the data as reported by the laboratory for non-detects and eliminate any missing values, since histograms cannot include missing data. With discrete variables, (e.g., family size) the intervals can be centered in between the variables. For the family size data, the intervals can span between 1.5 and 2.5, 2.5 and 3.5, and so on. Then the whole numbers that relate to the family size can be centered within the box. Directions for generating a histogram are contained in **Example 9-3**.

►EXAMPLE 9-3

Construct a histogram using the trichloroethene groundwater data in **Table 9-1** for each well. Examine the histogram to assess how each well is distributed (normal, lognormal, skewed, symmetric, etc.).

SOLUTION

Step 1. Import the data into data analysis software capable of producing histograms.

Step 2. Sort the data from smallest to largest results by well.

Step 3. With $n$ = 14 concentrations for Well 1, a rough estimate of the number of bins is $\sqrt{14}$ = 3.74 or 4 bins. Since the data from Well 1 range from 0.004 to 0.25, the suggested bin width is calculated as (maximum concentration – minimum concentration) / number of bins = (0.25 – 0.004) / 4 = 0.0615. Therefore, the bins for Well 1 are 0.004 to <0.0655, 0.0655 to <0.127, 0.127 to <0.1885, and 0.1885 to 0.25 mg/L.

Similarly, with $n$ = 13 concentrations for Well 2, the number of bins is $\sqrt{13}$ = 3.61 or 4 bins. Since the data from Well 2 range from 0.099 to 0.17, the suggested bin width is calculated as (maximum concentration – minimum concentration) / number of bins = (0.17 – 0.099) / 4 = 0.01775. Therefore, the bins for Well 2 are 0.099 to <0.11675, 0.11675 to <0.1345, 0.1345 to <0.15225, and 0.15225 to 0.17 mg/L.

Step 4. Construct a frequency table using the bins defined in Step 3. **Table 9-2** shows the frequency or number of observations within each bin defined in Step 3 for Wells 1 and 2. The third column shows the relative frequency which is the frequency divided by the sample size $n$. The final column of the table gives the densities or the relative frequencies divided by the bin widths calculated in Step 3.

Step 5. The horizontal axis for the data is from 0.004 to 0.25 mg/L for Well 1 and 0.099 to 0.17 for Well 2. The vertical axis for the histogram of frequencies is from 0 to 9 and the vertical axis for the histogram of relative frequencies is from 0% - 70%.

Step 6. The histograms of frequencies are shown in **Figure 9-3**. The histograms of relative frequencies or densities are shown in **Figure 9-4**. Note that frequency, relative frequency and density histograms all show the same shape since the scale of the vertical axis is divided by

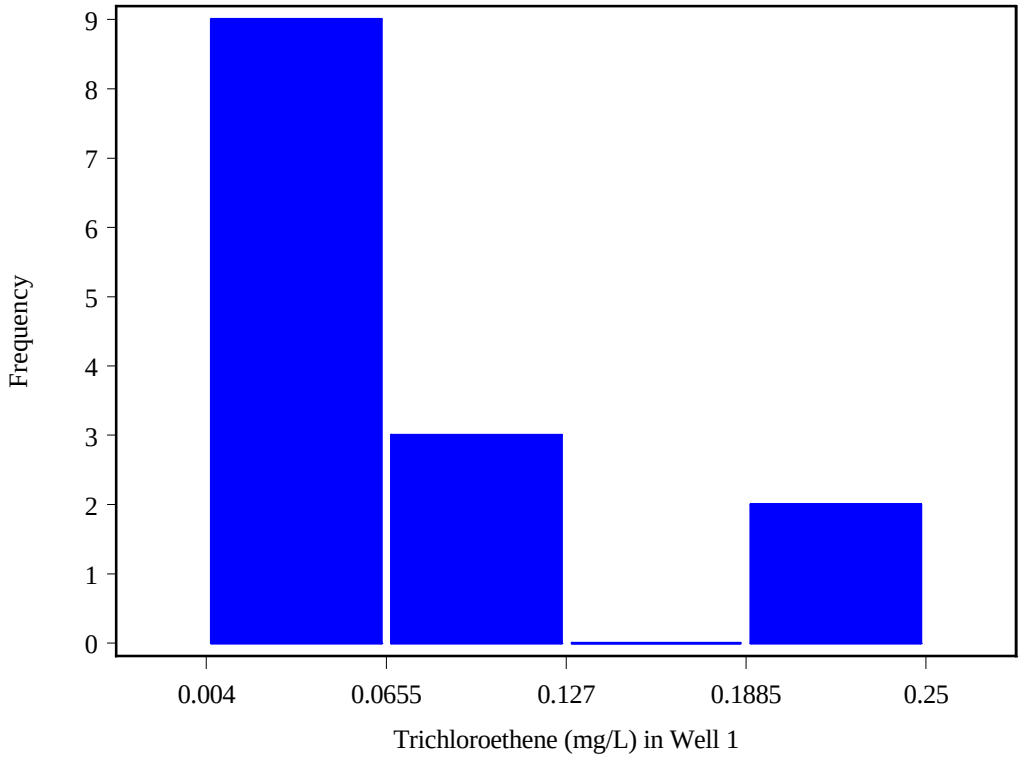
the sample size or the bin width. These histograms confirm the data are not normally distributed for either well, but are closer to lognormal.

Table 9-2. Histogram Bins for Trichloroethene Groundwater Data

| Bin | Frequency | Relative Frequency (%) | Density |
|---|---|---|---|
| Well 1 | | | |
| 0.0040 to <0.0655 mg/L | 9 | 64.3 | 10.5 |
| 0.0655 to <0.1270 mg/L | 3 | 21.4 | 3.5 |
| 0.1270 to <0.1885 mg/L | 0 | 0 | 0 |
| 0.1885 to 0.2500 mg/L | 2 | 14.3 | 2.3 |
| Well 2 | | | |
| 0.099 to <0.11675 mg/L | 8 | 61.5 | 34.7 |
| 0.11675 to <0.1345 mg/L | 3 | 23.1 | 13.0 |
| 0.1345 to <0.15225 mg/L | 1 | 7.7 | 4.3 |
| 0.15225 to 0.17 mg/L | 1 | 7.7 | 4.3 |

◄

Figure 9-3. Frequency Histograms of Trichloroethene by Well.



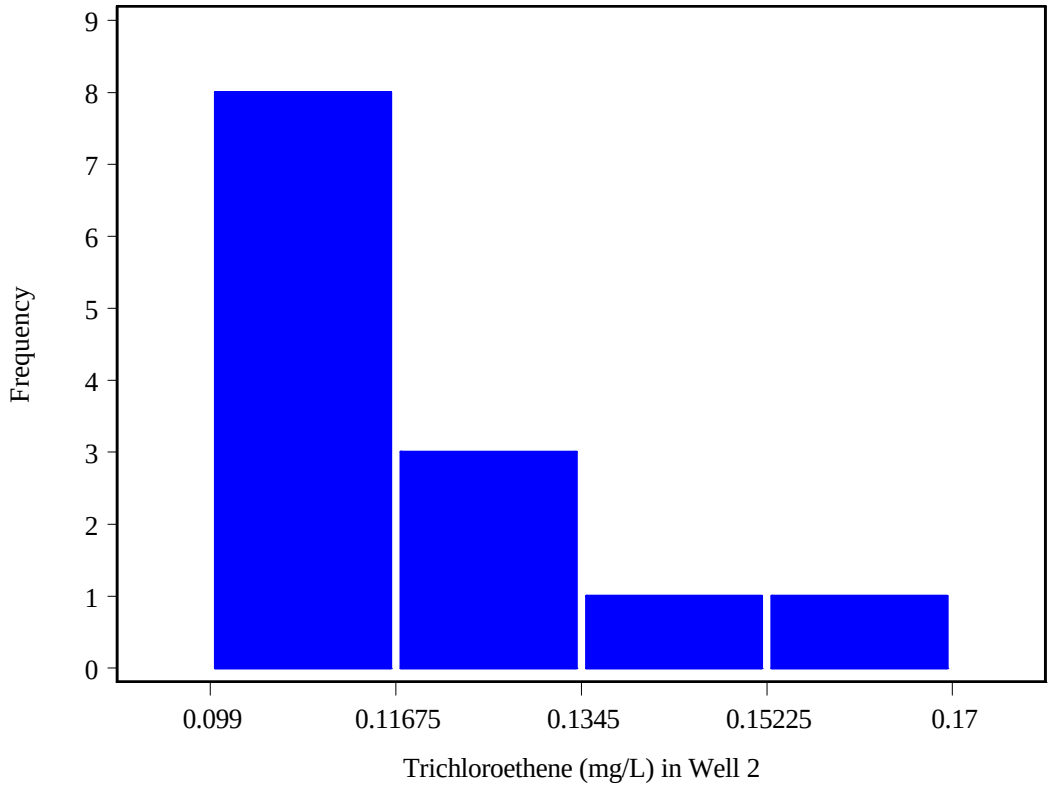
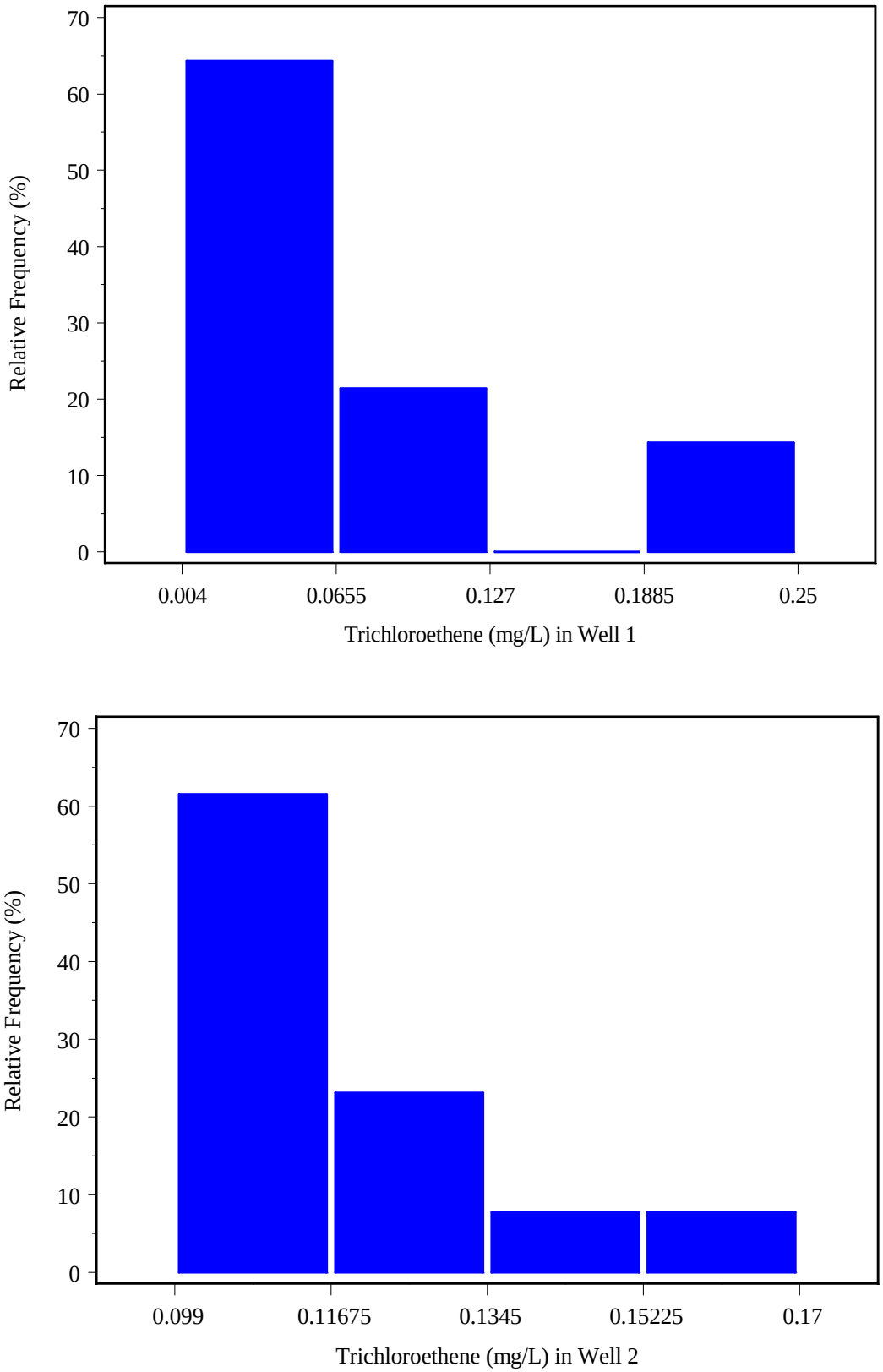



Figure 9-4. Relative Frequency Histograms of Trichloroethene by Well.



## 9.4 SCATTER PLOTS

For data sets consisting of multiple observations per sampling point, a scatter plot is one of the most powerful graphical tools for analyzing the relationship between two or more variables. Scatter plots are easy to construct for two variables, and many software packages can construct 3-dimensional scatter plots. A scatter plot can clearly show the relationship between two variables if the data range is sufficiently large. Truly linear relationships can always be identified in scatter plots, but truly nonlinear relationships may appear linear (or some other form) if the data range is relatively small. Scatter plots of linearly correlated variables cluster about a straight line.

As an example of a nonlinear relationship, consider two variables where one variable is approximately equal to the square of the other. With an adequate range in the data, a scatter plot of this data would display a partial parabolic curve. Other important modeling relationships that may appear are exponential or logarithmic. Two additional uses of scatter plots are the identification of potential outliers for a single variable or for the paired variables and the identification of clustering in the data. Directions for generating a scatter plot are contained in **Example 9-4**.

### ►EXAMPLE 9-4

Construct a scatter plot using the groundwater data in **Table 9-3** for arsenic and mercury from a single well collected approximately quarterly across time. Examine the scatter plot for linear or quadratic relationships between arsenic and mercury, correlation, and for potential outliers.

Table 9-3. Groundwater Concentrations from Well 3

| | **Arsenic** | | **Mercury** | | **Strontium** | |
|---|---|---|---|---|---|---|
| **Date Collected** | **Conc. (mg/L)** | **Data Qualifier** | **Conc. (mg/L)** | **Data Qualifier** | **Conc. (mg/L)** | **Data Qualifier** |
| 1/2/2005 | 0.01 | U | 0.02 | U | 0.10 | |
| 4/7/2005 | 0.01 | U | 0.03 | | 0.02 | U |
| 7/13/2005 | 0.02 | | 0.04 | U | 0.05 | U |
| 10/24/2005 | 0.04 | | 0.06 | | 0.11 | |
| 1/7/2006 | 0.01 | | 0.02 | | 0.05 | |
| 3/30/2006 | 0.05 | | 0.07 | | 0.07 | |
| 6/28/2006 | 0.09 | | 0.10 | | 0.03 | |
| 10/2/2006 | 0.07 | | 0.08 | | 0.04 | |
| 10/17/2006 | 0.10 | | NA | | 0.02 | U |
| 1/15/2007 | 0.02 | U | 0.03 | U | 0.15 | |
| 4/10/2007 | 0.15 | | 0.11 | | 0.03 | |
| 7/9/2007 | 0.12 | | 0.08 | | 0.10 | |
| 10/5/2007 | 0.10 | | 0.07 | | 0.09 | |
| 10/29/2007 | 0.30 | | 0.29 | | 0.05 | |
| 12/30/2007 | 0.25 | | 0.23 | | 0.22 | |

NA = Not available (missing data).

U denotes a non-detect.

SOLUTION

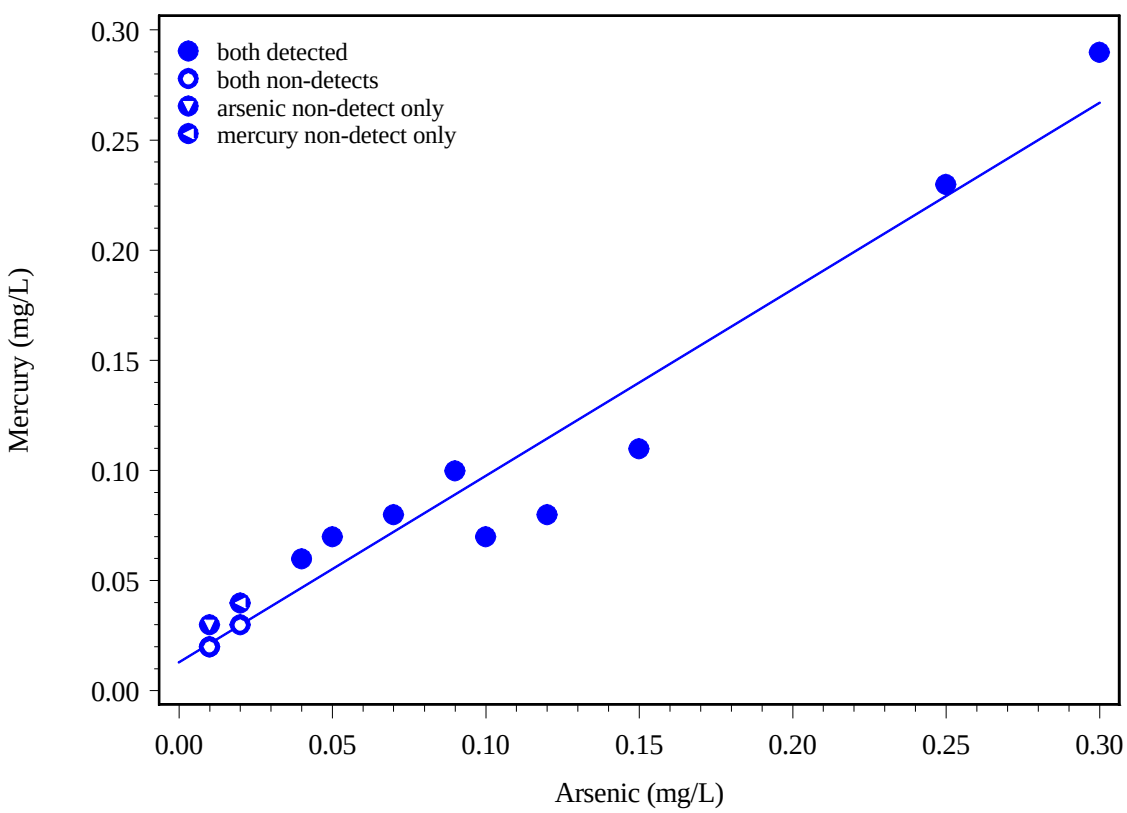
Step 1. Import the data into data analysis software capable of producing scatter plots.

Step 2. Sort the data by date collected.

Step 3. Calculate the range of concentrations for each constituent. If the range of both constituents are similar, then scale both the X and Y axes from the minimum to the maximum concentrations of both constituents. If the range of concentrations are very different (e.g., two or more orders of magnitude), then perhaps the scales for both axes should be logarithmic ($\log_{10}$). The data will be plotted as pairs from ($X_1$, $Y_1$) to ($X_n$, $Y_n$) for each sampling date, where $n$ = number of samples.

Step 4. Use separate symbols to distinguish detected from non-detected concentrations. Note that the concentration for one constituent may be detected, while the concentration for the other constituent may not be detected for the same sampling date. If the concentration for one constituent is missing, then the pair ($X_i$, $Y_i$) cannot be plotted since both concentrations are required. **Figure 9-5** shows a linear correlation between arsenic and mercury with two possible outliers. The Pearson correlation coefficient is 0.97, indicating a significantly high correlation. The linear regression line is displayed to show the linear correlation between arsenic and mercury.◀

Figure 9-5. Scatter Plot of Arsenic with Mercury from Well 3



Many software packages can extend the 2-dimensional scatter plot by constructing a 3-dimensional scatter plot for 3 constituents. However, with more than 3 variables, it is difficult to construct and interpret a scatter plot. Therefore, several graphical representations have been developed that extend the idea of a scatter plot for data consisting of more than 2 variables. The simplest of these graphical techniques is a coded scatter plot. All possible two-way combinations are given a symbol and the pairs of data are plotted on one 2-dimensional scatter plot. The coded scatter plot does not provide information on three way or higher interactions between the variables since only two dimensions are plotted. If the data ranges for the variables are comparable, then a single set of axes may suffice. If the data ranges are too dissimilar (e.g., at least two orders of magnitude), different scales may be required.

►EXAMPLE 9-5

Construct a coded scatter plot using the groundwater data in **Table 9-3** for arsenic, mercury, and strontium from Well 3 collected approximately quarterly across time. Examine the scatter plot for linear or quadratic relationships between the three inorganics, correlation, and for potential outliers.

SOLUTION

Step 1. Import the data into data analysis software capable of producing scatter plots.

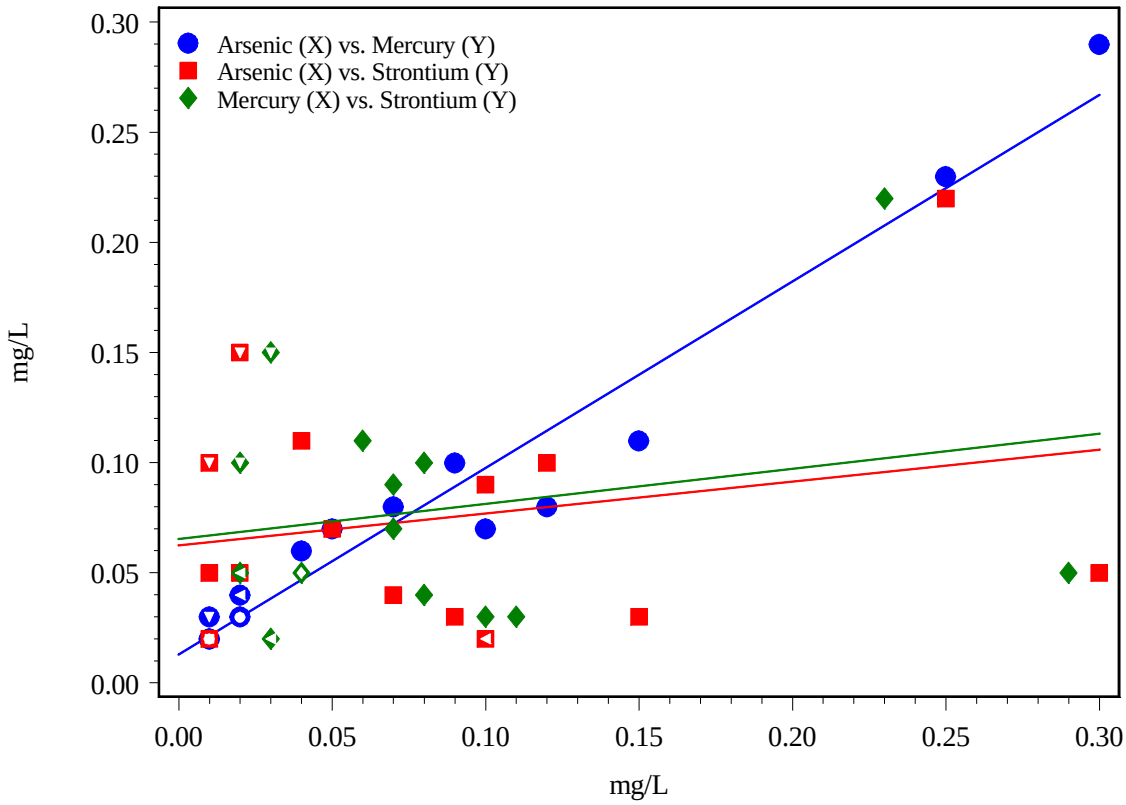
Step 2. Sort the data by date collected.

Step 3. Calculate the range of concentrations for each constituent. If the ranges of both constituents are similar, then scale both the X and Y axes from the minimum to the maximum concentrations of all three constituents. Since the ranges of concentrations are very similar, the minimum to the maximum concentrations of all three constituents will be used for both axes.

Step 4. Let each arsenic concentration be denoted by $X_i$, each mercury concentration be denoted by $Y_i$, and each strontium concentration be denoted by $Z_i$. The arsenic and mercury paired data will be plotted as pairs ($X_i$, $Y_i$) with solid blue circles for $1 \leq i \leq n$. The arsenic and strontium paired data will be plotted as pairs ($X_i$, $Z_i$) with solid red squares. The mercury and strontium paired data will be plotted as pairs ($Y_i$, $Z_i$) with solid green diamonds. If either concentration in each pair is a non-detect, then the non-detects will be displayed similar to **Figure 9-5**.

Step 5. Interpret the plot. **Figure 9-6** shows the linear correlation between arsenic and mercury with two possible outliers. The Pearson correlation coefficient is 0.97, indicating a significantly high correlation. The approximate 45º slope of the regression line indicates a strong correlation between arsenic and mercury. However, the nearly zero slope of the regression line between arsenic and strontium indicates little or no correlation between arsenic and strontium. There are two possible outliers for arsenic and strontium. Similarly, the nearly zero slope of the regression line between mercury and strontium indicates little or no correlation between mercury and strontium. There are also two possible outliers for mercury and strontium. The Pearson correlation coefficients for both arsenic with strontium and mercury with strontium are 0.23 which are not significantly different from zero.◄

Figure 9-6. Coded Scatter Plot of Well 3 Arsenic, Mercury, and Strontium




## 9.5 PROBABILITY PLOTS

A simple, but extremely useful visual assessment of normality is to graph the data as a probability plot. The *y*-axis is scaled to represent quantiles or *z*-scores from a standard normal distribution and the concentration measurements are arranged in increasing order along the *x*-axis. As each observed value is plotted on the *x*-axis, the *z*-score corresponding to the proportion of observations less than or equal to that measurement is plotted as the *y*-coordinate. Often, the *y*-coordinate is computed by the following formula:

$$y_i = \Phi^{-1}\left(\frac{i}{n+1}\right) \qquad [9.1]$$

where $\Phi^{-1}$ denotes the inverse of the cumulative standard normal distribution, $n$ represents the sample size, and $i$ represents the rank position of the $i^{th}$ ordered concentration. The plot is constructed so that, if the data are normal, the points when plotted will lie on a straight line. Visually apparent curves or bends indicate that the data do not follow a normal distribution.

Probability plots are particularly useful for spotting irregularities within the data when compared to a specific distributional model (usually, but not always, the normal). It is easy to determine whether departures from normality are occurring more or less in the middle ranges of the data or in the extreme

tails. Probability plots can also indicate the presence of possible outlier values that do not follow the basic pattern of the data and can show the presence of significant positive or negative skewness.

If a (normal) probability plot is constructed on the combined data from several wells and normality is accepted, it suggests — but does not prove — that all of the data came from the same normal distribution. Consequently, each subgroup of the data set (*e.g.*, observations from distinct wells) probably has the same mean and standard deviation. If a probability plot is constructed on the data residuals (each value minus its subgroup mean) and is not a straight line, the interpretation is more complicated. In this case, either the residuals are not normally-distributed, or there is a subgroup of the data with a normal distribution but a different mean or standard deviation than the other subgroups. The probability plot will indicate a deviation from the underlying assumption of a common normal distribution in either case. It would be prudent to examine normal probability plots by well on the same plot if the ranges of the data are similar. This would show how the data are distributed by well to determine which wells may depart from normality.

The same probability plot technique may be used to investigate whether a set of data or residuals follows a lognormal distribution. The procedure is generally the same, except that one first replaces each observation by its natural logarithm. After the data have been transformed to their natural logarithms, the probability plot is constructed as before. The only difference is that the natural logarithms of the observations are used on the *x*-axis. If the data are lognormal, the probability plot of the logged observations will approximate a straight line.

►EXAMPLE 9-6

Determine whether the dataset in **Table 9-4** is normal by using a probability plot.

SOLUTION

Step 1. After combining the data into a single group, list the measured nickel concentrations in order from lowest to highest.

Step 2. The cumulative probabilities, representing for each observation ($x_i$) the proportion of values less than or equal to $x_i$, are given in the third column of the table below. These are computed as $i / (n + 1)$ where $n$ is the total number of samples ($n = 20$).

Step 3. Determine the quantiles or *z*-scores from the standard normal distribution corresponding to the cumulative probabilities in Step 2. These can be found by successively letting $P$ equal each cumulative probability and then looking up the entry in **Table 10-1 (Appendix D)** corresponding to $P$. Since the standard normal distribution is symmetric about zero, for cumulative probabilities $P < 0.50$, look up the entry for $(1–P)$ and give this value a negative sign.

Step 4. Plot the normal quantile (*z*-score) versus the ordered concentration for each sample, as in the plot below (**Figure 9-7**). The curvature found in the probability plot indicates that there is evidence of non-normality in the data. ◄

Table 9-4. Nickel Concentrations from a Single Well

| Nickel Concentration (ppb) | Order ($i$) | Cumulative Probability [$i/(n+1)$] | Normal Quantile ($z$-score) |
|---|---|---|---|
| 1.0 | 1 | 0.048 | –1.668 |
| 3.1 | 2 | 0.095 | –1.309 |
| 8.7 | 3 | 0.143 | –1.068 |
| 10.0 | 4 | 0.190 | –0.876 |
| 14.0 | 5 | 0.238 | –0.712 |
| 19.0 | 6 | 0.286 | –0.566 |
| 21.4 | 7 | 0.333 | –0.431 |
| 27.0 | 8 | 0.381 | –0.303 |
| 39.0 | 9 | 0.429 | –0.180 |
| 56.0 | 10 | 0.476 | –0.060 |
| 58.8 | 11 | 0.524 | 0.060 |
| 64.4 | 12 | 0.571 | 0.180 |
| 81.5 | 13 | 0.619 | 0.303 |
| 85.6 | 14 | 0.667 | 0.431 |
| 151.0 | 15 | 0.714 | 0.566 |
| 262.0 | 16 | 0.762 | 0.712 |
| 331.0 | 17 | 0.810 | 0.876 |
| 578.0 | 18 | 0.857 | 1.068 |
| 637.0 | 19 | 0.905 | 1.309 |
| 942.0 | 20 | 0.952 | 1.668 |

### PROBABILITY PLOTS FOR LOG TRANSFORMED DATA

Step 1. List the natural logarithms of the measured nickel concentrations in **Table 9-4** in order from lowest to highest. These are shown in **Table 9-5**.

Step 2. The cumulative probabilities representing the proportion of values less than or equal to $x_i$ for each observation ($x_i$), are given in the third column of **Table 9-4**. These are computed as $i$ / ($n$ + 1) where $n$ is the total number of samples ($n$ = 20).

Step 3. Determine the quantiles or $z$-scores from the standard normal distribution corresponding to the cumulative probabilities in Step 2. These can be found by successively letting $P$ equal each cumulative probability and then looking up the entry in **Table 10-1 Appendix D** corresponding to $P$. Since the standard normal distribution is symmetric about zero, for cumulative probabilities $P < 0.50$, look up the entry for (1–$P$) and give this value a negative sign.

Table 9-5. Nickel Log Concentrations from a Single Well

| Order ($i$) | Log Nickel Concentration log(ppb) | Cumulative Probability [$i/(n+1)$] | Normal Quantile ($z$-score) |
|---|---|---|---|
| 1 | 0.00 | 0.048 | –1.668 |
| 2 | 1.13 | 0.095 | –1.309 |
| 3 | 2.16 | 0.143 | –1.068 |
| 4 | 2.30 | 0.190 | –0.876 |
| 5 | 2.64 | 0.238 | –0.712 |
| 6 | 2.94 | 0.286 | –0.566 |
| 7 | 3.06 | 0.333 | –0.431 |
| 8 | 3.30 | 0.381 | –0.303 |
| 9 | 3.66 | 0.429 | –0.180 |
| 10 | 4.03 | 0.476 | –0.060 |
| 11 | 4.07 | 0.524 | 0.060 |
| 12 | 4.17 | 0.571 | 0.180 |
| 13 | 4.40 | 0.619 | 0.303 |
| 14 | 4.45 | 0.667 | 0.431 |
| 15 | 5.02 | 0.714 | 0.566 |
| 16 | 5.57 | 0.762 | 0.712 |
| 17 | 5.80 | 0.810 | 0.876 |
| 18 | 6.36 | 0.857 | 1.068 |
| 19 | 6.46 | 0.905 | 1.309 |
| 20 | 6.85 | 0.952 | 1.668 |

Step 4. Plot the normal quantile ($z$-score) versus the ordered logged concentration for each sample, as in the plot below (**Figure 9-8**). The reasonably linear trend found in the probability plot indicates that the log-scale data closely follow a normal pattern, further suggesting that the original data closely follow a lognormal distribution.

Figure 9-7. Nickel Normal Probability Plot

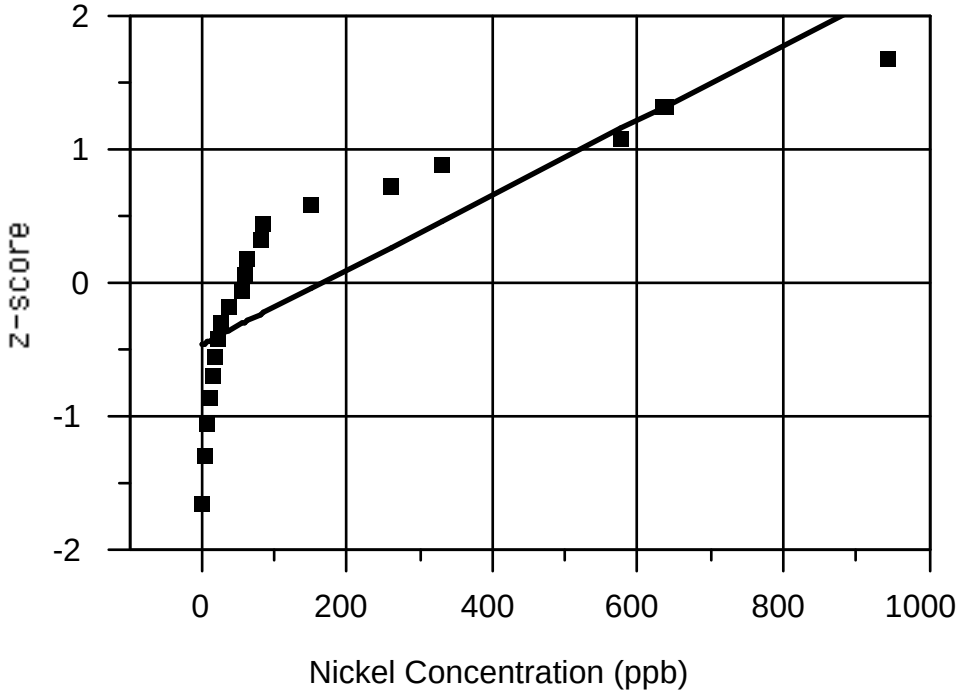
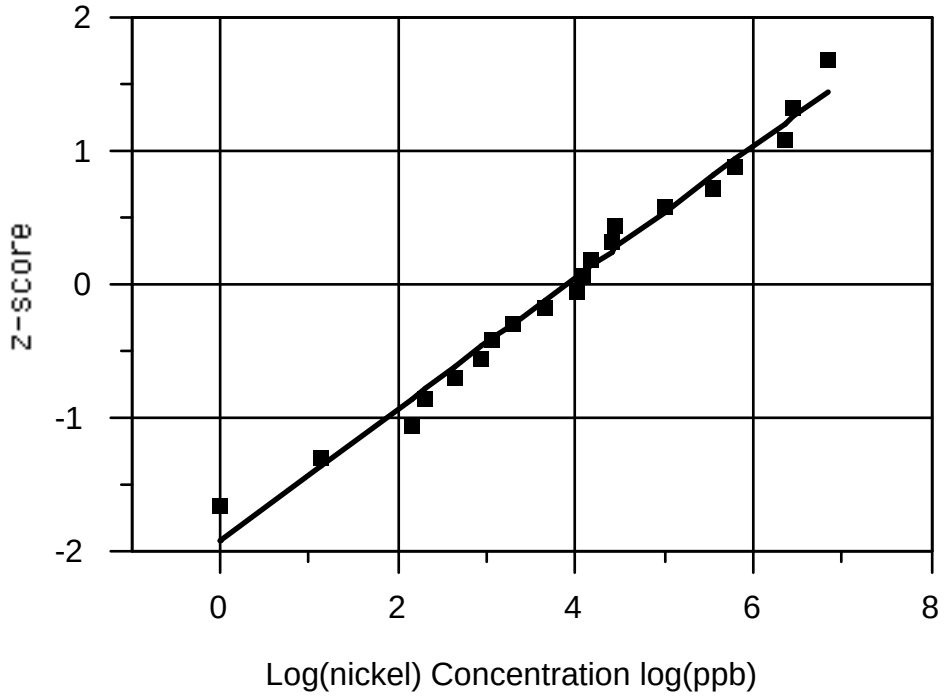



Figure 9-8. Probability Plot of Log Transformed Nickel Data



# CHAPTER 10. FITTING DISTRIBUTIONS

10.1 IMPORTANCE OF DISTRIBUTIONAL MODELS ..... 10-1
10.2 TRANSFORMATIONS TO NORMALITY ..... 10-3
10.3 USING THE NORMAL DISTRIBUTION AS A DEFAULT ..... 10-5
10.4 COEFFICIENT OF VARIATION AND COEFFICIENT OF SKEWNESS ..... 10-9
10.5 SHAPIRO-WILK AND SHAPIRO-FRANCÍA NORMALITY TESTS ..... 10-13
*10.5.1 Shapiro-Wilk Test (n ≤ 50) ..... 10-13*
*10.5.2 Shapiro-Francía Test (n > 50) ..... 10-15*
10.6 PROBABILITY PLOT CORRELATION COEFFICIENT ..... 10-16
10.7 SHAPIRO-WILK MULTIPLE GROUP TEST OF NORMALITY ..... 10-18

Because a statistical or mathematical model is at best an approximation of reality, all statistical tests and procedures require certain assumptions for the methods to be used correctly and for the results to be properly interpreted. Many tests make an assumption regarding the underlying distribution of the observed data; in particular, that the original or transformed sample measurements follow a normal distribution. Data transformations are discussed in **Section 10.2** while considerations as to whether the normal distribution should be used as a 'default' are explored in **Section 10.3**. Several techniques for assessing normality are also examined, including:

- The skewness coefficient (**Section 10.4**)
- The Shapiro-Wilk test of normality and its close variant, the Shapiro-Francía test (**Section 10.5**)
- Filliben's probability plot correlation coefficient test (**Section 10.6**)
- The Shapiro-Wilk multiple group test of normality (**Section 10.7**)

## 10.1 IMPORTANCE OF DISTRIBUTIONAL MODELS

As introduced in **Chapter 3**, all statistical testing relies on the critical assumption that the sample data are *representative* of the population from which they are selected. The statistical distribution of the *sample* is assumed to be similar to the distribution of the mostly unobserved *population* of possible measurements. Many *parametric* testing methods make a further assumption: that the form or type of the underlying population is at least approximately known or can be identified through diagnostic testing. Most of these parametric tests assume that the population is *normal* in distribution; the validity or accuracy of the test results may be in question if that assumption is violated.

Consequently, an important facet of choosing among appropriate test methods is determining whether a commonly-used statistical distribution such as the normal, adequately models the observed sample data. A large variety of possible distributional models exist in the statistical literature; most are not typically applied to groundwater measurements and often introduce additional statistical or mathematical complexity in working with them. So groundwater statistical models are usually confined to the gamma distribution, the Weibull distribution, or distributions that are normal or can be normalized via a transformation (e.g., the logarithmic or square root).

Although the Unified Guidance will occasionally reference procedures that assume an underlying gamma or Weibull distribution, the presentation in this guidance will focus on distributions that can be normalized and diagnostic tools for assessing normality. The principal reasons for limiting the discussion in this manner are: 1) the same tools useful for testing normality can be utilized with any distribution that can be normalized-- the only change needed is perform the normality test after first making a data transformation; 2) if no transformation works to adequately normalize the sample data, a non-parametric test can often be used as an alternative statistical approach; and 3) addressing more complicated scenarios is outside the scope of the guidance and may require professional statistical consultation.

Understanding the statistical behavior of groundwater measurements can be very challenging. The constituents of interest may occur at relatively low concentrations and frequently be left-censored because of current analytical method limitations. Sample data are often positively skewed and asymmetrical in distributional pattern, perhaps due to the presence of outliers, inhomogeneous mixing of contaminants in the subsurface, or spatially variable soils deposition affecting the local groundwater geochemistry. For some constituents, the distribution in groundwater is not stationary over time (*e.g.*, due to linear or seasonal trends) or not stationary across space (due to spatial variability in mean levels from well to well). A set of these measurements pooled over time and/or space may appear highly non-normal, even if the underlying population at any fixed point in time or space *is* normal.

Because of these complexities, fitting a distributional model to a set of sample data cannot be done in isolation from checks of other key statistical assumptions. The data must also be evaluated for outliers (**Chapter 12**), since the presence of even one extreme outlier may cause an otherwise recognizable distribution from being correctly identified. For data grouped across wells, the possible presence of spatial variability must be considered (**Chapter 13**). If identified, the Shapiro-Wilk multiple group test of normality may be needed to account for differing means and/or variances at distinct wells. Data pooled across sampling events (*i.e.*, over time) must be examined for the presence of trends or seasonal patterns (**Chapter 14**). A clearly identified pattern may need to be removed and the *data residuals* tested for normality, instead of the raw measurements.

A frequently encountered problem involves testing normality on data sets containing non-detect values. The best goodness-of-fit tests attempt to assess whether the sample data closely resemble the *tails* of the candidate distributional model. Since non-detects represent *left-censored observations* where the exact concentrations are unknown for the lower tail of the sample distribution, standard normality tests cannot be run without some estimate or *imputation* of these unknown values. For a small fraction of non-detects in a sample (10-15% or less) censored at a single reporting limit, it may be possible to apply a normality test by simply replacing each non-detect with an imputed value of half the RL. However, more complicated situations arise when there is a combination of multiple RLs (detected values intermingled with different non-detect levels), or the proportion of non-detects is larger. The Unified Guidance recommends different strategies in these circumstances.

Properly *ordering* the sample observations (*i.e.*, from least to greatest) is critical to *any* distributional goodness-of-fit test. Because the concentration of a non-detect measurement is only known to be in the range from zero to the RL, it is generally impossible to construct a full ordering of the

sample.[1] There are methods, however, to construct *partial orderings* of the data that allow the assignment of relative rankings to each of the detected measurements and which account for the presence of censored values. In turn, a partial ordering enables construction of an approximate normality test. This subject is covered in **Chapter 15**.

## 10.2 TRANSFORMATIONS TO NORMALITY

Guidance users will often encounter data sets indicating significant evidence of non-normality. Due to the presumption of most parametric tests that the underlying population is normal, a common statistical strategy for apparently non-normal observations is to search for a normalizing mathematical transformation. Because of the complexities associated with interpreting statistical results from data that have been transformed to another scale, some care must be taken in applying statistical procedures to transformed measurements. In questionable or disputable circumstances, it may be wise to analyze the same data with an equivalent non-parametric version of the same test (if it exists) to see if the same general conclusion is reached. If not, the data transformation and its interpretation may need further scrutiny.

Particularly with prediction limits, control charts, and some of the confidence intervals described in **Chapters 18**, **20,** and **21**, the parametric versions of these procedures are especially advantageous. Here, a transformation may be warranted to approximately normalize the statistical sample. Transformations are also often useful when combining or pooling intrawell background from several wells in order to increase the degrees of freedom available for intrawell testing (**Chapter 13**). Slight differences in the distributional pattern from well to well can skew the resulting pooled dataset, necessitating a transformation to bring about approximate normality and to equalize the variances.

The interpretation of transformed data is straightforward in the case of prediction limits for individual observations or when building a confidence interval around an upper percentile. An interval with limits constructed from the transformed data and then re-transformed (or *back-transformed*) to the original measurement domain will retain its original probabilistic interpretation. For instance, if the data are approximately normal under a square root transformation and a 95% confidence prediction limit is constructed on the square roots of the original measurements, *squaring* the resulting prediction limit allows for a 95% confidence level when applied to the original data.

The same ease of interpretation does not apply to prediction limits for a future arithmetic mean (**Chapter 18**) or to confidence intervals around an arithmetic mean compared to a fixed GWPS (**Chapter 21**). A back-transformed confidence interval constructed around the mean of log-transformed data (*i.e.*, the log-mean) corresponds to a confidence interval around the *geometric mean* of the raw (untransformed) data. For the lognormal distribution, the geometric mean is equal to the median, but it is *not* the same as the arithmetic mean. Using this back-transformation to bracket the location of the true arithmetic population mean will result in an incorrect interval.

For these particular applications, a similar problem of *scale bias* occurs with other potential normality transformations. Care is needed when applying and interpreting transformations to a data set

[1] Even when all the non-detects represent the lowest values in the sample, there is still no way to determine how this subset is internally ordered.

for which either a confidence interval around the mean or a prediction limit for a future mean is desired. The interpretation depends on which statistical parameter is being estimated or predicted. The geometric mean or median in some situations may be a satisfactory alternative as a central tendency parameter, although that decision must be weighed carefully when making comparisons against a GWPS.

Common normalizing transformations include the natural logarithm, the square root, the cube root, the square, the cube, and the reciprocal functions, as well as a few others. More generally, one might consider the "ladder of powers" (Helsel and Hirsch, 2002) technically known as the set of Box-Cox transformations (Box and Cox, 1964). The heart of these transformations is a power transformation of the original data, expressed by the equations:

$$y_\lambda = \begin{cases} \left(x^\lambda - 1\right)/\lambda & \text{for } \lambda \neq 0 \\ \log x & \text{for } \lambda = 0 \end{cases} \qquad [10.1]$$

The goal of a Box-Cox analysis is to find the value λ that best transforms the data to approximate normality, using a procedure such as maximum likelihood. Such algorithms are beyond the scope of this guidance, although an excellent discussion can be found in Helsel and Hirsch (2002). In practice, slightly different equation formulations can be used:

$$y_\lambda = \begin{cases} x^\lambda & \text{for } \lambda \neq 0 \\ \log x & \text{for } \lambda = 0 \end{cases} \qquad [10.2]$$

where the parameter λ can generally be limited to the choices 0, -1, 1/4, 1/3, 1/2, 1, 2, 3, and 4, except for unusual cases of more extreme powers.

As noted in **Section 10.1**, checking normality with transformed data does not require any additional tools. Standard normality tests can be applied using the transformed scale measurements. Only the interpretation of the test changes. A goodness-of-fit test can assess the normality of the raw measurements. Under a transformation, the same test checks for normality on the transformed scale. The data will still follow the non-normal distribution in the original concentration domain. So if a cube root transformation is attempted and the transformed data are found to be approximately normal, the original data are not normal but rather cube-root normal in distribution. If a log transformation is successfully used, the original measurements are not normal but lognormal instead. In sum, a series of non-normal distributions can be fitted to data with the goodness-of-fit tests described in this chapter without needing specific tests for other potential distributions.

Finding a reasonable transformation in practice amounts to systematically 'climbing' the "ladder of powers" described above. In other words, different choices of the power parameter λ would be attempted — beginning with λ = 0 and working upward from -1 toward more extreme power transformations — until a specific λ normalizes the data or all choices have been attempted. If no transformation seems to work, the user should instead consider a non-parametric test alternative.

## 10.3 USING THE NORMAL DISTRIBUTION AS A DEFAULT

Normal and lognormal distributions are frequently applied models in groundwater data because of their general utility. One or the other of these models might be chosen as a *default distribution* when designing a statistical approach, particularly when relatively little data has been collected at a site. Since the statistical behavior of these two models is very different and can lead to substantially different conclusions, the choice is not arbitrary. The type of test involved, the monitoring program, and the sample size can all affect the decision. For many data sets and situations, however, the normal distribution can be assumed as a default unless and until a better model can be pinpointed through specific *goodness-of-fit* testing provided in this chapter.

