APPENDIX D
Descriptive Statistics

D-1. Introduction. For most environmental sampling, the collected data for some measurement variable of interest constitute a small subset of its set of possible values. The data subset frequently consists of contaminant concentrations from the analysis of environmental (e.g., soil and groundwater) samples collected from the study area. In a statistical context, this subset is referred to as a *sample*. If it were possible to collect environmental observations from every portion of the study area (i.e., to exhaustively sample an entire site), the set of resulting values would constitute the *population*. As this is typically not possible, statistics calculated from the *sample* are used to describe or make inferences about the underlying *population*. For the environmental applications discussed herein, the statistical methods presented are implicitly for a sample, not the entire population. For more information on populations, the reader is referred to introductory statistical texts readily available in libraries and online.

D-1.1. Commonly used descriptive statistics for environmental data include measures of central tendency, such as mean, median, or mode; measures of relative standing, such as percentiles; measures of dispersion, such as range, variance, standard deviation, coefficient of variation, or interquartile range; measures of distribution symmetry or shape; and measures of association between two or more variables, such as correlation. These measures can also be used to test hypotheses regarding the populations from which the data were drawn.

D-1.2. In general, the sampling design influences how descriptive statistical quantities are calculated. The formulas presented in this monograph are for simple random sampling, simple random sampling with composite samples, and randomized systematic sampling. If more complex designs are used, such as a stratified design, then the formulas need to be adjusted. All of these designs are addressed in Appendix C.

D-1.3. In addition, the distribution of a data set may also influence how descriptive statistical quantities are calculated. Most of the discussion in this Appendix will be centered on normal populations. However, as detailed in Appendix F, it is not uncommon for environmental data to follow other distributions. The most commonly encountered alternative is the lognormal distribution. This Appendix will also present how to calculate the mean and quantiles of the population for a lognormally distributed data set. To estimate other parameters, the reader is urged to refer to any of the excellent texts available, including those referenced here.

D-1.4. The terminology used in presenting general formulas and calculations for this exercise are standard. Out of a total population N, let $x_1, x_2, \ldots, x_n$ represent the n data points, a sample set of n measurements. Additional information on calculating descriptive statistics for environmental applications can be found in the EPA/600/R-96/084, QA/G-9 and Gilbert (1987).

D-2. Measures of Central Tendency. Measures of central tendency characterize the center of a set of measured data values. The three most common estimates are the mean, median, and mode. These are described below, and examples of calculating each of them are presented in Paragraph D-2.2

D-2.1. *Mean*. The mean is the most commonly used measure of central tendency. The formula used to calculate the sample mean is a function of the sampling design. The sample mean $\bar{x}$ (arithmetic average) is the sum of the data points, $x_1, x_2, \ldots, x_n$, divided by the total number of data points (*n*):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{D-1}$$

where x_i denotes the value of the i^{th} point.

D-2.1.1. If distribution testing suggests that data are lognormally distributed, then the descriptive statistics are best calculated using the transformed data (for each value $y_i = Ln(x_i)$). Calculating the sample mean, $\bar{x}$, is possible, even for lognormally distributed data. Gilbert (1987) reports that $\bar{x}$ may be used when the population coefficient of variation is small (i.e., less than 1.2). Unfortunately, the sample mean is statistically biased for known lognormal conditions. It is highly sensitive to a few large data values, as is typical of lognormal data. There are alternatives for estimating the population mean that are not statistically biased, and these are preferred.

