Appendix A

Glossary

asymptotic normality—Refers to the condition in which the sampling distribution of a parameter estimate approaches that of a normal distribution as the sample size becomes "large." Depending on the estimator, large usually means 30 to 60 observations. When these conditions hold, the estimate is said to be asymptotically normal, and the normal approximation can be used to establish confidence limits for the parameters. One of the many desirable attributes of maximum likelihood estimators is that they are asymptotically normal under fairly simple but broadly applicable conditions.

Bayesian inference—A method that regards model parameters as random variables with prior probability distributions reflecting prior knowledge about the parameters. Bayesian inference is based, via Bayes Theorem, on the conditional (posterior) distribution of the parameters, given the data.

bootstrap estimation—A technique for estimating the variance and/or the bias of a sample estimate of a population parameter by repeatedly drawing (with replacement) a large number (e.g., 1,000) of new, "bootstrap" samples from the original sample. The sample size of each bootstrap sample is the same as the original sample. The variance and bias estimators are computed from the distribution of the bootstrap samples. This technique is most useful for cases where there is no known closed-form estimator for the population variance or in other situations where the usual estimators are not appropriate (e.g., for small sample sizes).

coefficient of variation (CV)—A dimensionless measure of dispersion, equal to the standard deviation divided by the mean, often expressed as a percentage.

complex sampling design—A sampling design in which individual population elements do not have equal probabilities of selection. Complex sample surveys generally incorporate stratification and/or clustering wherein the population members may be correlated. As a consequence, the iid assumption (see p. A-3) may not hold for the sampled population members.

confidence interval—The interval or region about a sample estimate within which the desired population parameter is expected to occur with some specified probability (i.e., the true value of the population parameter will lie within the interval or range for 95% of all samples).

continuous random variable—A random variable that may take on an infinite number of values. The cumulative distribution function of a continuous random variable is therefore a smooth function.

correlation coefficient—A scale-invariant measure of the association between two variables that takes on values between –1 and +1. The correlation coefficient has a value of +1 whenever an increase in one is accompanied by an increase in the other, zero when there is no relationship (i.e., the two variables are independent of one another), and -1 when there is an exact inverse relationship between them.

covariance—A scale-dependent measure of the tendency of the values of one variable to change with those of a second variable. Algebraically, the covariance is the expected value of the product of the deviations of two random variables from their respective means. When this product is zero, the two variables are said to be uncorrelated; otherwise, they will be correlated.

covariates—Random variables (discrete and/or continuous) that are specified as predictor variables in a multivariable model.

cumulative distribution function (CDF)—F(x) equals the probability that a randomly chosen member of a population has a value less than or equal to x for the variable of interest. With reference to a random variable X, the CDF of X, F(x), is the probability that the random variable X does not exceed the number x. Symbolically, $F(x) = P[X \# x]$.

degrees of freedom (df)—As used in statistics, df has several interpretations. A sample of *n* variate values is said to have *n* degrees of freedom, but if *k* functions of the sample values are held constant, the number of degrees of freedom is reduced by *k*. In this case, the number of degrees of freedom is conceptually the number of independent observations in the sample, given that *k* functions are held constant. By extension, the distribution of a statistic based on *n* independent observations is said to have *n-p* degrees of freedom, where *p* is the number of parameters of the distribution.

discrete random variable—A random variable that may take on only a finite number of values. The CDF of a discrete random variable is therefore a step function.

empirical distribution function (EDF)—The sample estimate of the CDF. For any value of $X=x_i$, it is the proportion of observations that are less than or equal to x_i. The graph of the EDF is a step function for which the value at $X=x_i$ is n_i/n, where n_i is the number of sample observations with values of $X \# x_i$ and n is the total number of observations in the sample. The plot is a series of steps ascending, left to right, from 0 to 1.

goodness-of-fit (GOF) test—Any of several statistical tests of the null hypothesis that the population distribution of the observations is a specified probability distribution or is in a specified set of probability distributions (e.g., the lognormal distribution). The tests evaluate whether or not the EDF is significantly different from the specified CDF.

iid assumption (independent and identically distributed assumption)—Assumes that the values of a random variable in a sample are not correlated with each other and that they share a common probability density function (PDF). This assumption will hold for data collected by simple random sampling but (usually) not for data from more complex (i.e., stratified and/or clustered) sampling designs.

kernel density estimation—A technique for estimating the probability density of a distribution by fitting a smooth curve to the underlying frequency histogram. The choice of the degree to which the distribution should be smoothed is crucial and usually is based on criteria that minimize the mean square error. Unlike the parametric methods that depend on the parameters of a known theoretical PDF, the kernel estimate is derived entirely from the attributes of the sample EDF.

key study—Designation used in the Exposure Factors Handbook (U.S. EPA, 1997a) to distinguish the studies that were regarded as the most useful (representative) for deriving a recommendation for an exposure factor.

likelihood ratio test (LRT)—A parametric test of a null hypothesis that uses as its test statistic (-2) times the natural logarithm of the ratio of two maximized likelihoods. The numerator likelihood is maximized under the constraint (condition) of the null hypothesis. The denominator likelihood does not

have this constraint. The test statistic is usually assumed to have a chi-squared distribution with degrees of freedom equal to the differences in dimensionality of the two parameter spaces.

maximum likelihood estimator (MLE)—The parameter estimates that maximize the probability of obtaining the sample observations.

meta-analysis—The process of using statistical techniques to combine the results of several different studies. Meta-analyses may permit stronger and/or broader inferences than were possible in any of the constituent studies.

model parameter—Numerical characteristic of a given population (e.g., the mean and variance of a normal population) that determines some response of interest in accordance with a specific mathematical formula. Such an expression is called a model. By convention, statistical model parameters are usually symbolized as Greek letters.