Assumptions of normality are most easily made with regard to naturally-occurring and measurable inorganic parameters, particularly under background conditions. Many ionic and other inorganic water quality analyte measurements exhibit decent symmetry and low variability within a given well data set, making these data amenable to assumptions of normality. Less frequently detected analytes (*e.g.*, certain colloidal trace elements) may be better fit either by a site-wide lognormal or another distribution that can be normalized, as well as evaluated with non-parametric methods.

Where contamination in groundwater is known to exist *a priori* (whether in background or compliance wells), default distributional assumptions become more problematic. At a given well, organic or inorganic contaminants may exhibit high or low variability, depending on local hydrogeologic conditions, the pattern of release from the source, the degree of solid phase absorption, degradability of a given constituent, and the variation in groundwater flow direction and depths. Non-steady state releases may result in a historical, occasionally non-linear pattern of trend increases or decreases. Such data might be fit by an apparent lognormal distribution, although removal of the trend may lead to normally-distributed residuals.

Sample size is also a consideration. With fewer than 8 samples in a data set, formal goodness-of-fit tests are often of limited value. Where larger sample sizes are available, goodness-of-fit tests should be conducted. The Shapiro-Wilk multiple group well test (**Section 10.7**) — even with small sample sizes — can sometimes be used to identify individual anomalous wells which might otherwise be presumed to meet the criterion of normality. Under compliance/assessment or corrective action monitoring, one might anticipate only four samples per well in the first year after instituting such monitoring. Under these conditions, a default assumption of normality for testing of the mean against a fixed standard is probably necessary. Aggregation of multi-year data when conducting compliance tests (see **Chapter 7**) may allow large enough sample sizes to warrant formal goodness-of-fit testing. With 8 (or more) samples, it may be possible to determine that a lognormal distribution is an appropriate fit for the data. Even in this latter approach, caution may be needed in applying Land's confidence interval for a lognormal mean (**Chapter 21**) if the sample variability is large and especially if the upper confidence limit is used in the comparison (*i.e.*, in corrective action monitoring).

The normal distribution may also serve as a reasonable default when it is not critical to ensure that sample data closely follow a specific distribution. For example, statistical tests on the mean are generally considered more *robust* with respect to departures from normality than procedures which involve upper or lower limits of an assumed distribution. Even if the data are not quite normal, tests on the mean such

as a Student's *t*-test will often still provide a valid result. However, one might need to consider transformations of the data for other reasons. Analysis of variance [ANOVA] can be run with small individual well samples (*e.g.*, $n = 4$), and as a comparison of means, it is fairly robust to departures from normality. A logarithmic or other transformation may be needed to stabilize or equalize the well-to-well variability (*i.e.*, achieve *homoscedasticity*), a separate and more critical assumption of the test.

Given their importance in statistical testing and the risks that sometimes occur in trying to interpret tests on other data transformation possibilities, it is useful to briefly consider the logarithmic transformation in more detail. As noted in **Section 10.1**, groundwater data can frequently be normalized using a logarithmic distribution model. Despite this, objections are sometimes raised that the log transformation is merely used to "make large numbers look smaller."

To better understand the log transformation, it should be recognized that logarithms are, in fact, exponents to some unit base. Given a concentration-scale variable $x$, re-expressed as $x = 10^y$ or $x = e^y$, the logarithm $y$ is the exponent of that base (10 or the natural base $e$). It is the behavior of the resultant $y$ *values* that is assessed when data are log-transformed. When data relationships are multiplicative in the original arithmetic domain ($x_1 \times x_2$), the relationships between exponents (*i.e.*, logarithms) are additive ($y_1 + y_2$). Since the logarithmic distribution by mathematical definition is normal in a log-transformed domain, working with the logarithms instead of the original concentration measurements may offer a sample distribution much closer to normal.

Similar to a unit scale transformation (ppm to ppb or Fahrenheit to Centigrade), the relative ordering of log-transformed measurements does not change. When non-parametric tests based on ranks (*e.g.*, the Wilcoxon rank-sum test) are applied to data transformed either to a different unit scale or by logarithms, the outcomes are identical. However, other relationships among the log-transformed data *do* change, so that the log-scale numerical 'spacing' between lower values is more similar to the log-scale spacing between higher values. While parametric tests like prediction limits, *t*-tests, *etc.*, are not affected by unit scale transformations, these tests may have different outcomes depending on whether raw concentrations or log-transformed measurements are used. The justification for utilizing log-transformed data is that the transformation helps to normalize the data so that these tests can be properly applied.

There is also a plausible physical explanation as to why pollutant concentrations often follow a logarithmic pattern (Ott, 1990). In Ott's model, pollutant sources are randomly dispersed through the subsurface or atmosphere in a multiplicative fashion through repeated dilutions when mixing with volumes of (uncontaminated) water or air, depending on the medium. Such random and repeated dilutions can mathematically lead to a lognormal distribution. In particular, if a final concentration ($c_0$) is the product of several random dilutions ($c_i$) as suggested by the following equation:

$$c_0 = \prod_{i=1}^{n} c_i = (c_1 \times c_2 \times \ldots \times c_n) \qquad [10.3]$$

the logarithm of this concentration is equivalent to the *sum* of the logarithms of the individual dilutions:

$$\log(c_0) = \sum_{i=1}^{n} \log(c_i) \qquad [10.4]$$

The Central Limit Theorem (**Chapter 3**) can be applied to conclude that the logged concentration in equation [10.4] should be approximately normal, implying that the original concentration ($c_0$) should be approximately lognormal in distribution. Contaminant fate-and-transport models more or less follow this same approach, using successive multiplicative dilutions (while accounting for absorption and degradation effects) across grids in time and space.

Despite the mathematical elegance of the Ott model, experience with groundwater monitoring data has shown that the lognormal model alone is not adequate to account for observed distribution patterns. While contaminant modeling might predict a lognormal contaminant distribution in space (and often in time at a fixed point during transient phases), individual well location points fixed in space and at rough contaminant equilibrium are more likely to be subject to a variety of local hydrologic and other factors, and the observed distributions can be almost limitless in form. Since most of the tests within the Unified Guidance presume a stationary population over time at a given well location (subject to identification and removal of trends), the resultant distributions may be other than lognormal in character. Individual constituents may also exhibit varying aquifer-related distributional characteristics.

A practical issue in selecting a default transformation is ease of use. Distributions like the lognormal usually entail more complicated statistical adjustments or calculations than the normal distribution. A confidence interval around the arithmetic mean of a lognormal distribution utilizes Land's *H*-factor, which is a function of both log sample data variability and sample size, and is only readily available for specific confidence levels. By contrast, a normal confidence interval around the sample mean based on the *t*-statistic can easily be defined for virtually any confidence level. As noted earlier, correct use of these confidence intervals depends on selecting the appropriate parameter and statistical measure (arithmetic mean versus the geometric mean).

While a transformation does not always necessitate using a different statistical formula to ensure unbiased results, use of a transformation *does* assume that the underlying population is non-normal. Since the true population will almost never be known with certainty, it may not be advantageous to simply default to a lognormal assumption for a variety of reasons. Under detection monitoring, the presumption is made that a statistically significant increase above background concentrations will trigger a monitoring exceedance. But the larger the prediction limit computed from background, the less *statistical power* the test will have for detecting true increases. An important question to answer is what the consequences are when incorrectly applying statistical techniques based on one distributional assumption (normal or lognormal), when the underlying distribution is in fact the other. More specifically, what is the impact on statistical power and accuracy of assuming the wrong underlying distribution? The general effects of violating underlying test assumptions can be measured in terms of false positive and negative error rates (and therefore power). These questions are particularly pertinent for prediction limit and control chart tests in detection monitoring. Similar questions could be raised regarding the application of confidence interval tests on the mean when compared against fixed standards.

To answer these questions, a series of Monte Carlo simulations was generated for the Unified Guidance to evaluate the impacts on prediction limit false positive error rates and statistical power of using normal and lognormal distributions (correctly and incorrectly applied to the underlying distributions). Detailed results of this study are provided in **Appendix C, Section C.1**.

The conclusions of the Monte Carlo study are summarized as follows:

- If an underlying population is truly normal, *treating the sample data as lognormal* in constructing a prediction limit can have significant consequences. With no retesting, the lognormal prediction limits were in every case considerably larger and thus less powerful than the normal prediction limits. Further, the lognormal limits consistently exhibited less than the expected (nominal) false positive rate, while the normal prediction limits tended to have slightly higher than nominal error rates.
- When retesting was added to the procedure, both types of prediction limits improved. While power uniformly improved compared to no retest, the normal limits were still on average about 13% shorter than the lognormal limits, leading again to a measurable loss of statistical power in the lognormal case.
- On balance, *misapplication* of logarithmic prediction limits to normally-distributed data consistently resulted in (often considerably) lower power and false positive rates that were lower than expected. The results argue *against* presuming the underlying data to be lognormal without specific goodness-of-fit testing.
- The highest penalties from misapplying lognormal prediction limits occurred for smaller background sizes. Since goodness-of-fit tests are least able to distinguish between normal and lognormal data with small samples, small background samples should not be presumed to be lognormal *as a default* unless other evidence from the site suggests otherwise. For larger samples, goodness-of-fit tests have much better discriminatory power, enabling a better indication of which model to use.
- If the underlying population is truly lognormal but the sample data *are treated as normal*, the penalty in overall statistical performance is substantial *only* if no retesting is conducted. With no retesting, the false positive rates of normal-based limits were often substantially higher than the expected rate. Under conditions of no retesting, *misapplying* normal prediction limits to lognormal data would result in an excessive site-wide false positive rate (SWFPR).
- If at least one retest was added, the achieved false positive rates for the misapplied normal limits tended to be *less* than the expected rates, especially for moderate to larger sample sizes. Except for highly skewed lognormal distributions, the power of the normal limits was comparable or greater than the power of the lognormal limits.

Overall, the Monte Carlo study indicated that adding a retest to the testing procedure significantly minimized the penalty of misapplying normal prediction limits to lognormal data, as long as the sample size was at least 8 and the distribution was not too skewed. Consequently, there is *less* penalty associated with making a default assumption of *normality* than in making a default assumption of *lognormality* under most situations. With highly skewed data, goodness-of-fit tests tend to better discriminate between the normal and lognormal models. The Unified Guidance therefore recommends that such diagnostic testing be done *explicitly* rather than simply assuming the data to be normal or lognormal.

The most problematic cases in the study occurred for very small background sample sizes, where a misapplication of prediction limits in either direction often resulted in poorer statistical performance, even with retesting. In some situations, compliance testing may need to be conducted on an interim basis until enough data has been collected to accurately identify a distributional model. The Unified Guidance does not recommend an automatic default assumption of lognormality.

In summary, during detection and compliance/assessment monitoring, data sets should be treated initially as normal in distribution unless a better model can be pinpointed through specific testing. The normal distribution is a fairly safe assumption for background distributions, particularly for naturally occurring, measurable constituents and when sample sizes are small. Goodness-of-fit tests provided in this chapter can be used to more closely identify the appropriate distributions for larger sample sizes. If the initial assumption of normality is not rejected, further statistical analyses should be performed on the raw observations. If the normal distribution *is* rejected by a goodness-of-fit test, one should generally test the normality of the logged data, in order to check for lognormality of the original observations. If this test also fails, one can either look for an alternate transformation to achieve approximate normality (**Section 10.2**) or use a non-parametric technique.

Since tests of normality have low power for rejecting the null hypothesis when the data are really lognormal but the sample size and degree of skewness are small, it is reassuring that a "wrong" default assumption of normality will infrequently lead to an incorrect statistical conclusion. In fact, the statistical power for detecting real concentration increases will generally be better than if the data were assumed to be lognormal. If the data *are* truly lognormal, there *is* a risk of greater-than-expected site-wide false positive error rates.

When the population is more skewed, normality tests in the Unified Guidance have much greater power for correctly rejecting the normal model in favor of the lognormal distribution. Consequently, an initial assumption of normality will not, in most cases, lead to an incorrect final conclusion, since the presumed normal model will tend to be rejected before further testing is conducted.

These recommendations do not apply to corrective action monitoring or other programs where it either known or reasonable to presume that groundwater is already impacted or has a non-normal distribution. In such settings, a default presumption of lognormality could be made, or a series of normalizing transformations could be attempted until a suitable fit is determined. Furthermore, even in detection monitoring, there are situations that often require the use of alternate transformations, for instance when pooling intrawell background across several wells to increase the degrees of freedom available for intrawell testing (**Chapter 13**).

Whatever the circumstance, the Unified Guidance recommends whenever possible that site-specific data be used to test the distributional presumption. If no data are initially available to do this, "referencing" may be employed to justify the use of, say, a normal or lognormal assumption in developing statistical tests at a particular site. Referencing involves the use of historical data or data from sites in similar hydrologic settings to justify the assumptions applied to the proposed statistical regimen. These initial assumptions should be checked when data from the site become available, using the procedures described in the Unified Guidance. Subsequent changes to the initial assumptions should be made if goodness-of-fit testing contradicts the initial hypothesis.

## 10.4 COEFFICIENT OF VARIATION AND COEFFICIENT OF SKEWNESS

### PURPOSE AND BACKGROUND

Because the normal distribution has a symmetric 'bell-shape,' the normal mean and median coincide and random observations drawn from a normal population are just as likely to occur below the mean as above it. More generally, in any symmetric distribution the distributional pattern below the

mean is a mirror-image of the pattern above the mean. By definition, such distributions have no degree of *skewness* or asymmetry.

Since the normal distribution has zero skewness, one way to look for non-normality is to estimate the degree of skewness. Non-zero values of this measure imply that the population is asymmetric and therefore something different from normal. Two exploratory screening tools useful for this task are the *coefficient of variation* and the *coefficient of skewness*.

The coefficient of variation [CV] is extremely easy to compute, but only indirectly offers an estimate of skewness and hence normality/non-normality. A more direct estimate can be determined via the coefficient of skewness. Furthermore, better, formal tests can be used instead of either coefficient to directly assess normality. Nevertheless, the CV provides a measure of intrinsic variability in positive-valued data sets. Although approximate, CVs can indicate the relative variability of certain data, especially with small sample sizes and in the absence of other formal tests (*e.g.*, see **Chapter 22**, when comparing confidence limits on the mean to a fixed standard in compliance monitoring).

The CV is also a valid measure of the multiplicative relationship between the population mean and the standard deviation for positively-valued random variables. Using sample statistics for the mean ($\bar{x}$) and standard deviation ($s$), the true CV for non-negative normal populations can be reasonably estimated as:

$$CV = s / \bar{x} \qquad [10.5]$$

In lognormal populations, the CV is also used in evaluations of statistical power. In this latter case, the population CV works out to be:

$$CV = \sqrt{\exp\left(\sigma_y^2\right) - 1} \qquad [10.6]$$

where $\sigma_y$ is the population log-standard deviation. Instead of a ratio between the original scale standard deviation and the mean, the lognormal CV is estimated with the equation:

$$CV = \sqrt{\exp\left(s_y^2\right) - 1} \qquad [10.7]$$

where $s_y$ is the sample log-standard deviation. The estimate in equation [10.7] is usually more accurate than the simple CV ratio of the arithmetic standard deviation-to-mean, especially when the underlying population coefficient of variation is high. Similar to using the normal CV as a formal indicator of normality, the lognormal coefficient of variation estimator in equation [10.7] will have little relevance ***as a test of lognormality*** of the data. Using it for that purpose is not recommended in the Unified Guidance. But it can provide a sense of how variable a data set is and whether a lognormal assumption might need to be tested.

While others have reported a ratio CV on logged measurements as $CV = s_y / \bar{y}$ for the transformation $y = \log x$, the result is essentially meaningless. The actual logarithmic CV in equations [10.6] and [10.7] is solely determined by the logarithmic variability of $\sigma_y$ or $s_y$. Negative logarithmic mean values are always possible, and the log ratio statistic is not invariant under a unit scale

transformation (*e.g.*, ppb to ppm or ppt). Similar problems in interpretation occur when CV estimators are applied to any variable which can be negatively valued, such as following a *z*-transformation to a standard normal distribution. This log ratio statistic is not recommended for any application in the guidance.

The coefficient of skewness ($\gamma_1$) directly indicates to what degree a dataset is skewed or asymmetric with respect to the mean. Sample data from a normal distribution will have a skewness coefficient near zero, while data from an asymmetric distribution will have a positive or negative skewness depending on whether the right- or left-hand tail of the distribution is longer and skinnier than the opposite tail.

Since groundwater monitoring concentrations are inherently non-negative, such data often exhibit skewness. A small degree of skewness is not likely to affect the results of statistical tests that assume normality. However, if the skewness coefficient is larger than 1 (in absolute value) and the sample size is small (*e.g.*, $n < 25$), past research has shown that standard normal theory-based tests are much less powerful than when the absolute skewness is less than 1 (Gayen, 1949).

Calculating the skewness coefficient is useful and only slightly more difficult than computing the CV. It provides a quick indication of whether the skewness is minimal enough to assume that the data are roughly symmetric and hopefully normal in distribution. If the original data exhibit a high skewness coefficient, the normal distribution will provide a poor approximation to the dataset. In that case — and unlike the CV — $\gamma_1$ can be computed on the log-transformed data to test for symmetry of the logged measurements, or similarly for other transformations.

## PROCEDURE

The CV is calculated simply by taking the ratio of the sample standard deviation to the sample mean, $CV = s/\bar{x}$ or its corresponding logarithmic version $CV = \sqrt{\exp\left(s_y^2\right) - 1}$.

The skewness coefficient may be computed using the following equation:

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3 \Bigg/ \left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]^{3/2} = n^{1/2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3 \Big/ \left(n-1\right)^{3/2} s^3 \qquad [10.8]$$

where the numerator represents the average cubed residual after subtracting the sample mean.

## ►EXAMPLE 10-1

Using the following data, compute the CVs and the coefficient of skewness to test for approximate symmetry.

| | Nickel Concentration (ppb) | | | |
|---|---|---|---|---|
| Month | Year 1 | Year 2 | Year 3 | Year 4 |
| Jan | 58.8 | 19 | 39 | 3.1 |
| Mar | 1.0 | 81.5 | 151 | 942 |
| Jun | 262 | 331 | 27 | 85.6 |
| Aug | 56 | 14 | 21.4 | 10 |
| Oct | 8.7 | 64.4 | 578 | 637 |

SOLUTION

Step 1. Compute the mean, standard deviation ($s$), and sum of the cubed residuals for the nickel concentrations:

$$\bar{x} = \frac{1}{20}(58.8 + 1 + \ldots + 637) = 169.52 \ ppb$$

$$s = \sqrt{\frac{1}{19}\left[(58.8 - 169.52)^2 + (1 - 169.52)^2 + \ldots + (637 - 169.52)^2\right]} = 259.7175 \ ppb$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^3 = \left[(58.8 - 169.52)^3 + \ldots + (637 - 169.52)^3\right] = 5.97845791 \times 10^8 \ ppb^3$$

Step 2. Compute the arithmetic normal coefficient of variation following equation [10.5]: $CV = 259.7175/169.52 = 1.53$

Step 3. Calculate the coefficient of skewness using equation [10.8]:

$$\gamma_1 = (20)^{1/2}\left(5.97845791 \times 10^8\right) / (19)^{3/2}(259.7175)^3 = 1.84$$

Both the CV and the coefficient of skewness are much larger than 1, so the data appear to be significantly positively skewed. Do not assume that the underlying population is normal.

Step 4. Since the original data evidence a high degree of skewness, one can instead compute the skewness coefficient and corresponding sample CV with equation [10.7] on the logged nickel concentrations. The logarithmic CV equals 4.97, a much more variable data set than suggested by the arithmetic CV. The skewness coefficient works out to be $|\gamma_1| = 0.24 < 1$, indicating that the logged data values are slightly skewed but not enough to clearly reject an assumption of normality in the logged data. In other words, the original nickel values may be lognormally distributed. ◄

# 10.5 SHAPIRO-WILK AND SHAPIRO-FRANCÍA NORMALITY TESTS

## 10.5.1 SHAPIRO-WILK TEST (N ≤ 50)

### PURPOSE AND BACKGROUND

The Shapiro-Wilk test is based on the premise that if a data set is normally distributed, the ordered values should be highly correlated with corresponding *quantiles* (*z*-scores) taken from a normal distribution (Shapiro and Wilk, 1965). In particular, the Shapiro-Wilk test gives substantial weight to evidence of non-normality in the tails of a distribution, where the robustness of statistical tests based on the normality assumption is most severely affected. A variant of this test, the Shapiro-Francía test, is useful for sample sizes greater than 50 (see **Section 10.5.2**).

The Shapiro-Wilk test statistic (*SW*) will tend to be large when a probability plot of the data indicates a nearly straight line. Only when the plotted data show significant bends or curves will the test statistic be small. The Shapiro-Wilk test is considered one of the best tests of normality available (Miller, 1986; Madansky, 1988).

### PROCEDURE

Step 1. Order and rank the dataset from least to greatest, labeling the observations as $x_i$ for rank $i$ = 1…$n$. Using the notation $x_{(i)}$, let the $i$th rank statistic from a data set represent the $i$th smallest value.

Step 2. Compute differences $\left[x_{(n-i+1)} - x_{(i)}\right]$ for each $i$ = 1…$n$. Then determine $k$ as the greatest integer less than or equal to ($n$/2).

Step 3. Use **Table 10-2** in **Appendix D** to determine the Shapiro-Wilk coefficients, $a_{n-i+1}$ , for $i$ = 1…$k$. Note that while these coefficients depend only on the sample size ($n$), the order of the coefficients must be preserved when used in Step 4. The coefficients can be determined for any sample size from $n$ = 3 up to $n$ = 50.

Step 4. Compute the quantity $b$ given by the following equation:

$$b = \sum_{i=1}^{k} b_i = \sum_{i=1}^{k} a_{n-i+1}(x_{(n-i+1)} - x_{(i)}) \qquad [10.9]$$

Note that the values $b_i$ are simply intermediate quantities represented by the terms in the sum of the right-hand expression in equation [10.9].

Step 5. Calculate the standard deviation ($s$) of the dataset. Then compute the Shapiro-Wilk test statistic using the equation:

$$SW = \left[\frac{b}{s\sqrt{n-1}}\right]^2 \qquad [10.10]$$

Step 6. Given the significance level ($\alpha$) of the test, determine the critical point of the Shapiro-Wilk test with $n$ observations using **Table 10-3** in **Appendix D**. To maximize the utility and power of the test, choose $\alpha = .10$ for very small data sets ($n < 10$), $\alpha = .05$ for moderately sized data sets ($10 \le n < 20$), and $\alpha = .01$ for larger sized data sets ($n \ge 20$). Compare the $SW$ against the critical point ($sw_c$). If the test statistic exceeds the critical point, accept normality as a reasonable model for the underlying population. However, if $SW < sw_c$, reject the null hypothesis of normality at the $\alpha$-level and decide that another distributional model might provide a better fit.

►EXAMPLE 10-2

Use the nickel data of **Example 10-1** to compute the Shapiro-Wilk test of normality.

SOLUTION

Step 1. Order the data from smallest to largest, rank in ascending order and list, as shown in columns 1 and 2 of the table below. Next list the data in reverse order in a third column.

| i | $x_{(i)}$ | $x_{(n-i+1)}$ | $x_{(n-i+1)} - x_{(i)}$ | $a_{n-i+1}$ | $b_i$ |
|---|---|---|---|---|---|
| 1 | 1.0 | 942.0 | 941.0 | .4734 | 445.47 |
| 2 | 3.1 | 637.0 | 633.9 | .3211 | 203.55 |
| 3 | 8.7 | 578.0 | 569.3 | .2565 | 146.03 |
| 4 | 10.0 | 331.0 | 321.0 | .2085 | 66.93 |
| 5 | 14.0 | 262.0 | 248.0 | .1686 | 41.81 |
| 6 | 19.0 | 151.0 | 132.0 | .1334 | 17.61 |
| 7 | 21.4 | 85.6 | 64.2 | .1013 | 6.50 |
| 8 | 27.0 | 81.5 | 54.5 | .0711 | 3.87 |
| 9 | 39.0 | 64.4 | 25.4 | .0422 | 1.07 |
| 10 | 56.0 | 58.8 | 2.8 | .0140 | 0.04 |
| 11 | 58.8 | 56.0 | −2.8 | | b = 932.88 |
| 12 | 64.4 | 39.0 | −25.4 | | |
| 13 | 81.5 | 27.0 | −54.5 | | |
| 14 | 85.6 | 21.4 | −64.2 | | |
| 15 | 151.0 | 19.0 | −132.0 | | |
| 16 | 262.0 | 14.0 | −248.0 | | |
| 17 | 331.0 | 10.0 | −321.0 | | |
| 18 | 578.0 | 8.7 | −569.3 | | |
| 19 | 637.0 | 3.1 | −633.9 | | |
| 20 | 942.0 | 1.0 | −941.0 | | |

Step 2. Compute the differences $\left[x_{(n-i+1)} - x_{(i)}\right]$ in column 4 of the table by subtracting column 2 from column 3. Since the total sample size is $n = 20$, the largest integer less than or equal to ($n$/2) is $k = 10$.

Step 3. Look up the coefficients $a_{n-i+1}$ from **Table 10-2** in **Appendix D** and list in column 4.

Step 4. Multiply the differences in column 3 by the coefficients in column 4 and add the first $k$ products ($b_i$) to get quantity $b$, using equation [10.9].

$$b = [.4734(941.0) + .3211(633.9) + \ldots + .0140(2.8)] = 932.88$$

Step 5. Compute the standard deviation of the sample, $s$ = 259.72. Then use equation [10.10] to calculate the $SW$:

$$SW = \left[\frac{932.88}{259.72\sqrt{19}}\right]^2 = 0.679$$

Step 6. Use **Table 10-3** in **Appendix D** to determine the 0.01-level critical point for the Shapiro-Wilk test when $n$ = 20. This gives $sw_c$ = 0.868. Then compare the observed value of $SW$ = 0.679 to the 1% critical point. Since $SW$ < 0.868, the sample shows significant evidence of non-normality by the Shapiro-Wilk test. The data should be transformed using logarithms or another transformation on the ladder of powers and re-checked using the Shapiro-Wilk test before proceeding with further statistical analysis. ◀

### 10.5.2 SHAPIRO-FRANCÍA TEST (N > 50)

The Shapiro-Wilk test of normality can be used for sample sizes up to 50. When $n$ is larger than 50, a slight modification of the procedure called the Shapiro-Francía test (Shapiro and Francia, 1972) can be used instead. Like the Shapiro-Wilk test, the Shapiro-Francía test statistic ($SF$) will tend to be large when a probability plot of the data indicates a nearly straight line. Only when the plotted data show significant bends or curves will the test statistic be small.

To calculate the test statistic $SF$, one can use the following equation:

$$SF = \left[\sum_{i=1}^{n} m_i x_{(i)}\right]^2 \Bigg/ \left[(n-1)s^2 \sum_{i=1}^{n} m_i^2\right] \qquad [10.11]$$

where $x_{(i)}$ represents the $i$th ranked value of the sample and where $m_i$ denotes the approximate expected value of the $i$th rank normal quantile (or $z$-score). The values for $m_i$ are approximately equal to

$$m_i = \Phi^{-1}\left(\frac{i}{n+1}\right) \qquad [10.12]$$

where $\Phi^{-1}$ denotes the inverse of the standard normal distribution with zero mean and unit variance. These values can be computed by hand using the normal distribution in **Table 10-1** of **Appendix D** or via simple commands found in many statistical computer packages.

Normality of the data should be rejected if the Shapiro-Francía statistic is too low when compared to the critical points provided in **Table 10-4** of **Appendix D**. Otherwise one can assume the data are approximately normal for purposes of further statistical analysis.

# 10.6 PROBABILITY PLOT CORRELATION COEFFICIENT

## BACKGROUND AND PURPOSE

Another test for normality that is essentially equivalent to the Shapiro-Wilk and Shapiro-Francía tests is the *probability plot correlation coefficient* test described by Filliben (1975). This test meshes perfectly with the use of probability plots, because the essence of the test is to compute the usual *correlation coefficient* for points on a probability plot. Since the correlation coefficient is a measure of the linearity of the points on a scatterplot, the probability plot correlation coefficient, like the *SW* test statistic, will be high when the plotted points fall along a straight line and low when there are significant bends and curves in the probability plot. Comparison of the Shapiro-Wilk and probability plot correlation coefficient tests has indicated very similar statistical power for detecting non-normality (Ryan and Joiner, 1990).

It should be noted that although some statistical software may not compute Filliben's test directly, the usual Pearson's correlation coefficient computed on the data pairs used to construct a probability plot will provide a very close approximation to the Filliben statistic. Some users may find this latter correlation easier to compute or more accessible in their software.

## PROCEDURE

Step 1. List the observations in order from smallest to largest, denoting $x_{(i)}$ as the $i$th smallest rank statistic in the data set. Then let $n$ = sample size and compute the sample mean ($\bar{x}$) and the standard deviation ($s$).

Step 2. Consider a random sample drawn from a standard normal distribution. The $i$th rank statistic of this sample is fixed once the sample is drawn, but beforehand it can be considered a random variable, denoted as $X_{(i)}$. Likewise, by considering all possible datasets of size $n$ that might be drawn from the normal distribution, one can think of the sampling distribution of the statistic $X_{(i)}$. This sampling distribution has its own mean and variance, and, of importance to the probability plot correlation coefficient, its own *median*, which can be denoted $M_i$.

To compute the median of the $i$th rank statistic, first compute intermediate probabilities $m_i$ for $i = 1 \ldots n$ using the equation:

$$m_i = \begin{cases} 1-(.5)^{1/n} & \text{for } i=1 \\ (i-.3175)/(n+.365) & \text{for } 1<i<n \\ (.5)^{1/n} & \text{for } i=n \end{cases} \qquad [10.13]$$

Then compute the medians $M_i$ as the standard normal quantiles or $z$-scores associated with the intermediate probabilities $m_i$. These can be determined from **Table 10-1** in **Appendix D** or computed according to the following equation, where $\Phi^{-1}$ represents the inverse of the standard normal distribution:

$$M_i = \Phi^{-1}(m_i) \qquad [10.14]$$

Step 3. With the rank statistic medians in hand, calculate the arithmetic mean of the $M_i$'s, denoted $\bar{M}$, and the intermediate quantity $C_n$, given by the equation:

$$C_n = \sqrt{\sum_{i=1}^{n} M_i^2 - n\bar{M}^2} \qquad [10.15]$$

Note that when the dataset is "complete" (meaning it contains no non-detects, ties, or censored values), the mean of the order statistic medians reduces to $\bar{M} = 0$. This in turn reduces the calculation of $C_n$ to:

$$C_n = \sqrt{\sum_{i=1}^{n} M_i^2} \qquad [10.16]$$

Step 4. Finally compute Filliben's probability plot correlation coefficient:

$$r = \frac{\sum_{i=1}^{n} x_{(i)} M_i - n\bar{x}\bar{M}}{C_n \cdot s\sqrt{n-1}} \qquad [10.17]$$

When the dataset is complete, the equation for the probability plot correlation coefficient also has a simplified form:

$$r = \sum_{i=1}^{n} x_{(i)} M_i \Big/ \left[ C_n \cdot s\sqrt{n-1} \right] \qquad [10.18]$$

Step 5. Given the level of significance ($\alpha$), determine the critical point ($r_{cp}$) for Filliben's test with sample size $n$ from **Table 10-5** in **Appendix D**. Compare the probability plot correlation coefficient ($r$) against the critical point ($r_{cp}$). If $r \geq r_{cp}$, conclude that normality is a reasonable model for the underlying population at the $\alpha$-level of significance. If, however, $r < r_{cp}$, reject the null hypothesis and conclude that another distributional model would provide a better fit.

►EXAMPLE 10-3

Use the data of **Example 10-1** to compute Filliben's probability plot correlation coefficient test at the $\alpha$ = .01 level of significance.

SOLUTION

Step 1. Order and rank the nickel data from smallest to largest and list, as in the table below. The sample size is $n$ = 20, with sample mean $\bar{x}$ = 169.52 and the standard deviation $s$ = 259.72.

Step 2. Compute the intermediate probabilities $m_i$ from equation [10.13] for each $i$ in column 3 and the rank statistic medians, $M_i$, in column 4 by applying the inverse normal transformation to column 3 using equation [10.14] and **Table 10-1** of **Appendix D**.

Step 3. Since this sample contains no non-detects or ties, the simplified equations for $C_n$ in equation [10.16] and for $r$ in equation [10.18] may be used. First compute $C_n$ using the squared order statistic medians in column 5:

$$C_n = \sqrt{[3.328 + 1.926 + \ldots + 3.328]} \quad = \quad 4.138$$

Step 4. Next compute the products $x_{(i)} \times M_i$ in column 6 and sum to get the numerator of the correlation coefficient (equal to 3,836.81 in this case). Then compute the final correlation coefficient:

$$r = 3{,}836.81 \Big/ \left[ 4.138 \times 259.72\sqrt{19} \right] = 0.819$$

| i | $x_{(i)}$ | $m_i$ | $M_i$ | $(M_i)^2$ | $x_{(i)} \times M_i$ |
|---|---|---|---|---|---|
| 1 | 1.0 | .03406 | −1.8242 | 3.328 | −1.824 |
| 2 | 3.1 | .08262 | −1.3877 | 1.926 | −4.302 |
| 3 | 8.7 | .13172 | −1.1183 | 1.251 | −9.729 |
| 4 | 10.0 | .18082 | −0.9122 | 0.832 | −9.122 |
| 5 | 14.0 | .22993 | −0.7391 | 0.546 | −10.347 |
| 6 | 19.0 | .27903 | −0.5857 | 0.343 | −11.129 |
| 7 | 21.4 | .32814 | −0.4451 | 0.198 | −9.524 |
| 8 | 27.0 | .37724 | −0.3127 | 0.098 | −8.444 |
| 9 | 39.0 | .42634 | −0.1857 | 0.034 | −7.242 |
| 10 | 56.0 | .47545 | −0.0616 | 0.004 | −3.448 |
| 11 | 58.8 | .52455 | 0.0616 | 0.004 | 3.621 |
| 12 | 64.4 | .57366 | 0.1857 | 0.034 | 11.959 |
| 13 | 81.5 | .62276 | 0.3127 | 0.098 | 25.488 |
| 14 | 85.6 | .67186 | 0.4451 | 0.198 | 38.097 |
| 15 | 151.0 | .72097 | 0.5857 | 0.343 | 88.445 |
| 16 | 262.0 | .77007 | 0.7391 | 0.546 | 193.638 |
| 17 | 331.0 | .81918 | 0.9122 | 0.832 | 301.953 |
| 18 | 578.0 | .86828 | 1.1183 | 1.251 | 646.376 |
| 19 | 637.0 | .91738 | 1.3877 | 1.926 | 883.941 |
| 20 | 942.0 | .96594 | 1.8242 | 3.328 | 1718.408 |

Step 5. Compare Filliben's test statistic of $r = 0.819$ to the 1% critical point for a sample of size 20 in **Table 10-5** of **Appendix D**, namely $r_{cp} = 925$. Since $r < 0.925$, the sample shows significant evidence of non-normality by the probability plot correlation coefficient. The data should be transformed and the correlation coefficient re-calculated before proceeding with further statistical analysis. ◀

## 10.7 SHAPIRO-WILK MULTIPLE GROUP TEST OF NORMALITY

### BACKGROUND AND PURPOSE

The main purpose for including the multiple group test normality (Wilk and Shapiro, 1968) in the Unified Guidance is to serve as a check for normality when using a Student's *t*-test (**Chapter 16**) or

when assessing the joint normality of multiple intrawell data sets. The multiple group test is an extension of the Shapiro-Wilk procedure for assessing the joint normality of several independent samples. Each sample may have a different mean and/or variance, but as long as the underlying distribution of each group is normal, the multiple group test statistic will tend to be non-significant. Conversely, the multiple group test is designed to identify when at least one of the groups being tested is definitely non-normal.

This test extends the Shapiro-Wilk procedure for a single sample, using individual *SW* test statistics computed separately for each group or sample. Then the individual *SW* statistics are transformed and combined into an overall or "omnibus" statistic ($G$). Like the single sample procedure — where non-normality is indicated when the test statistic *SW* is too low — non-normality in one or more groups is indicated when $G$ is too low. However, instead of a special table of critical points, $G$ is constructed to follow a standard normal distribution under the null hypothesis of normality. The value of $G$ can simply be compared to an $\alpha$-level $z$-score or normal quantile to decide whether the null or alternative hypothesis is better supported.

Since it may be unclear which one or more of the groups is actually non-normal when the $G$ statistic is significant, Wilk and Shapiro recommend that a probability plot (**Chapter 9**) be examined on the intermediate quantities, $G_i$ (at least for the case where several groups are being simultaneously tested). One of these statistics is computed for each separate sample/group and is designed to follow a standard normal distribution under $H_0$. Because of this, the $G_i$ statistics for non-normal groups will tend to look like outliers on a normal probability plot (see **Chapter 12**).

The multiple group test can also be used to check normality when performing Welch's *t*-test, a two-sample procedure in which the underlying data of both groups are assumed to be normal, but no assumption is made that the means or variances are the same. This is different from either the pooled variance *t*-test or the one-way analysis of variance [ANOVA], both of which assume *homoscedasticity* (*i.e.*, equal variances across groups). If the group variances can be shown to be equal, the single sample Shapiro-Wilk test can be run on the combined residuals, where the residuals of each group are formed by subtracting off the group mean from each of the individual measurements. However, if the group variances are possibly different, testing the residuals as a single group using the *SW* statistic may give an inaccurate or misleading result. Consequently, since a test of homoscedasticity is not required for Welch's *t*-test, it is suggested to first use the multiple group test to check normality.

Although the Shapiro-Wilk multiple group method is an attractive procedure for accommodating several groups of data at once, the user is cautioned against indiscriminate use. While many of the methods described in the Unified Guidance assume underlying normality, they also assume homoscedasticity. Other parametric multi-sample methods recommended for detection monitoring — prediction limits in **Chapter 18** and control charts in **Chapter 20** — all assume that each group has the same variance. Even if normality of the joint data can be demonstrated using the Shapiro-Wilk multiple group test, it says nothing about whether the assumption of equal variances is also satisfied. Generally speaking, except for Welch's *t*-test, a separate test of homoscedasticity may also be needed. Such tests are described in **Chapter 11**.

## PROCEDURE

Step 1. Assuming there are $K$ groups to be tested, let the sample size of the $i$th group be denoted $n_i$. Then compute the $SW_i$ test statistic for each of the $K$ groups using equation [10.10].

Step 2. Transform the $SW_i$ statistics to the intermediate quantities ($G_i$). If the sample size ($n_i$) of the $i$th group is at least 7, compute $G_i$ with the equation:

$$G_i = \gamma + \delta \ln\left(\frac{SW_i - \varepsilon}{1 - SW_i}\right) \qquad [10.19]$$

where the quantities γ, δ, and ε can be found in **Table 10-6** of **Appendix D** for $7 \le n_i \le 50$. If the sample size ($n_i$) is less than 7, determine $G_i$ directly from **Table 10-7** in **Appendix D** by first computing the intermediate value

$$u_i = \ln\left(\frac{SW_i - \varepsilon}{1 - SW_i}\right) \qquad [10.20]$$

(obtaining ε from the top of **Table 10-7**), and then using linear interpolation to find the closest value $G_i$ associated with $u_i$.

Step 3. Once the $G_i$ statistics are derived, compute the Shapiro-Wilk multiple group statistic with the equation:

$$G = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} G_i \qquad [10.21]$$

Step 4. Under the null hypothesis that all $K$ groups are normally-distributed, $G$ will follow a standard normal distribution. Given the significance level (α), determine an α-level critical point from **Table 10-1** of **Appendix D** as the *lower* $\alpha \times 100th$ normal quantile ($z_\alpha$). Then compare $G$ to $z_\alpha$. If $G < z_\alpha$, there is significant evidence of non-normality at the α level. Otherwise, the hypothesis of normality cannot be rejected.

►EXAMPLE 10-4

The previous examples in this chapter pooled the data of **Example 10-1** into a single group before testing for normality. This time, treat each well separately and compute the Shapiro-Wilk multiple group test of normality at the α = .05 level.

SOLUTION

Step 1. The nickel data in **Example 10-1** come from $K$ = 4 wells with $n_i$ = 5 observations per well. Using equation [10.10], the $SW_i$ individual well test statistics are calculated as:

Well 1: $SW_1 = 0.6062$

Well 2: $SW_2 = 0.5917$

Well 3: $SW_3 = 0.5652$

Well 4: $SW_4 = 0.6519$

Step 2. Since $n_i$ = 5 for each well, use **Table 10-7** of **Appendix D** to find ε = .5521. First calculating $u_1$ with equation [10.20]:

$$u_1 = \ln\left(\frac{.6062 - .5521}{1 - .6062}\right) = -1.985$$

The performing this step for each well group and using linear interpolation on $u$ in **Table 10-7**, the approximate $G_i$ statistics are:

Well 1: $u_1 = -1.985$ $G_1 = -3.238$

Well 2: $u_2 = -2.333$ $G_2 = -3.488$

Well 3: $u_3 = -3.502$ $G_3 = -4.013$ (taking the last and closest entry)

Well 4: $u_4 = -1.249$ $G_4 = -2.755$

Step 3. Compute the multiple group test statistic using equation [10.21]:

$$G = \frac{1}{\sqrt{4}}\left[(-1.985) + (-2.333) + (-4.013) + (-2.755)\right] = -5.543$$

Step 4. Since α = 0.05, the lower α × 100*th* critical point from the standard normal distribution in **Table 10-1** of **Appendix D** is $z_{.05} = -1.645$. Clearly, $G < z_{.05}$ ; in fact $G$ is smaller than just about any α-level critical point that might be selected. Thus, there is significant evidence of non-normality in at least one of these wells (and probably all of them). ◄

►EXAMPLE 10-5

The data in **Example 10-1** showed significant evidence of non-normality. In this example, use the same nickel data applying the coefficient of skewness, Shapiro-Wilk and the Probability Plot Correlation Coefficient tests to determine whether the combined well measurements better follow a lognormal distribution by first log-transforming the measurements. Computing the natural logarithms of the data gives the table below:

| | **Logged Nickel Concentrations log(ppb)** | | | |
|---|---|---|---|---|
| Month | Well 1 | Well 2 | Well 3 | Well 4 |
| 1 | 4.07 | 2.94 | 3.66 | 1.13 |
| 2 | 0.00 | 4.40 | 5.02 | 6.85 |
| 3 | 5.57 | 5.80 | 3.30 | 4.45 |
| 4 | 4.03 | 2.64 | 3.06 | 2.30 |
| 5 | 2.16 | 4.17 | 6.36 | 6.46 |

SOLUTION

METHOD 1. COEFFICIENT OF SKEWNESS

Step 1. Compute the log-mean ($\bar{y}$), log-standard deviation ($s_y$), and sum of the cubed residuals for the logged nickel concentrations ($y_i$):

$$\bar{y} = \frac{1}{20}(4.07 + 0.00 + \ldots + 6.46) \quad = \quad 3.918 \log(ppb)$$

$$s_y \quad = \quad \sqrt{\frac{1}{19}\left[(4.07 - 3.918)^2 + (0.00 - 3.918)^2 + \ldots + (6.46 - 3.918)^2\right]} \quad = \quad 1.8014 \log(ppb)$$

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^3 \quad = \quad \left[(4.07 - 3.918)^3 + \ldots + (6.46 - 3.918)^3\right] = -26.528 \log^3(ppb)$$

Step 2. Calculate the coefficient of skewness using equation [10.8] with Step 1 values as:

$$\gamma_1 = (20)^{1/2}(-26.528) \Big/ (19)^{3/2}(1.8014)^3 = -0.245$$

Since the absolute value of the skewness is less than 1, the data do not show evidence of significant skewness. Applying a normal distribution to the log-transformed data may therefore be appropriate, but this model should be further checked. The logarithmic CV of 4.97 computed in Example 10-1 was also suggestive of a highly skewed distribution, but can be difficult to interpret in determining if measurements, in fact, follow a logarithmic distribution.

METHOD 2. SHAPIRO-WILK TEST

Step 1. Order and rank the data from smallest to largest and list, as in the table below. List the data in reverse order alongside the first column. Denote the *i*th logged observation by $y_i = \log(x_i)$.

Step 2. Compute differences $\left[y_{(n-i+1)} - y_{(i)}\right]$ in column 4 of the table by subtracting column 2 from column 3. Since $n = 20$, the largest integer less than or equal to $(n/2)$ is $k = 10$.

Step 3. Look up the coefficients $a_{n-i+1}$ from **Table 10-2** of **Appendix D** and list in column 5.

Step 4. Multiply the differences in column 4 by the coefficients in column 5 and add the first $k$ products ($b_i$) to get quantity $b$, using equation [10.9].

$$b = [.4734(6.85) + .3211(5.33) + \ldots + .0140(.04)] = 7.77$$

| i | $y_{(i)}$ | $y_{(n-i+1)}$ | $y_{(n-i+1)} - y_{(i)}$ | $a_{n-i+1}$ | $b_i$ |
|---|---|---|---|---|---|
| 1 | 0.00 | 6.85 | 6.85 | .4734 | 3.24 |
| 2 | 1.13 | 6.46 | 5.33 | .3211 | 1.71 |
| 3 | 2.16 | 6.36 | 4.20 | .2565 | 1.08 |
| 4 | 2.30 | 5.80 | 3.50 | .2085 | 0.73 |
| 5 | 2.64 | 5.57 | 2.93 | .1686 | 0.49 |
| 6 | 2.94 | 5.02 | 2.08 | .1334 | 0.28 |
| 7 | 3.06 | 4.45 | 1.39 | .1013 | 0.14 |
| 8 | 3.30 | 4.40 | 1.10 | .0711 | 0.08 |
| 9 | 3.66 | 4.17 | 0.51 | .0422 | 0.02 |
| 10 | 4.03 | 4.07 | 0.04 | .0140 | 0.00 |
| 11 | 4.07 | 4.03 | −0.04 | | b = 7.77 |
| 12 | 4.17 | 3.66 | −0.51 | | |
| 13 | 4.40 | 3.30 | −1.10 | | |
| 14 | 4.45 | 3.06 | −1.39 | | |
| 15 | 5.02 | 2.94 | −2.08 | | |
| 16 | 5.57 | 2.64 | −2.93 | | |
| 17 | 5.80 | 2.30 | −3.50 | | |
| 18 | 6.36 | 2.16 | −4.20 | | |
| 19 | 6.46 | 1.13 | −5.33 | | |
| 20 | 6.85 | 0.00 | −6.85 | | |

Step 5. Compute the log-standard deviation of the sample, $s_y = 1.8014$. Then use [10.10] to calculate the *SW* test statistic:

$$SW = \left[\frac{7.77}{1.8014\sqrt{19}}\right]^2 = 0.979$$

Step 6. Use **Table 10-3** of **Appendix D** to determine the .01-level critical point for the Shapiro-Wilk test when $n = 20$. This gives $sw_{cp} = 0.868$. Then compare the observed value of $SW = 0.979$ to the 1% critical point. Since $SW > 0.868$, the sample shows no significant evidence of non-normality by the Shapiro-Wilk test. Proceed with further statistical analysis using the log-transformed data or by assuming the underlying population is lognormal.