D-2.1.2. The preferred method for estimating the population mean of a lognormal population is calculated by:

$$\hat{\mu}_1 = e^{\bar{y}} \Psi_n(t) \tag{D-2}$$

where

$\bar{y}$ = sample mean of the log-transformed data

n = number of data points

s_y = sample standard deviation of the log-transformed data

$\Psi_n(t)$ (with $t = s_y^2/2$) = the following infinite series

$$\Psi_n(t) = 1 + \frac{(n-1)t}{n} + \frac{(n-1)^3 t^2}{2!\, n^2 (n+1)} + \frac{(n-1)^5 t^3}{3!\, n^3 (n+1)(n+3)} + \frac{(n-1)^7 t^4}{4!\, n^4 (n+1)(n+3)(n+5)} + \ldots$$

D-2.1.3. This is the minimum variance unbiased estimate of the population mean. Likewise, the unbiased estimator of the variance of the mean is:

$$s^2(\hat{\mu}_1) = \exp(2\bar{y})\left\{\left[\Psi_n(t)\right]^2 - \Psi_n\left[t'\right]\right\} \quad \text{(D-3)}$$

where

$$t' = \frac{s_y^2(n-2)}{n-1}$$

D-2.1.4. The infinite series may be evaluated on a computer or estimated from tables referenced in Gilbert (1987). This method produces the minimum unbiased variance estimator (statistically unbiased and smallest sampling error variance) of the mean for a lognormal population.

D-2.1.5. Performing this calculation obviously can be laborious. There is a simpler method for estimating the mean and variance of a lognormal population that arises in Gilbert and in EPA guidance documentation. This method uses the formulas:

$$\hat{\mu} = \exp\left(\bar{y} + \frac{s_y^2}{2}\right)$$

$$\hat{\sigma}^2 = \hat{\mu}^2\left[\exp\left(s_y^2\right) - 1\right] \quad \text{(D-4)}$$

D-2.1.6. However, the approach can produce poor (biased high) estimates of mean and variance for small data sets and is not recommended unless n is large (e.g., $n > 50$). Paragraph D-2.2 presents an example calculation for the mean of a lognormal population using the three methods.

D-2.1.7. For complex sampling designs, such as stratification, the sample mean is a weighted arithmetic average of the sample means of the L strata. Because a stratified sampling plan weights the number of samples unequally among areas, the weights for each area are incorporated into the calculation of the average. A weighted average is very similar to the arithmetic average, where an arithmetic average weights each sample result equally (with a weight of $1/n$). A weighted arithmetic average is calculated by:

$$\bar{x} = \sum_{i=1}^{L} w_i \bar{x}_i \quad \text{(D-5)}$$

where:

w_i = weight for the i^{th} stratum
$\bar{x}_i$ = sample mean of the i^{th} stratum
L = number of strata
$\sum_{i=1}^{L} w_i = 1$

D-2.1.8. For example, consider a stratified sampling plan that collects a total of n = 20 samples from a site with L = 2 sub-groups, where 8 samples, x_{1i} i = 1,…8, are collected in sub-group 1, and 12 samples, x_{2i} i = 1,…12, are collected in subgroup 2. If the average for the site is required and the two strata are assumed to be of equal area or volume, then the weights for the weighted average are ½ for the sample mean from subgroup 1 and ½ for the sample mean from subgroup 2 so that

$$\sum_{i=1}^{L} w_i = \frac{1}{2} + \frac{1}{2} = 1$$

and the overall mean is

$$\bar{x} = \sum_{i=1}^{L} w_i \bar{x}_i = \frac{1}{2}\bar{x}_1 + \frac{1}{2}\bar{x}_2 = \frac{1}{2}\frac{\sum_{i=1}^{8} x_{1i}}{8} + \frac{1}{2}\frac{\sum_{i=1}^{12} x_{2i}}{12}.$$

D-2.1.9. Careful examination will show that each observation in subgroup 1 is weighted by 1/16 in the overall mean and each observation in subgroup 2 is weighted by 1/24 in the overall mean.

D-2.1.10. The mean is the "center of gravity." The mean is very sensitive to extreme values because each measurement, x_i, is used to calculate the mean. Note that the *sample mean*, $\bar{x}$, is distinguished from the corresponding *population parameter*, the *population mean*, μ. The population mean could hypothetically be calculated using Equation D-1 if it were possible to exhaustively sample the entire population. The number of all possible data points from the population, N, would appear in the denominator of Equation D-1. Typically, the number of data points in the sample data set, $n << N$ and the sample mean, $\bar{x}$, is a "best" estimate of μ. As previously stated, this section of the document focuses on sample statistics that are ultimately used to estimate the corresponding parameters.