Monte Carlo methods—Methods used to investigate the properties of an inferential procedure by applying it to computer-generated data that serve as a surrogate for "real data" collected by random sampling.

multivariate parametric distribution—The joint theoretical probability distributions of two or more random variables. The component univariate distributions can be of the same kind (e.g., three lognormal distributions), or they may be combinations of several different kinds (e.g., lognormal, Weibull, and exponential). Typically, the component variables are correlated; thus, correlations comprise additional parameters of multivariate distributions.

P-P (probability-probability) plot—A graph used to subjectively assess GOF. For any given $X=x_i$, the value of the CDF for the theoretical distribution of interest is plotted on one axis, and the observed value of the EDF for $X=x_i$ is plotted on the other axis. P-P plots that closely approximate a diagonal line through the origin indicate a good fit between the EDF and the theoretical CDF.

***p*-value—**A value between 0 and 1 that is often regarded as a measure of the belief in a statistical null hypothesis (H_0). In the Frequentist view (vs. the Bayesian view), a test statistic has a specific parametric distribution when H_0 is true. A test statistic is computed from sample data and compared with its

expected parametric distribution. The *p*-value is the probability that a value of the test statistic, as extreme or more extreme than the observed test statistic, came from the null distribution. If the *p*-value is less than or equal to the significance level (") of the test, the null hypothesis is rejected. A major objection of Bayesian statisticians to the Frequentist approach is that the specification of the " value (usually 0.05) is arbitrary.

percent error plots—A graphical GOF test. The difference between the observed and hypothesized quantile values, expressed as a percent of the hypothesized value, is plotted on the vertical axis versus the observed quantiles on the horizontal axis. Because the points on the plot are compared with a horizontal reference line (i.e., percent difference=0) and because of the relative nature of the differences being displayed, lack-of-fit is more apparent than in P-P or Q-Q plots.

point mass at zero—A positive probability that the observed value of the random variable is zero (e.g., the probability that the amount of tap water consumed by an infant per day is zero). If the distribution of a random variable, X, is a continuous parametric distribution, the probability of observing a value in the interval from X=a to X=b is equal to the area under the probability density function (the derivative of the CDF) between a and b. By definition, a single point (e.g., X=0) does not occupy any space and, hence, has probability zero of occurring exactly. Therefore, the distribution for some exposure factors may be a composite probability distribution that includes a positive probability of observing X=0 and a continuous parametric distribution (e.g., lognormal) for positive values of X.

probability density function (PDF)—The PDF of a continuous random variable X is the first derivative of its CDF, $f(x) = F^{N}(x)$. The probability that a # X # b is found by integrating f(x) from a to b.

Q-Q plot—A graph used to subjectively assess GOF. For any given probability value p_i, the value of the random variable, X, for which the theoretical CDF is p_i is plotted on one axis, and the observed value of $X=x_i$, for which the EDF is p_i is plotted on the other axis. Q-Q plots that closely approximate a diagonal line through the origin indicate a good fit between the EDF and the theoretical CDF. Depending on observed patterns of deviation from the diagonal, lack-of-fit due specifically to differences in location (i.e., mean or median) and/or scale (i.e., variance) can be diagnosed.

quantile—The q-1 partition values of a random variable that divide a sample or population into q subdivisions, each of which contain an equal proportion of the sample or population. For example, when

q=4, the three resulting values are the first, second (=median), and third quartiles that collectively divide the data into four equal parts.

random variable—A numeric event whose values change from one sampling unit or one experimental unit to the next. Random variable values may be either discrete or continuous.

relevant study—Designation in the Exposure Factors Handbook (U.S. EPA, 1997a) to distinguish the studies that were applicable or pertinent, but not necessarily the most important for making a recommendation for an exposure factor.

representative sample—A sample that captures the essence of the population from which it was drawn; one which is typical with respect to the characteristics of interest, regardless of the manner in which it was chosen. While representativeness in this sense cannot be completely assured, randomly selected samples are more likely to be representative than are haphazard or convenience samples. This is true because only in random sampling will every population element have an equal probability of selection.

residence time—The time in years between a person moving into a residence and the time the person moves out or dies.

risk assessment—Qualitative or quantitative estimation of the probability of adverse health or environmental effects due to exposure to specific behavioral, dietary, environmental, occupational, or social factors.

sensitivity analysis—The process of varying one or more model parameters while leaving the others constant to determine their effect on the model predictions. The results help to identify the variables that have the greatest effect on model estimates and may be useful for fine-tuning the model or identifying problems for additional research.

simple random sampling design—A sampling design in which every member of the target population has an equal probability (p=1/N) of selection to the sample. Random variables measured on observations from a simple random sample satisfy the iid assumption.

tap water—Water consumed directly from the tap as a beverage or used in preparation of foods and beverages (coffee, tea, frozen juices, soups, etc.).

uncertainty analysis—Identification of the components of variability of risk that are due to model uncertainty or parameter uncertainty, that is, to uncertainty in the type of model (e.g., gamma vs. lognormal vs. Weibull) or uncertainty in the values of model parameters. Parameter uncertainty can be built into risk assessment simulations by randomly drawing population parameters from appropriate distributions before selecting individuals from the population. Model uncertainty can be addressed by sensitivity analysis, using separate simulations for different viable competing models.

univariate parametric distribution—A theoretical probability distribution for a random variable whose CDF is described by a mathematical function of population parameters (e.g., the population mean and variance), such as a normal or lognormal distribution.