### METHOD 3. PROBABILITY PLOT CORRELATION COEFFICIENT

Step 1. Order and rank the logged nickel data from smallest to largest and list, as in the table below. Again let the *i*th logged value be denoted by $y_i = \log(x_i)$. The sample size is $n = 20$, the log-mean is $\bar{y} = 3.918$, and the log-standard deviation is $s_y = 1.8014$.

Step 2. Compute the intermediate probabilities $m_i$ from equation [10.13] for each *i* in column 3 and the rank statistic medians, $M_i$ , in column 4 by applying the inverse normal transformation to column 3 using equation [10.14] and **Table 10-1** of **Appendix D**.

| i | $y_{(i)}$ | $m_i$ | $M_i$ | $(M_i)^2$ | $y_{(i)} \times M_i$ |
|---|---|---|---|---|---|
| 1 | 0.00 | .03406 | −1.8242 | 3.328 | 0.000 |
| 2 | 1.13 | .08262 | −1.3877 | 1.926 | −1.568 |
| 3 | 2.16 | .13172 | −1.1183 | 1.251 | −2.416 |
| 4 | 2.30 | .18082 | −0.9122 | 0.832 | −2.098 |
| 5 | 2.64 | .22993 | −0.7391 | 0.546 | −1.951 |
| 6 | 2.94 | .27903 | −0.5857 | 0.343 | −1.722 |
| 7 | 3.06 | .32814 | −0.4451 | 0.198 | −1.362 |
| 8 | 3.30 | .37724 | −0.3127 | 0.098 | −1.032 |
| 9 | 3.66 | .42634 | −0.1857 | 0.034 | −0.680 |
| 10 | 4.03 | .47545 | −0.0616 | 0.004 | −0.248 |
| 11 | 4.07 | .52455 | 0.0616 | 0.004 | 0.251 |
| 12 | 4.17 | .57366 | 0.1857 | 0.034 | 0.774 |
| 13 | 4.40 | .62276 | 0.3127 | 0.098 | 1.376 |
| 14 | 4.45 | .67186 | 0.4451 | 0.198 | 1.981 |
| 15 | 5.02 | .72097 | 0.5857 | 0.343 | 2.940 |
| 16 | 5.57 | .77007 | 0.7391 | 0.546 | 4.117 |
| 17 | 5.80 | .81918 | 0.9122 | 0.832 | 5.291 |
| 18 | 6.36 | .86828 | 1.1183 | 1.251 | 7.112 |
| 19 | 6.46 | .91738 | 1.3877 | 1.926 | 8.965 |
| 20 | 6.85 | .96594 | 1.8242 | 3.328 | 12.496 |

Step 3. Since this sample contains no non-detects or ties, the simplified equations for $C_n$ in [10.16] and for $r$ in [10.18] may be used. First compute $C_n$ using the squared order statistic medians in column 5:

$$C_n = \sqrt{[3.328 + 1.926 + \ldots + 3.328]} \quad = \quad 4.138$$

Step 4. Next compute the products $y_{(i)} \times M_i$ in column 6 and sum to get the numerator of the correlation coefficient (equal to 32.226 in this case). Then compute the final correlation coefficient:

$$r = 32.226 \Big/ \left[ 4.138 \times 1.8014\sqrt{19} \right] = 0.992$$

Step 5. Compare the Filliben's test statistic of $r$ = 0.992 to the 1% critical point for a sample of size 20 in **Table 10-5** in **Appendix D**, namely $r_{cp}$ = 925. Since $r$ > 0.925, the sample shows no significant evidence of non-normality by the probability plot correlation coefficient test. Therefore, lognormality of the original data can be assumed in subsequent statistical procedures.

Note: the Shapiro-Wilk and Filliben's Probability Plot Correlation Coefficient tests for normality on a single data set perform quite comparably. Only one of these tests need be run in routine applications. ◄

# CHAPTER 11. TESTING EQUALITY OF VARIANCE

11.1 BOX PLOTS ..... 11-2
11.2 LEVENE'S TEST ..... 11-4
11.3 MEAN-STANDARD DEVIATION SCATTER PLOT ..... 11-8

Many of the methods described in the Unified Guidance assume that the different groups under comparison have the same variance (*i.e.*, are *homoscedastic*). This chapter covers procedures for assessing homoscedasticity and its counterpart, *heteroscedasticity* (*i.e.*, unequal variances). Equality of variance is assumed, for instance, when using prediction limits to make either upgradient-to-downgradient or intrawell comparisons. In the former case, the method assumes that the upgradient variance is equal to the variance in each downgradient well. In the latter case, the presumption is that the well variance is stable over time (*i.e.*, stationary) when comparing intrawell background versus more recent measurements.

If a prediction limit is constructed on a single new measurement at each downgradient well, it isn't feasible to test the variance equality assumption prior to each statistical evaluation. Homoscedasticity *can* be tested after several new rounds of compliance sampling by pooling collected compliance measurements within a well. The Unified Guidance recommends periodic testing of the presumption of equal variances by comparing newer data to historical background (**Chapter 6**).

Equality of variance between different groups (*e.g.*, different wells) is also an important assumption for an analysis of variance [ANOVA]. If equality of variance does not hold, the power of the *F*-test (its ability to detect differences among the group means) is reduced. Mild differences in variance are generally acceptable. But the effect becomes noticeable when the largest and smallest group variances differ by a ratio of about 4, and becomes quite severe when the ratio is 10 or more (Milliken and Johnson, 1984).

Three procedures for assessing or testing homogeneity of variance are described in the Unified Guidance, two of which that are more robust to departures from normality (*i.e.*, less sensitive to non-normality). These include:

1. The box plot (**Chapter 9**), a graphical method useful not only for checking equality of variance but also as an exploratory tool for visualizing the basic statistical characteristics of data sets. It can also provide a rough indication of differences in mean or median concentration levels across several wells;
2. Levene's test (**Section 11.2**), a formal ANOVA-type procedure for testing variance inequality; and
3. The mean-standard deviation scatter plot (**Chapter 9** and **Section 11.3**), a visual tool for assessing whether the degree of variability in a set of data groups or wells is correlated with the mean levels for those groups. This could potentially indicate whether a *variance stabilizing transformation* might be needed.

# 11.1 BOX PLOTS

## PURPOSE AND BACKGROUND

Box plots are described in **Chapter 9**. In the context of variance testing, one can construct a box plot for each well group and compare the boxes to see if the assumption of equal variances is reasonable. The comparison is not a formal test procedure, but is easier to perform and is often sufficient for checking the group variance assumption.

Box plots for each data group simultaneously graphed side-by-side provide a direct visual comparison of the dispersion in each group. As a rule of thumb, if the box length for each group is less than 1.5–2 times the length of the shortest box, the sample variances may be close enough to assume equal group variances. If the box length for any group is greater than 1.5–2 times the length of the box for another group, the variances may be significantly different. A formal test such as Levene's might be needed to more accurately decide. Sample data sets with unequal variances may need a *variance stabilizing transformation*, *i.e.*, one in which the transformed measurements have approximately equal variances.

Most statistical software packages will calculate the statistics needed to draw a box plot, and many will construct side-by-side box plots directly. Usually a box plot will also be shown with two "whiskers" extending from the edges of the box. These lines indicate either the positions of extreme minimum or maximum values in the data set. In Tukey's original formulation (Tukey, 1977), they indicate the most extreme lower and upper data points outside the box but falling within a distance of 1.5 times the interquartile range (that is, the length of the box) from either edge. The whiskers should generally *not* be used to approximate the overall variance under either formulation.

A convenient tactic when using box plots to screen for heteroscedasticity is to plot the *residuals* of each data group rather than the measurements themselves. This will line the boxes up at roughly a common level (close to zero), so that a visual comparison of box lengths is easier.

## REQUIREMENTS AND ASSUMPTIONS

The requirements and assumptions for box plots are discussed in **Section 9.2.**

## PROCEDURE

Step 1. For each of $j$ wells or data groups, compute the sample mean of that group $\bar{x}_j$. Then compute the residuals ($r_{ij}$) for each group by subtracting the group mean from each individual measurement: $r_{ij} = x_{ij} - \bar{x}_j$.

Step 2. Use the procedure outlined in **Section 9.2** to create side-by-side box plots of the residuals formed in Step 1. Then compare the box lengths to check for possibly unequal variances.

### ►EXAMPLE 11-1

Construct box plots on the residuals for each of the following well groups to check for homoscedasticity.

| | Arsenic Concentration (ppb) | | | | | |
|---|---|---|---|---|---|---|
| Month | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| 1 | 22.9 | 2.0 | 2.0 | 7.8 | 24.9 | 0.3 |
| 2 | 3.1 | 1.2 | 109.4 | 9.3 | 1.3 | 4.8 |
| 3 | 35.7 | 7.8 | 4.5 | 25.9 | 0.8 | 2.8 |
| 4 | 4.2 | 52 | 2.5 | 2.0 | 27 | 1.2 |

SOLUTION

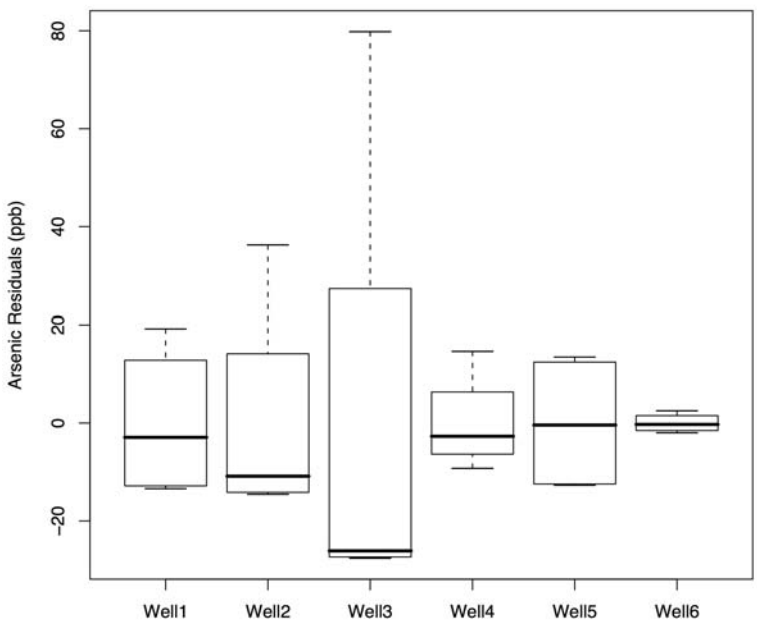
Step 1. Form the residuals for each well by subtracting the sample well mean from each observation, as shown in the table below.

| | Arsenic Residuals (ppb) | | | | | |
|---|---|---|---|---|---|---|
| Month | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| 1 | 6.43 | −13.75 | −27.6 | −3.45 | 11.4 | −1.98 |
| 2 | −13.38 | −14.55 | 79.8 | −1.95 | −12.2 | 2.52 |
| 3 | 19.22 | −7.95 | −25.1 | 14.65 | −12.7 | 0.52 |
| 4 | −12.28 | 36.25 | −27.1 | −9.25 | 13.5 | −1.08 |
| Mean | 16.48 | 15.75 | 29.6 | 11.25 | 13.5 | 2.28 |

Step 2. Follow the procedure in **Section 9.2** to compute a box plot of the residuals for each well. Line these up side by side on the same graph, as in **Figure 11-1**.

Step 3. Compare the box lengths. Since the box length for Well 3 is more than three times the box lengths of Wells 4 and 6, there is informal evidence that the population group variances may be different. These data should be further checked using a formal test and perhaps a variance stabilizing transformation attempted. ◄

Figure 11-1. Side-by-Side Box Plots of Arsenic Residuals




## 11.2 LEVENE'S TEST

### PURPOSE AND BACKGROUND

Levene's test is a formal procedure for testing homogeneity of variance that is fairly robust (i.e., not overly sensitive) to non-normality in the data. It is based on computing the new variables:

$$z_{ij} = \left| x_{ij} - \bar{x}_{i\bullet} \right| \qquad [11.1]$$

where $x_{ij}$ represents the *j*th sample value from the *i*th group (*e.g.*, well) and $\bar{x}_{i\bullet}$ is the *i*th group sample mean. The symbol $(\bullet)$ in the notation for the group sample mean represents an averaging over subscript *j*. The values $z_{ij}$ then represent the absolute values of the ***residuals***. Levene's test involves running a standard one-way ANOVA (**Chapter 17**) on the variables $z_{ij}$. If the *F*-test is significant, reject the hypothesis of equal group variances and perhaps seek a variance stabilizing transformation. Otherwise, proceed with analysis of the original $x_{ij}$'s.

Levene's test is based on a one-way ANOVA and contrasts the means of the groups being tested. This implies a comparison between averages of the form:

$$\bar{z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left|x_{ij} - \bar{x}_{i\bullet}\right| \qquad [11.2]$$

Such averages of the $z_{ij}$'s are very similar to the standard deviations of the original data groups, given by the formula:

$$s_i = \sqrt{\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2} \qquad [11.3]$$

In both cases, the statistics are akin to an average absolute residual. Therefore, the comparison of means in Levene's test is closely related to a direct comparison of the group standard deviations, the underlying aim of any test of variance equality.

## REQUIREMENTS AND ASSUMPTIONS

The requirements and assumptions for Levene's test are essentially the same as the one-way ANOVA in **Section 17.1**, but applied to the absolute residuals instead of the raw measurements.

## PROCEDURE

Step 1. Suppose there are $p$ data groups to be compared. Because there may be different numbers of observations per well, denote the sample size of the $i$th group by $n_i$ and the total number of data points across all groups by $N$.

Denote the observations in the $i$th group by $x_{ij}$ for $i = 1\ldots p$ and $j = 1\ldots n_i$. The first subscript then designates the well, while the second denotes the $j$th value in the $i$th well. After computing the sample mean ($\bar{x}_i$) for each group, calculate the absolute residuals ($z_{ij}$) using equation [11.1].

Step 2. Utilizing the absolute residuals — and not the original data — compute the mean of each group along with the overall (grand) mean of the combined data set using the formula:

$$\bar{z}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{n_i} z_{ij} \qquad [11.4]$$

Step 3. Compute the sum of squares of differences between the group means and the grand mean, denoted $SS_{grps}$:

$$SS_{grps} = \sum_{i=1}^{p} n_i\left(\bar{z}_{i\bullet} - \bar{z}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{p} n_i\bar{z}_{i\bullet}^2 - N\bar{z}_{\bullet\bullet}^2 \qquad [11.5]$$

The formula on the far right is usually the most convenient for calculation. This sum of squares has ($p$–1) degrees of freedom associated with it and is a measure of the variability *between* groups. It constitutes the numerator of the $F$-statistic.

Step 4. Compute the corrected total sum of squares, denoted by $SS_{\text{total}}$:

$$SS_{total} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(z_{ij} - \bar{z}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i} z_{ij}^2 - N\bar{z}_{\bullet\bullet}^2 \qquad [11.6]$$

Again, the formula on the far right is usually the most computationally convenient. This sum of squares has ($N$–1) associated degrees of freedom.

Step 5. Compute the sum of squares of differences between the absolute residuals and the group means. This is known as the within-groups component of the total sum of squares or, equivalently, as the sum of squares due to error. It is easiest to obtain by subtracting $SS_{\text{grps}}$ from $SS_{\text{total}}$ and is denoted $SS_{\text{error}}$:

$$SS_{error} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(z_{ij} - \bar{z}_{i\bullet}\right)^2 = SS_{total} - SS_{grps} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} z_{ij}^2 - \sum_{i=1}^{p} n_i\bar{z}_{i\bullet}^2 \qquad [11.7]$$

$SS_{\text{error}}$ is associated with ($N$–$p$) degrees of freedom and is a measure of the variability *within* groups. This quantity goes into the denominator of the $F$-statistic.

Step 6. Compute the mean sum of squares for both the between-groups and within-groups components of the total sum of squares, denoted by $MS_{\text{grps}}$ and $MS_{\text{error}}$. These quantities are obtained by dividing each sum of squares by its corresponding degrees of freedom:

$$MS_{grps} = SS_{grps} \big/ \left(p - 1\right) \qquad [11.8]$$

$$MS_{error} = SS_{error} \big/ \left(N - p\right) \qquad [11.9]$$

Step 7. Compute the $F$-statistic by forming the ratio between the mean sum of squares for wells and the mean sum of squares due to error, as in **Figure 11-2** below. This layout is known as the one-way parametric ANOVA table and illustrates each sum of squares component of the total variability, along with the corresponding degrees of freedom, the mean squares components, and the final $F$-statistic calculated as $F = MS_{\text{grps}}/MS_{\text{error}}$. Note that the first two rows of the one-way table sum to the last row.

Step 8. **Figure 11-2** is a generalized ANOVA table for Levene's test. To test the hypothesis of equal variances across all $p$ well groups, compare the $F$-statistic in **Figure 11-2** to the $\alpha$-level critical point found from the $F$-distribution with ($p$–1) and ($N$–$p$) degrees of freedom in **Appendix D Table 17-1**. When testing variance equality, only severe levels of difference typically impact test performance in a substantial way. For this reason, the Unified Guidance recommends setting $\alpha$ = .01 when screening multiple wells and/or constituents using Levene's test. In that case, the needed critical point equals the upper 99th percentage point of the $F$-distribution. If the observed $F$-statistic exceeds the critical point ($F_{.99,\text{p}-1,\text{N}-\text{p}}$), reject the hypothesis of equal group population variances. Otherwise, conclude that there is insufficient evidence of a significant difference between the variances.

Figure 11-2. ANOVA Table for Levene's Test

| Source of Variation | Sums of Squares | Degrees of Freedom | Mean Squares | *F*-Statistic |
|---|---|---|---|---|
| Between Wells | $SS_{grps}$ | $p-1$ | $MS_{grps} = SS_{grps}/(p-1)$ | $F = MS_{grps}/MS_{error}$ |
| Error (within wells) | $SS_{error}$ | $N-p$ | $MS_{error} = SS_{error}/(N-p)$ | |
| Total | $SS_{total}$ | $N-1$ | | |

## ►EXAMPLE 11-2

Use the data from **Example 11-1** to conduct Levene's test of equal variances at the α = 0.01 level of significance.

### SOLUTION

Step 1. Calculate the group arsenic mean for each well ($\bar{x}_{i\bullet}$):

Well 1 mean = 16.47 ppm  Well 4 mean = 11.26 ppm

Well 2 mean = 15.76 ppm  Well 5 mean = 13.49 ppm

Well 3 mean = 29.60 ppm  Well 6 mean = 2.29 ppm

Then compute the absolute residuals $z_{ij}$ in each well using equation [11.1] as in the table below.

| | **Absolute Arsenic Residuals ($z_{ij}$)** | | | | | |
|---|---|---|---|---|---|---|
| Month | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| 1 | 6.43 | 13.76 | 27.6 | 3.42 | 11.41 | 1.95 |
| 2 | 13.38 | 14.51 | 79.8 | 1.96 | 12.19 | 2.49 |
| 3 | 19.23 | 7.96 | 25.1 | 14.64 | 12.74 | 0.56 |
| 4 | 12.29 | 36.24 | 27.1 | 9.26 | 13.51 | 1.09 |
| Well Mean ($\bar{z}_{i\bullet}$) | 12.83 | 18.12 | 39.9 | 7.32 | 12.46 | 1.52 |
| Overall Mean ($\bar{z}_{\bullet\bullet}$) | 15.36 | | | | | |

Step 2. Compute the mean absolute residual ($\bar{z}_{i\bullet}$) in each well and then the overall grand mean using equation [11.4]. These results are listed above.

Step 3. Compute the between-groups sum of squares for the absolute residuals using equation [11.5]:

$$SS_{grps} = \left[4(12.83)^2 + 4(18.12)^2 + \ldots + 4(1.52)^2\right] - 24 \cdot (15.36)^2 = 3{,}522.90$$

Step 4. Compute the corrected total sum of squares using equation [11.6]:

$$SS_{total} = \left[(6.43)^2 + (13.38)^2 + \ldots + (1.09)^2\right] - 24 \cdot (15.36)^2 = 6{,}300.89$$

Step 5. Compute the within-groups or error sum of squares using equation [11.7]:

$$SS_{error} = 6{,}300.89 - 3{,}522.90 = 2{,}777.99$$

Step 6. Given that the number of groups is $p$ = 6 and the total sample size is $N$ = 24, calculate the mean squares for the between-groups and error components using formulas [11.8] and [11.9]:

$$MS_{grps} = 3{,}522.90 \big/ (6-1) = 704.58$$

$$MS_{error} = 2{,}777.99 \big/ (24-6) = 154.33$$

Step 7. Construct an ANOVA table following **Figure 11-2** to calculate the $F$-statistic. The numerator degrees of freedom [$df$] is computed as ($p$–1) = 5, while the denominator $df$ is equal to ($N$–$p$) = 18.

| Source of Variation | Sums of Squares | Degrees of Freedom | Mean Squares | $F$-Statistic |
|---|---|---|---|---|
| Between Well Grps | 3,522.90 | 5 | 704.58 | 4.56 |
| Error (within grps) | 2,777.99 | 18 | 154.33 | |
| Total | 6,300.89 | 23 | | |

Step 8. Determine the .01-level critical point for the $F$-test with 5 and 18 degrees of freedom from **Table 17-1**. This gives $F_{.99,5,18}$ = 4.25. Since the $F$-statistic of 4.56 exceeds the critical point, the assumption of equal variances should be rejected. Since the original concentration data are used in this example, a transformation such as the natural logarithm might be tried and the transformed data retested. ◄

## 11.3 MEAN-STANDARD DEVIATION SCATTER PLOT

### BACKGROUND AND PURPOSE

The mean-standard deviation scatter plot is described in **Chapter 9**. It is useful as an exploratory tool for multiple groups of data (*e.g.*, wells) to aid in identifying relationships between mean levels and variability. It is also helpful in providing a visual assessment of variance homogeneity across data groups. Like side-by-side box plots, the mean-standard deviation scatter plot graphs a measure of variability for each well. In the latter, however, the standard deviation is plotted rather than the interquartile range, so a more direct assessment of variance equality can be made. Since standard

deviations (and consequently variances) are often positively correlated with sample mean levels in skewed populations, the observed pattern on the mean-standard deviation scatter plot can offer valuable clues as to what sort of variance stabilizing transformation if any might work.

## REQUIREMENTS AND ASSUMPTIONS

The requirements for the mean-standard deviation scatter plot are listed in **Section 9.4.**

## PROCEDURE

See **Section 9.4.**

## ►EXAMPLE 11-3

Use the data from **Example 11-1** to construct a mean-standard deviation scatter plot.

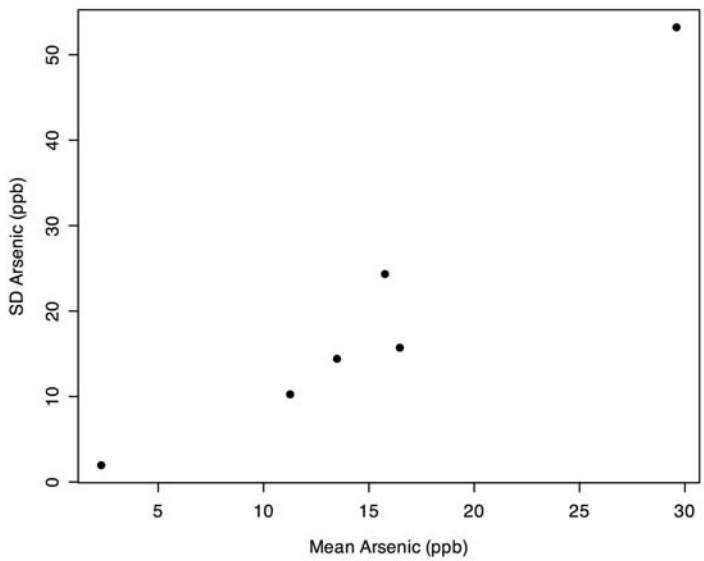
### SOLUTION

Step 1. First compute the sample mean ( $\bar{x}$ ) and standard deviation ($s$) of each well, as listed below.

| Well | Mean | Std Dev |
|---|---|---|
| 1 | 16.468 | 15.718 |
| 2 | 15.762 | 24.335 |
| 3 | 29.600 | 53.211 |
| 4 | 11.260 | 10.257 |
| 5 | 13.488 | 14.418 |
| 6 | 2.292 | 1.958 |

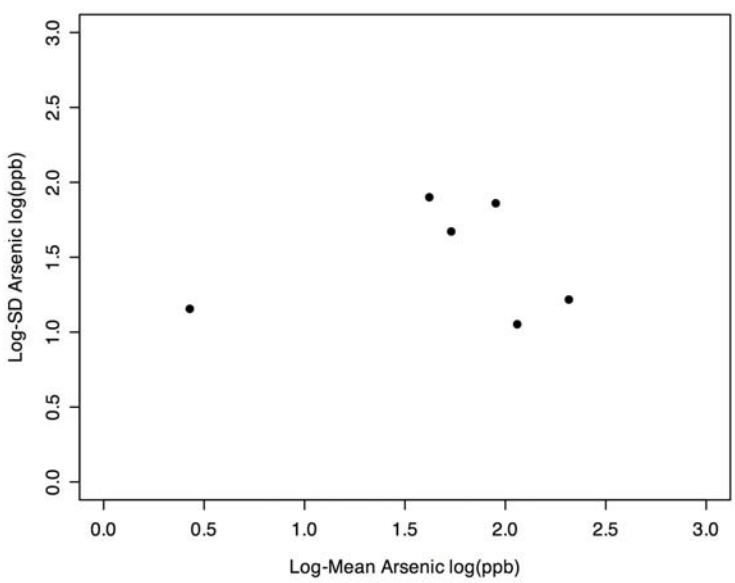
Step 2. Plot the well means versus the standard deviations as in **Figure 11-3** below. Note the roughly linear relationship between the magnitude of the standard deviations and their corresponding means. The data suggest unequal variances among the wells, as indicated by the large range in the standard deviations.

**Figure 11-3. Arsenic Mean-Standard Deviation Plot**




Step 3. Because lognormal data groups will tend to show a linear association between the sample means and standard deviations, apply a log transformation to the original arsenic measurements and reconstruct the mean-standard deviation scatter plot on the log scale. Computing the log-means and log-standard deviations and then re-plotting gives **Figure 11-4**. Now the apparent trend between the means and standard deviations is gone. Further, on the log scale, the standard deviations are much more similar in magnitude, all with values between 1 and 2. The log transformation thus appears to roughly stabilize the arsenic variances. ◄

**Figure 11-4. Log(Arsenic) Mean-Standard Deviation Plot**



# CHAPTER 12. IDENTIFYING OUTLIERS

12.1 SCREENING WITH PROBABILITY PLOTS ..... 12-1
12.2 SCREENING WITH BOX PLOTS ..... 12-5
12.3 DIXON'S TEST ..... 12-8
12.4 ROSNER'S TEST ..... 12-10

This chapter discusses screening tools and formal tests for identifying statistical outliers. Two screening tools are first presented: probability plots (**Section 12.1**) and box plots (**Section 12.2**). These are followed by two formal outlier tests:

- Dixon's test (**Section 12.3**) for a single outlier in smaller data sets, and
- Rosner's test (**Section 12.4**) for up to five separate outliers in larger data sets.

A statistical determination of one or more statistical outliers does not indicate *why* the measurements are discrepant from the rest of the data set. The Unified Guidance does not recommend that outliers be removed *solely* on a statistical basis. The outlier tests can provide supportive information, but generally a reasonable rationale needs to be identified for removal of suspect outlier values (usually limited to background data**).** At the same time there must be some level of confidence that the data are representative of ground water quality. A number of factors and considerations in removing outliers from potential background data are discussed in **Section 5.2.3.**

## 12.1 SCREENING WITH PROBABILITY PLOTS

### BACKGROUND AND PURPOSE

Probability plots (**Chapter 9**) are helpful in identifying outliers in at least two ways. First, since the straightness of the plot indicates how closely the data fit the pattern of a normal distribution, values that appear "out of line" with the remaining data can be visually identified as possible outliers. Secondly, the two formal outlier tests presented in the Unified Guidance assume that the underlying population minus the suspected outlier(s) is normal. Probability plots can provide visual evidence for this assumption. Data that appear non-normal after the suspected outliers have been removed from the probability plot may need to be transformed (e.g., via the natural logarithm) and re-examined on the transformed scale to see if potential outliers are still apparent.

As an aid to the interpretation of a given probability plot, the Unified Guidance recommends computation of the probability plot correlation coefficient, using either Filliben's procedure (**Chapter 10**) or the simple (Pearson) correlation (**Chapter 3**) between the numerical pairs plotted on the graph. The higher the correlation, the more linear the pattern is on the probability plot and therefore a better fit to normality. Note that while the Filliben correlation coefficient can be compared to critical points derived for that test of normality (**Chapter 10**), a low correlation may be related to other causes of non-normality besides the presence of outliers. The correlation coefficient is not a substitute for a formal outlier test, but can be useful as a screening tool.

## REQUIREMENTS AND ASSUMPTIONS

Probability plots are primarily a tool to assess normality, and not to identify outliers *per se*. It is critical that the remaining data without potential outliers is either normal in distribution or can be normalized via a transformation. Otherwise, the probability plot may appear non-linear and non-normal for reasons unrelated to the presence of outliers. Right-skewed lognormal distributions can appear to have one or more outliers on a probability plot unless the original data are first log-transformed. As a general rule, probability plots should be constructed on the original (or raw) measurements and one or more transformed data sets (*e.g.*, log or square root), in order to avoid mistaking inherent data skewness for outliers.

If the raw and transformed-data probability plots both indicate one or more values inconsistent with the pattern of the remaining values, continue with a second level of screening by temporarily removing the suspected outlier(s) and re-constructing the probability plots. If the raw-scale plot is reasonably linear, consider running a formal outlier test on the original measurements. On the other hand, if the raw-scale plot is skewed but the transformed-scale plot is linear, consider conducting a formal outlier test on the transformed measurements.

A related difficulty occurs when sample data includes censored or non-detect values. If simple substitution is used to estimate a value for each non-detect prior to plotting, the resulting probability plot may appear non-linear simply because the censored observations were not properly handled. In this case, a censored probability plot (**Chapter 15**) should be constructed instead of an uncensored, complete sample plot (**Chapter 9**). The same caveats apply to normalizing the sample data, perhaps by attempting at least one transformation. The only difference is that each probability plot constructed must appropriately account for the observed censoring in the sample.

## PROCEDURE

Step 1. After identifying one or more possible outliers (*e.g.*, values much higher in concentration than the remaining measurements), construct a probability plot on the entire sample using the procedure described in **Section 9.5**. Construct a *censored probability plot* from **Section 15.3** if the sample contains non-detects. If the data including the suspected outlier(s) follow a reasonably linear pattern, a formal outlier test is probably unnecessary. However, if one or more values are out of line compared to the pattern of the remaining data, construct a similar probability plot after applying one or more transformations. If one or more suspected outliers is still inconsistent, proceed to Step 2.

Step 2. Compute a probability plot correlation coefficient for each plot constructed in Step 1. Use these correlations as an aid to interpreting the degree of linearity in each probability plot.

Step 3. Reconstruct the probability plots from Step 1 after removing the suspected outlier(s). Recompute the correlation coefficients from Step 2 on this reduced sample.

Step 4. If the 'outlier-deleted' probability plot on the raw concentration scale indicates a linear pattern with high correlation, consider running a formal outlier test on the original measurements. When the pattern is distinctly non-linear but the corresponding probability plot on the transformed-scale is fairly linear (and higher in correlation), conduct the outlier test on the transformed values.

## ►EXAMPLE 12-1

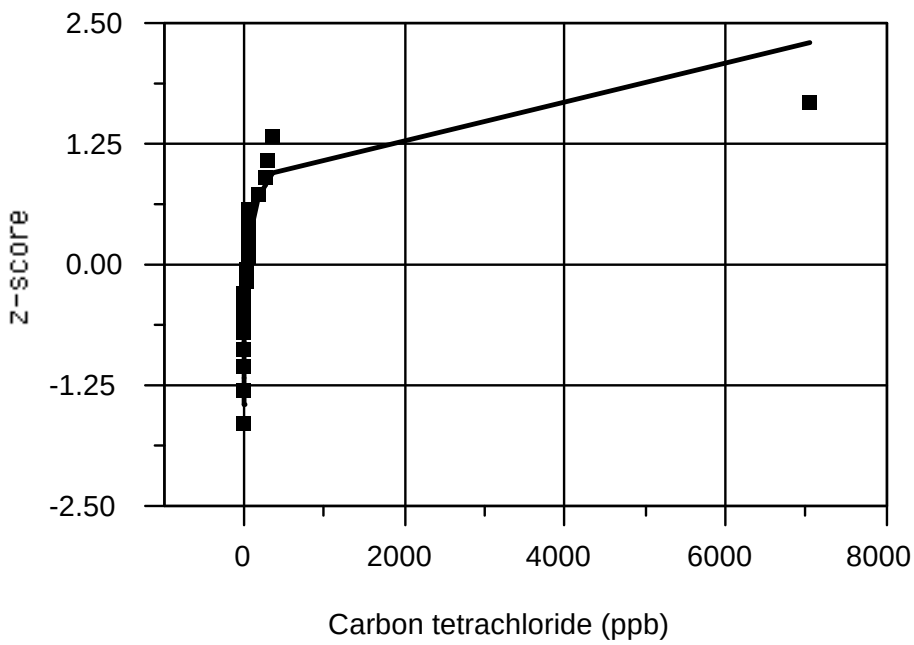
The table below contains data from five background wells measured over a four month period. The value 7,066 is found in the second month at Well 3. Use probability plots on the combined sample to determine whether or not a formal outlier test is warranted.

| **Carbon Tetrachloride Concentrations (ppb)** | | | | |
|---|---|---|---|---|
| Well 1 | Well 2 | Well 3 | Well 4 | Well 5 |
| 1.7 | 302 | 16.2 | 199 | 275 |
| 3.2 | 35.1 | 7066 | 41.6 | 6.5 |
| 7.3 | 15.6 | 350 | 75.4 | 59.7 |
| 12.1 | 13.7 | 70.1 | 57.9 | 68.4 |

### SOLUTION

Step 1. Examine the probability plots of the entire sample first using the raw measurements and then log-transformed values (**Figures 12-1** and **12-2**). Both these plots indicate that the suspected outlier does not follow the pattern of the remaining observations, but seems 'out of line.' The Pearson correlation coefficients for these probability plots are, respectively, $r$ = 0.513 and 0.975, indicating that the fit to normality overall is much closer using log-transformed measurements.

Figure 12-1. Probability Plot on Raw Concentrations ($r$ = .513)




Step 2. Next remove the suspected outlier and reconstruct the probability plots on both the original and logged observations (**Figures 12-3** and **12-4**). The plot on the original scale indicates heavy positive (or right-) skewness and a non-linear pattern, while the plot on the log-scale exhibits a fairly linear pattern. The respective correlation coefficients now become $r$ = 0.854 and 0.985, again favoring the log-transformed sample. On the basis of these plots, the

underlying data should be modeled as lognormal and the observations logged prior to running a formal outlier test. ◄

Figure 12-2. Probability Plot on Logged Observations (*r* = .975)

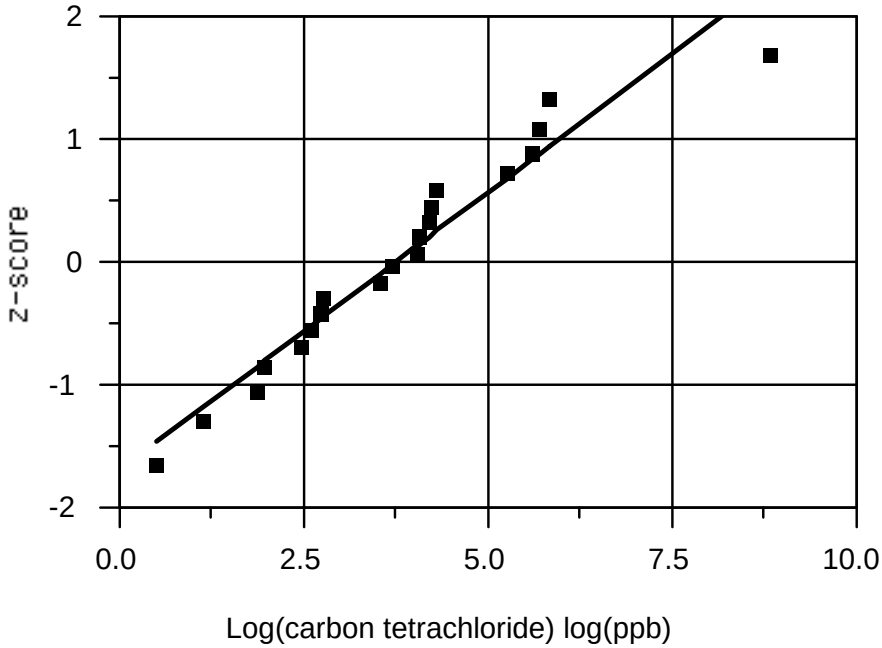



Figure 12-3. Outlier-Deleted Probability Plot on Original Scale (*r* = .854)

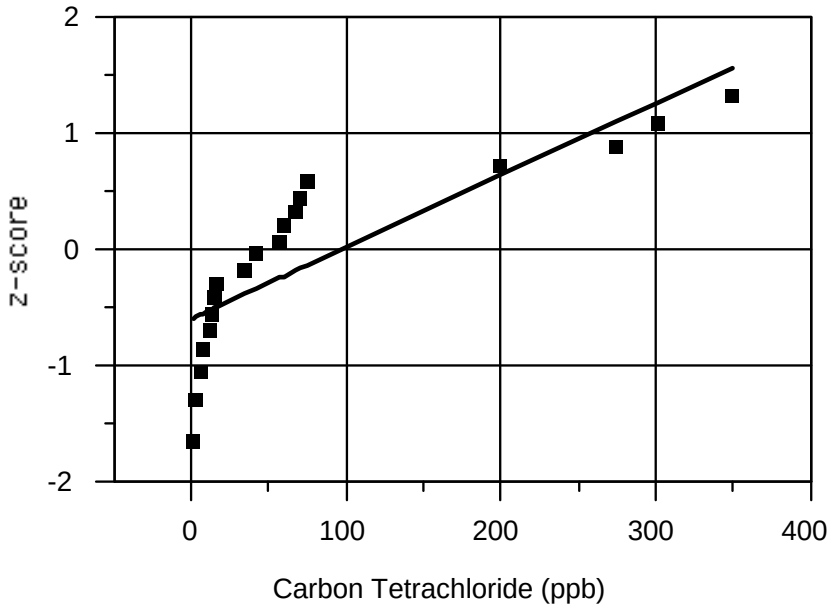
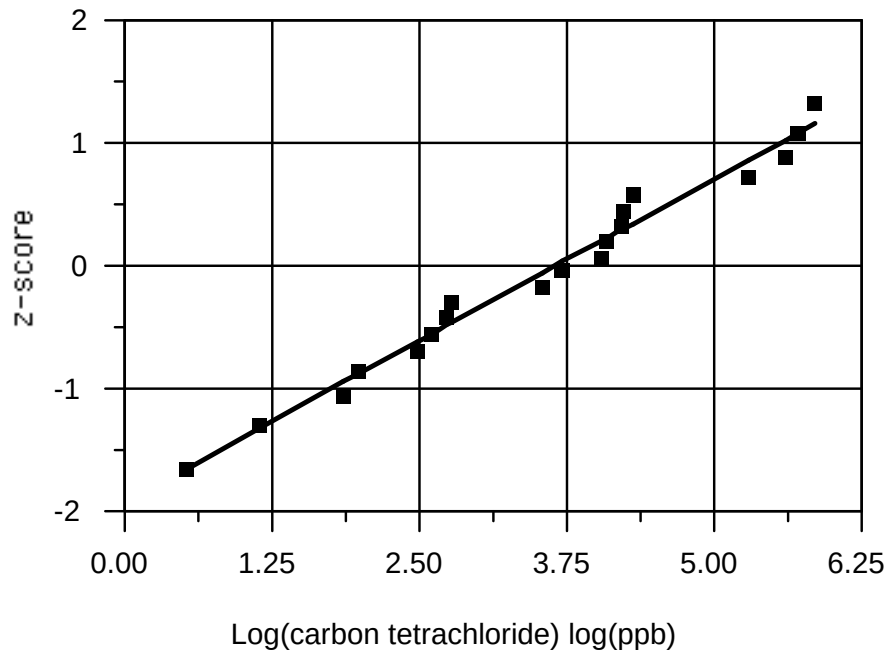


Figure 12-4. Outlier-Deleted Probability Plot on Logarithmic Scale ($r$ = .985)




## 12.2 SCREENING WITH BOX PLOTS

### BACKGROUND AND PURPOSE

Probability plots as described in **Section 12.1** require the remaining observations following removal of one or more suspected outliers to be either approximately normal or normalized via transformation. Box plots (**Chapter 9**) provide an alternate method to perform outlier screening, one not dependent on normality of the underlying measurement population. Instead of looking for points inconsistent with a linear pattern on a probability plot, the box plot flags as possible outliers values that are located in either or both of the ***extreme tails*** of the sample.

To define the extreme tails, Tukey (1977) proposed the concept of 'hinges' that would 'swing' off either end of a box plot, defining the range of concentrations consistent with the bulk of the data. Data points outside this concentration range could then be identified as potential outliers. Tukey defined the hinges, i.e., the lower and upper edges of the box plot, essentially as the lower and upper quartiles of the data set. Then multiples of the interquartile range [IQR] (***i.e.***, the range represented by the middle half of the sample) were added to or subtracted from these hinges as potential outlier boundaries. Any observation from 1.5 × IQR to 3 × IQR below the lower edge of the box plot was labeled a 'mild' low outlier; any value more than 3 × IQR below the lower edge of the box plot was labeled an 'extreme' low outlier. Similarly, values greater than the upper edge of the box plot in the range of 1.5 to 3 times the IQR were labeled 'mild' higher outliers, and 'extreme' high outliers if more than 3 times the IQR beyond the upper box plot edge.

### REQUIREMENTS AND ASSUMPTIONS

By using hinges and multiples of the interquartile range, Tukey's box plot method utilizes statistics (***i.e.***, the lower and upper quartiles) that are generally not or minimally affected by one or a few outliers

in the sample. Consequently, it isn't necessary to first delete possible outliers before constructing the box plot.

Screening for outliers with box plots is a very simple technique. Since no assumption of normality is needed, Tukey's procedure can be considered quasi-non-parametric. But note that rough symmetry of the underlying distribution is implicitly assumed. Legitimate observations from highly skewed distributions could be flagged as potential outliers on a box plot if no transformation of the data is first attempted. It may be necessary to first conduct multiple data transformations in order to achieve approximate symmetry before applying and evaluating potential outliers with box plots.

## PROCEDURE

Step 1. Construct a box plot on the sample using the method given in **Section 9.2**. Using the IQR from that calculation, along with the lower and upper quartiles ( $\tilde{x}_{.25}$ and $\tilde{x}_{.75}$), compute the first pair of lower and upper boundaries as:

$$LB_1 = \tilde{x}_{.25} - 1.5 \times IQR \tag{12.1}$$

$$UB_1 = \tilde{x}_{.75} + 1.5 \times IQR \tag{12.2}$$

Step 2. Construct the second pair of lower and upper boundaries as:

$$LB_2 = \tilde{x}_{.25} - 3 \times IQR \tag{12.3}$$

$$UB_2 = \tilde{x}_{.75} + 3 \times IQR \tag{12.4}$$

Step 3. Label any sample measurement lower than the first lower boundary ($LB_1$) but no less than the second lower boundary ($LB_2$) as a mild low outlier. Label any measurement greater than the first upper boundary ($UB_1$) but no greater than the second upper boundary ($UB_2$) as a mild high outlier.

Step 4. Label any sample measurement lower than the second lower boundary ($LB_2$) as an extreme low outlier. Label any value higher than the second upper boundary ($UB_2$) as an extreme high outlier.

### ►EXAMPLE 12-2

Use the carbon tetrachloride data from **Example 12-1** to screen for possible outliers using Tukey's box plot.

#### SOLUTION

Step 1. Using the procedure described in **Section 9.2**, the upper and lower quartiles of carbon tetrachloride sample are found to be $\tilde{x}_{.25} = 12.9$ and $\tilde{x}_{.75} = 137.2$, leading to an IQR = 124.3.

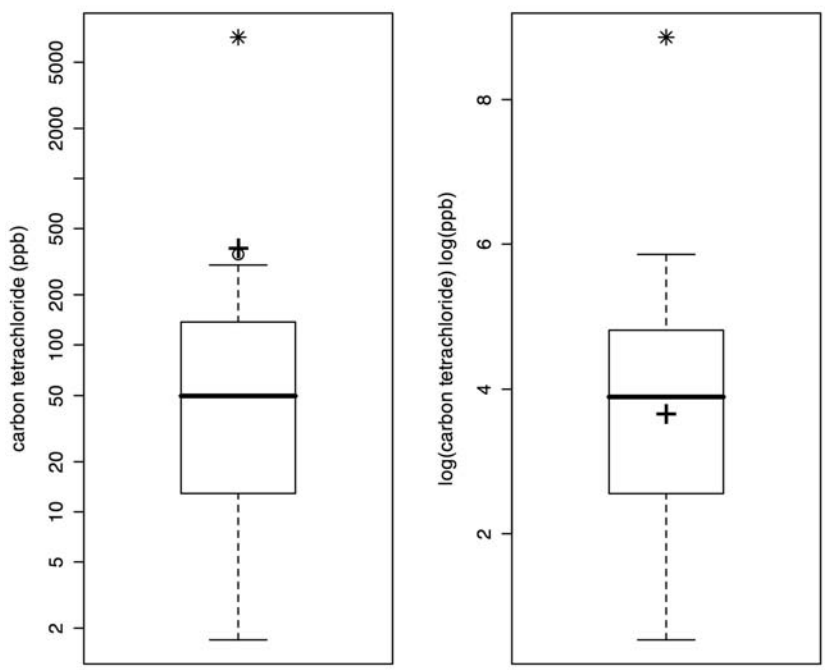
Step 2. Compute the two pairs of lower and upper boundaries using equations (12.1), (12.2), (12.3), and (12.4):

$$LB_1 = 12.9 - 1.5 \times 124.3 = -173.55$$
$$UB_1 = 137.2 + 1.5 \times 124.3 = 323.65$$
$$LB_2 = 12.9 - 3 \times 124.3 = -360$$
$$UB_2 = 137.2 + 3 \times 124.3 = 510.1$$

Step 3. Scan the list of carbon tetrachloride measurements and compare against the boundaries of **Step 2**. It can be seen that the value of 350 from Well 3 is greater than $UB_1$ but lower than $UB_2$, thus qualifying as a mild high outlier. Also, the measurement 7,066 from the same well is higher than $UB_2$ and so qualifies as an extreme high outlier.

Step 4. Because the box plot outlier screening method assumes roughly symmetric data, recompute the box plot on the log-transformed measurements (as shown in **Figure 12-5** alongside a similar box plot of the raw concentrations). Transforming the sample to the log-scale does result in much greater symmetry compared to the original measurement scale. This can be seen in the close similarity between the mean and median on the log-scale box plot. With a more symmetric data set, the mild high outlier from Step 3 disappears, but the extreme high value is still classified as an outlier. ◄

Figure 12-5. Comparative Carbon Tetrachloride Box Plots Indicating Outliers



# 12.3 DIXON'S TEST

## BACKGROUND AND PURPOSE

Dixon's test is helpful for documenting statistical outliers in smaller data sets (*i.e.*, $n \leq 25$). The test is particularly designed for cases where there is only a single high or low outlier, although it can also be adapted to test for multiple outliers. The test falls in the general class of tests for *discordancy* (Barnett and Lewis, 1994). The test statistic for such procedures is generally a ratio: the numerator is the difference between the suspected outlier and some summary statistic of the data set, while the denominator is always a measure of spread within the data. In this version of Dixon's test, the summary statistic in the numerator is an order statistic nearby to the potential outlier (*e.g.*, the second or third most extreme value). The measure of spread is essentially the observed sample range.