D-2.2. *Example of Lognormal Mean Calculations*. A group of arsenic measurements in soil were found to be lognormally distributed. The sample analytical results (in mg/kg) are:

SB1	SB2	SB3	SB4	SB5	SB6	SB7	SB8	SB9	SB10
12.461	13.451	13.056	11.502	10.835	30.06	17.72	17.11	12.02	13.73

D-2.2.1. *Method 1*. Using the simple (albeit biased) population average method, the sample mean of these data is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 15.19 \text{ mg/kg arsenic in soil.}$$

The sample variance $s^2 = 32.3$. Shapiro-Wilk testing (Appendix F) suggests that the lognormal distribution cannot be rejected. Also, the sample variance is high. These data would be better treated as lognormal.

D-2.2.2. *Method 2*. To calculate the minimum unbiased variance estimator of the mean, we first take the natural logarithm of the data set and calculate the following:

$$\bar{y} = 2.674\,, \quad s_y^2 = 0.09060\,, \quad t = \frac{s_y^2}{2} = 0.0453\,.$$

Using the minimum unbiased variance estimator, we see that the mean is 15.17 mg/kg. Method 1 above, which does not account for the lognormality, is biased high slightly.

D-2.2.3. *Method 3*. Others may choose to use the simpler Gilbert/EPA estimating method described above. This alternative also yields a sample mean of about 15.17 mg/kg. This result is low relative to the simple averaging method, but in this case is nearly identical to the minimum unbiased variance estimator. This is largely attributable to the low value of *t* in this example.

D-2.2.4. *Summary*. Ideally, with a computer, the method for minimum unbiased variance estimator of the mean for a lognormal population could be used. In cases of large *n*, it is suitable to use the third, and relatively simpler, method.

D-2.3. *Median*. The sample median ($\tilde{x}$) is the second most common measure of central tendency. When measurements are ranked from lowest to highest, the median is the middle of the data set. Half of the data are less than the sample median, and half of the data are greater than the sample median.

D-2.3.1. To compute the sample median, list the data from smallest to largest and label these points:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

So that $x_{(1)}$ is the smallest, $x_{(2)}$ is the second smallest, and so on, where $x_{(n)}$ is the largest.

D-2.3.2. The determination of the sample median depends upon whether the sample size *n* is odd or even:

$$\tilde{x} = \begin{cases} x_{[(n+1)/2]}, & n = 1,3,5.... \\ \dfrac{x_{(n/2)} + x_{[(n/2)+1]}}{2}, & n = 2,4,6... \end{cases}$$

D-2.3.3. The median is also referred to as the 50th percentile, the value greater than or equal to 50 percent of the measurements. Unlike the mean, the median is not influenced by extreme values. The median is also more robust than the mean for censored data (when non-detected results occur). When data are symmetrical, the mean and median of the data are very similar. If data are slightly skewed to higher values, the mean tends to be larger then the median because the mean is more influenced by these higher values than the median. Likewise, when data are skewed to lower values, the mean tends to be lower than the median.

D-2.4. *Mode*. The third method of measuring the center of the data is the mode. The mode is the value of the sample that occurs with the greatest frequency. To find the mode, count the number of times each value occurs. As this value may not always exist, or if it does, it may not be unique, mode is the least commonly used measure of central tendency; however, it is useful for qualitative data.

D-2.5. *Examples for Calculating the Measures of Central Tendency*. Consider estimating the measures of central tendency for the subsurface soil background chromium results (in mg/kg) as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84.

D-2.5.1. *Sample Mean*. The sample mean (in mg/kg) is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{4.60 + 5.29 + 4.26 + 5.28 + 4.53 + 5.74 + 5.86 + 3.84}{8} = 4.93 .$$

(Note that the mean is reported as three significant figures to reflect the minimum number of significant figures in the original data set.)

D-2.5.2. *Sample Median*. The data, from smallest to largest, are:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} = 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, 5.86 .$$

As there are eight points (*n* is even), the median (in mg/kg) is:

$$\tilde{x} = \frac{x_{(n/2)} + x_{[(n/2)+1]}}{2} = \frac{x_{(4)} + x_{(5)}}{2} = \frac{4.60 + 5.28}{2} = 4.94 \quad .$$

D-2.5.3. *Sample Mode*. In this example, mode does not exist since no value is repeated multiple times.

D-3. Measures of Relative Standing. Sometimes the analyst is interested in knowing the relative position of one of several observations in relation to all of the observations. Percentiles or quantiles are one such measure of relative standing that may also be useful for summarizing data.

- The **percentile** is the data value that is greater than or equal to a given percentage of the data values.

- The **quantile** is an alternative name for percentile when speaking in fractions (proportions) rather than in percents.

D-3.1. Just as the mean is a measure of location at the center of data, percentiles and quantiles are measures of location at various positions of the data. For a continuous variable *X*, the $p100^{th}$ percentile or *p* quantile, x_p, is the data point that is greater than or equal to 100*p*% of the data points and is less than or equal to (1 – *p*)100% of the data points. For example, if *x* is the 95% percentile (0.95 quantile), then it has the property that 95% (a proportion 0.95) of the observations lie at or below x_p and 5% (a proportion 0.05) of the data points lie at or above x_p.

D-3.2. The percentile and quantile for a discrete variable (i.e., a variable that may assume only a finite number of values) is defined somewhat differently than for a continuous variable. For a discrete variable *X*, X_p is the *p* quantile of *X* if

$$P(X < X_p) \leq p$$

and

$$P(X > X_p) \leq 1 - p$$

or equivalently,

$$P(X \leq X_p) \geq p.$$

D-3.3. To calculate percentiles or quantiles for a set of *n* sample points (x_1, x_2, ..., x_n), first list the data points from smallest to largest (x_1, x_2, ..., x_n). Multiply the sample size, *n*, by *p*. Di-

vide the result into the integer part and the fractional part, i.e., let $np = j + g$ where j is the integer part and g is the fraction part. The $p100^{th}$ percentile, x_p, is calculated by:

$$x_p = \begin{cases} \dfrac{x_{(j)} + x_{(j+1)}}{2}, & g = 0 \\ x_{(j+1)}, & g \neq 0 \end{cases}$$

D-3.4. One example of a percentile is the median. The median is the 50th percentile because half the results fall below this value and half of the results fall above this value. A sample percentile may fall between a pair of observations. For example, the 75th percentile of a data set of 10 observations is not uniquely defined.

D-3.5. Important percentiles usually reviewed are the quartiles of the data. The most common quartiles are 25th, 50th, and 75th percentiles. The 25th and 75th percentiles can be used to estimate the dispersion of a data set (see Paragraph D-4). Quartiles are discussed further in Paragraph D-4 to explain the dispersion of the data.

D-3.6. Also important for environmental data are the 90th, 95th, and 99th percentiles, where a decision-maker would like to be sure that 90, 95, or 99% of the contamination levels are below a fixed risk level. Directions and examples for calculating the measures of relative standing are presented below in Paragraph D-4.

D-3.7. Estimating quantiles in lognormal populations arises frequently in environmental applications. Of course, a probability plot may be used to estimate the quantiles, after the data are transformed and plotted. Alternatively, a mathematical method is recommended in Gilbert (1987). Simply,

$$\hat{x}_p = \exp\left(\bar{y} + Z_p s_y\right) \tag{D-6}$$

where Z_p is the value of the cumulative normal distribution for the p^{th} quantile. For the data in the preceding example (Paragraph D-2.2), the 99th quantile of the data is

$$\hat{x}_{0.95} = \exp\left(2.67 + 2.326 \times 0.301\right) = 29.1 \text{ mg/kg}.$$

D-4. Calculating the Measures of Relative Standing (Percentiles). The 95th, 75th, and 25th percentiles will be computed for the eight subsurface soil background chromium results (in mg/kg), ordered from lowest to highest, as follows: 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, and 5.86.