If there is more than one outlier in the data set, Dixon's test can be vulnerable to *masking*, at least for very small samples. Masking in the statistical literature refers to the problem of an extreme outlier being missed because one or more additional extreme outliers are also present. For instance, if the data consist of the values {2, 4, 10, 12, 15, 18, 19, 22, 200, 202}, identification of the maximum value (202) as an outlier might fail since the maximum by itself is not extreme with respect to the next highest value (200). However, both of these values are clearly much higher than the rest of the data set and might jointly be considered outliers.

If more than one outlier is suspected, the user is encouraged to consider Rosner's test (**Section 12.4**) as an alternative to Dixon's test, at least if the sample size is 20 or more. If the data set is smaller, Dixon's test should be modified so that the *least extreme* of the suspected outliers is tested first. This will help avoid the risk of masking. The same equations given below can be used, but the data set and sample size should be temporarily reduced to exclude any suspected outliers that are *more* extreme than the one being tested. If a less extreme value is found to be an outlier, then that observation and any more extreme values can also be regarded as outliers. Otherwise, add back the next most extreme value and test it in the same way.

## REQUIREMENTS AND ASSUMPTIONS

Dixon's test is only recommended for sample sizes $n \leq 25$. It assumes that the data set (minus the suspected outlier) is normally-distributed. This assumption should be checked prior to running Dixon's test using a goodness-of-fit technique such as the probability plots described in **Section 12.2**.

## PROCEDURE

Step 1. Order the data set and label the ordered values, $x_{(i)}$.

Step 2. If a "low" outlier is suspected (*i.e.*, $x_{(1)}$), compute the test statistic $C$ using the appropriate equation [12.5] depending on the sample size ($n$):

$$C = \begin{cases} \left(x_{(2)} - x_{(1)}\right) \Big/ \left(x_{(n)} - x_{(1)}\right) & \text{for } 3 \le n \le 7 \\ \left(x_{(2)} - x_{(1)}\right) \Big/ \left(x_{(n-1)} - x_{(1)}\right) & \text{for } 8 \le n \le 10 \\ \left(x_{(3)} - x_{(1)}\right) \Big/ \left(x_{(n-1)} - x_{(1)}\right) & \text{for } 11 \le n \le 13 \\ \left(x_{(3)} - x_{(1)}\right) \Big/ \left(x_{(n-2)} - x_{(1)}\right) & \text{for } 14 \le n \le 25 \end{cases} \qquad [12.5]$$

Step 3. If a "high" outlier is suspected (*i.e.*, $x_{(n)}$), and again depending on the sample size ($n$), compute the test statistic $C$ using the appropriate equation [12.6] as:

$$C = \begin{cases} \left(x_{(n)} - x_{(n-1)}\right) \Big/ \left(x_{(n)} - x_{(1)}\right) & \text{for } 3 \le n \le 7 \\ \left(x_{(n)} - x_{(n-1)}\right) \Big/ \left(x_{(n)} - x_{(2)}\right) & \text{for } 8 \le n \le 10 \\ \left(x_{(n)} - x_{(n-2)}\right) \Big/ \left(x_{(n)} - x_{(2)}\right) & \text{for } 11 \le n \le 13 \\ \left(x_{(n)} - x_{(n-2)}\right) \Big/ \left(x_{(n)} - x_{(3)}\right) & \text{for } 14 \le n \le 25 \end{cases} \qquad [12.6]$$

Step 4. In either case, given the significance level ($\alpha$), determine a critical point for Dixon's test with $n$ observations from **Table 12-1** in **Appendix D**. If $C$ exceeds this critical point, the suspected value should be declared a statistical outlier and investigated further (see discussion in **Chapter 6**).

### ►EXAMPLE 12-3

Use the data from **Example 12-1** in Dixon's test to determine if the anomalous high value is a statistical outlier at an $\alpha = 0.05$ level of significance.

#### SOLUTION

Step 1. In **Example 12-1**, probability plots of the carbon tetrachloride data indicated that the highest value might be an outlier, but that the distribution of the measurements was more nearly lognormal than normal. Since the sample size $n = 20$, Dixon's test can be used on the logged observations. Logging the values and ordering them leads to the following table:

| Order | Concentration (ppb) | Logged Concentration |
|---|---|---|
| 1 | 1.7 | 0.531 |
| 2 | 3.2 | 1.163 |
| 3 | 6.5 | 1.872 |
| 4 | 7.3 | 1.988 |
| 5 | 12.1 | 2.493 |
| 6 | 13.7 | 2.617 |
| 7 | 15.6 | 2.747 |
| 8 | 16.2 | 2.785 |
| 9 | 35.1 | 3.558 |
| 10 | 41.6 | 3.728 |
| 11 | 57.9 | 4.059 |
| 12 | 59.7 | 4.089 |
| 13 | 68.4 | 4.225 |
| 14 | 70.1 | 4.250 |
| 15 | 75.4 | 4.323 |
| 16 | 199.0 | 5.293 |
| 17 | 275.0 | 5.617 |
| 18 | 302.0 | 5.710 |
| 19 | 350.0 | 5.878 |
| 20 | 7066.0 | 8.863 |

Step 2. Because a high outlier is suspected and $n$ = 20, use the last option of equation [12.6] to calculate the test statistic $C$:

$$C = \frac{8.863 - 5.710}{8.863 - 1.872} = 0.451$$

Step 3. With $n$ = 20 and $\alpha$ = .05, the critical point from **Table 12-1** in **Appendix D** is equal to 0.450. Since the test statistic $C$ exceeds this critical point, the extreme high value can be declared a statistical outlier. Before excluding this value from further analysis, however, a valid explanation for this unusually high value should be sought. Otherwise, the outlier may need to be treated as an extreme but valid concentration measurement. ◄

## 12.4 ROSNER'S TEST

### BACKGROUND AND PURPOSE

Rosner's test (Rosner, 1975) is a useful method for identifying multiple outliers in moderate to large-sized data sets. The approach developed in Rosner's method is known as a *block*-style test. Instead of testing for outliers one-by-one in a consecutive manner from most extreme to least extreme (*i.e.*, most to least suspicious), the data are examined first to identify the total number of possible outliers, $k$. Once $k$ is determined, the set of possible outliers is tested together as a block. If the test is significant, all $k$ measurements are regarded as statistical outliers. If not, the set of possible outliers is reduced by one and the test repeated on the smaller block. This procedure is iterated until either a set of outliers is identified

or none of the observations are labeled an outlier. By testing outliers in blocks instead of one-by-one, Rosner's test largely avoids the problem of *masking* of one outlier by another (as discussed in **Section 12.3** regarding Dixon's test).

Although Rosner's test avoids the problem of masking when multiple outliers are present in the same data set, it is not immune to the related problem of *swamping*. A good discussion is found in Barnett and Lewis, 1994, Outliers in Statistical Data (3rd Edition), p. 236. Swamping refers to a block of measurements all being *labeled* as outliers even though only *some* of the observations are actually outliers. This can occur with Rosner's test especially if all the outliers tend to be at one end of the data set (*e.g.*, as upper extremes). The difficulty is in properly identifying the total number of possible outliers ($k$), which can be low outliers, high outliers, or some combination of the two extremes. If $k$ is made too large, swamping may occur. Again, the user is reminded to always do a preliminary screening for outliers via box plots (**Section 12.2**) and probability plots (**Section 12.1**).

## REQUIREMENTS AND ASSUMPTIONS

Rosner's test is recommended when the sample size ($n$) is 20 or larger. The critical points provided in **Table 12-2** in **Appendix D** can be used to identify from 2 to 5 outliers in a given data set. Like Dixon's test, Rosner's method assumes the underlying data set (minus any outliers) is normally distributed. If a probability plot of the data exhibits significant bends or curves, the data should first be transformed (*e.g.*, via a logarithm) and then re-plotted. The formal test for outliers should only be performed on (outlier-deleted) data sets that have been approximately normalized.

A potential drawback of Rosner's test is that the user must first identify the maximum number of potential outliers ($k$) prior to running the test. Therefore, this requirement makes the test ill-advised as an automatic outlier screening tool, and somewhat reliant on the user to identify candidate outliers.

## PROCEDURE

Step 1. Order the data set and denote the ordered values $x_{(i)}$. Then by simple inspection, identify the maximum number of possible outliers, $r_0$.

Step 2. Compute the sample mean and standard deviation of all the data; denote these values by $\bar{x}^{(0)}$ and $s^{(0)}$. Then determine the measurement furthest from $\bar{x}^{(0)}$ and denote it $y^{(0)}$. Note that $y^{(0)}$ could be either a potentially low or a high outlier.

Step 3. Delete $y^{(0)}$ from the data set and compute the sample mean and standard deviation from the remaining observations. Label these new values $\bar{x}^{(1)}$ and $s^{(1)}$. Again find the value in this reduced data set furthest from $\bar{x}^{(1)}$ and label it $y^{(1)}$.

Step 4. Delete $y^{(1)}$, recompute the mean and standard deviation, and continue this process until all $r_0$ potential outliers have been removed. At this point, the following set of statistics will be available:

$$\left[\bar{x}^{(0)}, s^{(0)}, y^{(0)}\right], \left[\bar{x}^{(1)}, s^{(1)}, y^{(1)}\right], \ldots, \left[\bar{x}^{(r_0-1)}, s^{(r_0-1)}, y^{(r_0-1)}\right] \qquad [12.7]$$

Step 5. Now test for $r$ outliers (where $r \le r_0$) by iteratively computing the test statistic:

$$R_{r-1} = \left| y^{(r-1)} - \bar{x}^{(r-1)} \right| \Big/ s^{(r-1)} \qquad [12.8]$$

First test for $r_0$ outliers. If the test statistic $R_{r_0-1}$ in equation [12.8] exceeds the first critical point from **Table 12-2** in **Appendix D** based on sample size ($n$) and the Type I error ($\alpha$), conclude there are $r_0$ outliers. If not, test for $r_0$–1 outliers in the same fashion using the next critical point, continuing until a certain number of outliers have either been identified or Rosner's test finds no outliers at all.

►EXAMPLE 12-4

Consider the following series of 25 background napthalene measurements (in ppb). Use Rosner's test to determine whether any of the values should be deemed statistical outliers.

| | **Naphthalene Concentrations (ppb)** | | | | |
|---|---|---|---|---|---|
| Qtr | BW-1 | BW-2 | BW-3 | BW-4 | BW-5 |
| 1 | 3.34 | 5.59 | 1.91 | 6.12 | 8.64 |
| 2 | 5.39 | 5.96 | 1.74 | 6.05 | 5.34 |
| 3 | 5.74 | 1.47 | 23.23 | 5.18 | 5.53 |
| 4 | 6.88 | 2.57 | 1.82 | 4.43 | 4.42 |
| 5 | 5.85 | 5.39 | 2.02 | 1.00 | 35.45 |

SOLUTION

Step 1. Screening with probability plots of the combined data indicates a less than linear fit with both the raw measurements and log-transformed data (see **Figures 12-6 and 12-7**); two points appear rather discrepant from the rest. Correlation coefficients for these plots are 0.740 on the concentration scale and 0.951 on the log-scale. Re-plotting after removing the two possible outliers gives a substantially improved correlation on the concentration scale of 0.958 but reduces the log-scale correlation to 0.929. Normality appears to be a slightly better default distribution for the outlier-deleted data set. Run Rosner's test on the original data with $k$ = 2 possible outliers.

Step 2. Compute the mean and standard deviation of the complete data set. Then identify the observation farthest from the mean. These results are listed, along with the ordered data, in the table below. After removing the farthest value (35.45), recompute the mean and standard deviation on the remaining values and again identify the most discrepant observation (23.23). Repeat this process one more time so that both suspected outliers have been removed (see table below).

Step 3. Now test for 2 joint outliers by computing Rosner's statistic on subset $SS_{k-1} = SS_1$ using equation [12.8]:

$$R_1 = \frac{23.23 - 5.23}{4.326} = 4.16$$

Figure 12-6. Napthalene Probability Plot

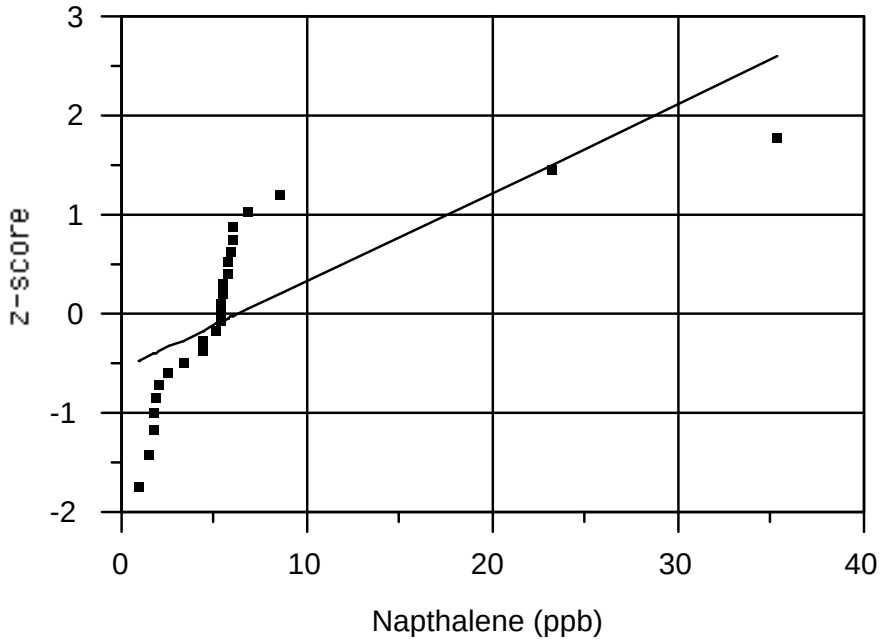
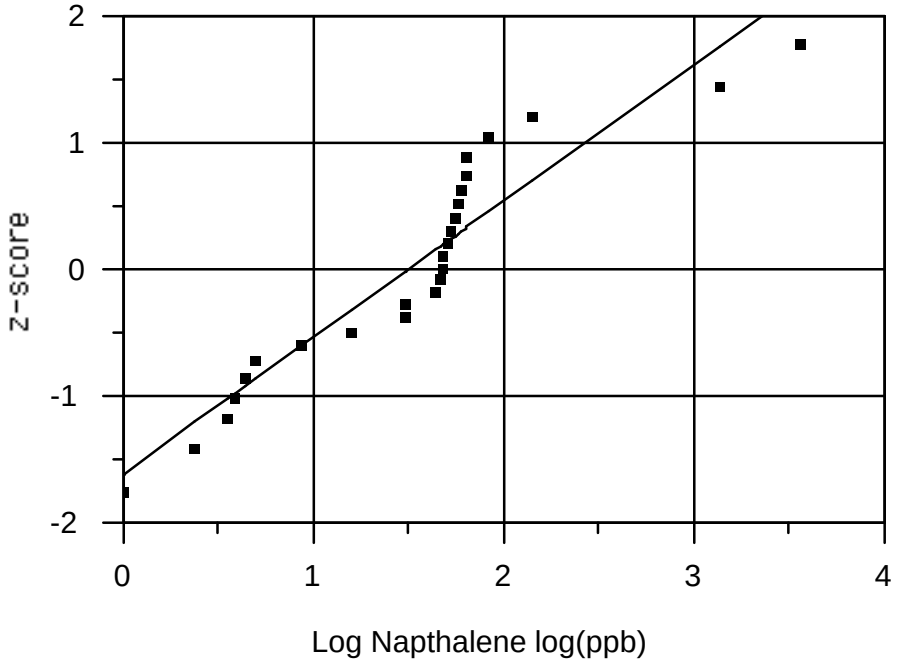



Figure 12-7. Log Napthalene Probability Plot



| Successive Naphthalene Subsets ($SS_i$) | | |
|---|---|---|
| $SS_0$ | $SS_1$ | $SS_2$ |
| 1.00 | 1.00 | 1.00 |
| 1.47 | 1.47 | 1.47 |
| 1.74 | 1.74 | 1.74 |
| 1.82 | 1.82 | 1.82 |
| 1.91 | 1.91 | 1.91 |
| 2.02 | 2.02 | 2.02 |
| 2.57 | 2.57 | 2.57 |
| 3.34 | 3.34 | 3.34 |
| 4.42 | 4.42 | 4.42 |
| 4.43 | 4.43 | 4.43 |
| 5.18 | 5.18 | 5.18 |
| 5.34 | 5.34 | 5.34 |
| 5.39 | 5.39 | 5.39 |
| 5.39 | 5.39 | 5.39 |
| 5.53 | 5.53 | 5.53 |
| 5.59 | 5.59 | 5.59 |
| 5.74 | 5.74 | 5.74 |
| 5.85 | 5.85 | 5.85 |
| 5.96 | 5.96 | 5.96 |
| 6.05 | 6.05 | 6.05 |
| 6.12 | 6.12 | 6.12 |
| 6.88 | 6.88 | 6.88 |
| 8.64 | 8.64 | 8.64 |
| 23.23 | 23.23 | |
| 35.45 | | |
| $\bar{x}_0$ = 6.44 | $\bar{x}_1$ = 5.23 | $\bar{x}_2$ = 4.45 |
| $s_0$ = 7.379 | $s_1$ = 4.326 | $s_2$ = 2.050 |
| $y_0$ = 35.45 | $y_1$ = 23.23 | $y_2$ = 8.64 |

Step 4. Given $\alpha = 0.05$, a sample size of $n = 25$, and $k = 2$, the first critical point in **Table 12-2** in **Appendix D** equals 2.83 for $n = 20$ and 3.05 for $n = 30$. The value $R_1$ in Step 3 is larger than either of these critical points, so both suspected values may be declared statistical outliers by Rosner's test at the 5% significance level. Before excluding these values from further analysis, however, a valid explanation for them should be found. Otherwise, treat the outliers as extreme but valid concentration measurements.

Note: had $R_1$ been *less* than these values, a test could still be run for a *single* outlier using the second critical point for each sample size (or a critical point interpolated between them). ◄

The guidance considers Dixon's and Rosner's outlier evaluation methods preferable for groundwater monitoring data situations, when assumptions of normality are reasonable and data are quantified. We did not include the older method found in the 1989 guidance based on ASTM paper E178-75, which can still be used as an alternative. Where data do not appear to be fit by a normal or transformably normal distribution, other robust outlier evaluation methods can be considered from the wider statistical literature. The literature will also need to be consulted when data contains non-detect values along with potential outliers.

# CHAPTER 13. SPATIAL VARIABILITY

13.1 INTRODUCTION TO SPATIAL VARIATION ..... 13-1
13.2 IDENTIFYING SPATIAL VARIABILITY ..... 13-2
*13.2.1 Side-by-Side Box Plots ..... 13-2*
*13.2.2 One-Way Analysis of Variance for Spatial Variability ..... 13-5*
13.3 USING ANOVA TO IMPROVE PARAMETRIC INTRAWELL TESTS ..... 13-8

This chapter discusses a type of ***statistical dependence*** in groundwater monitoring data known as ***spatial variability***. Spatial variability exists when the distribution or pattern of concentration measurements changes from well location to well location (most typically in the form of differing mean levels). Such variation may be natural or ***synthetic***, depending on whether it is caused by natural or anthropogenic factors. Methods for ***identifying*** spatial variation are detailed via the use of box plots (**Section 13.2.1**) and analysis of variance [ANOVA] (**Section 13.2.2**). Once identified, ANOVA can sometimes be employed to construct more powerful intrawell background limits. This topic is addressed in **Section 13.3**.

## 13.1 INTRODUCTION TO SPATIAL VARIATION

***Spatial dependence***, ***spatial variation*** or ***variability***, and ***spatial correlation*** are closely related concepts. All refer to the notion of measurement levels that vary in a structured way as a function of the ***location*** of sampling. Although spatial variation can apply to any statistical characteristic of the underlying population (including the population variance or upper percentiles), the usual sense in groundwater monitoring is that ***mean*** levels of a given constituent vary from one well to the next.

Standard geostatistical models posit that an area exhibits positive spatial correlation if any two sampling locations share a greater similarity in concentration level the closer the distance between them, and more dissimilarity the further apart they are. Such models have been applied to both groundwater and soil sampling problems, but are not applicable in all geological configurations. It may be, for instance, that mean concentration levels differ across wells but vary in a seemingly random way with no apparent connection to the distance between the sampling points. In that case, the concentrations between pairs of wells are not correlated with distance, yet the measurements within each well are strongly associated with the mean level at that particular location, whether due to a change in soil composition, hydrological characteristics or some other factor. In other words, spatial ***variation*** may exist even when spatial ***correlation*** does not.

Spatial variation is important in the guidance context since substantial differences in mean concentration levels between different wells can ***invalidate*** interwell, upgradient-to-downgradient comparisons and point instead toward ***intrawell*** tests (**Chapter 6**). Not all spatial variability is natural. Average concentration levels can vary from well to well for a variety of reasons.

In this guidance, a distinction is occasionally made between ***natural*** versus ***synthetic*** spatial variation. Natural spatial variability refers to a pattern of changing mean levels in groundwater associated with normal geochemical behavior unaffected by human activities. Natural spatial variability

is not an indication of groundwater contamination, even if concentrations at one or more compliance wells exceed (upgradient) background. In contrast, synthetic spatial variability is related to human activity. Sources can include recent releases affecting compliance wells, migration of contaminants from off-site sources, or historic contamination at certain wells due to past industrial activity or pre-RCRA waste disposal. Whether natural or synthetic, techniques and test methods for dealing with spatial variation will still be identical from a purely statistical standpoint. It is interpreting the testing outcomes which will necessitate a consideration of why the spatial variation occurs.

The goal of groundwater analysis is not simply to identify significant concentration differences among monitoring wells at compliance point locations. It is also to determine *why* those differences exist. Especially with prior groundwater contamination, regulatory decisions outside the scope of this guidance need to address the problem. In some cases, compliance/assessment monitoring or remedial action may be warranted. In other cases, chronic contamination from offsite sources may simply have to be considered as the current background condition at a given location. At least the ability to attribute certain mean differences to *natural* spatial variation allows the range of potential concerns to be somewhat narrowed. Of course, deciding that an observed pattern of spatial variation is natural and not synthetic may not be easy. Ultimately, expert judgment and knowledge concerning site hydrology, geology and geochemistry are important in providing more definitive answers.

One statistical approach to use when a site has multiple, non-impacted background wells is to conduct a one-way ANOVA for inorganic constituents on those wells. Substantial differences among the mean levels at a set of *uncontaminated* sampling locations are suggestive of natural spatial variability. At a true 'greenfield' site, ANOVA can be run on all the wells — both background and compliance — after a few preliminary sampling rounds have been collected.

The Unified Guidance offers two basic tools to explore and test for spatial correlation. The first, side-by-side box plots (**Section 13.2.1**), provides a quick screen for possible spatial variation. When multiple well data are plotted on the same concentration axis, noticeably staggered boxes are often an indication of significantly different mean levels.

A more formal test of spatial variation is the one-way ANOVA (**Section 13.3.2**). When significant spatial variation exists and an intrawell test strategy is pursued, one-way ANOVA can also be used to adjust the standard deviation estimate used in forming intrawell prediction and control chart limits, and to increase the effective sample size of the test, via the *degrees of freedom*. This is discussed in **Section 13.3**.

## 13.2 IDENTIFYING SPATIAL VARIABILITY

### 13.2.1 SIDE-BY-SIDE BOX PLOTS

#### BACKGROUND AND PURPOSE

Box plots for graphing side-by-side statistical summaries of multiple wells were introduced in **Chapter 9**. They are also discussed in **Chapter 11** as an initial screen for differences in population variances and as a tool to check the assumption of equal variances in ANOVA. They can further be employed to screen for possible spatial variation in mean levels. While variability in a sample from a given well is roughly indicated by the length of the box, the average concentration level is indicated by the position of the box relative to the concentration axis. Many standard box plot software routines

display both the sample median value and the sample mean on each box, so these values may be compared from well to well. A high degree of staggering in the box positions is then indicative of potentially significant spatial variation.

Since side-by-side box plots provide a picture of the variability at each well, the extent to which apparent differences in mean levels seem to be real rather than chance fluctuations can be examined. If the boxes are staggered but there is substantial overlap between them, the degree of spatial variability may not be significant. A more formal ANOVA might still be warranted as a follow-up test, but side-by-side box plots will offer a initial sense of how spatially variable the groundwater data appear.

## REQUIREMENTS, ASSUMPTIONS AND PROCEDURE

Requirements, assumptions and the procedure for box plots are outlined in **Chapter 9, Section 9.2**.

## ►EXAMPLE 13-1

Quarterly dissolved iron concentrations measured at each of six upgradient wells are listed below. Construct side-by-side box plots to initially screen for the presence of spatial variability.

| | **Iron Concentrations (ppm)** | | | | | |
|---|---|---|---|---|---|---|
| Date | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| Jan 1997 | 57.97 | 46.06 | 100.48 | 34.12 | 60.95 | 83.10 |
| Apr 1997 | 54.05 | 76.71 | 170.72 | 93.69 | 72.97 | 183.09 |
| Jul 1997 | 29.96 | 32.14 | 39.25 | 70.81 | 244.69 | 198.34 |
| Oct 1997 | 46.06 | 68.03 | 52.98 | 83.10 | 202.35 | 160.77 |
| Mean | 47.01 | 55.71 | 90.86 | 70.43 | 145.24 | 156.32 |
| Median | 55.06 | 57.04 | 76.73 | 76.96 | 137.66 | 171.93 |

### SOLUTION

Step 1. Determine the median, mean, lower and upper quartiles of each well. Then plot these against a concentration axis to form side-by-by side box plots (**Figure 13-1**) using the procedure in **Section 9.2**. .

Step 2. From this plot, the means and medians at the last two wells (Wells 5 and 6) appear elevated above the rest. This is a possible indication of spatial variation. However, the variances as represented by the box lengths also appear to differ, with the highest means associated with the largest boxes. A transformation of the data should be attempted and the data re-plotted. Spatial variability is only a significant problem if it is apparent on the scale of the data actually used for statistical analysis.

Step 3. Take the logarithm of each measurement as in the table below. Recompute the mean, median, lower and upper quartiles, and then re-construct the box plot as in **Figure 13-2**.

| Date | Log Iron Concentrations log(ppm) | | | | | |
|---|---|---|---|---|---|---|
| | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| Jan 1997 | 4.06 | 3.83 | 4.61 | 3.53 | 4.11 | 4.42 |
| Apr 1997 | 3.99 | 4.34 | 5.14 | 4.54 | 4.29 | 5.21 |
| Jul 1997 | 3.40 | 3.47 | 3.67 | 4.26 | 5.50 | 5.29 |
| Oct 1997 | 3.83 | 4.22 | 3.97 | 4.42 | 5.31 | 5.08 |
| Mean | 3.82 | 3.96 | 4.35 | 4.19 | 4.80 | 5.00 |
| Median | 3.70 | 4.02 | 4.29 | 4.34 | 4.80 | 5.14 |

Figure 13-1. Side-by-Side Iron Box Plots

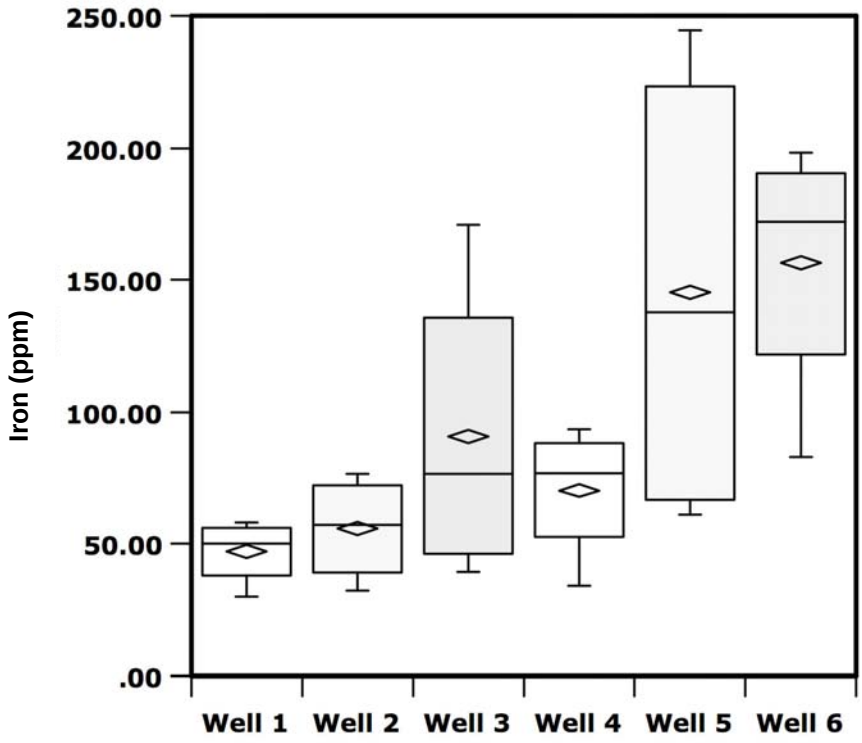
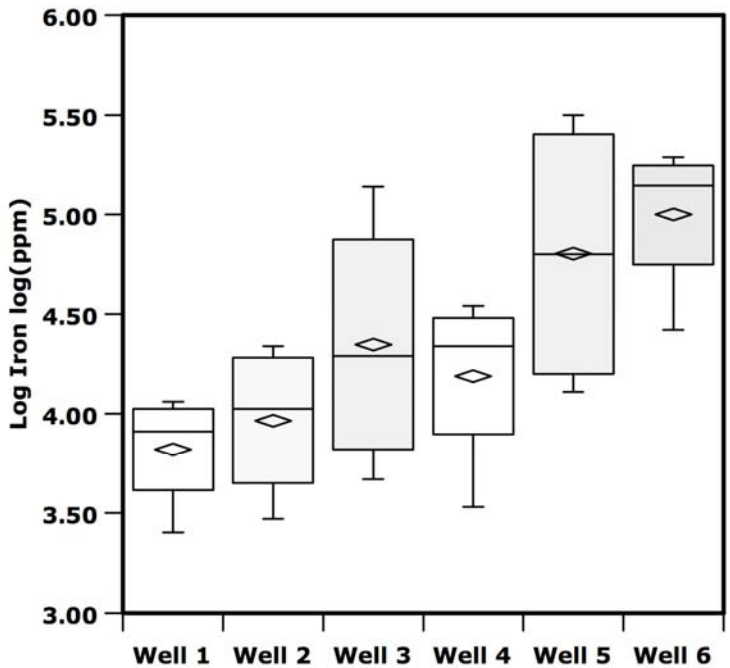


Figure 13-2. Side-by-Side Log(Iron) Box Plots




Step 4. While more nearly similar on the log-scale, the means and medians are still elevated in Wells 5 and 6. Since the differences in box lengths are much less on the log-scale, the log transformation has worked to somewhat stabilize the variances. These data should be tested formally for significant spatial variation using an ANOVA, probably on the log-scale. ◄

## 13.2.2 ONE-WAY ANALYSIS OF VARIANCE FOR SPATIAL VARIABILITY

### PURPOSE AND BACKGROUND

**Chapter 17** presents Analysis of Variance [ANOVA] in greater detail. When using ANOVA to check for spatial variability, the observations from each well are taken as a single group. Significant differences between data groups represent monitoring wells with different mean concentration levels. The lack of significant well mean differences may afford an opportunity to pool the data for larger background sizes and conduct interwell detection monitoring tests.

ANOVA used for this purpose should be performed either on a set of multiple non-impacted upgradient wells, or on historically uncontaminated compliance and upgradient background wells. If significant mean differences exist among naturally occurring constituent data at upgradient wells, natural spatial variability is the likely reason. Synthetic consitituents in upgradient wells might also exhibit spatial differences if affected by an offsite- plume. Presumably, if the flow gradient has been correctly

assessed and no migration of contaminants from off-site has occurred, differences in mean levels across upgradient wells ought to signal the influence of factors not attributable to a monitored release. A similar, but potentially weaker, argument can be made if spatial differences exist between uncontaminated historical data at compliance wells. The lack of spatial differences between uncontaminated compliance and upgradient background well data, may again allow for even larger background sample sizes.

## REQUIREMENTS AND ASSUMPTIONS

The basic assumptions and data requirements for one-way ANOVA are presented in **Section 17.1**. If the assumption that the observations are statistically independent over time is not met, both identifying spatial variability using ANOVA as well as improving intrawell prediction limits and control charts can be impacted. It is usually difficult to verify that the measurements are temporally independent with only a limited number of observations per well. This potential problem can be somewhat minimized by collecting samples far enough apart in time to guard against autocorrelation. Another option is to construct a parallel time series plot (**Chapter 14**) to look for time-related effects or dependencies occurring simultaneously across the set of wells.

If a significant temporal dependence or autocorrelation exists, the one-way ANOVA can still identify well-to-well mean level differences. But the power of the test to do so is lessened. If a parallel time series plot indicates a potentially strong time-related effect, a two-way ANOVA including temporal effects can be performed to test and correct for a significant temporal factor. This slightly more complicated procedure is discussed in Davis (1994).

Another key assumption of parametric ANOVA is that the residuals are normal or can be normalized. If a normalizing transformation cannot be found, a test for spatial variability can be made using the Kruskal-Wallis non-parametric ANOVA (**Chapter 17**). As long as the measurements can be ranked, average ranks that differ significantly across wells provide evidence of spatial variation.

## PROCEDURE

Step 1. Assuming there are $p$ distinct wells to test, designate the measurements from each well as a separate group for purposes of computing the ANOVA. Then follow **Steps 1** through **7** of the procedure in **Section 17.1.1** to compute the overall $F$-statistic and the quantities of the ANOVA table in **Figure 13-3** below.

Figure 13-3. One-Way Parametric ANOVA Table

| Source of Variation | Sums of Squares | Degrees of Freedom | Mean Squares | *F*-Statistic |
|---|---|---|---|---|
| Between Wells | $SS_{wells}$ | $p-1$ | $MS_{wells} = SS_{wells}/(p-1)$ | $F = MS_{wells}/MS_{error}$ |
| Error (within wells) | $SS_{error}$ | $n-p$ | $MS_{error} = SS_{error}/(n-p)$ | |
| Total | $SS_{total}$ | $n-1$ | | |

Step 2. To test the hypothesis of equal means for all $p$ wells, compare the $F$-statistic from Step 1 to the $\alpha$-level critical point found from the $F$-distribution with ($p$–1) and ($n$–$p$) degrees of freedom in **Table 17-1** of **Appendix D**. Usually $\alpha$ is taken to be 5%, so that the needed comparison value equals the upper 95th percentage point of the $F$-distribution. If the observed $F$-statistic exceeds the critical point ($F_{.95,p-1,n-p}$), reject the hypothesis of equal well population means and conclude there is significant spatial variability. Otherwise, the evidence is insufficient to conclude there are significant differences between the means at the $p$ wells.

►EXAMPLE 13-2

The iron concentrations in **Example 13-1** show evidence of spatial variability in side-by-side box plots. Tested for equal variances and normality, these same data are best fit by a lognormal distribution. The statistics for natural logarithms of the iron measurements are shown below; individual log data are provided in the **Example 13-1** second table. Compute a one-way parametric ANOVA to determine whether there is significant spatial variation at the $\alpha$ = .05 significance level.

| | **Log Iron Concentration Statistics log(ppm)** | | | | | |
|---|---|---|---|---|---|---|
| Date | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| N | 4 | 4 | 4 | 4 | 4 | 4 |
| Mean | 3.820 | 3.965 | 4.348 | 4.188 | 4.802 | 5.000 |
| SD | 0.296 | 0.395 | 0.658 | 0.453 | 0.704 | 0.396 |
| | | | Grand Mean = 4.354 | | | |

SOLUTION

Step 1. With 6 wells and 4 observations per well, $n_i$ = 4 for all the wells. The total sample size is $n$ = 24 and $p$ = 6. Compute the (overall) grand mean and the sample mean concentrations in each of the well groups using equations [17.1] and [17.2]. These values are listed (along with each group's standard deviation) in the above table.

Step 2. Compute the sum of squares due to well-to-well differences using equation [17.3]:

$$SS_{wells} = \left[4(3.820)^2 + 4(3.965)^2 + \ \ldots \ + 4(5.000)^2\right] - 24(4.354)^2 = 4.294$$

This quantity has (6 – 1) = 5 degrees of freedom.

Step 3. Compute the corrected total sum of squares using equation [17.4] with ($n$ – 1) = 23 df:

$$SS_{total} = \left[(4.06)^2 + \ \ldots \ + (5.08)^2\right] - 24(4.354)^2 = 8.934$$

Step 4. Obtain the within-well or error sum of squares by subtraction using equation [17.5]:

$$SS_{error} = 8.934 - 4.294 = 4.640$$

This quantity has ($n$ – $p$) = 24–6 = 18 degrees of freedom.

Step 5. Compute the well and error mean sum of squares using equations [17.6] and [17.7]:

$$MS_{wells} = 4.294/5 = 0.859$$

$$MS_{error} = 4.640/18 = 0.258$$

Step 6. Construct the *F*-statistic and the one-way ANOVA table, using **Figure 13-3** as a guide:

| Source of Variation | Sums of Squares | Degrees of Freedom | Mean Squares | *F*-Statistic |
|---|---|---|---|---|
| Between Wells | 4.294 | 5 | 0.859 | *F* = 0.859/0.258=3.33 |
| Error (within wells) | 4.640 | 18 | 0.258 | |
| Total | 8.934 | 23 | | |

Step 7. Compare the observed *F*-statistic of 3.33 against the critical point taken as the upper 95th percentage point from the *F*-distribution with 5 and 18 degrees of freedom. Using **Table 17-1** of **Appendix D**, this gives a value of $F_{.95,5,18}$ = 2.77. Since the *F*-statistic exceeds the critical point, the null hypothesis of equal well means can be rejected, suggesting the presence of significant spatial variation. ◄

## 13.3 USING ANOVA TO IMPROVE PARAMETRIC INTRAWELL TESTS

### BACKGROUND AND PURPOSE

Constituents that exhibit significant spatial variability usually should be formally tested with intrawell procedures such as a prediction limit or control chart. Historical data from each compliance well are used as background for these tests instead of from upgradient wells. At an early stage of intrawell testing, there may only be a few measurements per well which can be designated as background. Depending on the number of statistical tests that need to be performed across the monitoring network, available intrawell background at individual compliance wells may not provide sufficient statistical power or meet the false positive rate criteria (**Chapter 19**).

One remedy first suggested by Davis (1998) can increase the *degrees of freedom* of the test by using one-way ANOVA results (**Section 13.2**) from a number of wells to provide an alternate estimate of the average intrawell variance. In constructing a parametric intrawell prediction limit for a single compliance well, the intrawell background of sample size *n* is used to compute a well-specific sample mean ($\bar{x}$). The intrawell standard deviation (*s*) is replaced by the root mean squared error [RMSE] component from an ANOVA of the intrawell background associated with a series of compliance and/or background wells.[1] This raises the degrees of freedom from (*n*–1) to (*N*–*p*), where *N* is the total sample size across the group of wells input to the ANOVA and *p* is the number of distinct wells.

[1] RMSE is another name for the square root of the mean error sum of squares ($MS_{error}$) in the ANOVA table of **Figure 13-3**.

As an example of the difference this adjustment can make, consider a site with 6 upgradient wells and 15 compliance wells. Assuming $n$ = 6 observations per well that have been collected over the last year, a total of 36 potential background measurements are available to construct an interwell test. If there is significant natural spatial variation in the mean levels from well to well, an interwell test is probably not appropriate. Switching to an intrawell method is the next best solution, but with only six observations per compliance well, either the power of an intrawell test to identify contaminated groundwater is likely to be quite low (even with retesting) or the site-wide false positive rate [SWFPR] will exceed the recommended target.

If the six upgradient wells were tested for spatial variability using a one-way ANOVA (presuming that the equal variance assumption is met), the degrees of freedom [*df*] associated with the mean error sum of squares term is (6 wells × 5 *df* per well) = 30 *df* (see **Section 13.2**). Thus by substituting the RMSE in place of each compliance well's intrawell standard deviation ($s$), the degrees of freedom for the modified intrawell prediction or control chart limit is 30 instead of 5.

ANOVA can be usefully employed in this manner since the RMSE is very close to being a weighted average of the individual well sample standard deviations. As such, it can be considered a measure of average within-well variability across the wells input to the ANOVA. Substituting the RMSE for $s$ at an individual well consequently provides a better estimate of the typical within-well variation, since the RMSE is based on levels of fluctuation averaged across several wells. In addition, the number of observations used to construct the RMSE is much greater than the $n$ values used to compute the intrawell sample standard deviation ($s$). Since both statistical measures are estimates of within-well variation, the RMSE with its larger degrees of freedom is generally a superior estimate if certain assumptions are met.

## REQUIREMENTS AND ASSUMPTIONS

Using ANOVA to bolster parametric intrawell prediction or control chart limits will not work at every site or for every constituent. Replacement of the well-specific, intrawell sample standard deviation (s) by the RMSE from ANOVA assumes that the true within-well variability is *approximately the same* at all the wells for which an intrawell background limit (*i.e.*, prediction or control chart) will be constructed, and not just those wells tested in the ANOVA procedure. This last assumption can be difficult to verify if the ANOVA includes only background or upgradient wells. But to the extent that uncontaminated intrawell background measurements from compliance point wells can be included, the ANOVA should be run on all or a substantial fraction of the site's wells (excluding those which might already be contaminated). Whatever mix of upgradient and downgradient wells are included in the ANOVA, the purpose of the procedure is *not* to identify groundwater contamination, but rather to compute a better and more powerful estimate of the average intrawell standard deviation.

For the ANOVA to be valid and the RMSE to be a reasonable estimate of average within-well variability, a formal check of the equal variance assumption should be conducted using **Chapter 11** methods. A spatially variable constituent will often exhibit well-specific standard deviations that increase with the well-specific mean concentration. Equalizing the variances in these cases will require a data transformation, with an ANOVA conducted on the transformed data. Ultimately, any transformation applied to the wells in the ANOVA *also* need to be applied to intrawell background before computing intrawell prediction or control chart limits. The *same transformation* has to be appropriate for *both* sets

of data (*i.e.*, wells included in ANOVA *and* intrawell background at wells for which background limits are desired).

Even when the ANOVA procedure described in this section is utilized, the resulting intrawell limits should also be designed to incorporate retesting. When intrawell background is employed to estimate both a well-specific background mean ($\bar{x}$) and well-specific standard deviation ($s$), the **Appendix D** tables associated with **Chapters 19** and **20** can be used to look up the intrawell sample size ($n$) and number of wells ($w$) in the network in order to find a prediction or control chart multiplier that meets the targeted SWFPR and has acceptable statistical power. However, these tables implicitly assume that the degrees of freedom [$df$] associated with the test is equal to ($n$–1). The ANOVA method of this section results in a much larger $df$, and more importantly, in a $df$ that does not 'match' the intrawell sample size ($n$).

Consequently, the parametric multipliers in the **Appendix D** tables cannot be directly used when constructing prediction or control chart limits with retesting. Instead, a multiplier must be computed for the specific combination of $n$ and $df$ computed as a result of the ANOVA. Tabulating all such possibilities would be prohibitive. For prediction limits, the Unified Guidance recommends the free-of-charge, open source **R** statistical computing environment. A pre-scripted program is included in **Appendix C** that can be run in **R** to calculate appropriate prediction limit multipliers, once the user has supplied an intrawell sample size ($n$), network size ($w$), and type of retesting scheme.

If guidance users are unable to utilize the **R-script** approach, the following approximation for the well-specific prediction limit κ-factors is suggested based on EPA Region 8 Monte Carlo evaluations. Given a per- test confidence level of $1-\alpha$, $r$ total tests of $w \cdot c$ well-constituents, an individual well size $n_i$, a pooled variance sample size of $n_{df} = df + 1$, and $\kappa_{ndf,1-\alpha}$ obtained from annual intrawell Unified Guidance tables, the individual well $\kappa_{ni,1-\alpha}$ factor can be estimated using the following equation:

$$\kappa_{n_i,1-\alpha} = \kappa_{n_{df},1-\alpha} \cdot \left[ \frac{(n_i + \mu) \cdot n_{df}}{n_i \cdot (n_{df} + \mu)} \right]^{m^* = \frac{A \cdot n_i^b}{r^c}}$$

where $\mu$ =1 for future $1{:}m$ observations or $\mu$ is the size of a future mean. The value of $m^*$ is specific to each of the nine parametric prediction limit tests and is a function of the three coefficients $A$, $b$ and $c$, individual well sample size $n_i$ and $r$ tests. For a 1:1 test of future means or observations, the equation is exact; for higher order 1:$m$ tests, the results are approximate.[2] The equation is also useful in

[2] For each of the nine prediction limit tests, the following coefficients ($A$, $b$ & $c$) are recommended: a 1:2 future values test (1.01, .0524 & .0158); a 1:3 test (1.63, .108 & .0407); a 1:4 test (2.41, .157 & .0668); the modified California plan (1.36, .103 & .0182); a 1:1 mean size 2 test (.5, 0 & 0); a 1:2 mean size 2 test (.898, .0856 & .0172); a 1:3 mean size 2 test (1.27, .168 & .0363); a 1:1 mean size 3 test (.5, 0 & 0); and a 1:2 mean size 3 test (.817, .108 & .0158). %. The coefficients were obtained from regression analysis; approximation values were compared with R-script values for κ-factors. In 1260 comparisons of the seven tests using repeat values ($m > 1$), 86% of the approximations lay within or equal to ± 1% of the true value and 96% within or equal to ± 2%. The 1:4 test had the greatest variability, but all values lay within ± 4%. 81% of the values lay within or equal to ± .01 κ-units and 93% less than or equal to ± .02 units.

gauging **R-script** method results. Another virtue of this equation is that it can be readily applied to different individual well sample sizes based on the common $\kappa_{ndf,1-\alpha}$ for pooled variance data.

A less elegant solution is available for intrawell control charts. Currently, an appropriate multiplier needs to be simulated via Monte Carlo methods. The approach is to simulate separate normally-distributed data sets for the background mean based on $n$ measurements, and the background standard deviation based on $df + 1$ measurements. Statistical independence of the sample mean ($\bar{x}$) and standard deviation (s) for normal populations allows this to work. With the background mean and standard deviation available, a series of possible multipliers ($h$) can be investigated in simulations of control chart performance. The multiplier which meets the targeted SWFPR and provides acceptable power should be selected. Further detail is presented in **Chapter 20**. **R** can also be used to conduct these simulations.

### ►EXAMPLE 13-3

The logged iron concentrations from **Example 13-2** showed significant evidence of spatial variability. Use the results of the one-way ANOVA to compute adjusted intrawell prediction limits (without retesting) for each of the wells in that example and compare them to the unadjusted prediction limits.