D-4.1. For the 95th percentile, $p = 0.95$ and

$np = (8)(0.95) = 7.6$

Therefore:

$$np = j + g$$

$$7.6 = 7 + 0.6$$

So: $j = 7$ and $g = 0.6$.

D-4.2 Since $g \neq 0$, $x_{(p)} = x_{(j+1)}$. The 95th percentile of this data set is:

$$x_{0.95} = x_{(7+1)} = x_{(8)} = 5.86 \text{ mg/kg}$$

Note that 100% of the data points (8 out of 8 values) rather than 95% of the measurements are less than or equal to the 95th percentile. The 95th percentile is being calculated for the set of eight measured chromium values and not for the set of all possible values of chromium. The set of measured chromium concentrations is a discrete variable (there are only eight possible values for chromium). If a larger number of measurements were made, nearly (or precisely) 95% of the measurements would be less than or equal to the 95th percentile.

D-4.3. For the 75th percentile, $p = 0.75$ and

$np = (8)(0.75) = 6$.

Therefore:

$$np = j + g$$

$$6 = 6 + 0.0$$

So: $j = 6$ and $g = 0$.

D-4.4 The 75th percentile of these data is:

$$x_{0.75} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{5.29 + 5.74}{2} = 5.52 \text{ mg/kg.}$$

Note that 6 out of the 8 measured values (0.75 of the total number of observations) are less than or equal to the 75th percentile 5.52 mg/kg.

D-4.5. For the 25th percentile, $p = 0.25$ and

$$np = (8)(0.25) = 2$$

Therefore:

$$np = j + g$$

$$2 = 2 + 0.0$$

So: $j = 2$ and $g = 0$.

D-4.6. The 25th percentile of these data is:

$$x_{0.25} = \frac{x_{(2)} + x_{(3)}}{2} = \frac{4.26 + 4.53}{2} = 4.40 \text{ mg/kg.}$$

D-5. Measures of Dispersion.

D-5.1. *Introduction*. Measures of central tendency are more meaningful if accompanied by information on how the data spread out from the center. Measures of dispersion or variability in a data set include the sample range, variance, standard deviation, coefficient of variation, and the interquartile range. Directions for calculating these measures of dispersion follow, and examples are presented in Paragraph D-6.

D-5.1.1. *Range*. This is the difference between the largest and smallest result from the data set.

D-5.1.2. *Variance*. This is a measurement of the dispersion or deviation of results from the mean of a data set.

D-5.1.3. *Standard Deviation*. This is the square root of the sample variance, it has the same unit of measure as the original data.

D-5.1.4. *Coefficient of Variation (CV).* This is sometimes called the relative standard deviation (RSD), a unitless measure equal to the standard deviation divided by the mean.

D-5.1.5. *Interquartile Range*. This is the difference between the 75th and 25th percentiles, it measures the central 50% of the results in the data set.

D-5.2. *Sample Range*. The simplest measure of dispersion to compute is the sample range. The sample range (R) is the difference between the largest value and the smallest value of the sample:

$$R = x_{(n)} - x_{(1)} \tag{D-7}$$

where:

$x_{(n)}$ = largest ordered value
$x_{(1)}$ = smallest ordered value

For small samples, the range is easy to interpret and may adequately represent the dispersion of the data. For large samples, the range is not very informative because it only considers (and is greatly influenced by) extreme values.

D-5.3. *Sample Variance*. The sample variance measures the dispersion or deviation of results from the mean of a data set.

D-5.3.1. To find the sample variance (s^2), compute:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{D-8}$$

D-5.3.2. If the variance is being manually calculated, a simpler version of this calculation is the following:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} \tag{D-9}$$

D-5.3.3. However, this version should not be used when calculating the variance with a computer because too much rounding error is introduced into this calculation.