#### SOLUTION

Step 1. Summary statistics by well for the logged iron measurements are listed in the table below. With $n$ = 4 measurements per well, use equation [13.1] and $t_{1-\alpha,n-1} = t_{.99,3}$ = 4.541 from **Table 16-1** in **Appendix D** to compute at each well an unadjusted 99% intrawell prediction limit for the next single measurement, based on lognormal data:

$$PL_{1-\alpha} = \exp\left[\bar{y} + s_y t_{1-\alpha,n-1}\sqrt{1+\frac{1}{n}}\right] \qquad [13.1]$$

| | **Unadjusted 99% Prediction Limits for Iron (ppm)** | | | | | |
|---|---|---|---|---|---|---|
| | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| Log-mean | 3.820 | 3.965 | 4.348 | 4.188 | 4.802 | 5.000 |
| Log-SD | 0.296 | 0.395 | 0.658 | 0.453 | 0.704 | 0.396 |
| $n$ | 4 | 4 | 4 | 4 | 4 | 4 |
| $t_{.99,3}$ | 4.541 | 4.541 | 4.541 | 4.541 | 4.541 | 4.541 |
| 99% PL | 204.9 | 391.6 | 2183.0 | 657.0 | 4341.5 | 1108.1 |

Step 2. Use the RMSE (*i.e.*, square root of the mean error sum of squares [$MS_{error}$] component) of the ANOVA in **Example 13-2** as an estimate of the adjusted, pooled standard deviation, giving $\sqrt{MS_{error}} = \sqrt{.258} = .5079$. The degrees of freedom ($df$) associated with this pooled standard deviation is $p(n-1) = 6(3) = 18$, the same as listed in the ANOVA table of **Example 13-2**.

Step 3. Use equation [13.2], along with the adjusted pooled standard deviation and its associated *df*, to compute an adjusted 99% prediction limit for each well, as given in the table below. Note that the adjusted t-value based on the larger *df* is $t_{1-\alpha,\mathrm{df}} = t_{.99,18} = 2.552$.

$$PL_{1-\alpha} = \exp\left[\bar{y} + t_{1-\alpha,df}\sqrt{MS_{error}\left(1+\frac{1}{n}\right)}\right] \qquad [13.2]$$

| | **Adjusted 99% Prediction Limits for Iron (ppm)** | | | | | |
|---|---|---|---|---|---|---|
| | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
| Log-mean | 3.820 | 3.965 | 4.348 | 4.188 | 4.802 | 5.000 |
| RMSE | 0.5079 | 0.5079 | 0.5079 | 0.5079 | 0.5079 | 0.5079 |
| *df* | 18 | 18 | 18 | 18 | 18 | 18 |
| $t_{.99,18}$ | 2.552 | 2.552 | 2.552 | 2.552 | 2.552 | 2.552 |
| 99% PL | 194.3 | 224.6 | 329.4 | 280.7 | 518.7 | 632.3 |

Step 4. Compare the adjusted and unadjusted lognormal prediction limits. By estimating the average intrawell standard deviation using ANOVA, the adjusted prediction limits are significantly lower and thus more powerful than the unadjusted limits, especially at Wells 3, 5, and 6.

In this example, use of the **R**-script approach was unnecessary, since the corresponding $\kappa$-multiple used in 1-of-1 prediction limit tests can be directly derived analytically. ◄

# CHAPTER 14. TEMPORAL VARIABILITY

14.1 TEMPORAL DEPENDENCE ..... 14-1
14.2 IDENTIFYING TEMPORAL EFFECTS AND CORRELATION ..... 14-3
*14.2.1 Parallel Time Series Plots ..... 14-3*
*14.2.2 One-Way Analysis of Variance for Temporal Effects ..... 14-6*
*14.2.3 Sample Autocorrelation Function ..... 14-12*
*14.2.4 Rank von Neumann Ratio Test ..... 14-16*
14.3 CORRECTING FOR TEMPORAL EFFECTS AND CORRELATION ..... 14-19
*14.3.1 Adjusting the Sampling Frequency and/or Test Method ..... 14-19*
*14.3.2 Choosing a Sampling Interval Via Darcy's Equation ..... 14-20*
*14.3.3 Creating Adjusted, Stationary Measurements ..... 14-28*
*14.3.4 Identifying Linear Trends Amidst Seasonality: Seasonal Mann-Kendall Test ..... 14-37*

This chapter discusses the importance of ***statistical independence*** in groundwater monitoring data with respect to ***temporal variability***. Temporal variability exists when the distribution of measurements varies with the times at which sampling or analytical measurement occurs. This variation can be caused by seasonal fluctuations in the groundwater itself, changes in the analytical method used, the re-calibration of instruments, anomalies in sampling method, ***etc***.

Methods to ***identify*** temporal variability are discussed for both groups of wells (parallel time series plots; one-way analysis of variance [ANOVA] for temporal effects) and single data series (sample autocorrelation function; rank von Neumann ratio). Procedures are also presented for correcting or accommodating temporal effects. These include guidance on adjusting the sampling frequency to avoid temporal correlation, choosing a sampling interval using the Darcy equation, removing seasonality or other temporal dependence, and finally testing for trends with seasonal data.

## 14.1 TEMPORAL DEPENDENCE

A key assumption underlying most statistical tests is that the sample data are independent and identically distributed [*i.i.d.*] (**Chapter 3**). In part, this means that measurements collected over a period of time should not exhibit a clear ***time dependence*** or significant ***autocorrelation***. Time dependence refers to the presence of trends or cyclical patterns when the observations are graphed on a time series plot. The closely related concept of autocorrelation is essentially the degree to which measurements collected later in a series can be predicted from previous measurements. Strongly autocorrelated data are highly predictable from one value to the next. Statistically independent values vary in a random, unpredictable fashion.

While temporal independence is a complex topic, there are several common types of temporal dependence. Some of these include: 1) correlation across wells over time in the concentration pattern of a single constituent (***i.e.***, concentrations tending to jointly rise or fall at each of the wells on common sampling events); 2) correlation across multiple constituents over time in their concentration patterns (***i.e.***, a parallel rise or fall in concentration across several parameters on common sampling events); 3) seasonal cycles; 4) trends, linear or otherwise; and 5) serial dependence or autocorrelation (***i.e.***, greater correlation between sampling events more closely spaced in time).

Any of these patterns can invalidate or weaken the results of statistical testing. In some cases, a statistical method can be chosen that specifically accounts for temporal dependence (e.g., seasonal Mann-Kendall trend test). In other instances, the sample data need to be adjusted for the dependence. Future data might also need to be collected in a manner that avoids temporal correlation. The goal of this chapter is to present straightforward tools that can be used to first identify temporal dependence and then to adjust for this correlation.

To better understand why most statistical tests depend on the assumption of statistical independence, consider a hypothetical series of groundwater measurements exhibiting an obvious pattern of seasonal fluctuation (**Figure 14-1**). These data demonstrate regular and repeated cycles of higher and lower values. Even though fluctuating predictably and highly dependent, the characteristics of the entire groundwater population will be observed over a long period of monitoring. This provides an estimate of the full range of concentrations and an accurate gauge of total variability.

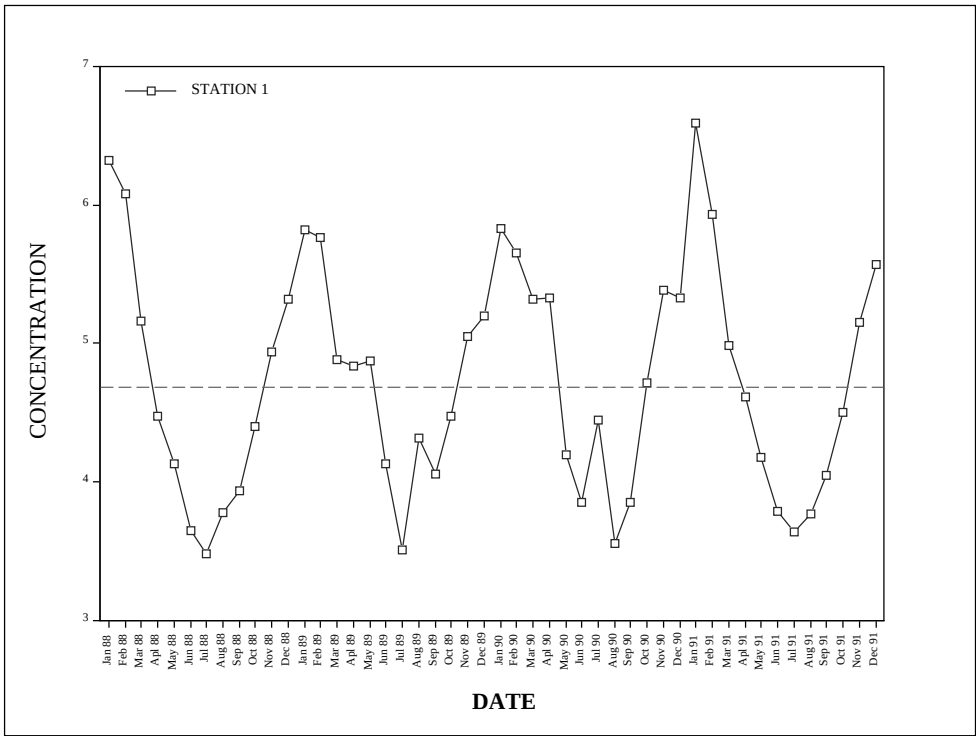
The same is not true for data collected from the same population over a much shorter span, say in five to six months. A much narrower range of sample concentrations would be observed due to the cyclical pattern. Depending on when the sampling was conducted, the average concentration level would either be much higher or much lower than the overall average; no single sampling period is likely to accurately estimate either the true population mean or its variance.

From this example, an important lesson can be drawn about temporally dependent data. Variance estimates in a sample of dependent, positively autocorrelated data are likely to be biased low. This is important because the guidance methods require and assume that an accurate and unbiased estimate of the sample standard deviation be available. A case in point was the practice of using aliquot replicates of a single physical sample for comparison with other combined replicate aliquot samples from a number of individual physical water quality samples (*e.g.*, in a Student-*t* test). Aliquot replicate values are much more similar to each other than to measurements made on physically discrete groundwater samples. Consequently, the estimate of variance was too low and the *t*-test frequently registered false positives.

Using physically discrete samples is not always sufficient. If the sampling interval ensures that discrete volumes of groundwater are being sampled on consecutive sampling events, the observations can be described as *physically independent*. However, they are not necessarily *statistically* independent. Statistical independence is based not on the physical characteristics of the sample data, but rather on the statistical pattern of measurements.

Temporally dependent and autocorrelated data generally contain both a truly random and non-random component. The relative strength of the latter effect is a measured by one or more correlation techniques. The degree of correlation among dependent sample measurements lies on a continuum. Sample pairs can be mildly correlated or strongly correlated. Only strong correlations are likely to substantially impact the results of further statistical testing.

Figure 14-1. Seasonal Fluctuations




# 14.2 IDENTIFYING TEMPORAL EFFECTS AND CORRELATION

## 14.2.1 PARALLEL TIME SERIES PLOTS

### BACKGROUND AND PURPOSE

Time series plots were introduced in **Chapter 9**. A time series plot such as **Figure 14-1** is a simple graph of concentration versus time of sample collection. Such plots are useful for identifying a variety of temporal patterns. These include identifying a trend over time, one or more sampling events that may signal contaminant releases, measurement outliers resulting in anomalous 'spikes' due to field handling or analytical problems, cyclical and seasonal fluctuations, as well as the presence of other time-related dependencies.

Time series plots can be used in two basic ways to identify temporal dependence. By graphing single constituent data from multiple wells together on a time series plot, potentially significant temporal components of variability can be identified. For example, seasonal fluctuations can cause the mean concentration levels at a number of wells to vary with the time of sampling events. This dependency will show up in the time series plot as a pattern of *parallel traces*, in which the individual wells will tend to rise and fall together across the sequence of sampling dates. The parallel pattern may be the result of the measurement process such as mid-stream changes in field handling or sample collection procedures, periodic re-calibration of analytical instrumentation, and changes in laboratory or analytical methods. It could also be the result from significant autocorrelation present in the groundwater population itself. Hydrologic factors such as drought, recharge patterns or regular (*e.g.*, seasonal) water table fluctuations may be responsible. In these cases, it may be useful to test for the presence of a significant temporal

effect by first constructing a parallel time series plot and then running a formal one-way ANOVA for temporal effects (**Section 14.2.2**).

The second way time series plots can be helpful is by plotting multiple constituents over time for the same well, or averaging values for each constituent across wells on each sampling event and then plotting the averages over time. In either case, the plot can signify whether the general concentration pattern over time is simultaneously observed for different constituents. If so, it may indicate that a group of constituents is highly correlated in groundwater or that the same artifacts of sampling and/or lab analysis impacted the results of several monitoring parameters.

## REQUIREMENTS AND ASSUMPTIONS

The requirements for time series plots were discussed in **Chapter 9**. Two very useful recommendations follow from that discussion. First, a different plot symbol should be used to display any non-detect measurements (*e.g.*, solid symbols for detected values, hollow symbols for non-detects). This can help prevent mistaking a change over time in reporting limits as a trend, since detected and non-detected data are clearly distinguished on the plot. It also allows one to determine whether non-detects are more prevalent during certain portions of the sample record and less prevalent at other times. Secondly, when multiple constituents are plotted on the same graph, it may be necessary to *standardize* each constituent prior to plotting to avoid trying to simultaneously visualize high-valued and low-valued traces on the same *y*-axis (*i.e.*, concentration axis). The goal of such a plot is to identify parallel concentration patterns over time. This can be done most readily by subtracting each constituent's sample mean ($\bar{x}$) from the measurements for that constituent and dividing by the standard deviation ($s$), so that every constituent is plotted on roughly the same scale.

## PROCEDURE FOR MULTIPLE WELLS, ONE CONSTITUENT

Step 1. For each well to be plotted, form data pairs by matching each concentration value with its sampling date.

Step 2. Graph the data pairs for each well on the same set of axes, the horizontal axis representing time and the vertical axis representing concentration. Connect the points for each individual well to form a 'trace' for that well.

Step 3. Look for parallel movement in the traces across the wells. Even if all the well concentrations tend to rise on a given sampling event, but not to the same magnitude or degree, this is evidence of a possible temporal effect.

## PROCEDURE FOR MULTIPLE CONSTITUENTS, ONE OR MANY WELLS

Step 1. For each constituent to be plotted, compute the constituent-specific sample mean ($\bar{x}$) and standard deviation ($s$). Form standardized measurements ($z_i$) by subtracting the mean from each concentration ($x_i$) and dividing by the standard deviation, using the equation:

$$z_i = \frac{x_i - \bar{x}}{s} \qquad [14.1]$$

Form data pairs by matching each standardized concentration with its sampling event.

Step 2. If correlation is suspected in a group of wells, average the standardized concentrations for each given constituent *across wells* for each specific sampling event. Otherwise, form a multi-constituent time series plot separately for each well.

Step 3. Graph the data pairs for each constituent on the same set of axes, the horizontal axis representing time and the vertical axis representing standardized concentrations. Connect the points for each constituent to form a trace for that parameter.

Step 4. Look for parallel movement in the traces across the constituents. A strong degree of parallelism indicates a high degree of correlation among the monitoring parameters.

►EXAMPLE 14-1

The following well sets of manganese measurements were collected over a two-year period. Construct a time series plot of these data to check for possible temporal effects.

| | **Manganese Concentrations (ppm)** | | | |
|---|---|---|---|---|
| Qtr | BW-1 | BW-2 | BW-3 | BW-4 |
| 1 | 28.14 | 31.41 | 27.15 | 30.46 |
| 2 | 29.33 | 30.27 | 30.24 | 30.60 |
| 3 | 30.45 | 32.57 | 29.14 | 30.96 |
| 4 | 32.42 | 32.77 | 30.59 | 30.70 |
| 5 | 34.37 | 33.03 | 34.88 | 32.71 |
| 6 | 33.25 | 32.18 | 30.53 | 31.76 |
| 7 | 31.02 | 28.85 | 30.33 | 31.85 |
| 8 | 28.50 | 32.88 | 30.42 | 29.58 |

SOLUTION

Step 1. Graph each well's concentrations versus sampling event on the same set of axes to construct the following time series plot (**Figure 14-2**).

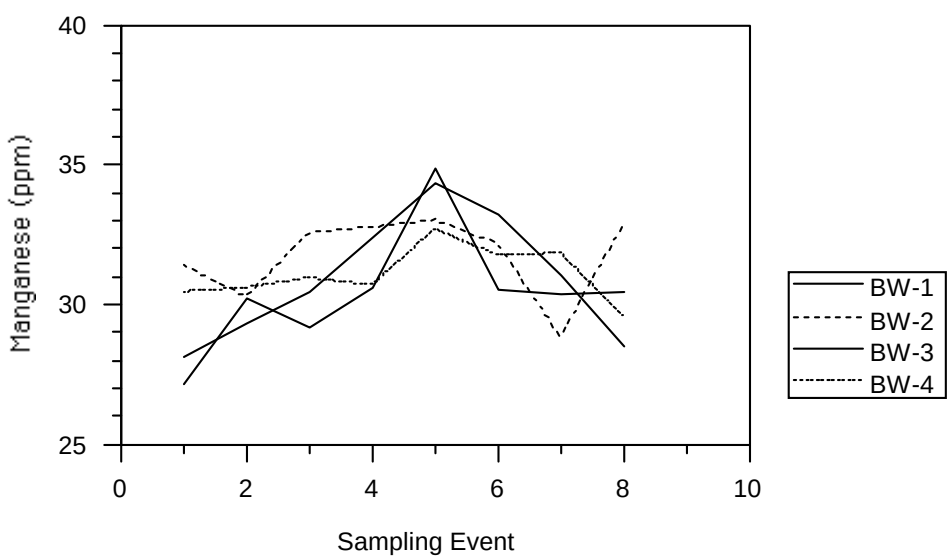



Figure 14-2. Manganese Parallel Time Series Plot

Step 2. Examining the traces on the plot, there is some degree of parallelism in the pattern over time. Particularly for the fifth quarter, there is an across-the-board increase in the manganese level, followed by a general decline the next two quarterly events. Note, however, that there is little evidence of differences in mean levels by *well location.* ◄

## 14.2.2 ONE-WAY ANALYSIS OF VARIANCE FOR TEMPORAL EFFECTS

### PURPOSE AND BACKGROUND

Parametric ANOVA is a comparison of means among a set of populations. The one-way ANOVA for temporal effects is no exception. A one-way ANOVA for spatial variation (**Chapter 13**) uses well data sets to represent *locations* as the statistical factor of interest. In contrast, a one-way ANOVA for temporal effects considers multiple well data sets for individual *sampling events* or *seasons* as the relevant statistical factor. A significant temporal factor implies that the average concentration depends to some degree on *when* sampling takes place.

Three common examples of temporal factors include: 1) an irregular, but consistent shift of average concentrations over time perhaps due to changes in laboratories or analytical method interferences; 2) cyclical seasonal patterns; or 3) parallel upward or downward trends. These can occur in both upgradient and downgradient well data.

If event-specific analytical differences or seasonality appear to be an important temporal factor, the one-way ANOVA for temporal effects can be used to formally identify seasonality, parallel trends, or changes in lab performance that affect other temporal effects. Results of the ANOVA can also be used to create temporally *stationary residuals*, where the temporal effect has been 'subtracted from' the original measurements. These stationary residuals may be used to replace the original data in subsequent statistical testing.

The one-way ANOVA for a temporal factor described below can be used for an additional purpose when interwell testing is appropriate. For this situation, *there can be no significant spatial variability*. If a group of upgradient or other background wells indicates a significant temporal effect, an interwell prediction limit can be designed which properly accounts for this temporal dependence. A more powerful interwell test of upgradient-to-downgradient differences can be developed than otherwise would be possible. This can occur because the ANOVA separates or 'decomposes' the overall data variation into two sources: a) temporal effects and b) random variation or statistical error. It also estimates how the background mean is changing from one sampling event to the next. The final prediction limit is formed by computing the background mean, using the separate structural and random variation components of the ANOVA to better estimate the standard deviation, and then adjusting the effective sample size (via the degrees of freedom) to account for these factors.

### REQUIREMENTS AND ASSUMPTIONS

Like the one-way ANOVA for spatial variation (**Chapter 13**), the one-way ANOVA for temporal effects assumes that the data groups are normally-distributed with constant variance. This requirement means that the group residuals should be tested for normality (**Chapter 10**) and also for equality of

variance (**Chapter 11**). It is also assumed that for each of a series of background wells, measurements are collected at each well on sampling events or dates *common* to all the wells.

Two variations in the basic procedure are described below. For cases of temporal effects *excluding* seasonality, each sampling event is treated as a separate population. The ANOVA residuals are grouped and tested *by sampling event* to test for equality of variance. In cases of apparent seasonality, each *season* is treated as a distinct population. The difference is that seasons contain multiple sampling events across a span of multiple years, with sampling events collected at the same time of year assigned to one of the seasons (*e.g.*, all January or first quarter measurements). Here, the ANOVA residuals are grouped by season to test for homoscedasticity.

If the assumption of equal variances or normal residuals is violated, a data transformation should be considered. This should be followed by testing of the assumptions on the transformed scale. The one-way ANOVA for a non-seasonal effect should include a minimum of four wells and at least 4 observations (*i.e.*, distinct sampling dates) per well. In the seasonal case, there should be a minimum of 3-4 sampling events per distinct season, with the events thus spanning at least three years (*i.e.*, one per year per season). Larger numbers of both wells and observations are preferable. Sampling dates should also be approximately the *same* for each well if a temporal effect is to be tested.

If the data cannot be normalized, a similar test for a temporal or seasonal effect can be performed using the Kruskal-Wallis test (**Chapter 17**). The only difference from the procedure outlined in **Section 17.1.2** is that the roles of wells/groups and sampling events have to be reversed. That is, each sampling event should be treated as a separate 'well,' while each well is treated as a separate 'sampling event.' Then the same equations can be applied to the reversed data set to test for a significant temporal dependence. If testing for a seasonal effect, the wells in the notation of **Section 17.1.2** become the groups of common sampling events from different years, while the sampling events are again the distinct wells.

Even when a temporal effect exists and is apparent on a time series plot, the variation between well locations (*i.e.*, spatial variability) may overshadow the temporal variability. This could result in a non-significant one-way ANOVA finding for the temporal factor. In these cases, a two-way ANOVA can be considered where both well location and sampling event/season are treated as statistical factors. This procedure is described in Davis (1994). Evidence for a temporal effect can be documented using this last technique, although the two-way ANOVA isn't necessary if the goal is simply to construct temporally stationary residuals. That can be accomplished with a one-way ANOVA even when significant spatial variability exists.

## PROCEDURE

Step 1. Given a set of $W$ wells and measurements from each of $T$ sampling events at each well on each of $K$ years, label the observations as $x_{ijk}$, for $i$ = 1 to $W$, $j$ = 1 to $T$, and $k$ = 1 to $K$. Then $x_{ijk}$ represents the measurement from the $i$th well on the $j$th sampling event during the $k$th year.

Step 2. When testing for a *non-seasonal* temporal effect, form the set of event means ($x_{\bullet jk}$) and the grand mean ($x_{\bullet\bullet\bullet}$) using equations [14.2] and [14.3] respectively:

$$x_{\bullet jk} = \frac{1}{W}\sum_{i=1}^{W} x_{ijk} \quad \text{for } j = 1 \text{ to } T \text{ and } k = 1 \text{ to } K \qquad [14.2]$$

$$x_{\bullet\bullet\bullet} = \sum_{i=1}^{W}\sum_{j=1}^{T}\sum_{k=1}^{K} x_{ijk} \Big/ WTK \qquad [14.3]$$

Step 2a. If testing for a seasonal effect common to all wells, form the seasonal means ($x_{\bullet j\bullet}$) instead of the event means of **Step 2**, using the equation:

$$x_{\bullet j\bullet} = \frac{1}{WK}\sum_{i=1}^{W} x_{ijk} \quad \text{for } j = 1 \text{ to } T \qquad [14.4]$$

Step 3. Compute the set of residuals for each sampling event or season using either equation [14.5] or equation [14.6] respectively:

$$r_{ijk} = x_{ijk} - x_{\bullet jk} \quad \text{for } i = 1 \text{ to } W \qquad [14.5]$$

$$r_{ijk} = x_{ijk} - x_{\bullet j\bullet} \quad \text{for } i = 1 \text{ to } W \text{ and } k = 1 \text{ to } K \qquad [14.6]$$

Step 4. Test the residuals for normality (**Chapter 10**). If significant non-normality is evident, consider transforming the data and re-doing the computations in **Steps 1** through **4** on the transformed scale.

Step 5. Test the sets of residuals grouped either by sampling event or season for equal variance (**Chapter 11**). If the variances are significantly different, consider transforming the data and re-doing the computations in **Steps 1** through **5** on the transformed data.

Step 6. If testing for a non-seasonal temporal effect, compute the mean error sum of squares term ($MS_E$) using equation:

$$MS_E = \sum_{i=1}^{W}\sum_{j=1}^{T}\sum_{k=1}^{K} r_{ijk}^2 \Big/ TK\left(W-1\right) \qquad [14.7]$$

This term is associated with $TK(W–1)$ degrees of freedom. Also compute the mean sum of squares for the temporal effect ($MS_T$) with degrees of freedom ($TK–1$), using equation:

$$MS_T = W\sum_{j=1}^{T}\sum_{k=1}^{K}\left(x_{\bullet jk} - x_{\bullet\bullet\bullet}\right)^2 \Big/ \left(TK-1\right) \qquad [14.8]$$

Step 6a. If testing for a seasonal effect, compute the mean error sum of squares ($MS_E$) using equation:

$$MS_E = \sum_{i=1}^{W}\sum_{j=1}^{T}\sum_{k=1}^{K} r_{ijk}^2 \Big/ T\left(WK-1\right) \qquad [14.9]$$

This term is associated with $T(WK-1)$ degrees of freedom. Also compute the mean sum of squares for the seasonal effect ($MS_T$) with degrees of freedom ($T-1$), using equation:

$$MS_T = WK\sum_{j=1}^{T}\left(x_{\bullet j \bullet} - x_{\bullet\bullet\bullet}\right)^2 \Big/ \left(T-1\right) \qquad [14.10]$$

Step 7. Test for a significant event-to-event or seasonal effect by computing the ratio of the mean sum of squares for time and the mean error sum of squares:

$$F_T = MS_T \big/ MS_E \qquad [14.11]$$

Step 8. If testing for a non-seasonal temporal effect, the test statistic $F_T$ under the null hypothesis (*i.e.*, of no significant time-related variability among the sampling events) will follow an $F$-distribution with $(TK-1)$ and $TK(W-1)$ degrees of freedom. Therefore, using a significance level of $\alpha = 0.05$, compare $F_T$ against the critical point $F_{.05,\ TK-1,TK(W-1)}$ taken from the $F$-distribution in **Table 17-1** in **Appendix D**. If the critical point is exceeded, conclude there is a significant temporal effect.

Step 8a. If testing for a seasonal effect, the test statistic $F_T$ under the null hypothesis (*i.e.*, of no seasonal pattern) will follow an $F$-distribution with $(T-1)$ and $T(WK-1)$ degrees of freedom. Therefore, using a significance level of $\alpha = 0.05$, compare $F_T$ against the critical point $F_{.05,\ T-1,T(WK-1)}$ taken from the $F$-distribution in **Table 17-1** of **Appendix D**. If the critical point is exceeded, conclude there is a significant seasonal pattern.

Step 9. If there is no spatial variability but a significant temporal effect exists among a set of background wells, compute an appropriate *interwell* prediction or control chart limit as follows. First replace the background sample standard deviation ($s$) with the following estimate built from the one-way ANOVA table:

$$\hat{\sigma} = \sqrt{\frac{1}{W}\left[MS_T + (W-1)\,MS_E\right]} \qquad [14.12]$$

Then calculate the effective sample size for the prediction limit as:

$$n^* = 1 + \left\{\left[T\cdot(T-1)\cdot(F_T + W - 1)^2\right] \Big/ \left[T\cdot F_T^2 + (T-1)\cdot(W-1)\right]\right\} \qquad [14.13]$$

►EXAMPLE 14-2

Some parallelism was found in the time series plot of **Example 14-1**. Test those same manganese data for a significant, non-seasonal temporal effect using a one-way ANOVA at the 5% significance level.

| Qtr | Manganese Concentrations (ppm) | | | | |
|---|---|---|---|---|---|
| | Event Mean | BW-1 | BW-2 | BW-3 | BW-4 |
| 1 | 29.290 | 28.14 | 31.41 | 27.15 | 30.46 |
| 2 | 30.110 | 29.33 | 30.27 | 30.24 | 30.60 |
| 3 | 30.780 | 30.45 | 32.57 | 29.14 | 30.96 |
| 4 | 31.620 | 32.42 | 32.77 | 30.59 | 30.70 |
| 5 | 33.747 | 34.37 | 33.03 | 34.88 | 32.71 |
| 6 | 31.930 | 33.25 | 32.18 | 30.53 | 31.76 |
| 7 | 30.513 | 31.02 | 28.85 | 30.33 | 31.85 |
| 8 | 30.345 | 28.50 | 32.88 | 30.42 | 29.58 |
| | Grand mean = 31.042 | | | | |

SOLUTION

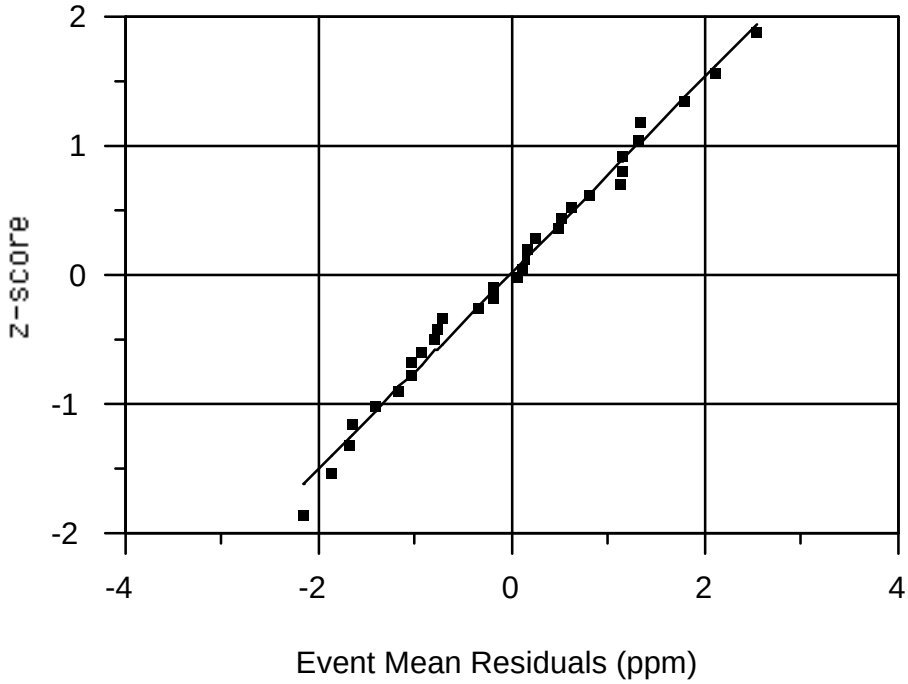
Step 1. First compute the means for each sampling event and the grand mean of all the data. These values are listed in the table above. With four wells and eight quarterly events per well, $W = 4$, $T = 4$, and $K = 2$.

Step 2. Determine the residuals for each sampling event by subtracting off the event mean. These values are listed in the table below.

| Qtr | Manganese Event Residuals (ppm) | | | |
|---|---|---|---|---|
| | BW-1 | BW-2 | BW-3 | BW-4 |
| 1 | -1.150 | 2.120 | -2.140 | 1.170 |
| 2 | -0.780 | 0.160 | 0.130 | 0.490 |
| 3 | -0.330 | 1.790 | -1.640 | 0.180 |
| 4 | 0.800 | 1.150 | -1.030 | -0.920 |
| 5 | 0.622 | -0.718 | 1.132 | -1.038 |
| 6 | 1.320 | 0.250 | -1.400 | -0.170 |
| 7 | 0.508 | -1.662 | -0.182 | 1.338 |
| 8 | -1.845 | 2.535 | 0.075 | -0.765 |

Step 3. Test the residuals for normality. A probability plot of these residuals is given in **Figure 14-3**. An adequate fit to normality is suggested by Filliben's probability plot correlation coefficient test.

Figure 14-3. Probability Plot of Manganese Sampling Event Residuals




Step 4. Next, test the groups of residuals for equal variance across sampling events. Levene's test (**Chapter 11**) gives an *F*-statistic of 1.30, well below the 5% critical point with 7 and 24 degrees of freedom of $F_{.95,7,24} = 2.42$. Therefore, the group variances test out as adequately homogeneous.

Step 5. Compute the mean error sum of squares term using equation [14.7]:

$$MS_E = \left[(-1.150)^2 + (-.780)^2 + \ldots + (1.338)^2 + (-.765)^2\right] / (4 \cdot 2)(7) = 1.87$$

Step 6. Compute the mean sum of squares term for the time effect using equation [14.8]:

$$MS_T = 4\left[(29.290 - 31.042)^2 + (30.11 - 31.042)^2 + \ldots + (30.345 - 31.042)^2\right] / 7 = 7.55$$

Step 7. Test for a significant temporal effect, computing the *F*-statistic in equation [14.11]:

$$F_T = 7.55 / 1.87 = 4.04$$

The degrees of freedom associated with the numerator and denominator respectively are $(T–1) = 7$ and $T(W–1) = 24$. Just as with Levene's test run earlier, the 5% level critical point for the test is $F_{.95,7,24} = 2.42$. Since $F_T$ exceeds this value, there is evidence of a significant temporal effect in the manganese background data.

Step 8. Assuming a lack of spatial variation, the presence of a temporal effect can be used to compute a standard deviation estimate and effective background sample size appropriate for an

interwell prediction limit test, using equations [14.12] and [14.13] respectively. The adjusted standard deviation becomes:

$$\hat{\sigma} = \sqrt{\frac{1}{4}\left[7.55 + 7(1.87)\right]}$$

while the effective sample size is:

$$n^* = 1 + \left\{\left[8 \cdot 7 \cdot (4.04 + 4 - 1)^2\right] \Big/ \left[8 \cdot (4.04)^2 + 7 \cdot 3\right]\right\} = 19.31 \approx 19$$

If the background data had simply been pooled together and the sample standard deviation computed, $s$ = 1.776 ppm with a sample size of $n$ = 32. So the adjustments based on the temporal effect alter the final prediction limit by enlarging it to account for the additional component of variation. ◀

## 14.2.3 SAMPLE AUTOCORRELATION FUNCTION

### BACKGROUND AND PURPOSE

The sample autocorrelation function enables a test of temporal autocorrelation in a single data series (*e.g.*, from a single well over time). When a time-related dependency affects several wells simultaneously, parallel time series plots (**Section 14.2.1**) and one-way ANOVA for temporal effects (**Section 14.2.2**) should be considered. But when a longer data series is to be used for an intrawell test such as a prediction limit or control chart, the autocorrelation function does an excellent job of identifying temporal dependence.

Given a sequence of consecutively-collected measurements, $x_1$, $x_2$,…, $x_n$, form the set of overlapping pairs ($x_i$, $x_{i+1}$) for $i$ = 1,…, $n$–1. The approximate first-order sample autocorrelation coefficient is then computed from these pairs as (Chatfield, 2004):

$$r_1 = \frac{\sum_{i=1}^{n-1}\left(x_i - \bar{x}\right)\left(x_{i+1} - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad [14.14]$$

Equation [14.14] estimates the *first-order autocorrelation*, that is, the correlation between pairs of sample measurements collected one event apart (*i.e.*, consecutive events). The number of sampling events separating each pair is called the *lag*, representing the temporal distance between the pair measurements.

Autocorrelation can also be computed at other lags. The general approximate equation for the *k*th lag is given by:

$$r_k = \frac{\sum_{i=1}^{n-k}\left(x_i - \bar{x}\right)\left(x_{i+k} - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad [14.15]$$

which estimates the $k$th-order autocorrelation for pairs of measurements separated in time by $k$ sampling events. Note that the number of pairs used to compute $r_k$ decreases with increasing $k$ due to the fact that fewer and fewer sample pairs can be formed which are separated by that many lags.

By computing the first few sample autocorrelation coefficients and defining $r_0 = 1$, the sample autocorrelation function can be formed by plotting $r_k$ against $k$. Since the autocorrelation coefficients are approximately normal in distribution with zero mean and variance equal to $1/n$, a test of significant autocorrelation at the 95% significance level can be made by examining the sample autocorrelation function to see if any coefficients exceed $2/\sqrt{n}$ in absolute value ($\pm 2/\sqrt{n}$ represent approximate upper and lower confidence limits).

The sample autocorrelation function is a valuable visual tool for assessing different types of autocorrelation (Chatfield, 2004). For instance, a stationary (*i.e.*, stable, non-trending) but non-random series of measurements will often exhibit a large value of $r_1$ followed by perhaps one or two other significantly non-zero coefficients. The remaining coefficients will be progressively smaller and smaller, tending towards zero. A series with a clear seasonal pattern will exhibit a seasonal (i.e., approximately sinusoidal) autocorrelation function. If the series tends to alternate between high and low values, the autocorrelation function will also alternate, with $r_1$ being negative to reflect that consecutive measurements tend to be on 'opposite sides' of the sample mean. Finally, if the series contains a trend, the sample autocorrelation function will not drop to zero as the lag $k$ increases. Rather, there will a persistent autocorrelation even at very large lags.

## REQUIREMENTS AND ASSUMPTIONS

The approximate distribution of the sample autocorrelation coefficients is predicated on the sample measurements following a normal distribution. A test for significant autocorrelation may therefore be inaccurate unless the sample measurements are roughly normal. Non-normal data series should be tested for temporal autocorrelation using the non-parametric rank Von Neumann ratio (**Section 14.2.4**).

Outliers can drastically affect the sample autocorrelation function (Chatfield, 2004). Before assessing autocorrelation, check the sample for possible outliers, removing those that are identified. A series of at least 10-12 measurements is minimally recommended to construct the autocorrelation function. Otherwise, the number of lagged data pairs will be too small to reliably estimate the correlation, especially for larger lags. Sampling events should be regularly spaced so that pairs lagged by the same number of events ($k$) represent the same approximate time interval.

## PROCEDURE

Step 1. Given a series of $n$ measurements, $x_1,\ldots, x_n$, form sets of lagged data pairs ($x_i$, $x_{i+k}$), $i = 1,\ldots, n–k$, for $k \leq [n/3]$, where the notation $[c]$ represents the largest integer no greater than $c$. For longer series, computing lags to a maximum of $k = 15$ is generally sufficient.

Step 2. For each set of lagged pairs from **Step 1**, compute the sample autocorrelation coefficient, $r_k$, using equation [14.15]. Also define $r_0 = 1$.

Step 3. Graph the sample autocorrelation function by plotting $r_k$ versus $k$ for $k = 0,\ldots,[n/3]$, generally up to a maximum lag of 15. Also plot horizontal lines at levels equal to: $\pm 2/\sqrt{n}$ .

Step 4. Examine the sample autocorrelation function. If any coefficient $r_k$ exceeds $2/\sqrt{n}$ in absolute value, conclude that the sample has significant autocorrelation.

►EXAMPLE 14-3

The following series of monthly total alkalinity measurements were collected from leachate at a solid waste landfill during a four and a half year period. Use the sample autocorrelation function to test for significant temporal dependence in this series.

| Date | Total Alkalinity (mg/L) | Date | Total Alkalinity (mg/L) | Date | Total Alkalinity (mg/L) |
|---|---|---|---|---|---|
| 01/26/96 | 1400 | 07/01/97 | 2400 | 01/15/99 | 1350 |
| 02/20/96 | 1700 | 08/15/97 | 3500 | 02/02/99 | 1560 |
| 03/19/96 | 1900 | 09/15/97 | 3100 | 03/02/99 | 1220 |
| 04/22/96 | 1800 | 10/15/97 | 3300 | 04/15/99 | 1390 |
| 05/22/96 | 1300 | 11/15/97 | 2100 | 05/04/99 | 1940 |
| 06/24/96 | 2000 | 12/15/97 | 2100 | 06/02/99 | 2160 |
| 07/15/96 | 2300 | 01/15/98 | 1500 | 07/07/99 | 1990 |
| 08/21/96 | 2500 | 02/15/98 | 710 | 08/03/99 | 2540 |
| 09/15/96 | 1700 | 03/15/98 | 1100 | 09/02/99 | 2250 |
| 10/15/96 | 1600 | 04/15/98 | 1900 | 10/07/99 | 1630 |
| 11/11/96 | 1400 | 05/08/98 | 2100 | 11/02/99 | 1710 |
| 12/10/96 | 1600 | 06/15/98 | 2000 | 12/07/99 | 1210 |
| 01/22/97 | 1800 | 07/15/98 | 2500 | 01/06/00 | 1170 |
| 02/11/97 | 1000 | 08/15/98 | 2700 | 02/02/00 | 1330 |
| 03/04/97 | 720 | 09/02/98 | 2400 | 03/02/00 | 1540 |
| 04/07/97 | 1400 | 10/06/98 | 3000 | 04/04/00 | 1670 |
| 05/01/97 | 1600 | 11/03/98 | 2700 | 05/02/00 | 1520 |
| 06/09/97 | 990 | 12/15/98 | 2680 | 06/06/00 | 2080 |

SOLUTION

Step 1. Create a time series plot of the $n = 54$ alkalinity measurements, as in **Figure 14-4**. The series indicates an apparent seasonal fluctuation.

Step 2. Form lagged data pairs from the alkalinity series for each lag k = 1,…, [n/3] = 18. The first two pairs for k = 1 (i.e., first order lag) are (1400, 1700) and (1700, 1900). For k = 2, the first two pairs are (1400, 1900) and (1700, 1800), etc.

Step 3. At each lag ($k$), compute the sample autocorrelation coefficient using equation [14.15]. Note that the denominator of this equation equals $(n–1)s^2$. For the alkalinity data, the sample mean and variance are $\bar{x} = 1865.93$ and $s^2 = 392349.1$ respectively. The lag-1 autocorrelation is thus:

$$r_1 = \frac{(1400-1865.93)\cdot(1700-1865.93)+\ldots+(1520-1865.93)\cdot(2080-1865.93)}{(54-1)\cdot 392349.1} = .64$$

Other lags are computed similarly.

Step 4. Plot the sample autocorrelation function as in **Figure 14-5**. Overlay the plot with 95% confidence limits (dotted lines) shown at $\pm 2/\sqrt{n} = \pm 2/\sqrt{54} = 0.27$ .

Step 5. The autocorrelation function indicates coefficients at several lags that lie outside the 95% confidence limits, confirming the presence of temporal dependence. Further, the shape of autocorrelation function is sinusoidal, suggesting a strong seasonal fluctuation in the alkalinity levels. ◄

Figure 14-4. Time Series Plot of Total Alkalinity (mg/L)

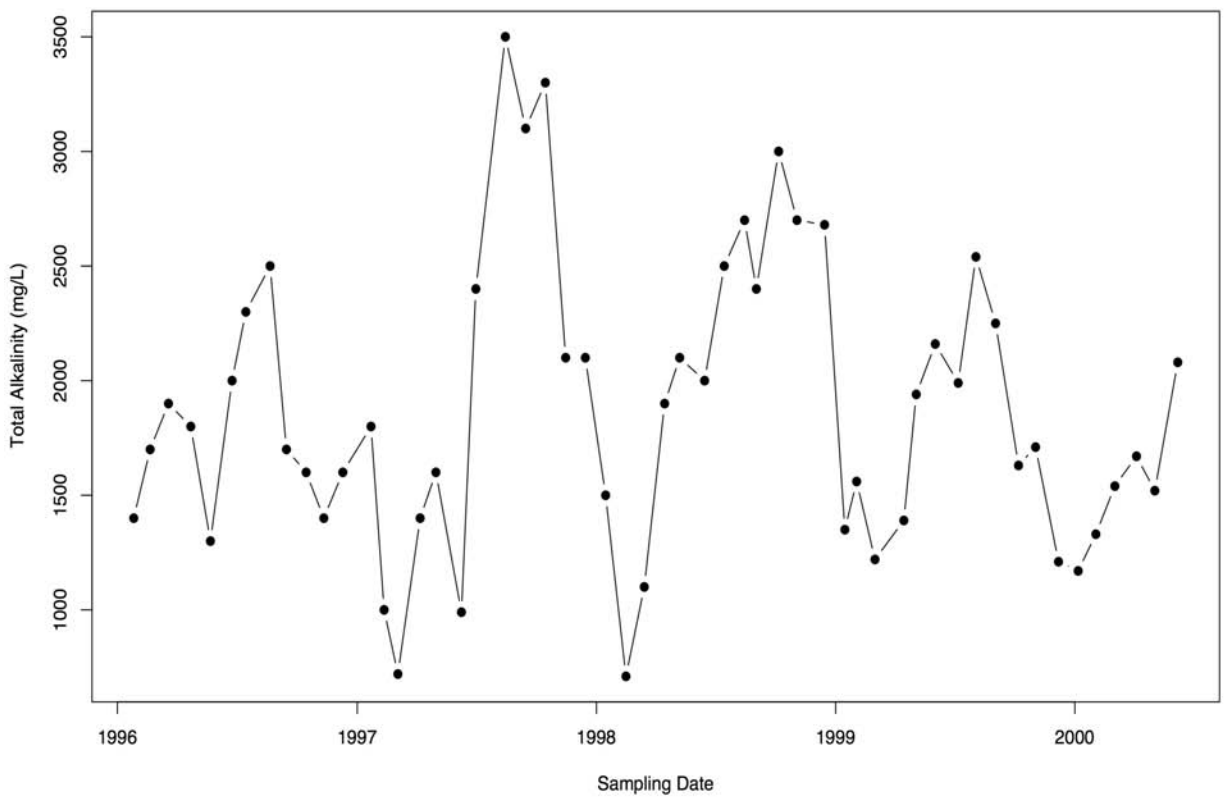
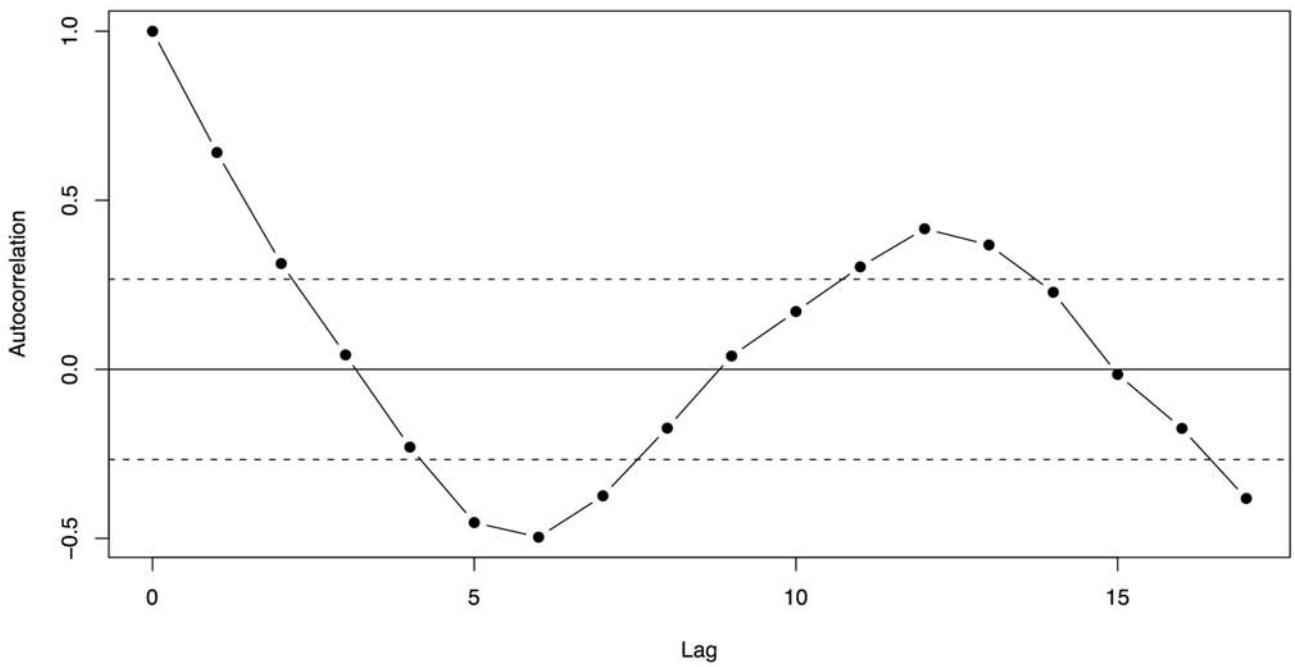


Figure 14-5. Sample Autocorrelation Function for Total Alkalinity




## 14.2.4 RANK VON NEUMANN RATIO TEST

### BACKGROUND AND PURPOSE

The rank von Neumann ratio is a non-parametric test of first-order temporal autocorrelation in a single data series (*e.g.*, from a single well over time). It can be used as an alternative to the sample autocorrelation function (**Section 14.2.3**) for non-normal data, and is both easily computed and effective.

The rank von Neumann ratio is based on the idea that a truly independent series of data will vary in an unpredictable fashion as the list is examined sequentially. The first order or lag-1 autocorrelation will be approximately zero. By contrast, the first-order autocorrelation in *dependent* data will tend to be positive (or negative), implying that lag-1 data pairs in the series will tend to be more similar (or dissimilar) in magnitude than would expected by chance.

Not only will the concentrations of lag-1 data pairs tend to be similar (or dissimilar) when the series is autocorrelated, but the *ranks* of lag-1 data pairs will share that similarity or dissimilarity. Although the test is non-parametric and rank-based, the ranks of non-independent data still follow a discernible pattern. Therefore, the rank von Neumann ratio is constructed from the sum of differences between the *ranks* of lag-1 data pairs. When these differences are *small*, the ranks of consecutive data measurements need to be fairly similar, implying that the pattern of observations is somewhat predictable. Given the relative position and magnitude of one observation, the approximate relative position and magnitude of the next sample measurement can be predicted. Low values of the rank von Neumann ratio are therefore indicative of temporally dependent data series.

Compared to other tests of statistical independence, the rank von Neumann ratio has been shown to be more powerful than non-parametric methods such as the Runs up-and-down test (Madansky, 1988). It is also a reasonable test when the data follow a normal distribution. In that case, the efficiency of the test is always close to 90 percent when compared to the von Neumann ratio computed on concentrations instead of the ranks. Thus, very little effectiveness is lost by using the ranks in place of the original measurements. The rank von Neumann ratio will correctly detect dependent data and do so over a variety of underlying data distributions. The rank von Neumann ratio is also fairly robust to departures from normality, such as when the data derive from a skewed distribution like the lognormal.

## REQUIREMENTS AND ASSUMPTIONS

An unresolved problem with the rank von Neumann ratio test is the presence of a substantial fraction of tied observations. Like the Wilcoxon rank-sum test (**Chapter 16**), Bartels (1982) recommends replacing each tied value by its mid-rank (*i.e.*, the average of all the ranks that would have been assigned to that set of ties). However, no explicit adjustment of the ratio for ties has been developed. The rank von Neumann critical points may not be appropriate (or at best very approximate) when a large portion of the data consists of non-detects or other tied values. Especially in the case of frequent non-detects, too much information is lost regarding the pattern of variability to use the rank von Neumann ratio as an accurate indication of autocorrelation. In fact, no test is likely to provide a good estimate of temporal correlation, whether non-parametric or parametric.

While the rank von Neumann ratio test is recommended in the Unified Guidance for its ease of use and robustness when applied to either normal or non-normal distributions, the literature on time series analysis and temporal correlation is extensive with respect to other potential tests. Many other tests of autocorrelation are available, especially when either the original measurements or the residuals of the data are normally distributed after a trend has been removed. Chatfield (2004) and (Madansky, 1988) are two good references for some of these alternate tests.

## PROCEDURE

Step 1. Order the sample from least to greatest and assign a unique rank to each measurement. If some data values are tied, replace tied values with their mid-ranks as in the Wilcoxon rank-sum test (**Chapter 16**). Then list the observations and their corresponding ranks in the order that they were collected (*i.e.*, by sampling event or time order).

Step 2. Using the list of ranks, $R_i$, for the sampling events $i = 1 \ldots n$, compute the rank von Neumann ratio with the equation:

$$v = \sum_{i=2}^{n} \left(R_i - R_{i-1}\right)^2 \Big/ \left[ n\left(n^2 - 1\right)/12 \right] \qquad [14.16]$$

Step 3. Given sample size ($n$) and desired significance level ($\alpha$), find the lower critical point of the rank von Neumann ratio in **Table 14-1** of **Appendix D**. In most cases, a choice of $\alpha = .01$ should be sufficient, since only substantial non-independence is likely to affect subsequent statistical testing. If the computed ratio, $v$, is *smaller* than this critical point, conclude that the data series is strongly autocorrelated. If not, there is insufficient evidence to reject the

hypothesis of independence; treat the data as temporally independent in subsequent statistical testing.

►EXAMPLE 14-4

Use the rank von Neumann ratio test on the following series of 16 quarterly measurements of arsenic (ppb) to determine whether or not the data set should be treated as temporally independent in subsequent tests. Compute the test at the $\alpha$ = .01 level of significance.

| Sample Date | Arsenic (ppb) | Rank ($R_i$) |
|---|---|---|
| Jan 1990 | 4.0 | 5 |
| Apr 1990 | 7.2 | 15 |
| Jul 1990 | 3.1 | 2 |
| Oct 1990 | 3.5 | 3 |
| Jan 1991 | 4.4 | 8 |
| Apr 1991 | 5.1 | 9 |
| Jul 1991 | 2.2 | 1 |
| Oct 1991 | 6.3 | 13 |
| Jan 1992 | 6.5 | 14 |
| Apr 1992 | 7.5 | 16 |
| Jul 1992 | 5.8 | 11 |
| Oct 1992 | 5.9 | 12 |
| Jan 1993 | 5.7 | 10 |
| Apr 1993 | 4.1 | 6 |
| Jul 1993 | 3.8 | 4 |
| Oct 1993 | 4.3 | 7 |

SOLUTION

Step 1. Assign ranks to the data values as in the table above. Then list the data in chronological order so that each rank value occurs in the order sampled.

Step 2. Compute the von Neumann ratio using the set of ranks in column 3 using equation [14.16], being sure to take squared differences of successive, *overlapping* pairs of rank values:

$$v = \frac{\left[(15-5)^2 + (2-15)^2 + \ldots + (7-4)^2\right]}{16 \cdot \left(16^2 - 1\right)/12} = 1.67$$

Step 3. Look up the lower critical point ($v_{cp}$) for the rank von Neumann ratio in **Table 14-1** of **Appendix D**. For $n$ = 16 and $\alpha$ = .01, the lower critical point is equal to 0.93. Since the test statistic $v$ is larger than $v_{cp}$, there is insufficient evidence of autocorrelation at the $\alpha$ = .01 level of significance. Therefore, treat these data as statistically independent in subsequent testing. ◄

# 14.3 CORRECTING FOR TEMPORAL EFFECTS AND CORRELATION

## 14.3.1 ADJUSTING THE SAMPLING FREQUENCY AND/OR TEST METHOD

If a data series is temporally correlated, a simple remedy (if allowable under program rules) is to change the sampling frequency and/or statistical method used to analyze the data. In some cases, increasing the sampling interval will effectively eliminate the statistical dependence exhibited by the series. This may happen because the longer time between sampling events allows more groundwater to flow through the well screen, further differentiating measurements of consecutive volumes of groundwater and lessening the impact of seasonal fluctuations or other time-dependent patterns in the underlying concentration distribution.

Many authors including Gibbons (1994a) and ASTM (1994) have recommended that sampling be conducted no more often than quarterly to avoid temporal dependence. If the sampling frequency is reduced, there are obviously fewer measurements available for statistical analysis during any given evaluation period. A *t*-test or ANOVA cannot realistically be run with fewer than four measurements per well. A prediction limit for a future mean requires at least two new observations, and a prediction limit for a future median requires at least three measurements, not counting any resamples. Depending on the length of the evaluation period (i.e., quarterly, semi-annual, annual), a change of statistical method may also be necessary when groundwater measurements are autocorrelated.

When sufficient background data have been collected over a longer period of time, a prediction limit test for future values can be run with as few as one or two new measurements per compliance well. The same is true for control charts. Therefore, if a low groundwater flow velocity and/or evidence of statistical dependence suggest a reduction in sampling frequency, certain prediction limits and control charts should be strongly considered as alternate statistical procedures.

### RUNNING A PILOT STUDY

An optional approach to adjusting the sampling frequency is to run a site-specific *pilot study* of autocorrelation. Such a study can be conducted in several ways, but perhaps the easiest is to pick two or three wells from the network (perhaps one background well and one or two compliance wells) and then conduct weekly sampling at these wells over a one year period. For each well in the study, construct the sample autocorrelation function (**Section 14.2.3**) for a variety of constituents, and determine from these graphs the smallest lagged interval at which the autocorrelation coefficient becomes insignificantly different from zero for most of the study constituents.

Since an autocorrelation of zero is equivalent to temporal independence for practical purposes, finding the smallest lag between sampling events with no correlation indicates the minimum sampling frequency needed to approximately ensure statistical independence. If the sample autocorrelation function does not drop down to zero with increasing lag ($k$), there may be a strong seasonal component or a trend involved. In these circumstances, lengthening the sampling frequency may do little to lessen the temporal dependence. A seasonal pattern may need to be estimated instead and regularly removed from the data prior to statistical testing. Likewise, any apparent trends should be investigated to determine if there is evidence of increasing concentration levels indicative of a possible release.

### 14.3.2 CHOOSING A SAMPLING INTERVAL VIA DARCY'S EQUATION

Another strategy for determining an appropriate sampling interval is to use Darcy's equation. The goal of this approach is to calculate groundwater flow velocity and the time needed to ensure that physically independent or distinct volumes of groundwater are collected on each sampling trip. As noted in **Chapter 6**, physical independence does not guarantee statistical independence. However, statistical independence may be more likely if the same general volume of groundwater is not re-sampled on multiple occasions.

This section discusses the important hydrological parameters to consider when choosing a sampling interval. The *Darcy equation* is used to determine the horizontal component of the average linear velocity of ground water for confined, semi-confined, and unconfined aquifers. This value provides a good estimate of travel time for most soluble constituents in groundwater, and can be used to determine a minimal sampling interval. Example calculations are provided to further assist the reader. Alternative methods should be employed to determine a sampling interval in groundwater environments where Darcy's law is invalid. Karst, cavernous basalt, fractured rocks, and other 'pseudo-karst' terranes usually require specialized monitoring approaches.

Section 264.97(g) of 40 CFR Part 264 Subpart F allows the owner or operator of a RCRA facility to choose a sampling procedure that will reflect site-specific concerns. It specifies that the owner or operator shall obtain a sequence of at least four samples from each well collected at least semi-annually. The interval is determined after evaluating the uppermost aquifer's effective porosity, hydraulic conductivity, and hydraulic gradient, and the fate and transport characteristics of potential contaminants. The intent of this provision is to set a sampling frequency that allows sufficient time between sampling events to ensure, to the greatest extent technically feasible, that independent groundwater observations are taken from each well.

The sampling frequency required in Part 264 Subpart F can be based on estimates using the average linear velocity of ground water. Two forms of the Darcy equation stated below relate groundwater velocity ($V$) to effective porosity ($Ne$), hydraulic gradient ($i$), and hydraulic conductivity ($K$):

$$V_h = \left(K_h \cdot i\right) / Ne \qquad [14.17]$$

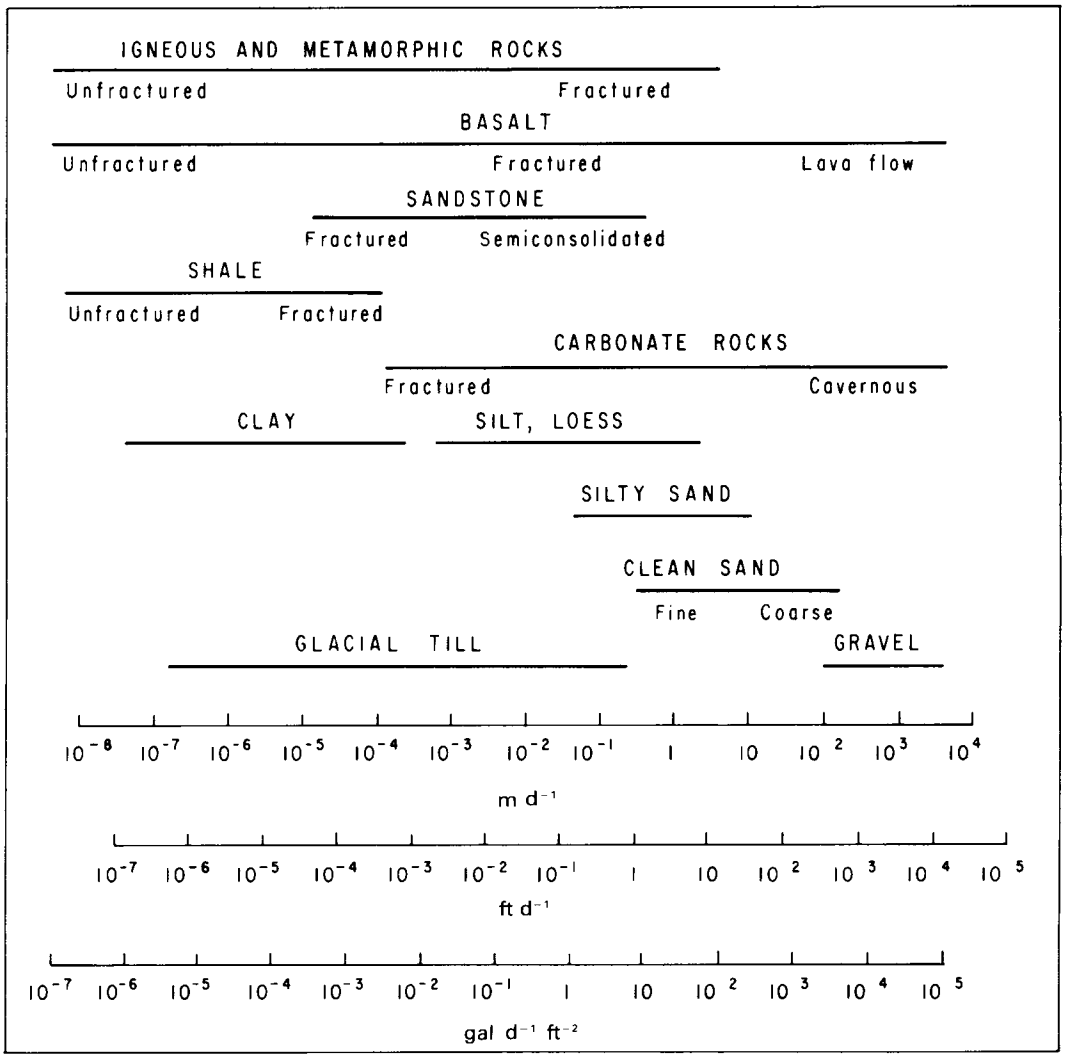
$$V_v = \left(K_v \cdot i\right) / Ne \qquad [14.18]$$

where $V_h$ and $V_v$ are the horizontal and vertical components of the average linear velocity of groundwater, respectively; $K_h$ and $K_v$ are the horizontal and vertical components of hydraulic conductivity, respectively; $i$ is the head gradient; and $Ne$ is the effective porosity.

In applying these equations to ground-water monitoring, the horizontal component of the average linear velocity ($V_h$) can be used to determine an appropriate sampling interval. Usually, field investigations will yield bulk values for hydraulic conductivity. In most cases, the bulk hydraulic conductivity determined by a pump test, tracer test, or a slug test will be sufficient for these calculations. The vertical component ($V_v$), however, should be considered in estimating flow velocities in areas with significant components of vertical velocity such as recharge and discharge zones.

To apply the Darcy equation to groundwater monitoring, the parameters *K*, *i*, and *Ne* need to be determined. The hydraulic conductivity, *K*, is the volume of water at the existing kinematic viscosity that will move in unit time under a unit hydraulic gradient through a unit area measured at right angles to the direction of flow. "[E]xisting kinematic viscosity" refers to the fact that hydraulic conductivity is not only determined by the media (aquifer), but also by fluid properties (groundwater or potential contaminants). Thus, it is possible to have several hydraulic conductivity values for different chemical substances present in the same aquifer. The lowest velocity value calculated using the Darcy equation should be used to determine sampling intervals, ensuring physical independence of consecutive sample measurements.

Figure 14-6. Hydraulic Conductivity of Selected Rocks

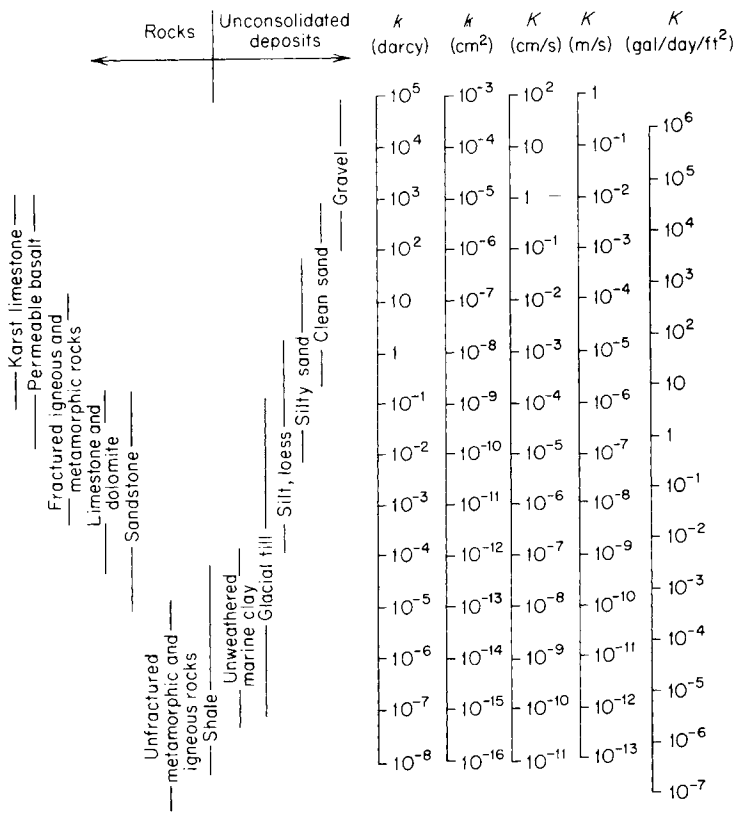



Source: Heath, R.C. 1987. Basic Ground-Water Hydrology. U.S. Geological Survey Water Supply Paper, 2220, 13 pp.

A range of hydraulic conductivities (the transmitted fluid is water) for various aquifer materials is given in **Figures 14-6** and **14-7**. The conductivities are given in several units. **Figure 14-8** lists conversion factors to change between various permeability and hydraulic conductivity units.

The hydraulic gradient, *i*, is the change in hydraulic head per unit of distance in a given direction. It can be determined by dividing the difference in head between two points on a potentiometric surface map by the orthogonal distance between those two points (see calculation in **Example 14-5**). Water level measurements are normally used to determine the natural hydraulic gradient at a facility. However, the effects of mounding in the event of a release may produce a steeper local hydraulic gradient in the vicinity of the monitoring well. These local changes in hydraulic gradient should be accounted for in the velocity calculations.

Figure 14-7. Range of Values of Hydraulic Conductivity and Permeability




Source: Freeze, R.A., and J.A. Cherry. 1979. Ground Water. Prentice Hall, Inc., Englewood Cliffs, New Jersey. p. 29.

Figure 14-8. Conversion Factors for Permeability and Hydraulic Conductivity Units

| | Permeability, *k** | | | Hydraulic conductivity, *K* | | |
|---|---|---|---|---|---|---|
| | $cm^2$ | $ft^2$ | darcy | m/s | ft/s | $gal/day/ft^2$ |
| $cm^2$ | 1 | $1.08\times10^{-3}$ | $1.01\times10^{8}$ | $9.80\times10^{2}$ | $3.22\times10^{3}$ | $1.85\times10^{9}$ |
| $ft^2$ | $9.29\times10^{2}$ | 1 | $9.42\times10^{10}$ | $9.11\times10^{5}$ | $2.99\times10^{6}$ | $1.71\times10^{12}$ |
| darcy | $9.87\times10^{-9}$ | $1.06\times10^{-11}$ | 1 | $9.66\times10^{-6}$ | $3.17\times10^{-5}$ | $1.82\times10^{1}$ |
| m/s | $1.02\times10^{-3}$ | $1.10\times10^{-6}$ | $1.04\times10^{5}$ | 1 | 3.28 | $2.12\times10^{6}$ |
| ft/s | $3.11\times10^{-4}$ | $3.35\times10^{-7}$ | $3.15\times10^{4}$ | $3.05\times10^{-1}$ | 1 | $6.46\times10^{5}$ |
| $gal/day/ft^2$ | $5.42\times10^{-10}$ | $5.83\times10^{-13}$ | $5.49\times10^{-2}$ | $4.72\times10^{-7}$ | $1.55\times10^{-6}$ | 1 |

*To obtain *k* in $ft^2$, multiply *k* in $cm^2$ by $1.08\times10^{-3}$

Source: Freeze, R.A., and J.A. Cherry (1979). *Ground Water*. Prentice Hall, Inc., Englewood Cliffs, New Jersey, p. 29.

The effective porosity, *Ne*, is the ratio, usually expressed as a percentage, of the total volume of voids available for fluid transmission to the total volume of the porous medium de-watered. It can be estimated during a pump test by dividing the volume of water removed from an aquifer by the total volume of aquifer dewatered (see calculation in **Example 14-5**). **Figure 14-9** presents approximate effective porosity values for a variety of aquifer materials. In cases where the effective porosity is unknown, specific yield may be substituted into the equation. Specific yields of selected rock units are given in **Figure 14-10**. In the absence of measured values, drainable porosity is often used to approximate effective porosity. **Figure 14-11** illustrates representative values of drainable porosity and total porosity as a function of aquifer particle size. If available, field measurements of effective porosity are preferred.

## Figure 14-9. Default Values of Effective Porosity (*Ne*) For Travel Time Analyses

| Soil textural classes | Effective porosity of saturation[a] |
|---|---|
| *Unified soil classification system* | |
| GS, GP, GM, GC, SW, SP, SM, SC | 0.20 (20%) |
| ML, MH | 0.15 (15%) |
| CL, OL, CH, OH, PT | 0.01 (1%)[b] |
| *USDA soil textural classes* | |
| Clays, silty clays, sandy clays | 0.01 (1%)b |
| Silts, silt loams, silty clay loams | 0.10 (10%) |
| All others | 0.20 (20%) |
| *Rock units (all)* | |
| Porous media (non-fractured rocks such as sandstone and some carbonates) | 0.15 (15%) |
| Fractured rocks (most carbonates, shales, granites, etc.) | 0.0001 (0.01%) |

Source: Barari, A., and L. S. Hedges (1985). Movement of Water in Glacial Till. *Proceedings of the 17th International Congress of the International Association of Hydrogeologists*, pp. 129-134.

[a]These values are estimates and there may be differences between similar units. For example, recent studies indicate that weathered and unweathered glacial till may have markedly different effective porosities (Barari and Hedges, 1985; Bradbury et al., 1985).

[b]Assumes *de minimus* secondary porosity. If fractures or soil structure are present, effective porosity should be 0.001 (0.1%).

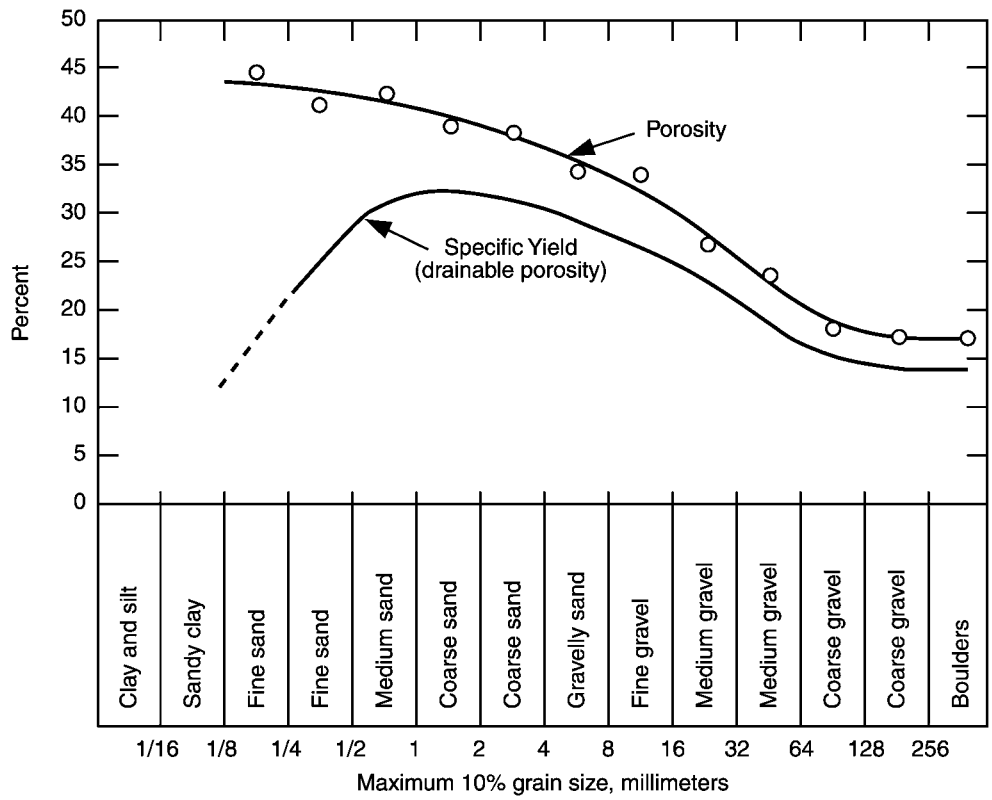
## Figure 14-10. Specific Yield Values for Selected Rock Types

| Rock Type | Specific Yield (%) |
|---|---|
| Clay | 2 |
| Sand | 22 |
| Gravel | 19 |
| Limestone | 18 |
| Sandstone (*semi-consolidated*) | 6 |
| Granite | 0.09 |
| Basalt (*young*) | 8 |

Source: Heath, R.C. (1983). *Basic Ground-Water Hydrology*. U.S. Geological Survey, Water Supply Paper 2220, 84 pp.

Once the values for *K*, *i*, and *Ne* are determined, the horizontal component of average linear groundwater velocity can be calculated. Using the Darcy equation [14.17], the time required for groundwater to pass through the complete monitoring well diameter can be determined by dividing the well diameter by the horizontal component of the average linear groundwater velocity. If considerable exchange of water occurs during well purging, the diameter of the filter pack may be used rather than the well diameter. This value represents the ***minimum*** time interval required between sampling events yielding a physically independent (*i.e.*, distinct) ground-water sample. Note that three-dimensional mixing of groundwater in the vicinity of the monitoring well is likely to occur when the well is purged before sampling. Partly for that reason, this method can only provide an estimated travel time.

Figure 14-11. Total Porosity and Drainable Porosity for Typical Geologic Materials




(The grain size in which the cumulative total, beginning with the coarsest material, reaches 10% of the total sample.)

Source: Todd, D.K. 1980. Ground Water Hydrology. John Wiley and Sons, New York, 534 pp.

In determining these sampling intervals, many chemical compounds do not travel at the same velocity as groundwater. Chemical characteristics such as adsorptive potential, specific gravity, and molecular size influence the way chemicals travel in the subsurface. Large molecules, for example, tend to travel slower than the average linear groundwater velocity because of matrix interactions. Compounds that exhibit a strong adsorptive potential undergo a similar fate that dramatically changes time of travel predictions using the Darcy equation. In some cases chemical interaction with the matrix material alters the matrix structure and its associated hydraulic conductivity and may result in an increase in contaminant mobility. This effect has been observed with certain organic solvents in clay units (see Brown and Andersen, 1981). Contaminant fate and transport models may be useful in determining the influence of these effects on movement in the subsurface.

### ►EXAMPLE 14-5

Compute the effective porosity, *Ne*, expressed as a percent (%), using results obtained during a pump test.

SOLUTION

Step 1. Compute the effective porosity using the following equation:

$$Ne = 100\% \times \text{volume of water removed} / \text{volume of aquifer dewatered} \qquad [14.19]$$

Step 2. Based on a pumping rate of 50 gal/min and a pumping duration of 30 min, compute the volume of water removed as:

$$\text{volume of water removed} = 50 \text{ gal/min} \times 30 \text{ min} = 1{,}500 \text{ gal}$$

Step 3. To calculate the volume of aquifer de-watered, use the equation:

$$V = \frac{1}{3}\pi h r^2 \qquad [14.20]$$

where $r$ is the radius (in ft) of the area affected by pumping and $h$ (ft) is the drop in the water level. If, for example, $h$ = 3 ft and $r$ = 18 ft, then:

$$V = \frac{1}{3}\left(3.14 \times 3 \times 18^2\right) = 1{,}018 \text{ ft}^3$$

Next, converting cubic feet of water to gallons of water,

$$V = 1{,}018 \text{ ft}^3 \times 7.48 \text{ gal/ft}^3 = 7{,}615 \text{ gal}$$

Step 4. Finally, substitute the two volumes from **Step 3** into equation [14.19] to obtain the effective porosity:

$$Ne = 100\% \times \left(1{,}500 \text{ gal}/7{,}615 \text{ gal}\right) = 19.7\% \quad ◄$$

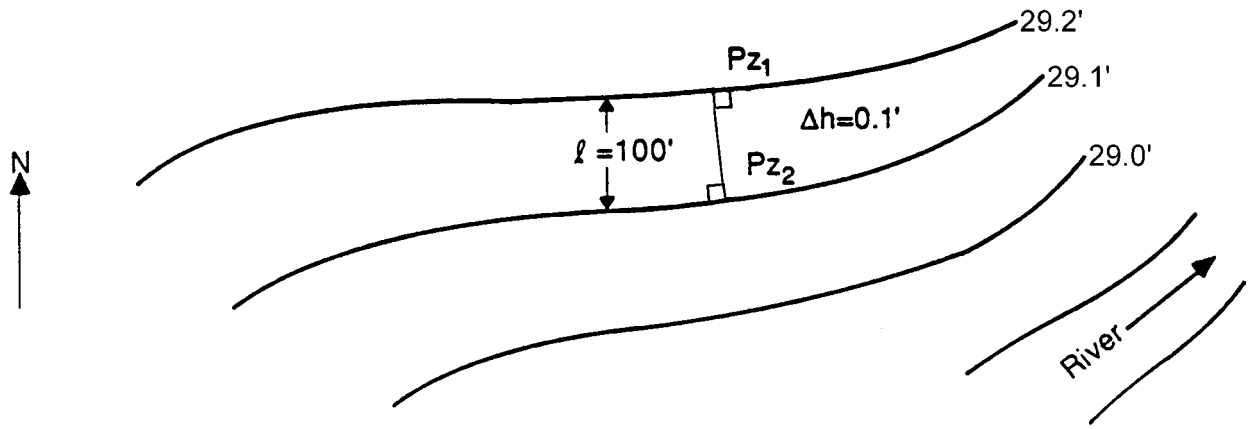
### ►EXAMPLE 14-6

Determine the hydraulic gradient, *i*, from a potentiometric surface map.

SOLUTION

Step 1. Consider the potentiometric surface map in **Figure 14-12**. The hydraulic gradient can be constructed as $i = \Delta h/l$, where $\Delta h$ is the difference measured in the gradient at piezometers $Pz_1$ and $Pz_2$,and $l$ is the orthogonal distance between the two piezometers.

Figure 14-12. Potentiometric Surface Map for Computation of Hydraulic Gradient




Step 2. Using the values given in **Figure 14-12**, the hydraulic gradient is computed as:

$$i = \Delta h / l = (29.2 \text{ ft} - 29.1 \text{ ft}) / 100 \text{ ft} = 0.001 \text{ ft/ft}$$

Step 3. Note that this method provides only a very general estimate of the natural hydraulic gradient existing in the vicinity of the two piezometers. Chemical gradients are known to exist and may override the effects of the hydraulic gradient. A detailed study of the effects of multiple chemical contaminants may be necessary to determine the actual average linear groundwater velocity (horizontal component) in the vicinity of the monitoring wells. ◄

►EXAMPLE 14-7

Determine the horizontal component of the average linear groundwater velocity ($V_h$) at a land disposal facility which has monitoring wells screened in an unconfined silty sand aquifer.

SOLUTION

Step 1. Slug tests, pump tests, and tracer tests conducted during a hydrologic site investigation have revealed that the aquifer has a horizontal hydraulic conductivity ($K_h$) of 15 ft/day and an effective porosity ($Ne$) of 15%. Using a potentiometric map (as in **Example 14-6**), the regional hydraulic gradient ($i$) has been determined to be 0.003 ft/ft.

Step 2. To estimate the minimum time interval between sampling events enabling the collection of physically independent samples of ground water, calculate the horizontal component of the average linear groundwater velocity ($V_h$) using Darcy's equation [14.17]. With $K_h$ = 15 ft/day, $Ne$ = 15%, and $i$ = 0.003 ft/ft, the velocity becomes:

$$V_h = (15 \text{ ft/day} \times 0.003 \text{ ft/ft}) / 15\% = 0.3 \text{ ft/day or } 3.6 \text{ in/day}$$

Step 3. Based on these calculations, the horizontal component of the average linear groundwater velocity, $V_h$, is equal to 3.6 in/day. Since monitoring well diameters at this particular facility are 4 inches, the minimum time interval between sampling events enabling a physically

independent groundwater sample can be computed by dividing the horizontal component into the monitoring well diameter:

$$\text{Minimum time interval} = (4\ \text{in})/(3.6\ \text{in/day}) = 1.1\ \text{days}$$

As a result, the facility could theoretically sample every other day. However, this may be unwise because velocity can seasonally vary with recharge rates. It is also emphasized that *physical* independence does not guarantee *statistical* independence. **Figure 14-13** gives results for common situations. The overriding point is that it may not be necessary to set the minimum sampling frequency to quarterly at every site. Some hydrologic environments may allow for more frequent sampling, some less. ◀

Figure 14-13. Typical Darcy Equation Results in Determining a Sampling Interval

| Unit | $K_h$ (ft/day) | *Ne* (%) | $V_h$ (in/mo) | Sampling Interval |
|---|---|---|---|---|
| Gravel | $10^4$ | 19 | $9.6\times10^4$ | Daily |
| Sand | $10^2$ | 22 | $8.3\times10^2$ | Daily |
| Silty Sand | 10 | 14 | $1.3\times10^2$ | Weekly |
| Till | $10^{-3}$ | 2 | $9.1\times10^{-2}$ | Monthly |
| Silty Sand (semi-consolidated) | 1 | 6 | 30 | Weekly |
| Basalt | $10^{-1}$ | 8 | 2.28 | Monthly |

## 14.3.3 CREATING ADJUSTED, STATIONARY MEASUREMENTS

When an existing data set exhibits temporal correlation or other variability, it is sometimes possible to *remove* the temporal pattern and thereby create a set of adjusted data which are uncorrelated and stationary over time in mean level. As long as the same temporal pattern seems to affect both background and the compliance point data to be tested, the effect (*e.g.*, regular seasonal fluctuation) can be estimated and removed from both data sets prior to statistical testing. Testing the adjusted data instead of the raw measurements in this way results in a more powerful and accurate test. An extraneous source of variation *not related* to identifying a contaminant release has been removed from the sample data.

The general topic of stationary, adjusted data is complex, contained within the extensive literature on time series. The Unified Guidance discusses two simple cases below: removing a seasonal pattern from a single well and creating adjusted data from a one-way ANOVA for temporal effects across several wells. More complicated situations may need professional consultation.

### 14.3.3.1 CORRECTING FOR SEASONAL PATTERN IN A SINGLE WELL

#### BACKGROUND AND PURPOSE

Sometimes an obvious cyclical seasonal pattern can be seen in a time series plot. Such data are not statistically independent. They do not fluctuate randomly but rather in a predictable way from one sampling event to the next. Data from such patterns can be adjusted to correct for or remove the seasonal fluctuation, but only if a longer series of data is available. This is also known as *deseasonalizing* the data. Seasonal correction should be done both to minimize the chance of mistaking a seasonal effect for

evidence of contaminated groundwater, and also to build more powerful background-to-compliance point tests.

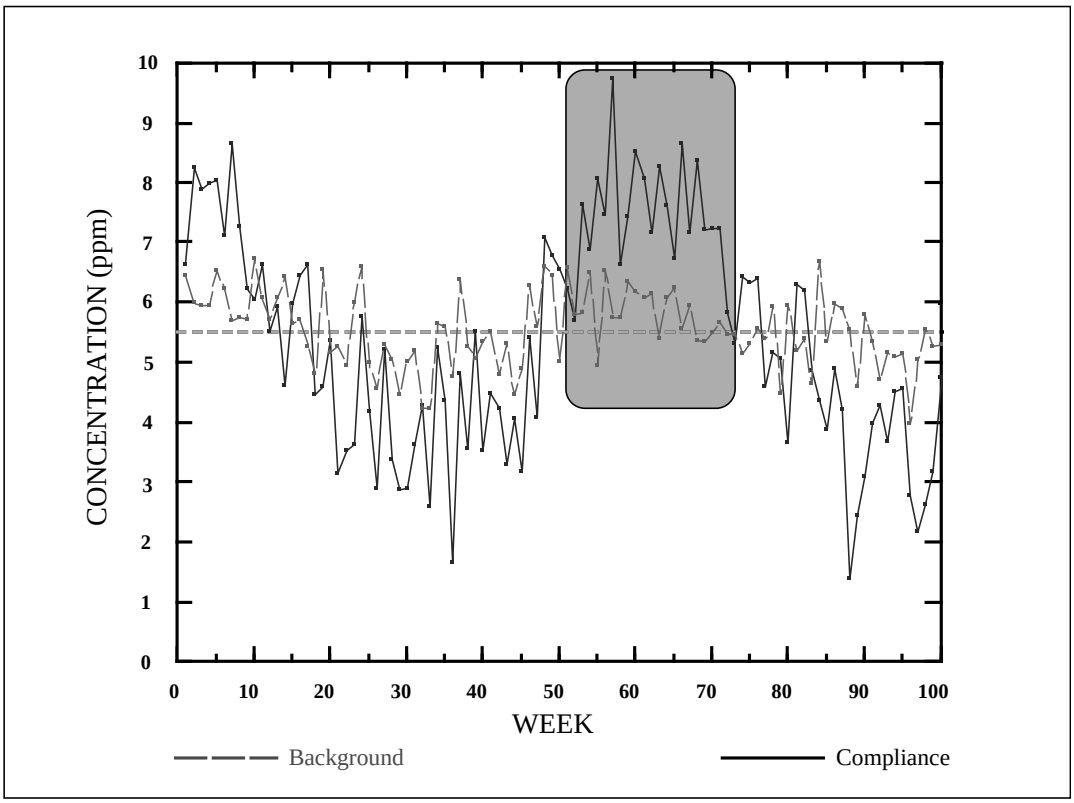
Problems can arise, for instance, from measurement variations associated with changing recharge rates during different seasons. Compliance point concentrations can exceed a groundwater protection standard [GWPS] for a portion of the year, but on average lie below. If the long-term average is of regulatory concern, the data should first be de-seasonalized before comparing it against a GWPS.

If *point-in-time, interwell comparisons* are being made between simultaneously collected background and downgradient data, a correction may not be necessary even when seasonal fluctuations exist. A temporal cycle may cover a period of several years so that both the background and downgradient values are observed on essentially the same parts of the overall cycle. In this case, the short-term averages in both data sets will be fairly stable and the seasonal or cyclical effect may equivalently impact both sets of data.

For intrawell tests, the data need to be collected sequentially at each well, with background formed from the earliest measurements in the series. The point-in-time argument would not apply and the presence of seasonality should be checked and accounted for.

Even with interwell testing, it is sometimes difficult to verify whether or not a seasonal pattern is impacting upgradient and compliance point wells similarly. If the groundwater velocity is low, the lag between the time groundwater passes through a background well screen and then travels downgradient may create a noticeable shift as to when corresponding portions of the seasonal cycle are observed in compliance point locations. It also may be the case that differences in geochemistry from well to well may cause the same seasonal pattern to differentially impact concentration levels at distinct wells (**Figure 14-14**).

Figure 14-14. Differential Seasonal Effects: Background vs. Compliance Wells




If the timing of the cycle and the relative *magnitude* of the concentration swings are essentially the same in upgradient and downgradient wells, both data sets should be deseasonalized prior to statistical analysis. If the seasonal effects are ignored, real differences in mean levels between upgradient and downgradient well data may not be observed, simply because the short-term seasonal fluctuations add variability that can mask the difference being tested. In this case, the non-independent nature of the seasonal pattern adds unwanted noise to the observations, obscuring statistical evidence of groundwater contamination.

## REQUIREMENTS AND ASSUMPTIONS

Seasonal correction is only appropriate for wells where a cyclical pattern is clearly present and very regular over time. Many approaches to deseasonalizing data exist. If the seasonal pattern is highly regular, it may be modeled with a sine or cosine function. Often, moving averages and/or lag-based differences (of order 12 for monthly data, for example) are used. General time series models may include these and other more complicated methods for deseasonalizing the data.

The simple method described in the Unified Guidance has the advantage of being easy to understand and apply, and of providing natural estimates of the monthly or quarterly seasonal effects via the monthly or quarterly means. The method can be applied to any seasonal or recurring cycle-- perhaps an annual cycle for monthly or quarterly data or a longer cycle for certain kinds of geologic environments. In some cases, recharge rates are linked to drought cycles that may be on the order of

several years long. For these situations, the 'seasonal' cycle may not correspond to typical fluctuations over the course of a single year.

Corrections for seasonality should be used cautiously, as they represent extrapolation into the future. There should be a good physical explanation for the seasonal fluctuation as well as good empirical evidence for seasonality before corrections are made. Higher than average rainfall for two or three Augusts in a row does not justify the belief that there will never be a drought in August, and this idea extends directly to groundwater quality. At least three complete cycles of the seasonal pattern should be observed on a time series plot before attempting the adjustment below. If seasonality is suspected but the pattern is complicated, the user should seek the help of a professional statistician.

## PROCEDURE

Step 1. If a time series plot clearly shows at least 3 full cycles of the seasonal pattern, determine the length of time to complete one full cycle. Since the correction presumes a regular sampling schedule, determine the number of observations ($k$) in each full cycle (this number should be the *same* for each cycle). Then, assuming that $N$ complete cycles of data are available, let $x_{ij}$ denote the raw, unadjusted measurement for the $i$th sampling event during the $j$th complete cycle. Note that this could represent monthly data over an annual cycle, quarterly data over a biennial cycle, semi-annual data over a 10-year cycle, *etc.*

Step 2. Compute the mean concentration for sampling event $i$ over the $N$-cycle period:

$$\bar{x}_i = \left(x_{i1} + x_{i2} + \ldots + x_{iN}\right)/N \qquad [14.21]$$

This is the average of all observations taken in different cycles, but during the same sampling event. For instance, with monthly data over an annual cycle, one would calculate the mean concentrations for all Januarys, the mean for all Februarys, and so on for each of the 12 months.

Step 3. Calculate the grand mean, $\bar{x}$ , of all $N \times k$ observations:

$$\bar{x} = \sum_{i=1}^{k}\sum_{j=1}^{N}\frac{x_{ij}}{N \times k} = \sum_{i=1}^{k}\frac{\bar{x}_i}{k} \qquad [14.22]$$

Step 4. Compute seasonally-corrected, adjusted concentrations using the equation:

$$z_{ij} = x_{ij} - \bar{x}_i + \bar{x} \qquad [14.23]$$

Computing $x_{ij} - \bar{x}_i$ removes the average seasonal effect of sampling event $i$ from the data series. Adding back the overall mean, $\bar{x}$, gives the adjusted $z_{ij}$ values the same mean as the raw, unadjusted data. Thus, the overall mean of the corrected values, $\bar{z}$, equals the overall mean value, $\bar{x}$.

### ►EXAMPLE 14-8

Consider the monthly unadjusted concentrations of an analyte over a 3-year period graphed in **Figure 14-15** and listed in the table below. Given the clear and regular seasonal pattern, use the above method to produce a deseasonalized data set.

| | **Unadjusted Concentrations** | | | | **Adjusted Concentrations** | | |
|---|---|---|---|---|---|---|---|
| | 1983 | 1984 | 1985 | Monthly Average | 1983 | 1984 | 1985 |
| January | 1.99 | 2.01 | 2.15 | 2.05 | 2.11 | 2.13 | 2.27 |
| February | 2.10 | 2.10 | 2.17 | 2.12 | 2.14 | 2.14 | 2.21 |
| March | 2.12 | 2.17 | 2.27 | 2.19 | 2.10 | 2.15 | 2.25 |
| April | 2.12 | 2.13 | 2.23 | 2.16 | 2.13 | 2.14 | 2.24 |
| May | 2.11 | 2.13 | 2.24 | 2.16 | 2.12 | 2.14 | 2.25 |
| June | 2.15 | 2.18 | 2.26 | 2.20 | 2.12 | 2.15 | 2.23 |
| July | 2.19 | 2.25 | 2.31 | 2.25 | 2.11 | 2.17 | 2.23 |
| August | 2.18 | 2.24 | 2.32 | 2.25 | 2.10 | 2.16 | 2.24 |
| September | 2.16 | 2.22 | 2.28 | 2.22 | 2.11 | 2.17 | 2.23 |
| October | 2.08 | 2.13 | 2.22 | 2.14 | 2.10 | 2.15 | 2.24 |
| November | 2.05 | 2.08 | 2.19 | 2.11 | 2.11 | 2.14 | 2.25 |
| December | 2.08 | 2.16 | 2.22 | 2.16 | 2.09 | 2.17 | 2.23 |

Overall 3-year average = 2.17

SOLUTION

Step 1. From **Figure 14-15**, there are $N = 3$ full cycles represented, each lasting approximately a year. With monthly data, the number of sampling events per cycle is $k = 12$.

Step 2. Compute the monthly averages across the 3 years for each of the 12 months using equation [14.21]. These values are shown in the fifth column of the table above.