D-5.3.4. A large sample variance implies that there is a large spread among the data, that the data are not clustered tightly around the mean. A small sample variance implies that there is little spread among the data, and that most of the data are near the mean. Like the mean, the sample variance is affected by extreme values and by a large number of non-detected results. Note that the sample variance s^2 is distinguished from the corresponding population parameter, the population variance, σ^2.

D-5.4. *Sample Standard Deviation*. The sample standard deviation has the same unit of measure as the original data. The sample standard deviation (s) is the square root of the sample variance:

$$s = \sqrt{s^2} \tag{D-10}$$

Frequently, the sample standard deviation will not be an appropriate measure of dispersion unless the data are normally distributed.

D-5.5. *Sample Coefficient of Variation*. The CV or RSD is a unitless measure that allows the comparison of dispersion across several sets of data because it is scaled to the mean. The sample CV is the sample standard deviation divided by the sample mean:

$$\text{CV} = \frac{s}{\bar{x}} \tag{D-11}$$

The CV is often expressed as a percentage:

$$\%\text{RSD} = \frac{s}{\bar{x}} 100\% .$$

The CV is often used in environmental applications because variability (expressed as a standard deviation) is often proportional to the mean.

D-5.6. *Sample Interquartile Range (IQR).* When extreme values are present, the interquartile range may be more representative of dispersion in the data than the standard deviation. This range is not heavily influenced by extreme values because it measures the spread within the center portion of a data set, rather than include the most extreme values as does the range. As a result, it is useful when the data include a large number of non-detects. Use the directions in Paragraph D-6 to compute the 25th and 75th percentiles of the data ($x_{0.25}$ and $x_{0.75}$ respectively). Then,

$$\text{IQR} = x_{0.75} - x_{0.25} \tag{D-12}$$

D-6. Examples for Calculating the Measures of Dispersion. Consider estimating the measures of dispersion for subsurface soil chromium results (in mg/kg) as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84. The data are ordered as follows:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} = 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, 5.86 .$$

D-6.1. *Sample Range (R).* The sample range is simply:

$$R = x_{(n)} - x_{(1)} = 5.86 - 3.84 = 2.02$$

D-6.2. *Sample Variance (s^2).* Before the variance can be computed, the mean must be computed. The mean was computed in Paragraph D-2.2 and is 4.93 mg/kg. Both methods of calculating the variance are illustrated below:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$= \frac{(4.60 - 4.925)^2 + (5.29 - 4.925)^2 + (4.26 - 4.925)^2 + (5.28 - 4.925)^2}{8-1} +$$
$$\frac{(4.53 - 4.925)^2 + (5.74 - 4.925)^2 + (5.86 - 4.925)^2 + (3.84 - 4.925)^2}{8-1}$$

$$= 0.5255$$

or

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

$$= \frac{\left(4.60^2 + 5.29^2 + 4.26^2 + 5.28^2 + 4.53^2 + 5.74^2 + 5.86^2 + 3.84^2\right)}{8-1} - \frac{\left(8 \times 4.925^2\right)}{8-1}$$

$$= 0.5255$$

D-6.3. *Sample Standard Deviation (*s*).*

$$s = \sqrt{s^2} = \sqrt{0.5255} = 0.7249$$

D-6.4. *Sample Coefficient of Variation (CV).*

$$\text{CV} = \frac{s}{\bar{x}} = \frac{0.7249}{4.925} = 0.1472$$

D-6.5. *Sample Interquartile Range (IQR).* The 25^{th} and 75^{th} percentiles of the data, $x_{0.25}$ and $x_{0.75}$ respectively, were computed in Paragraph D-4. So:

$$\text{IQR} = x_{0.75} - x_{0.25} = 5.515 - 4.395 = 1.12$$

Note that the single data set presented above results in a number of different numerical values that all summarize dispersion:

Range	IQR	s	s^2	CV
2.0 mg/kg	1.1 mg/kg	0.72 mg/kg	0.52 mg^2/kg^2	0.15