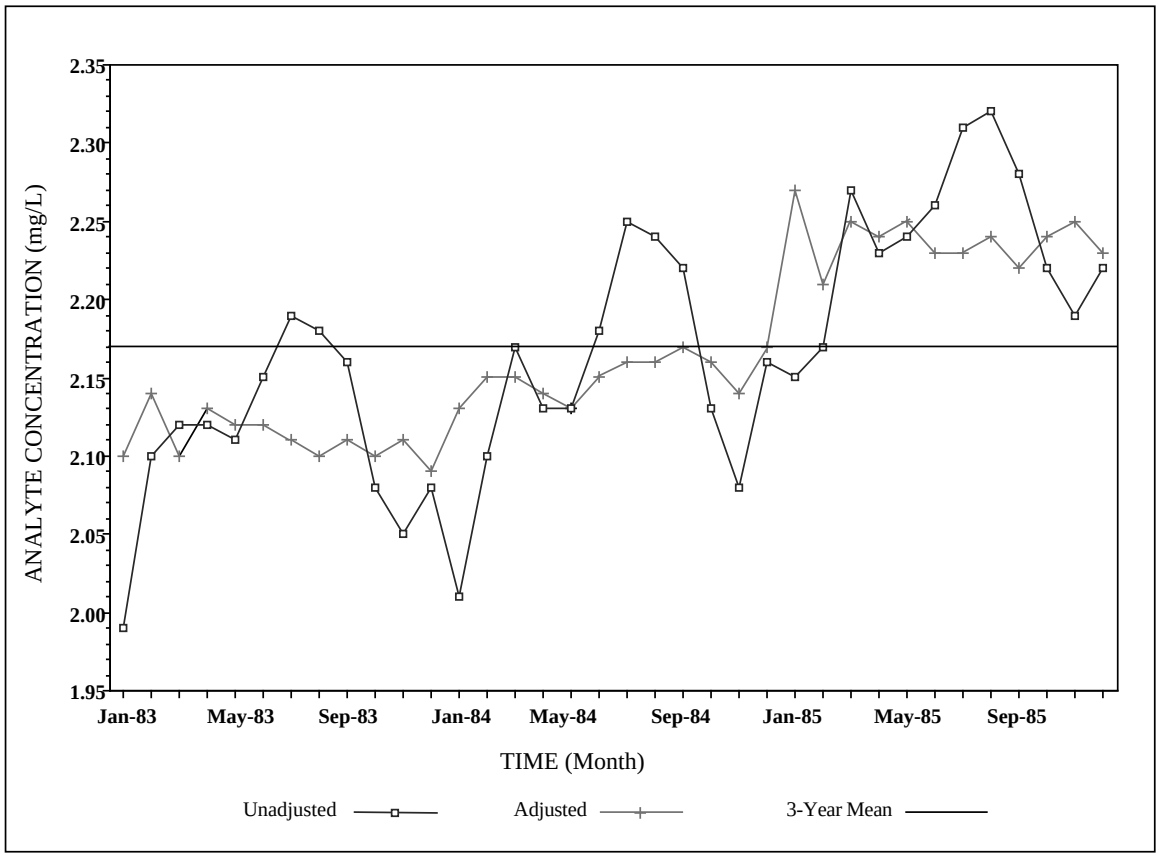
Step 3. Calculate the grand mean over the 3-year period using equation [14.22]:

$$\bar{x} = \frac{1}{3 \cdot 12}\left(1.99 + 2.01 + 2.15 + 2.10 \quad + \ldots + 2.22\right) = 2.17$$

Step 4. Within each month and year, subtract the average monthly concentration for that month and add-in the grand mean, using equation [14.23]. As an example, for January 1983, the adjusted concentration becomes:

$$z_{11} = 1.99 - 2.05 + 2.17 = 2.11$$

Figure 14-15. Seasonal Time Series Over a Three-Year Period




The adjusted concentrations are shown in the last three columns of the table above. The average of all 36 adjusted concentrations equals 2.17, the same as the mean *unadjusted* concentration. **Figure 14-15** shows the adjusted data superimposed on the unadjusted data. The raw data exhibit seasonality, as well as an upward trend. The adjusted data, on the other hand, no longer exhibit a seasonal pattern, although the upward trend still remains. From a statistical standpoint, the trend is much more easily identified by a trend test on the adjusted data than with the raw data. ◄

### 14.3.3.2 Correcting For A Temporal Effect Across Several Wells

#### BACKGROUND AND PURPOSE

When a significant temporal dependence or correlation is identified across a group of wells using one-way ANOVA for temporal effects (**Section 14.2.2**), results of the ANOVA can be used to create stationary adjusted data similar to the seasonal correction described in **Section 14.3.3.1**. The difference is that the adjustment is not applied to a data series at a single well, but rather simultaneously to several well sets.

The adjustment works in the same way as a correction for seasonality. First, the mean for each sampling event or season (averaged across wells) is computed along with the grand mean. Then each individual measurement is adjusted by subtracting off the event/seasonal mean and adding the overall or grand mean. In practice, this process is identical to adding the one-way ANOVA residual to the grand mean, so the already-computed results of the ANOVA can be used. By removing or correcting for a significant temporal effect, the adjusted data will have a temporally stationary mean and less overall variation. This allows for more powerful and accurate detection monitoring tests.

Temporal dependence (*e.g.*, seasonality) is sometimes observed as parallel traces on a time series plot across multiple wells (**Section 14.2.1**), although the one-way ANOVA for temporal effects is *non-significant*. This can occur due to the simultaneous presence of strong spatial variability (**Chapter 13**). Differences in mean levels from well to well can be large enough to 'swamp' the added variation due to the temporal dependence. The one-way ANOVA for temporal effects will not identify the dependence because the mean error sum of squares will then include the spatial variation component and not just random error.

Two remedies are possible when the ANOVA for temporal effects is non-significant. First, if a strong parallelism is evident on time series plots, the residuals from the ANOVA can still be used to create a set of adjusted, temporally-stationary measurements. The adjustment will not eliminate or remove any existing spatial variation, but it may not matter. Intrawell tests are needed anyway when such spatial variability is evident, and those tests assume temporal independence of the measurements collected at each well.

A second remedy is to perform a *two-way* ANOVA, testing for both spatial variation and temporal effects. This procedure is discussed in Davis (1994). Not only will a two-way ANOVA more readily identify a significant temporal effect even when there is simultaneous spatial variability, but the *F*-statistic used to test for the temporal dependence can be utilized to further adjust the appropriate degrees of freedom in intrawell background limits, such as prediction limits and control charts.

## REQUIREMENTS AND ASSUMPTIONS

The key requirement to correct for a temporal effect using ANOVA is that the same effect must be present in all wells to which the adjustment is applied. Otherwise, the adjustment will tend to skew or bias measurements at wells with no observable temporal dependence. Parallel time series plots (**Section 14.2.1**) should be examined to determine whether all the wells under consideration exhibit a similar temporal pattern.

The parametric one-way ANOVA assumes the data are normal or can be normalized. If the data cannot be normalized, a Kruskal-Wallis non-parametric ANOVA can be conducted to test for the presence of a temporal dependence. In this case, no residuals can be computed since the Kruskal-Wallis test employs ranks of the data rather than the measurements themselves. So the adjustment presented below is only applicable for data sets that can be normalized.

## PROCEDURE

Step 1. Given a set of $W$ wells and measurements from each of $T$ sampling events at each well on each of $K$ years, label the observations as $x_{ijk}$, for $i$ = 1 to $W$, $j$ = 1 to $T$, and $k$ = 1 to $K$. Then $x_{ijk}$ represents the measurement from the $i$th well on the $j$th sampling event during the $k$th year.

Step 2. Using the one-way ANOVA for temporal effects (**Section 14.2.2**), compute the sampling event or seasonal means (whichever is appropriate), along with the grand (overall) mean. Also construct the ANOVA residuals using either equation [14.5] or [14.6].

Step 3. Add each residual to the grand mean to form adjusted values $z_{ijk} = x_{\bullet\bullet\bullet} + r_{ijk}$. Use these adjusted values in subsequent statistical testing instead of the original measurements.

►EXAMPLE 14-9

The manganese data of **Examples 14-1** and **14-2** were found to have a significant temporal dependence using ANOVA for temporal effects. Adjust these data to remove the temporal pattern.

| | **Manganese Residuals (ppm)** | | | | |
|---|---|---|---|---|---|
| Qtr | Event Mean | BW-1 | BW-2 | BW-3 | BW-4 |
| 1 | 29.290 | -1.15 | 2.12 | -2.14 | 1.17 |
| 2 | 30.110 | -0.78 | 0.16 | 0.13 | 0.49 |
| 3 | 30.780 | -0.33 | 1.79 | -1.64 | 0.18 |
| 4 | 31.620 | 0.80 | 1.15 | -1.03 | -0.92 |
| 5 | 33.747 | 0.6225 | -0.7175 | 1.1325 | -1.0375 |
| 6 | 31.930 | 1.32 | 0.25 | -1.40 | -0.17 |
| 7 | 30.513 | 0.5075 | -1.6625 | -0.1825 | 1.3375 |
| 8 | 30.345 | -1.845 | 2.535 | 0.075 | -0.765 |
| | Grand mean = 31.042 | | | | |

SOLUTION

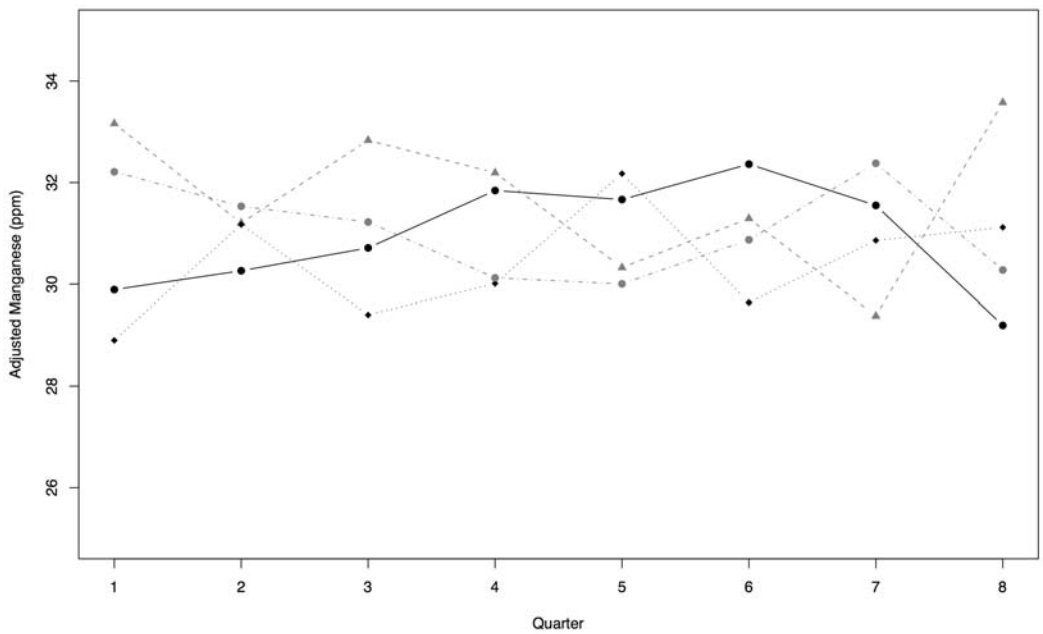
Step 1. The mean of each sampling event taken across the four background wells was computed in **Example 14-2**, along with the grand mean. These results are listed in the table above, along with the individual residuals which were also computed in that example.

Step 2. Add the grand mean to each residual to form the adjusted manganese concentrations, as in the table below.

| | **Adjusted Manganese (ppm)** | | | | |
|---|---|---|---|---|---|
| Qtr | Event Mean | BW-1 | BW-2 | BW-3 | BW-4 |
| 1 | 29.290 | 29.89 | 33.16 | 28.90 | 32.21 |
| 2 | 30.110 | 30.26 | 31.20 | 31.17 | 31.53 |
| 3 | 30.780 | 30.71 | 32.83 | 29.40 | 31.22 |
| 4 | 31.620 | 31.84 | 32.19 | 30.01 | 30.12 |
| 5 | 33.747 | 31.66 | 30.32 | 32.17 | 30.00 |
| 6 | 31.930 | 32.36 | 31.29 | 29.64 | 30.87 |
| 7 | 30.513 | 31.55 | 29.38 | 30.86 | 32.38 |
| 8 | 30.345 | 29.20 | 33.58 | 31.12 | 30.28 |
| | Grand mean = 31.042 | | | | |

Step 3. Plot a time series of the adjusted manganese values, as in **Figure 14-16**. The 'hump-like' temporal pattern evident in **Figure 14-2** is no longer apparent. Instead, the overall mean is stationary across the 8 quarters. ◄

Figure 14-16. Parallel Time Series Plot of Adjusted Manganese Concentrations




### 14.3.3.3 CORRECTING FOR LINEAR TRENDS

If a data series exhibits a linear trend, the sample will exhibit temporal dependence when tested via the sample autocorrelation function (**Section 14.2.3**), the rank von Neumann ratio (**Section 14.2.4**), or similar procedure. These data can be de-trended, much like the data in the previous example were deseasonalized. Probably the easiest way to de-trend observations with a linear trend is to compute a linear regression on the data (**Section 17.3.1**) and then use the regression *residuals* instead of the original measurements in subsequent statistical analysis.

But no matter how tempting it may be to automatically de-trend data of this sort, the user is strongly cautioned to consider what a linear trend may represent. Often, an upward trend is indicative of changing groundwater conditions at a site, perhaps due to the increasing presence of contaminants during a gradual waste release. The trend in this case may itself be statistically significant evidence of groundwater contamination, particularly if it occurs at compliance wells but not at upgradient background wells. The trend tests of **Chapter 17** are useful for such determinations. Trends in background may signal other important factors, including migration of contaminants from off-site sources, changes in the regional aquifer, or possible groundwater mounding.

The overriding point is that data should be deseasonalized when a cyclical pattern might obscure the random deviations around an otherwise stable average concentration level, or to more clearly identify an existing trend. However, a linear trend is inherently indicative of a changing mean level. Such data should not be de-trended before it is determined what the trend likely represents, and whether or not it is itself *prima facie* evidence of possible groundwater contamination.

A similar trend both in direction and slope may be exhibited by background wells *and* compliance wells, perhaps suggestive of sitewide changes in natural groundwater conditions. Residuals from a one-way ANOVA for temporal effects (**Section 14.2.2**) can be used to simultaneously create adjusted values across the well network (**Section 14.3.3.2**). Linear trends are just as easily identified and adjusted in this way as are parallel seasonal fluctuations or other temporal effects.

## 14.3.4 IDENTIFYING LINEAR TRENDS AMIDST SEASONALITY: SEASONAL MANN-KENDALL TEST

### BACKGROUND AND PURPOSE

Corrections for seasonality or other cyclical patterns over time in a single well are discussed in **Section 14.3.3.1**. These adjustments work best when the long-term mean at the well is stationary. In cases where a test for trend is desired and there are also seasonal fluctuations, **Chapter 17** tests may not be sensitive enough to detect a real trend due to the added seasonal variation.

One possible remedy is to use the seasonal correction in **Section 14.3.3.1** and illustrated in **Example 14-8**. The seasonal component of the trend is removed prior to conducting a formal trend test. A second option is the seasonal Mann-Kendall test (Gilbert, 1987).

The seasonal Mann-Kendall is a simple modification to the Mann-Kendall test for trend (**Section 17.3.2**) that accounts for apparent seasonal fluctuations. The basic idea is to divide a longer multi-year data series into subsets, each subset representing the measurements collected on a common sampling event (*e.g.*, all January events or all fourth quarter events). These subsets then represent different points along the regular seasonal cycle, some associated with peaks and others with troughs. The usual Mann-Kendall test is performed on each subset separately and a Mann-Kendall test statistic $S_i$ formed for each. Then the separate $S_i$ statistics are summed to get an overall Mann-Kendall statistic $S$.

Assuming that the same basic trend impacts each subset, the combined statistic $S$ will be powerful enough to identify a trend despite the seasonal fluctuations.

### REQUIREMENTS AND ASSUMPTIONS

The basic requirements of the Mann-Kendall trend test are discussed in **Section 17.3.2**. The only differences with the seasonal Mann-Kendall test are that 1) the sample should be a multi-year series with an observable seasonal pattern each year; 2) each 'season' or subset of the overall series should include at least three measurements in order to compute the Mann-Kendall statistic; and 3) a normal approximation to the overall Mann-Kendall test statistic must be tenable. This will generally be the case if the series has at least 10-12 measurements.

## PROCEDURE

Step 1. Given a series of measurements from each of $T$ sampling events on each of $K$ years, label the observations as $x_{ij}$, for $i$ = 1 to $T$, and $j$ = 1 to $K$. Then $x_{ij}$ represents the measurement from the $i$th sampling event during the $j$th year.

Step 2. For each distinct sampling event ($i$), form a seasonal subset by grouping together observations $x_{i1}, x_{i2},...., x_{iK}$. This results in $T$ separate seasons.

Step 3. For each seasonal subset, use the procedure in **Section 17.3.2** to compute the Mann-Kendall statistic $S_i$ and its standard deviation $SD[S_i]$. Form the overall seasonal Mann-Kendall statistic ($S$) and its standard deviation with the equations:

$$S = \sum_{i=1}^{T} S_i \qquad [14.24]$$

$$SD\left[S\right] = \sqrt{\sum_{i=1}^{T} SD^2\left[S_i\right]} \qquad [14.25]$$

Step 4. Compute the normal approximation to the seasonal Mann-Kendall statistic using the equation:

$$Z = \left(\left|S\right| - 1\right) \Big/ SD\left[S\right] \qquad [14.26]$$

Step 5. Given significance level, $\alpha$, determine the critical point $z_{cp}$ from the standard normal distribution in **Table 10-1** of **Appendix D**. Compare $Z$ against this critical point. If $Z > z_{cp}$, conclude there is statistically significant evidence at the $\alpha$-level of an increasing trend. If $Z < -z_{cp}$, conclude there is statistically significant evidence of a decreasing trend. If neither, conclude that the sample evidence is insufficient to identify a trend.

### ►EXAMPLE 14-10

The data set in **Example 14-8** replicated below indicated both clear seasonality and an apparent increasing trend. Use the seasonal Mann-Kendall procedure to test for a significant trend with $\alpha$ = 0.01 significance.

| | Analyte Concentrations | | | | |
|---|---|---|---|---|---|
| | 1983 | 1984 | 1985 | $S_i$ | $SD[S_i]$ |
| January | 1.99 | 2.01 | 2.15 | 3 | 1.915 |
| February | 2.10 | 2.10 | 2.17 | 2 | 1.633 |
| March | 2.12 | 2.17 | 2.27 | 3 | 1.915 |
| April | 2.12 | 2.13 | 2.23 | 3 | 1.915 |
| May | 2.11 | 2.13 | 2.24 | 3 | 1.915 |
| June | 2.15 | 2.18 | 2.26 | 3 | 1.915 |
| July | 2.19 | 2.25 | 2.31 | 3 | 1.915 |
| August | 2.18 | 2.24 | 2.32 | 3 | 1.915 |
| September | 2.16 | 2.22 | 2.28 | 3 | 1.915 |
| October | 2.08 | 2.13 | 2.22 | 3 | 1.915 |
| November | 2.05 | 2.08 | 2.19 | 3 | 1.915 |
| December | 2.08 | 2.16 | 2.22 | 3 | 1.915 |
| | | | | $S$ = 35 | $SD[S]$= 6.558 |

SOLUTION

Step 1. Form a seasonal subset for each month by grouping all the January measurements, all the February measurements, and so on, across the 3 years of sampling. This gives 12 seasonal subsets with $n$ = 3 measurements per season. Note there are no tied values in any of the seasons except for February.

Step 2. Use equations [17.30] and [17.31] in **Section 17.3.2** to compute the Mann-Kendall statistic ($S_i$) for each subset. These values are listed in the table above. Also compute their sum to form the overall seasonal Mann-Kendall statistic, giving $S$ = 35.

Step 3. Use equation [17.28] from **Section 17.3.2** for all months but February to compute the standard deviation of $S_i$. Since $n$ = 3 for each of these subsets, this gives

$$SD\left[S_i\right] = \sqrt{\frac{1}{18} n(n-1)(2n+5)} = \sqrt{\frac{1}{18} 3 \cdot 2 \cdot 11} = 1.915$$

For the month of February, one pair of tied values exists. Use equation [17.27] to compute the standard deviation for this subset:

$$SD\left[S_i\right] = \sqrt{\frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} t_j\left(t_j - 1\right)\left(2t_j + 5\right)\right]} = \sqrt{\frac{1}{18}\left[3 \cdot 2 \cdot 11 - 2 \cdot 1 \cdot 9\right]} = 1.633$$

List all the subset standard deviations in the table above. Then use equation [14.25] to compute the overall standard deviation:

$$SD\left[S\right] = \sqrt{\sum_{i=1}^{12} SD^2\left[S_i\right]} = \sqrt{11 \cdot (1.915)^2 + (1.633)^2} = 6.558$$

Step 4. Compute a normal approximation to $S$ with equation [17.29]:

$$Z = \left(35 - 1\right) / 6.558 = 5.18$$

Step 5. Compare $Z$ against the 1% critical point from the standard normal distribution in **Table 10-1** of **Appendix D**, $z_{.01}$ = 2.33. Since $Z$ is clearly larger than $z_{.01}$, the increasing trend evidence in **Figure 14-15** is highly significant. ◄

# CHAPTER 15. MANAGING NON-DETECT DATA

15.1 GENERAL CONSIDERATIONS FOR NON-DETECT DATA ..... 15-1
15.2 IMPUTING NON-DETECT VALUES BY SIMPLE SUBSTITUTION ..... 15-3
15.3 ESTIMATION BY KAPLAN-MEIER ..... 15-7
15.4 ROBUST REGRESSION ON ORDER STATISTICS ..... 15-13
15.5 OTHER METHODS FOR A SINGLE CENSORING LIMIT ..... 15-21
15.5.1 COHEN'S METHOD ..... 15-21
15.5.2 PARAMETRIC REGRESSION ON ORDER STATISTICS ..... 15-23
15.6 USE OF THE 15%/50% NON-DETECTS RULE ..... 15.24

This chapter considers strategies for accommodating non-detect measurements in groundwater data analysis. Five particular methods are described for incorporating non-detects into parametric statistical procedures. These include:

- Simple substitution (**Section 15.2**);
- Kaplan-Meier (**Section 15.3**);
- Robust Regression on Order Statistics (**Section 15.4**);
- Cohen's Method (**Section 15.5.1**); and
- Parametric Regression on Order Statistics (**Section 15.5.2**).

## 15.1 GENERAL CONSIDERATIONS FOR NON-DETECT DATA

Non-detects commonly reported in groundwater monitoring are statistically known as "left-censored" measurements, because the concentration of any non-detect either cannot be estimated or is not reported directly. Rather, it is known or assumed only to fall within a certain range of concentration values (*e.g.*, between zero and the *quantitation limit* [QL]). The direct estimate has been censored by the limitations of the measurement process or analytical technique, and is deemed too uncertain to be considered reliable. Groundwater non-detect data are censored on the low or left end of a sample concentration range. Other kinds of threshold data, particularly survival rates in the medical literature, are often reported as right-censored values.

Historically, there has been inconsistent treatment of non-detects in groundwater analysis. Often, easily applied techniques have been favored over more sophisticated methods of handling non-detects. This may primarily be due to the lack of familiarity and difficulties with software that can incorporate such methods. Even at present, most statistical packages include analysis routines for right-censored values but not left-censored ones (Helsel, 2005). Left-censored data needs to be converted to right-censored data for analysis and then back again. Despite these limitations, the more sophisticated methods are almost always superior to the methods of simple substitution.

The past twenty years has seen considerable research on statistical aspects of non-detect data analysis. Helsel (2005) provides a detailed summary of available methods for non-detects, and

concludes that simple substitution usually leads to greater statistical bias and inaccuracy than with better technical methods. Gibbons (1994b) and Gibbons & Coleman (2001) offer a broad review of some of the same research, not all of it directly relating to groundwater data. Both Gibbons and McNichols & Davis (1988) note that most of the existing studies focus on an *estimation of parameters* such as the mean and variance of an underlying population from which the censored and detected data originate. For these tasks, simple substitution methods tend to perform poorly, especially when the non-detect percentages are high (Gilliom & Helsel, 1986).

Much less attention has been given to how left-censored data impact the *results of statistical tests*, the actual data-based conclusions that are drawn when using detection, compliance, or corrective action monitoring tests. Closely estimating the true mean and variance of the underlying background population may be important, but does not directly answer how well a given test performs (in achieving the nominal false positive error rate and correctly identifying true significant differences). McNichols & Davis (1988) performed a limited study to address these concerns. They found that simple substitution methods were among the best performers in simulated prediction limit tests even with fairly high rates of censoring, *so long as* the prediction limit procedure incorporated a verification resample.

Gibbons (1994b; also Gibbons and Coleman, 2001) conducted a similar limited simulation of prediction limit testing performance incorporating a verification resample. They, too, found that a type of simple substitution was one of the best performers when either an average of 20% or 50% of the data was non-detect. The Gibbons study concluded that substituting zero for each non-detect worked better to keep the false positive rate low than by substituting half the method detection limit [MDL].

Both studies primarily focused on the achievable false positive rate when censored data are present, rather than the statistical power of these tests to identify contaminated groundwater. In addition, both only considered parametric prediction limits. For data sets with fairly low detection frequencies (*e.g.*, <50%), parametric prediction limits may not accurately accommodate left-censored measurements, with or without retesting. The McNichols & Davis study in particular found that *none* of the simpler methods for handling non-detects did well when the underlying data came from a skewed distribution and the non-detect percentage was over 50%.

On balance, there are four general strategies for handling non-detects: 1) employing a test specifically designed to accommodate non-detects, such as the Tarone-Ware two-sample alternative to the *t*-test (**Section 16.3**); 2) using a rank-based, non-parametric test such as the Kruskal-Wallis alternative to analysis of variance [ANOVA] (**Section 17.1.2**) when the non-detects and detects can be jointly sorted and ordered (except for tied values); 3) estimating the mean and standard deviation of samples containing non-detects by means of a *censored estimation technique*; and 4) *imputing an estimated value* for each non-detect prior to further statistical manipulation.

The first two strategies mentioned above are discussed in **Chapters 16** and **17** of the Unified Guidance as alternative testing procedures for evaluating left-censored data when parametric distribution assumptions cannot be made. Tests that can accommodate non-detects are typically non-parametric and thus carry both the advantages and disadvantages of non-parametric methods. The third and fourth strategies — presented in this chapter — are often employed as an *intermediate step* in parametric analyses. Estimates of the background mean and standard deviation are needed to construct parametric prediction and control chart limits, as well as confidence intervals. Imputed values for individual non-detects can be used as an alternate way to construct mean and standard deviation estimates, which are

needed to update the *cumulative sum* [CUSUM] portion of control charts or to compute the means of order *p* that get compared against prediction limits.

The guidance generally favors the use of the more sophisticated Kaplan-Meier or Robust ROS methods which can address the problem of multiple detection limits. Two older techniques-- Cohen's method and parametric ROS-- are also included as somewhat easier methods which can work in some circumstances. Applying any of the four estimation techniques as well as simple substitution does rely on a fundamental underlying assumption. Both the detectable and non-detect portions of a data set are assumed to arise from a single distribution, and in particular this underlying population is expected to be stable or *stationary* during the period of the sampling record.

However, if an underlying distribution is subject to a trend over time, applying any of these techniques including simple substitution is more problematic. If data indicating a decreasing trend also happen to contain multiple detection limit data (perhaps the result of improved analytical methods), it may be very difficult to determine whether there is truly a trend or analytical problems are the apparent cause of the observed decreases. None of the techniques provided in this chapter can directly address this issue. As discussed in **Chapter 5**, careful exploratory review of the historical data sets, particularly those which might serve as background, need to consider which data including non-detects are most representative of present or near-term future conditions. In some cases, removal of the older, less reliable data may also resolve multiple detection limit problems. If non-detect values higher than other quantified data at reasonable detection limits are included in a data set (especially if dictated by reporting policy rather than analytical considerations), these will almost invariably need to be removed. Even sophisticated multiple detection limit techniques cannot realistically address these particular information-limited data values. But presuming valid and reliable data are selected, the four estimation techniques are provided to address the management of non-detects.

A data set may also not be defined by a single distribution. If observed data are the result of two or more different generative processes and indicate one or more separate peaks, it is referred to as a mixture distribution. One example might be trace organics data in a release subject to changes in the flow direction of the aquifer, which can result in very high to absent values. The subject is a complex one and generally beyond the scope of this guidance. Aitchison's method can be used in limited situations where detectable data form one discrete distribution, and the remainder are non-detect. The following discussion also addresses when Aitchison's method might be appropriate. The non-detect data are simply considered as some single value, another form of simple substitution.

## 15.2 IMPUTING NON-DETECT VALUES BY SIMPLE SUBSTITUTION

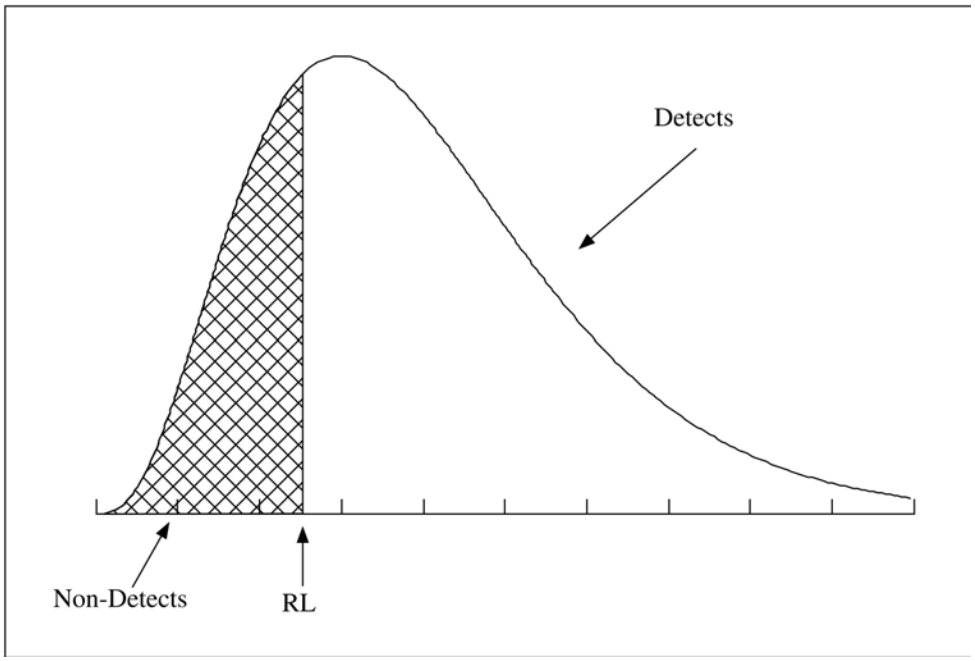
The simplest approach in managing non-detects is to substitute an *imputed value* for each prior to subsequent statistical analysis. The imputation is intended to be a ‘reasonable estimate’ of the true, but unknown concentration, usually a fraction (*e.g.*, 0, ½, 1) of the reporting limit [RL]. If non-detects represent an *absence* of the contaminant being measured, replacing a reported ‘less than’ value by zero may make sense. If the true concentration is completely unknown, but believed to be between zero and the RL, half the RL, or RL/2, may be a reasonable substitution, since this choice is the *maximum likelihood estimate* [MLE] of the mean or median for a population of measurement values *uniformly*

*distributed* along the interval [0, RL].[1] In other cases, a conservative choice might be made to maximize the *possible* concentration levels present in non-detects by selecting the RL itself as the imputation.

Any of these substitution choices is imperfect since they ignore two realities about left-censored measurements. First, non-detects are a product of *both* the underlying distribution of actual concentrations *and* the measurement process used to estimate these concentrations. In particular, the measurement technique may impart random or not so random bias to the 'true' concentration levels, causing the reported values to be 'shifted away from' the true values. As an example, simple substitution of zero for each non-detect ignores the fact that only the *measurements* can be observed and analyzed, not the actual concentration levels. Physical groundwater samples that are completely devoid of a given chemical may not receive *measurements* of zero, even if the actual amount is zero. Simple substitution by zero thereby ignores the *measurement distribution* in favor of an *a priori* assumption about what non-detects might represent.

A second reality is that non-detects must be considered with respect to other, detected measurements, as well as the physical process that generated the data. In many cases, the entire sample is drawn from a single statistical distribution (representing a common physical process) but some portion of the lower tail has been censored during measurement, as illustrated in **Figure 15-1**. In this situation, the overall distribution (and especially the shape of the lower tail) dictates how likely it is that a given non-detect would have an *uncensored* measurement close to zero or close to the RL. Substitution by half the RL or by the RL itself ignores the larger distributional pattern, especially since this distribution will rarely be uniform in the interval [0, RL].

Figure 15-1. Single Distributional Model For Detects and Non-Detects




[1] The uniform distribution places equal probability along every point of a finite concentration or measurement range. This model implies that a true value close to zero is just as likely as a true value close to RL or any other point along the interval.

These realities can lead to severe biases in statistical parameter estimates made from censored data when simple substitution methods are used (Helsel, 2005). Even if only 20% of the data are censored, Gibbons (1994b) found that the false positive rate of a prediction limit test was far above the *nominal* (*i.e.*, expected or targeted) rate of $\alpha = .05$ when a simple imputation strategy was employed. For that reason, the Unified Guidance recommends imputation by simple substitution only in select circumstances described below:

❖ ***When the sample size is too small to do anything else***.

With only a handful of measurements (*e.g.*, 5 or less), it will be almost impossible to accurately apply a censored estimation technique, such as those described in **Sections 15-3 to 15-5**. Instead, simple substitution of half the RL is recommended, perhaps until enough data has been collected to allow a more sophisticated analysis. Three situations where simple substitution might commonly be needed include:

1. Plotting cumulative sums [CUSUM] on control charts (**Chapter 20**). While there should be enough background data to allow for a more sophisticated estimate of the control limit, the CUSUM must be updated with each single new compliance observation ($n = 1$). If the new measurement is a non-detect, the value must be imputed for the CUSUM to be calculated.

2. Constructing future means for prediction limits (**Chapter 19**). Again, if censored data exist in background, the prediction limit for a future mean can be computed with the help of a censored estimation technique. But with only 2 or 3 new measurements per compliance well ($p = 2, 3$), the same strategy will not work for computing a mean of order $p$.

3. Construction of confidence intervals in compliance monitoring or corrective action. Especially in the early months or years after the onset of compliance monitoring or a corrective action plan, there may be too few compliance point measurements to allow for a statistically refined treatment of non-detects. Until more data has been collected that is representative of the conditions under which these phases of monitoring have been triggered, simple substitution of non-detects will probably be needed. Furthermore, if groundwater conditions are in a state of flux, it may be impossible — even with a larger sample size — to postulate a single, stationary distributional model (similar to **Figure 15-1**) on which to base a censored estimation technique.

❖ ***When non-detects comprise no more than 10-15% of the total sample***.

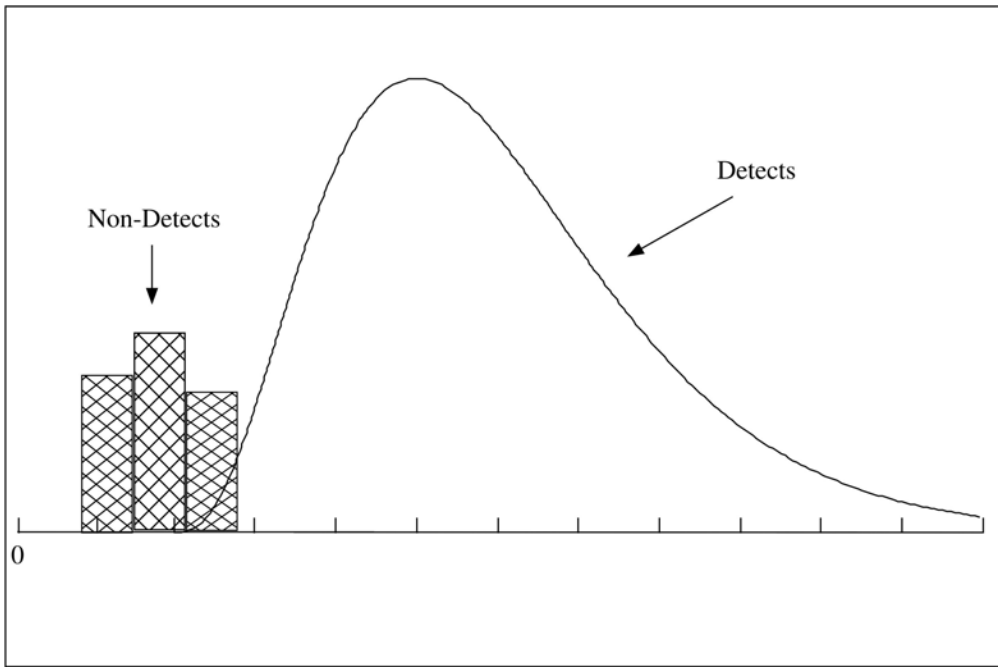
If the percentage of non-detects is small enough, results of parametric *t*-tests and ANOVA are usually not significantly affected if non-detects are first replaced by half their reporting limits [RLs]. A similar statement can be made for parametric prediction limits, tolerance limits, control charts, and confidence intervals. However, because *t*-tests and ANOVA involve a comparison of means utilizing multiple data points per mean estimate,[2] while prediction limits for individual observations, tolerance limits, and control charts focus on single measurements, it is important that *retesting* be included in the statistical procedure whenever simple substitution is utilized with these latter methods.

[2] Parametric confidence intervals around the mean also involve an estimate of the population average using multiple data points.

- ***When non-detects are generated by a different physical process than the detected values, and thus represent a distinct statistical distribution***.

One non-detect treatment recommended in past EPA guidance — Aitchison's method (1955), as applied to groundwater[3] — assumed that non-detects were actually *free* of the contaminant being measured, so that all non-detects could be regarded as zero concentrations. In some cases, if an analyte has been detected infrequently or not at all in background measurements, and/or all non-detects are qualified as "U" (*i.e.*, undetected) values, this assumption may be practical, even if it cannot be directly verified. Another example might be seasonal changes in groundwater elevation at wells located on the edges of a contaminant plume. Parameters detectable at certain times of the year may be non-detect during other seasons, even within the same well. Such non-detects may result from a different data-generating mechanism, due to seasonal changes in groundwater chemistry, and so may not follow the same distribution as detects.

Figure 15-2. Modified Delta Model For Mixture Distribution of Detects/Non-Detects




More generally, Aitchison's original model posited a 'spike' of zero-valued measurements, combined with a lognormal distribution governing the detected values. A modification to Aitchison's model known as the *modified delta method*[4] (USEPA, 1993) has been found to be more practical and realistic in many circumstances (**Figure 15-2**). Instead of assuming that all non-detects represent zero

[3] Aitchison's model was not originally applied to concentration data. More typical applications were in the fields of economics and demographics.

[4] The original Aitchison model was termed the *delta-lognormal*, so called because it presumed that the data consisted of a mixture of two distinct populations: 1) a lognormal distribution representing the observed continuous measurements, and 2) a 'spike' of values, known as a delta function, located at zero.

concentrations, the modified delta method assumes that non-detects constitute a separate, discrete distribution. When combined with the detected values, a *mixture distribution* is formed consisting of a continuous detected portion (usually the normal or lognormal distribution) and a discrete non-detect portion. Rather than assuming that all non-detects are zeros, the modified delta model assigns all non-detects at half the reporting limit [RL]. (Note: this might be a method detection limit [MDL], a quantitation limit [QL], or a contract RL). This method can accommodate multiple reporting limits since each non-detect is assigned half of its possibly sample-specific RL. It can also accommodate low-valued detects *intermingled* with the non-detects, since the non-detects and detects are modeled by distinct distributions.

## 15.3 ESTIMATION BY KAPLAN-MEIER

### BACKGROUND AND PURPOSE

When a sample contains both detects and non-detects generated by a common process and governed by a single underlying distribution (**Figure 15-1**), a more reliable strategy is to attempt to fit the sample to a known distribution (*e.g.*, normal, lognormal) and then to estimate the mean and standard deviation of this distribution via a *censored estimation technique*. These adjusted estimates can be input into standard equations for parametric prediction, tolerance, and control chart limits, as well parametric confidence intervals around the mean.

Two censored estimation methods which can address the multiple detection limit problem are discussed in the Unified Guidance: the Kaplan-Meier estimator and *robust regression on order statistics* [ROS] (**Section 15.4**). Both involve initially fitting a left-censored sample to a known distribution. After that, the procedures differ. The Kaplan-Meier creates an estimate of the population mean and standard deviation adjusted for data censoring, based on the fitted distributional model, whereas the Robust ROS uses the fitted model to construct a *model-based imputation* for each non-detect. Once the imputations are made, the adjusted mean and standard deviation are estimated using standard equations for the sample mean ($\bar{x}$) and standard deviation ($s$).

The key to either method is finding a single distributional model that adequately fits the joint sample of detects and non-detects. While each procedure does the fitting in a slightly different fashion, both utilize the notion of *partial ranking*. As discussed in **Section 16.2** on "Handling Non-Detects," the presence of left-censored measurements, particularly when there are multiple RLs and/or an *intermingling* of detects and non-detects, prevents a full and complete ranking of the sample. Both Kaplan-Meier and ROS construct a partial ranking of the data, accounting for the non-detects and assigning explicit ranks to each of the detected values. These detected values can then be graphed on a *censored probability plot* and fitted against a known distribution.

The Kaplan-Meier technique estimates the approximate proportion of concentrations below each observed level by sorting and ordering the distinct sample values, although the exact concentrations of non-detects are unknown. In particular, the probability of observing a concentration no greater than a given level ($x_i$) depends on the relative proportion of the sample greater than $x_i$. Any detects larger than $x_i$ obviously fall into this latter proportion, while non-detects with RLs of at most $x_i$ do not. On balance, the proportion of the sample greater than $x_i$ cannot be precisely calculated for every $x_i$, but it can be estimated.

The Kaplan-Meier estimator for left-censored data thus depends on a series of *conditional* probabilities, where the frequency of lower concentrations depends on how many larger concentrations have already been observed. The final result is an estimate of the *cumulative distribution function* [CDF] for each distinct concentration level in the sample.

In mathematical notation, suppose there are $m$ distinct values in the sample (out of a total of $n$ measurements), including distinct reporting limits. Order these values from least to greatest and denote them as $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$. Let $n_i$ for $i = 1$ to $m$ denote the 'risk set' associated with value $x_{(i)}$. The *risk set* represents the total number of measurements — both detects and non-detects — no greater than $x_{(i)}$. Since a non-detect with a RL larger than $x_{(i)}$ is *potentially* (but not necessarily) larger than $x_{(i)}$, non-detects with RL > $x_{(i)}$ are *not* included in $n_i$. A further term $d_i$ identifies the number of detected measurements exactly equal to $x_{(i)}$.

With these definitions in place and letting $X$ denote a random variable concentration from the true underlying distribution, the Kaplan-Meier estimator is constructed from the pair of probabilities:

$$\Pr\left(X \le x_{(m)}\right) = 1 \qquad [15.1]$$

$$\Pr\left(X \le x_{(i)} \middle| X \le x_{(i+1)}\right) = 1 - \frac{d_{i+1}}{n_{i+1}} \quad for\ \ i = 1\ \ to\ \ m \qquad [15.2]$$

where $x_{(m+1)} = +\infty$, $d_{m+1} = 0$, and $n_{m+1} = n$ all by definition. Equation [15.2] represents the conditional probability that the concentration does not exceed $x_{(i)}$ given that it does not exceed $x_{(i+1)}$. The final Kaplan-Meier CDF estimate ($F_{KM}$) for each $i = 1$ to $m$–1 (each distinct detected value) is given by a product of these conditional probabilities and can be expressed as:

$$F_{KM}\left(x_{(i)}\right) = \Pr\left(X \le x_{(i)}\right) = \left(1 - \frac{d_{i+1}}{n_{i+1}}\right) \times \left(1 - \frac{d_{i+2}}{n_{i+2}}\right) \times \ldots \times \left(1 - \frac{d_m}{n_m}\right) = \prod_{k=i}^{m-1}\left(1 - \frac{d_{k+1}}{n_{k+1}}\right) \qquad [15.3]$$

Once the CDF is estimated using equation [15.3], two additional steps are made possible. One is to use the distinct values ($x_{(i)}$) and their corresponding CDF values ($F_{KM}$) to construct censored probability plots. The other is to use the Kaplan-Meier CDF to estimate the population mean and standard deviation.

## REQUIREMENTS AND ASSUMPTIONS

The Kaplan-Meier estimator is a non-parametric procedure originally devised to estimate *survival probabilities* for right-censored samples (Kaplan and Meier, 1958), such as in medical studies of cancer treatments. Because it is non-parametric, there is no requirement that the underlying population be normal or transformable to normality. However, in adapting the technique to left-censored data (*i.e.*, samples containing non-detects), the Unified Guidance recommends that the Kaplan-Meier procedure be utilized to estimate the mean and variance of a normal or normalized distribution for use in *parametric* statistical tests.

The Kaplan-Meier assumes that all detected and non-detect data arise from the same population, but that non-detect values have been 'censored' at their RLs. This implies that the contaminant of

concern is *actually present* in non-detect samples, but that the analytical method cannot accurately measure, or is not sufficiently sensitive to, concentrations lower than the RL.

To construct a censored probability plot, a normal quantile or $z$-score needs to be computed for each value of the Kaplan-Meier CDF ($F_{KM}$). Doing so is straightforward except for the CDF value of the sample maximum, which is assigned a value of one. The $z$-score associated with a cumulative probability of one is infinite. To surmount this difficulty, the Unified Guidance recommends temporarily setting the CDF value for the sample maximum equal to $(n - .375)/(n + .25)$. This value is the Blom plotting position often utilized in standard probability plots (Helsel, 2005). It is close to one for large $n$, but allows for a finite $z$-score.

Estimation of the Kaplan-Meier mean and standard deviation using equations [15.4] and [15.5] below will tend to be slightly biased, typically with the mean on the high side and the standard deviation on the low side. This occurs because the Kaplan-Meier CDF levels corresponding to distinct RLs are treated as if they were known measurements rather than the upper bounds on possible values. As long as the total proportion of censored measurements is not too high, the degree of bias will tend to be small. Larger biases are more likely whenever the detection rate is less than 50%.

## PROCEDURE

Step 1. Given a sample of size $n$ containing left-censored measurements, identify and sort the $m < n$ distinct values, including distinct RLs. Label these as $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$.

Step 2. For each $i = 1$ to $m$, calculate the *risk set* ($n_i$) as the total number of detects and non-detects no greater than $x_{(i)}$. Also compute $d_i$ as the number of *detected* values exactly equal to $x_{(i)}$.

Step 3. Using equation [15.3], compute the Kaplan-Meier CDF estimate $F_{KM}\left(x_{(i)}\right)$ for $i = 1, \ldots, m-1$. Also let $F_{KM}\left(x_{(m)}\right) = 1$.

Step 4. Construct censored probability plots using the estimated CDF. First temporarily set $F_{KM}\left(x_{(m)}\right) = (n - .375)/(n + .25)$ so that a finite normal quantile (or $z$-score; see **Chapter 9**) can be associated with $x_{(m)}$. Then compute normal quantiles (*i.e.*, $z$-scores) for each value of $F_{KM}$ from **Step 3** as $z_{(i)} = \Phi^{-1}\left[F_{KM}\left(x_{(i)}\right)\right]$, where $\Phi^{-1}[\cdot]$ is the inverse of the standard normal distribution function as discussed in the construction of probability plots in **Chapter 9**. Plot the values $z_{(i)}$ against the unique detected concentrations $x_{(i)}$ to form a *normal* censored probability plot. Plot the $z_{(i)}$'s against a transformation of the $x_{(i)}$'s (*e.g.*, log, square root, inverse, *etc.*) to form a *normalized* censored probability plot.

Step 5. For each attempted transformation $f(\cdot)$ including the unchanged observations as one option, compute the *correlation coefficient* between the pairs $[f(x_{(i)}), z_{(i)}]$ (**Chapter 3**). The transformation with the highest correlation coefficient and also a linear appearance on the censored probability plot, is one that optimally normalizes the left-censored sample. Estimate the mean and standard deviation in Step 6 on the transformed scale and use these estimates in subsequent statistical analysis.

If no transformation results in an adequately linear censored probability plot, conclude that the sample cannot be normalized. Mean and standard deviation estimates of the original concentrations can still be computed, but they will not correspond to a known probability distribution.

Step 6. If the raw concentration data are approximately normal, compute mean and standard deviation estimates adjusted for censoring using the equations:

$$\hat{\mu}_{KM} = \sum_{i=1}^{m} x_{(i)} \cdot \left[F_{KM}\left(x_{(i)}\right) - F_{KM}\left(x_{(i-1)}\right)\right] \qquad [15.4]$$

$$\hat{\sigma}_{KM} = \sqrt{\sum_{i=1}^{m} \left(x_{(i)} - \hat{\mu}_{KM}\right)^2 \cdot \left[F_{KM}\left(x_{(i)}\right) - F_{KM}\left(x_{(i-1)}\right)\right]} \qquad [15.5]$$

where $x_{(0)} = 0$ and $F_{\mathrm{KM}}\left(x_{(0)}\right) = F_{\mathrm{KM}}\left(0\right) = 0$ by definition. Otherwise, compute the adjusted mean and standard deviation after applying the normalizing transformation $f(\cdot)$ with the equations:

$$\hat{\mu}_{KM} = \sum_{i=1}^{m} f\left(x_{(i)}\right) \cdot \left[F_{KM}\left(x_{(i)}\right) - F_{KM}\left(x_{(i-1)}\right)\right] \qquad [15.6]$$

$$\hat{\sigma}_{KM} = \sqrt{\sum_{i=1}^{m} \left(f\left(x_i\right) - \hat{\mu}_{KM}\right)^2 \cdot \left[F_{KM}\left(x_{(i)}\right) - F_{KM}\left(x_{(i-1)}\right)\right]} \qquad [15.7]$$

Estimates from equations [15.4] and [15.5] can then be used in place of the sample mean ($\bar{x}$) and standard deviation ($s$) in parametric equations for prediction and control limits, and for confidence intervals. If a normalizing transformation is required, equations [15.6] and [15.7] can be used to construct similar statistical limits and intervals on the *transformed* scale.

►EXAMPLE 15-1

Use the Kaplan-Meier technique on the following manganese concentration data to construct estimates of the population mean and standard deviation that are adjusted for censoring.

| | **Manganese Concentrations (ppb) in Background** | | | | |
|---|---|---|---|---|---|
| Sample | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 |
| 1 | <5.0 | <5.0 | <5.0 | 6.3 | 17.9 |
| 2 | 12.1 | 7.7 | 5.3 | 11.9 | 22.7 |
| 3 | 16.9 | 53.6 | 12.6 | 10.0 | 3.3 |
| 4 | 21.6 | 9.5 | 106.3 | <2.0 | 8.4 |
| 5 | <2.0 | 45.9 | 34.5 | 77.2 | <2.0 |

SOLUTION

Step 1. From the combined sample of $n$ = 25 measurements, identify and sort the 21 distinct values including distinct RLs as in the table below. Compute the risk set ($n_i$) for each distinct level ($x_{(i)}$) as the total number of detects and non-detects no greater than $x_{(i)}$. Also calculate the exact number of detects ($d_i$) equal to each level.

Step 2. Compute the Kaplan-Meier estimate of the CDF using equations [15.1] and [15.3], shown in column 5 of the table below. Two example calculations are given by:

$$F_{\mathrm{KM}}(22.7) = \left(1-\frac{1}{21}\right)\left(1-\frac{1}{22}\right)\left(1-\frac{1}{23}\right)\left(1-\frac{1}{24}\right)\left(1-\frac{1}{25}\right) = 0.8$$

$$F_{KM}(3.3) = \left(1-\frac{0}{7}\right)\cdot\left(1-\frac{1}{8}\right)\cdot\left(1-\frac{1}{9}\right)\cdot\ldots\cdot\left(1-\frac{1}{24}\right)\cdot\left(1-\frac{1}{25}\right) = 0.28$$

| i | $x_{(i)}$ | At Risk ($n_i$) | $d_i$ | CDF |
|---|---|---|---|---|
| 1 | <2.0 | 3 | 0 | 0.21 |
| 2 | 3.3 | 4 | 1 | 0.28 |
| 3 | <5.0 | 7 | 0 | 0.28 |
| 4 | 5.3 | 8 | 1 | 0.32 |
| 5 | 6.3 | 9 | 1 | 0.36 |
| 6 | 7.7 | 10 | 1 | 0.40 |
| 7 | 8.4 | 11 | 1 | 0.44 |
| 8 | 9.5 | 12 | 1 | 0.48 |
| 9 | 10.0 | 13 | 1 | 0.52 |
| 10 | 11.9 | 14 | 1 | 0.56 |
| 11 | 12.1 | 15 | 1 | 0.60 |
| 12 | 12.6 | 16 | 1 | 0.64 |
| 13 | 16.9 | 17 | 1 | 0.68 |
| 14 | 17.9 | 18 | 1 | 0.72 |
| 15 | 21.6 | 19 | 1 | 0.76 |
| 16 | 22.7 | 20 | 1 | 0.80 |
| 17 | 34.5 | 21 | 1 | 0.84 |
| 18 | 45.9 | 22 | 1 | 0.88 |
| 19 | 53.6 | 23 | 1 | 0.92 |
| 20 | 77.2 | 24 | 1 | 0.96 |
| 21 | 106.3 | 25 | 1 | 1.00 |

Step 3. Compute normal quantiles or z-scores for each value of $F_{\mathrm{KM}}$ in the above table. First re-set the last entry to ($n$ – .375)/($n$ + .25) = 0.9752 so that a finite quantile can be associated with the sample maximum.

Step 4. Plot the $z$-scores against the distinct manganese levels to form a normal censored probability plot (**Figure 15-3**). The probability plot correlation coefficient is $r$ = 0.902. The plot itself shows substantial curvature, suggesting that the sample is non-normal.

Step 5. Plot the *z*-scores against one or more transformations of the manganese levels. First attempt a log transformation, as shown in **Figure 15-4**. In this case, the correlation coefficient improves to $r = 0.989$ and the normalized censored probability plot looks fairly linear. Conclude that the sample is approximately normal on the log-scale, that is, the manganese concentrations are lognormal in distribution.

Step 6. Compute Kaplan-Meier log-mean ($\hat{\mu}_{y,KM}$) and log-standard deviation ($\hat{\sigma}_{y,KM}$) estimates for the manganese data using equations [15.6] and [15.7], taking $f(\cdot)$ as the natural logarithm. This gives for the log-mean:

$$\hat{\mu}_{y,KM} = \log(2)\cdot[.21-0] + \log(3.3)\cdot[.28-.21] + \ldots + \log(106.3)\cdot[1-.96] = 2.31\log(ppb)$$

and for the log-standard deviation:

$$\hat{\sigma}_{y,KM} = \sqrt{(\log(2)-2.31)^2\cdot[.21-0]+\ldots+(\log(106.3)-2.31)^2\cdot[1-.96]} = 1.18\ \log(ppb)$$

These adjusted mean and standard deviation estimates can then be used in place of the sample log-mean and log-standard deviation in parametric prediction and control limits, or in parametric confidence intervals. ◀

Figure 15-3. Censored Probability Plot of Manganese Concentrations

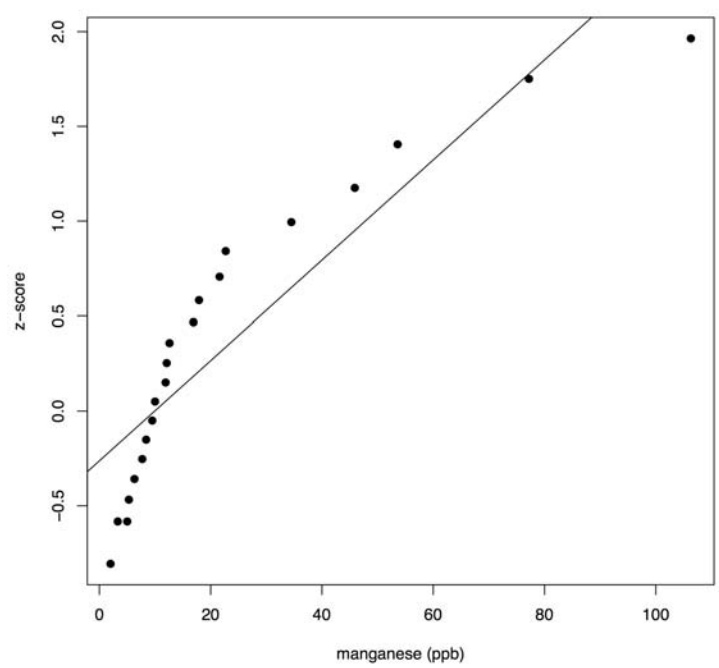
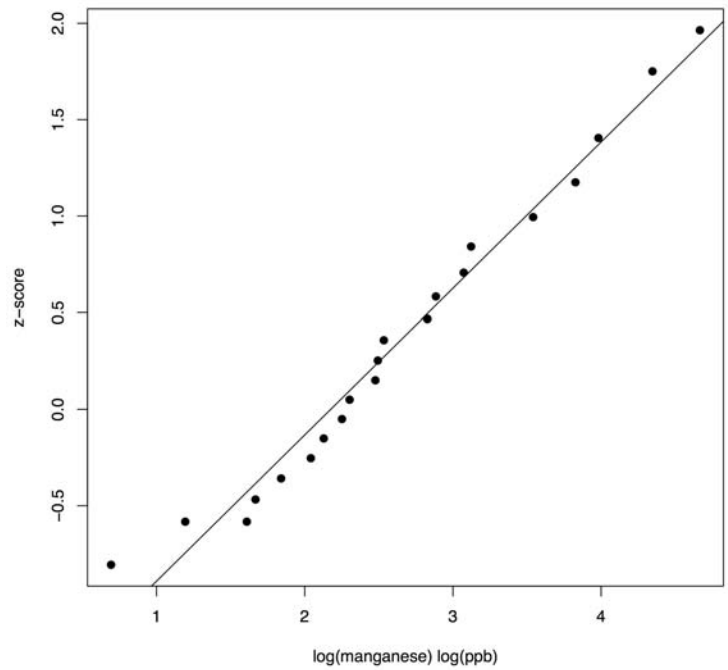


Figure 15-4. Censored Probability Plot of Logged Manganese Sample




## 15.4 ROBUST REGRESSION ON ORDER STATISTICS

### BACKGROUND AND PURPOSE

*Robust regression on order statistics* [ROS] differs from Kaplan-Meier in that it uses the fitted model to construct a ***model-based imputation*** for each non-detect. Once the imputations are made, the adjusted mean and standard deviation are estimated using standard equations for the sample mean ($\bar{x}$) and standard deviation ($s$).

The first step in using Robust ROS is to find a single distributional model that adequately fits the joint sample of detects and non-detects. Standard probability plots (**Chapter 9**) and normality tests (**Chapter 10**) rely on a full ranking or ordering of the sample in order to fit candidate distributions. With left-censored data, the true concentrations of non-detects are unknown, so only a *partial ranking* is possible. Like Kaplan-Meier, the Robust ROS technique constructs a partial ranking of the data, accounting for the non-detects and assigning explicit ranks to each of the detected values. These detected values can be graphed on a *censored probability plot* to check the fit of possible distributional models.

Once an adequate distribution is found, Robust ROS determines the approximate cumulative probability associated with each distinct RL. The method then arbitrarily distributes non-detects with a common RL so that each one accounts for an equal share of the estimated cumulative probability

assigned to that RL. Once non-detects are ranked in this manner, the fitted distributional model is used to impute a value for each non-detect. This last task is accomplished by conducting a linear regression (Chapter 17) between the detected values and the z-scores from the censored probability plot. The parameters of the regression line (i.e., intercept and slope) can be used to estimate the mean and standard deviation of the distributional model, which in turn will generate imputed values for the non-detects.

The mathematics behind Robust ROS can be expressed as follows. First suppose there are $k$ distinct RLs in the sample. Order these from least to greatest. Define $A_i$ as the number of detected values between the $i$th and ($i$+1)th RLs for $i$ = 1 to $k$–1. Let $A_k$ = number of detects above the highest RL, and take $A_0$ = number of detects below the lowest RL. Also define $B_i$ as the total number of observations, both detects and non-detects, with values below the $i$th RL. Define $B_0$ = 0. Then the number of non-detects below the $i$th RL can be written as:

$$C_i = B_i - B_{i-1} - A_{i-1} \qquad for\ i = 1\ to\ k \qquad [15.8]$$

With these definitions in place, exceedance probabilities can be assigned to each of the $k$ RLs, representing the proportion of the sample greater than or equal to each distinct RL. These probabilities can be written as:

$$pe_i = pe_{i+1} + \frac{A_i}{A_i + B_i}\left(1 - pe_{i+1}\right) \qquad [15.9]$$

where $pe_j$ denotes the proportion of the sample exceeding the $i$th RL. Equation [15.9] can be interpreted in the following manner. The exceedance probability associated with a given RL is equal to the exceedance probability assigned to the next highest RL combined with a fraction of the remaining, non-exceedance probability (*i.e.*, 1 – $pe_{i+1}$). The specific fraction depends on the relative occurrence of detects between the $i$th and ($i$+1)th RLs. When $i = k$, define $pe_{i+1}$ = 0; when $i$ = 0, define $pe_0$ = 1.

Once the exceedance probabilities are computed, plotting positions for the detects — *i.e.*, cumulative probabilities on a probability plot — can be calculated with the equation

$$pd_{ij} = \left(1 - pe_i\right) + \left(\frac{j}{A_i + 1}\right) \cdot \left(pe_i - pe_{i+1}\right) \qquad for\ j = 1\ to\ A_i;\quad and\ i = 0\ to\ k \qquad [15.10]$$

for each set of detected values falling between the $i$th and ($i$+1)th RLs. Note that this equation also applies to any detects below the lowest RL [$i$ = 0] or above the highest RL [$i$ = $k$]). Similarly, plotting positions for each group of non-detects can be written as:

$$pc_{ij} = \left(\frac{j}{C_i + 1}\right) \cdot \left(1 - pe_i\right) \qquad for\ j = 1\ to\ C_i;\ and\ i = 1\ to\ k \qquad [15.11]$$

With plotting positions for the detects, a normal quantile or $z$-score can be computed for each value of $pd_{ij}$. Then censored probability plots can be constructed using either the detected concentrations ($x_{ij}$) or some normalizing transformation of the detected values, say f($x_{ij}$). If a linear probability plot can be identified, a linear regression (**Chapter 17**) can be calculated for the pairs ($z_{ij}$, $f(x_{ij})$) and used to impute values for the non-detects in the sample.

## REQUIREMENTS AND ASSUMPTIONS

Robust ROS was originally devised to account for non-detects in water quality data (Helsel, 2005). Robust ROS is an extension of a technique termed *regression on order statistics [ROS]* (Gilliom and Helsel, 1986), described in **Section 15.5.** That procedure assumes the joint sample of detects and non-detects follows an underlying lognormal distribution. The fitted lognormal is used to estimate the population mean and standard deviation as a parametric technique. Robust ROS by contrast only relies on a parametric model to impute values for the non-detects. It can be applied to any normal or normalized distribution, rather than just the lognormal distribution. It may also be regarded as quasi-non-parametric since estimates for the sample are computed from the combined group of observed detects and imputed non-detects, rather than from the mean and standard deviation of the underlying distributional model, as in the original formulation.

In practice, because Robust ROS is not fully non-parametric, a known distribution must be fitted to the entire sample in order to construct imputed values for the non-detects. Closely related to this, Robust ROS assumes that both detected and non-detect data arise from the same population, with non-detect values censored at their respective RLs. Like Kaplan-Meier, this implies that the contaminant of concern is *present* in non-detect samples, but that the analytical method cannot accurately measure concentrations lower than the RL.

## PROCEDURE

Step 1. Given a left-censored sample with a total of $n$ measurements, identify and sort the $k$ distinct RLs. Following the discussion above, count the number of detected values below the lowest RL ($A_0$), the number of detected values at least as great as the highest RL ($A_k$), and the number of detects between the $i$th and ($i$+1)th RLs ($A_i$ for $i$ = 1 to $k$–1). Also let $B_0 = 0$ and count the total number of detects and non-detects below the $i$th RL ($B_i$ for $i$ = 1 to $k$). Then use equation [15.8] to calculate the number of non-detects ($C_i$ for $i$ = 1 to $k$) below the $i$th RL.

Step 2. Let $pe_0 = 1$ and $pe_{k+1} = 0$. For $i$ = 1 to $k$, compute the probability of exceeding the $i$th distinct RL ($pe_i$) using equation [15.9].

Step 3. With the exceedance probabilities from Step 2, sort each group of detects associated with $A_i$ and then compute plotting positions (*i.e.*, cumulative probabilities) for these detects — $pd_{ij}$ — using equation [15.10].

Step 4. Form normal quantiles (*i.e.*, $z$-scores) associated with the detected measurements and plotting positions $pd_{ij}$ by computing $z_{ij}^d = \Phi^{-1}\left(pd_{ij}\right)$, where $\Phi^{-1}\left(\cdot\right)$ is the inverse standard normal CDF.

Step 5. Construct censored probability plots using the $z$-scores from Step 4. Plot the values $z_{ij}^d$ against the detected concentrations $x_{ij}^d$ to form a *normal* censored probability plot. Plot the $z_{ij}^d$'s against a transformation of the $x_{ij}^d$'s (*e.g.*, log, square root, inverse, *etc.*) to form a *normalized* censored probability plot.

Step 6. For each attempted transformation $f(\cdot)$ including the unchanged observations as one option, compute the *correlation coefficient* between the pairs $\left[ f\left(x_{ij}^{d}\right), z_{ij}^{d} \right]$ (**Chapter 3**). The transformation with the highest correlation coefficient and also a linear appearance on the censored probability plot, is the one that optimally normalizes the left-censored sample. If no transformation results in an adequately linear censored probability plot, conclude that the sample cannot be normalized and that the Robust ROS may not provide reasonable imputations for the non-detects.

Step 7. If a normalizing transformation can be identified, compute a linear regression (**Chapter 17**) of the values $f\left(x_{ij}^{d}\right)$ on the *z*-scores, $z_{ij}^{d}$, to form the regression equation $f(X) = \hat{a} + \hat{b} \cdot Z$. The slope and intercept can be estimated using the equations

$$\hat{b} = \sum_{i=0}^{k} \sum_{j=1}^{A_i} \left(z_{ij}^{d} - \bar{z}_d\right) \cdot f\left(x_{ij}^{d}\right) \Big/ \left(n_d - 1\right) \cdot s_{z_d}^{2} \qquad [15.12]$$

$$\hat{a} = \bar{x}_d - \hat{b} \cdot \bar{z}_d \qquad [15.13]$$

where $\bar{z}_d$ is the mean of the *z*-scores associated with the detected values, $n_d$ = number of detects, $s_{z_d}^{2}$ is the sample variance of the detected *z*-scores, and $\bar{x}_d$ is the mean of the detected measurements. The regression intercept ($\hat{a}$) is an estimate of the population mean of the normalized distribution, while the slope ($\hat{b}$) is an estimate of the population standard deviation.

Step 8. Compute plotting positions ($pc_{ij}$) for the non-detects (*i.e.*, censored observations) associated with each distinct RL using equation [15.11]. Then form a second set of *z*-scores, this time associated with the non-detects, by computing $z_{ij}^{c} = \Phi^{-1}\left(pc_{ij}\right)$ for $j$ = 1 to $C_i$; and $i$ = 1 to $k$.

Step 9. Form imputed values $f\left(\hat{x}_{ij}^{c}\right) = \hat{a} + \hat{b} \cdot z_{ij}^{c}$ using the slope and intercept from **Step 7** and the censored z-scores from **Step 8**. Combine these (transformed) imputed values for the non-detects with the (transformed) detected measurements $f\left(x_{ij}^{d}\right)$ to get censored estimates of the population mean and standard deviation by computing the overall sample mean ($\hat{\mu} = \bar{x}$) and sample standard deviation ($\hat{\sigma} = s$).

These censored estimates can be used in place of the unadjusted sample mean ($\bar{x}$) and standard deviation ($s$) in parametric equations for prediction and control limits, and for confidence intervals. If a normalizing transformation $f(\cdot)$ is needed, the censored estimates should be used to construct statistical limits and intervals on the *transformed* scale.

►EXAMPLE 15-2

In **Example 15-1**, the Kaplan-Meier technique was used on a sample of background manganese concentrations to compute the log-mean and log-standard deviation, adjusted for the presence of non-detects. Apply Robust ROS to these same data to compare the estimates.

SOLUTION

Step 1. The n = 25 manganese observations include 2 distinct RLs (<2 and <5). Count the number of detected measurements below the lowest RL, above the highest RL, and between the two RLs, denoted by $A_i$ in the table below. Also count the total number of measurements — both detected and non-detect — below each RL, denoted below by $B_i$. Use equation [15.8] to count the number of non-detects associated with each RL, denoted below by $C_i$.

| i | RL | $A_i$ | $B_i$ | $C_i$ |
|---|---|---|---|---|
| 0 | | 0 | 0 | 0 |
| 1 | <2 | 1 | 3 | 3 |
| 2 | <5 | 18 | 7 | 3 |

Step 2. Compute the probability of exceeding each RL using equation [15.9] and noting that $pe_3 = 0$:

$$pe_2 = pe_3 + \frac{A_2}{A_2 + B_2}\left(1 - pe_3\right) = \frac{18}{18+7} = 0.72$$

$$pe_1 = pe_2 + \frac{A_1}{A_1 + B_1}\left(1 - pe_2\right) = 0.72 + \frac{1}{1+3}\left(1 - 0.72\right) = 0.79$$

Step 3. Sort the detects associated with each $A_i$ and compute plotting positions for these detects using equation [15.10], as listed in the table below. For instance, $A_1 = 1$, corresponding to the detected value 3.3. The plotting position for this observation equals

$$pd_{11} = \left(1 - pe_1\right) + \left(\frac{1}{A_1 + 1}\right)\left(pe_1 - pe_2\right) = 0.21 + 0.5\left(0.79 - 0.72\right) = 0.245$$

Also form the normal quantiles (i.e., z-scores) associated with the detected observations, as listed below:

| Detected Value (ppb) | Plotting Position | z-score |
|---|---|---|
| 3.3 | 0.245 | −0.690 |
| 5.3 | 0.318 | −0.474 |
| 6.3 | 0.356 | −0.370 |
| 7.7 | 0.394 | −0.270 |
| 8.4 | 0.432 | −0.172 |
| 9.5 | 0.469 | −0.077 |
| 10.0 | 0.507 | 0.018 |
| 11.9 | 0.545 | 0.114 |
| 12.1 | 0.583 | 0.210 |
| 12.6 | 0.621 | 0.308 |
| 16.9 | 0.659 | 0.410 |
| 17.9 | 0.697 | 0.515 |
| 21.6 | 0.735 | 0.627 |
| 22.7 | 0.773 | 0.748 |
| 34.5 | 0.811 | 0.880 |
| 45.9 | 0.848 | 1.030 |
| 53.6 | 0.886 | 1.207 |
| 77.2 | 0.924 | 1.434 |
| 106.3 | 0.962 | 1.776 |

Step 4. Plot the *z*-scores against the detected manganese levels to form a normal censored probability plot (**Figure 15-5)**. The probability plot correlation coefficient is $r = 0.901$, almost identical to the Kaplan-Meier censored probability plot constructed in **Example 15-1**. The plot also shows substantial curvature, suggesting that the sample is non-normal. Also plot the *z*-scores against a log transformation of the detected manganese values (**Figure 15-6**). Not only does the normalized probability plot appear linear, but the correlation coefficient increases to $r = 0.994$. Conclude as in **Example 15-1** that the sample is approximately normal on the log-scale, so that the manganese concentrations are lognormal in distribution.

Step 5. Compute a linear regression of the $n_d = 19$ logged manganese detects against their corresponding *z*-scores using equations [15.12] and [15.13]. The sample mean and variance of the detected *z*-scores are $\bar{z}_d = 0.3802$ and $s^2_{z_d} = 0.4577$. Also, the log-mean of the detected observations equals $\overline{\log(x_d)} = 2.80$. The slope and intercept of the resulting line are:

$$\hat{b} = \frac{1}{18 \times .4577}\left[1.194 \cdot (-.690 - .3802) + \ldots + 4.666(1.776 - .3802)\right] = 1.372$$

$$\hat{a} = \bar{x}_d - \hat{b} \cdot \bar{z}_d = 2.80 - 1.372 \times .3802 = 2.278$$

Figure 15-5. Robust ROS Censored Probability Plot of Manganese Concentrations

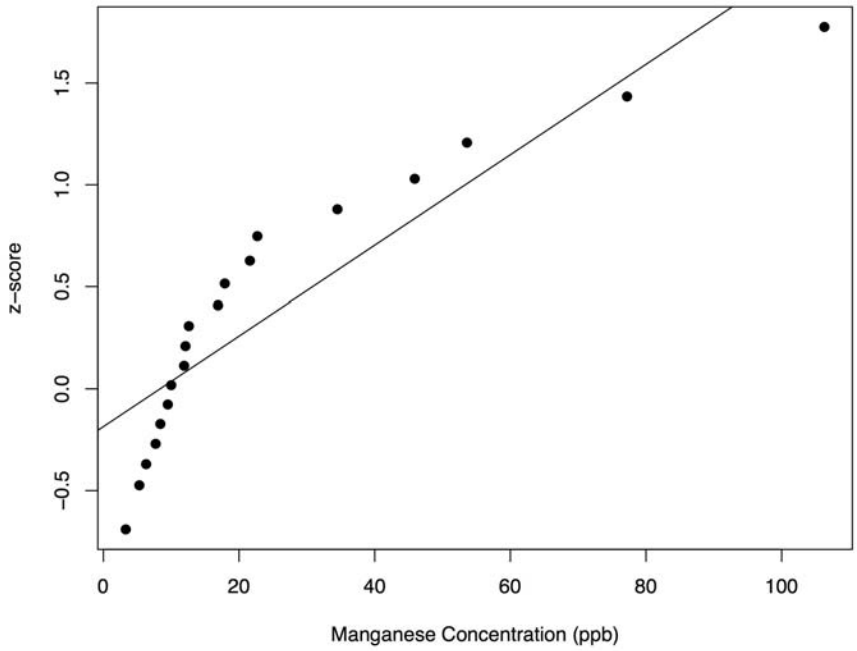
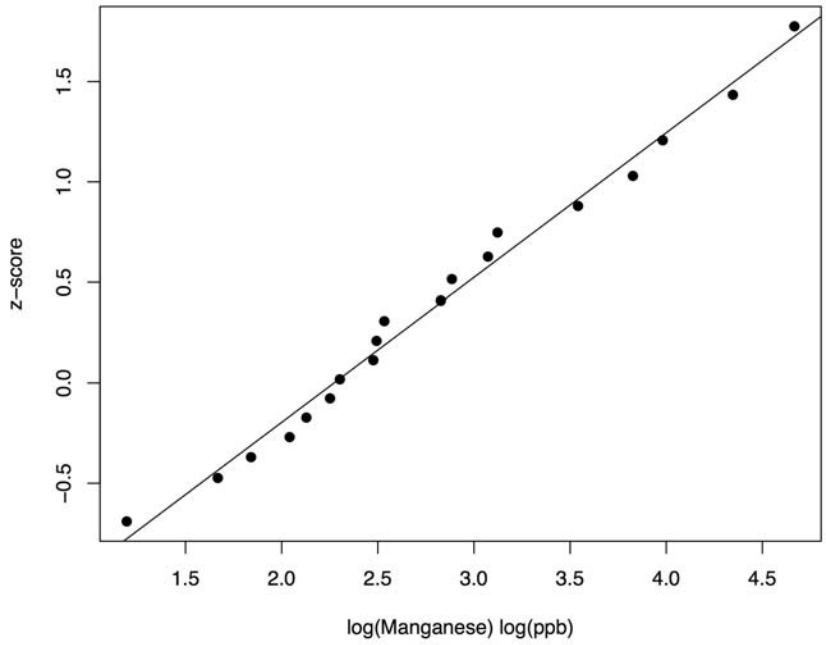



Figure 15-6. Robust ROS Censored Probability Plot of Logged Manganese



Step 6. Compute plotting positions for the non-detects (*i.e.*, censored observations) associated with each distinct RL using equation [15.11], listed in the table below. Form a second set of *z*-scores, this time associated with the non-detects, also listed below. Note that each non-detect is given a distinct plotting position, even though they cannot be ordered. This is done to 'fill in' the unknown portion of the underlying distribution, but should *not* be interpreted as a legitimate 'estimate' for any particular non-detect observation. The positions for the first pair of the 3 non-detects with RLs of 2 (*i.e.*, <2) are

$$pc_{11} = \left(\frac{1}{C_1+1}\right)\left(1-pe_1\right) = \left(\frac{1}{3+1}\right)\left(1-0.79\right) = 0.0525$$

$$pc_{12} = \left(\frac{2}{C_1+1}\right)\left(1-pe_1\right) = \left(\frac{2}{3+1}\right)\left(1-0.79\right) = 0.105$$

| RL | Plotting Position | z-score | Imputed Value |
|---|---|---|---|
| <2 | 0.0525 | −1.621 | 0.054 |
| <2 | 0.1050 | −1.254 | 0.558 |
| <2 | 0.1575 | −1.005 | 0.899 |
| <5 | 0.0700 | −1.476 | 0.253 |
| <5 | 0.1400 | −1.080 | 0.796 |
| <5 | 0.2100 | −0.806 | 1.172 |

Step 7. Form a second set of z-scores associated with the censored plotting positions from **Step 6**. These are listed in the table above. Then, using the regression parameters from **Step 5**, form a prediction for each non-detect using the equation $\log\left(x_{ij}^c\right) = \hat{\alpha} + \hat{\beta} \cdot z_{ij}^c$. Take these predictions as the imputed values for the set of non-detects, as listed above. The first two imputed values are computed as:

$$\log\left(x_{11}^c\right) = 2.278 + 1.372 \cdot \left(-1.621\right) = 0.054$$

$$\log\left(x_{12}^c\right) = 2.278 + 1.372 \cdot \left(-1.254\right) = 0.558$$

Step 8. Combine the logged detected manganese values with the imputed values from Step 7. Then compute the sample mean and standard deviation using the adjusted sample. These calculations give $\hat{\mu} = 2.28 \log(ppb)$ and $\hat{\sigma} = 1.26 \log(ppb)$. By comparison, the Kaplan-Meier method in **Example 15-1** gives very similar corresponding estimates of 2.31 log(ppb) and 1.18 log(ppb). ◄

# SECTION 15.5 OTHER METHODS FOR A SINGLE CENSORING LIMIT

The two preferred methods using Kaplan-Meier or Robust ROS provided above for multiple detection limits are computationally intensive. Helsel (2005) indicates that public software is available for the Robust ROS method. Although the more common situation encountered in evaluating data sets is the presence of multiple detection limits (hence the UG recommendations), two older techniques are still applicable in some situations. The Cohen method and the parametric ROS techniques are both simpler to apply, but depend on the use of a single censoring limit. One needs to evaluate the prospects before applying them. If detectable data sets are large enough (e.g., $n > 50$) and detection percentages near or greater than 50%, most of these methods will work comparably.

## 15.5.1 COHEN'S ADJUSTMENT

Cohen's adjustment (Cohen, 1959) can be useful when a significant fraction (up to 50%) of the observed measurements in a data set are reported as non-detects. The technique assumes that all the measurements, detects and non-detects alike, arise from a common population, but that the lowest valued observations have been *censored* at the QL. Using the censoring point (*i.e.*, QL) and the pattern in the detected values, Cohen's method attempts to reconstruct the key features of the original population, providing explicit estimates of the population mean and standard deviation. These in turn can be used in certain statistical interval estimates, where Cohen's adjusted estimates are used as replacements for the sample mean and sample standard deviation.

### REQUIREMENTS AND ASSUMPTIONS

Cohen's adjustment assumes that the common underlying population has a normal distribution. The technique should only be used when the observed sample data approximately fit a normal model including transformations to normality. Because the presence of a large fraction of non-detects will make explicit normality testing difficult, if not impossible, the most helpful diagnostic aid may be to construct a censored probability plot on the detected measurements. If the censored probability plot is clearly linear on the original measurement scale but not on the log-scale, assume normality for purposes of computing Cohen's adjustment. If, however, the censored probability plot is clearly linear on the log-scale, but not on the original scale, assume instead that the common underlying population is lognormal. Then compute Cohen's adjustment to the estimated mean and standard deviation on the log-scale measurements and construct the desired statistical interval using the algorithm for lognormally-distributed observations.

When the detection rate is less than 50%, the accuracy of Cohen's method worsens as the percentage of non-detects increases. The guidance does not generally recommend the use of Cohen's adjustment when more than half the data are non-detect. In such circumstances, one should consider an alternate statistical method, for instance a non-parametric interval or perhaps the Wilcoxon rank-sum test for small samples.

One other requirement of Cohen's original method is that there should be just a single censoring point. Data sets with multiple RLs will usually require a more sophisticated treatment such as Kaplan-Meier or Robust ROS methods or via maximum likelihood techniques (Cohen, 1963) or perhaps a multiply-censored probability plot technique (Helsel and Cohn, 1988). If only 2 or 3 RLs do not substantially differ and few detected intermingled data are lost, the censoring point (*QL*) can be set to

the highest RL. Cohen's method requires explicit definition of the censoring limit, and is somewhat sensitive to variation in this parameter.

## PROCEDURE

Step 1. Divide the data set into two groups, detects and non-detects. If the total sample size equals $n$, let $m$ represent the number of detects and ($n$–$m$) represent the number of non-detects. Denote the $i$th detected measurement by $x_i$. Then compute the mean and sample variance of the set of detects using the equations:

$$\bar{x}_d = \frac{1}{m}\sum_{i=1}^{m} x_i \quad \text{and} \quad s_d^2 = \frac{1}{m-1}\left[\sum_{i=1}^{m} x_i^2 - m\bar{x}_d^2\right]$$

Step 2. Denote the single censoring point by $QL$. Then compute the two intermediate quantities, $h$ and $\gamma$, necessary to derive Cohen's adjustment via the following equations:

$$h = 100 \cdot (n-m)/n = ND\% \quad \text{and} \quad \gamma = s_d^2 \Big/ \left(\bar{x}_d - QL\right)^2$$

Step 3. Use the intermediate quantities $h$ and $\gamma$ to determine Cohen's adjustment parameter $\lambda$ from the table below.

### Values of Lamba ($\lambda$) for Cohen's Adjustment

| $\gamma$\ND% | 1 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **.01** | .0102 | .0530 | .1111 | .1747 | .2443 | .3205 | .4043 | .4967 | .5989 | .7128 | .8403 |
| **.05** | .0105 | .0547 | .1143 | .1793 | .2503 | .3279 | .4130 | .5066 | .6101 | .7252 | .8540 |
| **.10** | .0110 | .0566 | .1180 | .1848 | .2574 | .3366 | .4233 | .5184 | .6234 | .7400 | .8703 |
| **.20** | .0116 | .0600 | .1247 | .1946 | .2703 | .3525 | .4422 | .5403 | .6483 | .7678 | .9012 |
| **.30** | .0122 | .0630 | .1306 | .2034 | .2819 | .3670 | .4595 | .5604 | .6713 | .7937 | .9300 |
| **.40** | .0128 | .0657 | .1360 | .2114 | .2926 | .3803 | .4755 | .5791 | .6927 | .8179 | .9570 |
| **.50** | .0133 | .0681 | .1409 | .2188 | .3025 | .3928 | .4904 | .5967 | .7129 | .8408 | .9826 |
| **.60** | .0137 | .0704 | .1455 | .2258 | .3118 | .4045 | .5046 | .6133 | .7320 | .8625 | 1.0070 |
| **.70** | .0142 | .0726 | .1499 | .2323 | .3206 | .4156 | .5180 | .6291 | .7502 | .8832 | 1.0303 |
| **.80** | .0146 | .0747 | .1540 | .2386 | .3290 | .4261 | .5308 | .6441 | .7676 | .9031 | 1.0527 |
| **.90** | .0150 | .0766 | .1579 | .2445 | .3370 | .4362 | .5430 | .6586 | .7844 | .9222 | 1.0743 |
| **1.00** | .0153 | .0785 | .1617 | .2502 | .3447 | .4459 | .5548 | .6725 | .8005 | .9406 | 1.0951 |
| **1.25** | .0162 | .0828 | .1705 | .2636 | .3627 | .4687 | .5825 | .7053 | .8385 | .9841 | 1.1443 |
| **1.50** | .0170 | .0868 | .1786 | .2758 | .3793 | .4897 | .6081 | .7357 | .8738 | 1.0245 | 1.1901 |
| **1.75** | .0177 | .0905 | .1861 | .2873 | .3948 | .5094 | .6321 | .7641 | .9069 | 1.0625 | 1.2332 |
| **2.00** | .0184 | .0940 | .1932 | .2981 | .4093 | .5279 | .6547 | .7909 | .9382 | 1.0984 | 1.2739 |
| **2.25** | .0191 | .0973 | .1999 | .3082 | .4231 | .5454 | .6761 | .8164 | .9679 | 1.1325 | 1.3127 |
| **2.50** | .0197 | .1005 | .2062 | .3179 | .4363 | .5621 | .6965 | .8407 | .9962 | 1.1651 | 1.3498 |
| **2.75** | .0203 | .1035 | .2123 | .3272 | .4489 | .5781 | .7161 | .8639 | 1.0234 | 1.1963 | 1.3854 |
| **3.00** | .0209 | .1063 | .2182 | .3361 | .4609 | .5935 | .7348 | .8863 | 1.0495 | 1.2264 | 1.4197 |
| **3.50** | .0219 | .1118 | .2292 | .3529 | .4838 | .6226 | .7704 | .9287 | 1.0990 | 1.2835 | 1.4847 |
| **4.00** | .0229 | .1168 | .2395 | .3687 | .5052 | .6498 | .8038 | .9685 | 1.1455 | 1.3371 | 1.5458 |
| **4.50** | .0239 | .1216 | .2492 | .3836 | .5253 | .6755 | .8353 | 1.0060 | 1.1895 | 1.3878 | 1.6037 |
| **5.00** | .0248 | .1262 | .2585 | .3977 | .5445 | .7000 | .8653 | 1.0418 | 1.2312 | 1.4359 | 1.6587 |
| **5.50** | .0256 | .1305 | .2673 | .4111 | .5628 | .7233 | .8938 | 1.0758 | 1.2711 | 1.4820 | 1.7113 |
| **6.00** | .0264 | .1346 | .2757 | .4240 | .5803 | .7456 | .9212 | 1.1085 | 1.3094 | 1.5262 | 1.7617 |

Step 4. Using the adjustment parameter λ found in Step 3, compute adjusted estimates of the population mean and standard deviation with the equations:

$$\hat{\mu} = \bar{x}_d - \lambda(\bar{x}_d - QL) \quad \text{and} \quad \hat{\sigma} = \sqrt{s_d^2 + \lambda \cdot (\bar{x}_d - QL)^2}$$

Step 5. Once the adjusted estimates for the population mean and standard deviation are derived, these values can be substituted for the sample mean and standard deviation in equations for the statistical intervals.

## 15.5.2 PARAMETRIC REGRESSION ON ORDER STATISTICS (ROS)

A second useful method (EPA, 2004) for estimating mean and standard deviation parameters for data sets with non-detect values censored at a single limit is a parametric Regression on Order Statistics (ROS). The same assumptions apply as with Cohen's method. Both the detected and non-detect portions of the data are presumed to arise from a single population. That population should either be normal or transformable to a normal distribution. The parametric ROS method performs similarly to Cohen's method, and offers two principal advantages. The procedure can easily be implemented on almost any statistical software, and the method is not sensitive to the exact censoring limit.

If variable $X$ originates from a normal distribution with mean $\mu$ and standard deviation $\sigma$ $[X \succ N(\mu,\sigma)]$ and $Z$ is the standard normal distribution $[Z \succ N(0,1)]$, statistical theory indicates that $X = \mu + \sigma \cdot Z$ when $X$ and $Z$ are at the same percentiles in their respective distributions. For a given observation or sample $x$ above a detection limit, the order statistic (i.e., the proportion of observations less than $x$) can be estimated. This order statistic is an estimate of the percentile. The corresponding $Z$-value can be obtained from reference tables or a computer algorithm. For a list of ordered observations above the detection limit $(x_1, x_2, \ldots\ldots \text{ to } x_m)$ of $m$ detectable samples out of a total $n$ and a corresponding set of Z-values $(Z_1, \quad Z_2, \ldots\ldots \text{ to } Z_m)$ at the same percentiles, regression analysis of $X$ against $Z$ will provide estimates of the mean and standard deviation of distribution $X$. The intercept is the mean estimate and the slope of the regression is the standard deviation estimate.

When sample data better fit a lognormal or other normal transformable distribution, the regression is performed on the transformed data. The mean and standard deviation estimates are also for the transformed data (e.g., logarithmic mean and standard deviation). One may also use the regression results to "fill in" or quantify the values below the detection limit. When the $Z$-distribution is developed for the full set of total $n$ sample values, the $Z$-values for the detectable portion are separated from those for the remaining $n - m$ non-detect percentiles. Estimates for the non-detect values are obtained from the equation $X = \hat{\mu} + \hat{\sigma} \cdot Z$, using $\hat{\mu}$ the intercept mean estimate, $\hat{\sigma}$ the slope standard deviation estimate and the non-detect Z-values. These can then be aggregated with the sample detectable values to obtain the overall mean and standard deviation estimate.

### PROCEDURE

Step 1. Determine the appropriate normal transformation and convert the data if necessary. Divide the data set into two groups, detects and non-detects. If the total sample size equals $n$, let $m$ represent the number of detects and $(n - m)$ represent the number of non-detects. Denote the $i$th detected measurement by $x_i$. Order the $m$ detected data from smallest to largest.

Step 2. Define the normal percentiles for the total n sample set as follows. For a set of *i* values from 1 to *n*, $p_i = (i - .375)/(n + .25)$. Then convert to Z-values using the inverse normal distribution $Z_i = \Phi^{-1}(p_i)$. Separate the $Z_i$ values into two groups: the larger *m* detected and *n - m* non-detected portions.

Step 3. Use linear regression of the ordered *m* data values against the corresponding Z-values. Obtain the intercept and slope of the regression as the estimated mean and standard deviation estimates, $\hat{\mu}$ and $\hat{\sigma}$. These can be used directly as the distributional parameter estimates or Step 4 can be followed.

Step 4. Using equation $X_{n-m} = \hat{\mu} + \hat{\sigma} \cdot Z_{n-m}$ with $\hat{\mu}$ the intercept mean estimate, $\hat{\sigma}$ the slope standard deviation estimate and the non-detect $Z_{n-m}$ values, calculate the remaining $x_{n-m}$ values and combine with the $x_m$ detected data. Use the combined direct sample mean and standard deviation calculations as the final parameter estimates:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

►EXAMPLE 15-3

Use Cohen's and the parametric ROS methods for the data in **Example 15-1** and compare the results to the Kaplan-Meier and Robust ROS Methods. A single overall logarithmic distribution can be assumed. In the example, it is possible to utilize the higher detection limit (<5) as the censoring limit, with the loss of only a single detected point of information. The detection frequency is still 72%.

For Cohen's method, $h$ = .28 and $\gamma$ = .465 for the logarithmic data. The adjustment parameter from the above table is interpolated as $\lambda$ = .445. The resulting mean and standard deviation estimates for the full data set are $\hat{\mu}$ = 2.32 log(ppb) and $\hat{\sigma}$ = 1.22 log(ppb).

Mean and standard deviation estimates for the parametric ROS method are $\hat{\mu}$ = 2.33 log(ppb) and $\hat{\sigma}$ = 1.21 log(ppb) following regression of the ordered detectable log values against the corresponding Z- values of the standard normal distribution. With such few non-detects near the lowest end of the sample distribution, the results are quite similar to the Robust ROS and Kaplan-Meier methods. For higher non-detect percentages and more heavily intermingled non-detect data, the results using these methods can differ considerably. ◄

## 15.6 USE OF THE 15% AND 50% NON-DETECT RULE

In this chapter and elsewhere in the Unified Guidance, it is recommended that imputing arbitrary values be limited to data sets with 10-15% or fewer non-detects and that parametric procedures be applied when there are 50% or fewer non-detects. The guidance continues to suggest this basic non-detects rule for both historical and conservative reasons. The same approach was found in both the earlier RCRA 1989 and 1992 RCRA statistical guidance documents, although it was recognized in the

first as a guideline "based on judgment". It was also noted that "there is no general procedure that is applicable in all cases." The 10-15% rule using direct substitution of arbitrary values is believed adequate for many applications, but one of the censoring estimation techniques provided in this chapter can be used instead. For a skewed distribution like the lognormal, the latter approach would be preferable. We have cited studies above by Davis and others indicating that parameter estimation and test performance can suffer when more than 50% of the data are non-detects. Most of the common parameters (i.e., mean, median, standard deviation, etc.) can be estimated with tolerable bias and error when no more than 50% of the values are originally non-detect and the superior non-detect fitting techniques used. Statistical test performance using these limitations appears to be reasonable for most applications. However, it should be recognized that they are only "rules of thumb", not absolute criteria.

Other authors (e.g., Helsel 2005) have suggested that certain tests will perform adequately even with higher non-detect rates in data. The criterion of non-detect percentage is not the only factor. For example with very large data sets (e.g., 100-300), quite reasonable fits can be made to the detectable portion using techniques found in **Chapter 15** even with non-detect percentages greater than 50%. Having a sufficient number of detectable data is also an important consideration, applying equally to small data sets. One should have a fairly good idea that the detect data themselves follow one or another parametric distributions. To do so, one should have a sufficiently large number of detected data points for comparison.

A second factor is the potential application for fitted non-detect data. As an example, fits of high non-detect percentage larger data sets using the lognormal distribution can provide decent parameter estimates (log mean and log standard deviation) for use with upper prediction limit detection monitoring tests. Generally, the fits accurately describe the upper portions of the observed data sets. At the same time, these estimated logarithmic parameters may result in considerably larger errors when estimating the true arithmetic mean and standard deviation (the bias problem in transformations), such as with compliance level tests. In this case, the 50% rule is best followed.

The guidance generally recommends non-parametric options when non-detect data exceed 50%. However, even this suggestion comes with caveats. For example, if a number of wells to be compared using Kruskal-Wallis non-parametric ANOVA had mostly or all well data sets greater than 50% non-detects, the outcome would be ambiguous. This is because the test involves comparisons of medians, which would lie below the detection limit. At very high non-detect percentages, fewer options are available. Upper non-parametric prediction limits can work with very few detectable values, but the assumption of any distributional pattern is increasingly tenuous. In some cases, a binomial test of proportions (found in the 1989 guidance) may be the only realistic option. As a final suggestion, we recommend that users take these factors into account and consider recommendations of other statistical literature in the field as well, when considering non-detect limitations to specific test procedures.

*This page intentionally left blank*