ATTACHMENT 1

ATTACHMENT 1: Glossary

Beta Distribution is a flexible, bounded PDF described by two shape parameters. It is commonly used when a range of the random variable is known. (p. A3-14)

Boxplot is a graphical representation showing the center and spread of a distribution, along with a display of outliers. (p. A3-10)

Central Limit Theorem says that for a relatively large sample size, the random variable x (the mean of the samples) is normally distributed, regardless of the population's distribution. (p. A3-14)

Coefficient of Variation (also Coefficient of Variance or Coefficient of Variability)* is an estimate of relative standard deviation. Equals the standard deviation divided by the mean. Results can be represented in percentages for comparison purposes. (p. A3-7)

Confidence Interval is the range within which one has a given level of confidence that the range includes the true value of the unknown parameter (e.g. a 95% confidence interval for a parameter means that 95% of the time the true value of that parameter will be within the interval).

Continuous Probability Distribution* is a probability distribution that describes a set of uninterrupted values over a range. In contrast to the Discrete distribution, the Continuous distribution assumes there are an infinite number of possible values.

Correlation, Correlation Analysis is an investigation of the measure of statistical association among random variables based on samples. Widely used measures include the linear correlation coefficient (also called the product-moment correlation coefficient or Pearson correlation coefficient), and such non-parametric measures as Spearman rank-order correlation coefficient, and Kendall's tau. When the data are nonlinear, non-parametric correlation is generally considered to be more robust than linear correlation.

Correlation Coefficient* is a number between -1 and 1 that specifies mathematically the degree of positive or negative correlation between assumption cells. A correlation of 1 indicates a perfect positive correlation, minus 1 indicates a perfect negative correlation, and 0 indicates there is no correlation.

Cumulative Distribution Function (CDF) is alternatively referred to in the literature as the distribution function, cumulative frequency function, or the cumulative probability function. The cumulative distribution function, F(x), expresses the probability the random variable X assumes a value less than or equal to some value x, $F(x)=\text{Prob}(X \leq x)$. For continuous random variables, the cumulative distribution function is obtained from the probability density function by integration. In the case of discrete random variables, it is obtained by summation.

Cumulative Frequency Distribution is a chart that shows the number or proportion (or percentage) of values less than or equal to a given amount.

Deterministic Model, as opposed to a stochastic model, is one which contains no random elements.

Discrete Probability Distribution* is a probability distribution that describes distinct values, usually integers, with no intermediate values. In contrast, the continuous distribution assumes there are an infinite number of possible values.

Distribution is the pattern of variation of a random variable.

Frequency (also Frequency Count)* is the number of times a value recurs in a group interval.

Frequency Distribution* is a chart that graphically summarizes a list of values by subdividing them into groups and displaying their frequency counts.

Goodness-of-Fit is a set of mathematical tests performed to find the best fit between a standard probability distribution and a data set.

Goodness-of-Fit Test is a formal way to verify that the chosen distribution is consistent with the sample data.

Group Interval is a subrange of a distribution that allows similar values to be grouped together and given a frequency count.

Histogram is a plot of the range of values of a variable into intervals and displays only the count of the observations that fall into each interval. (p. A3-9)

Interquartile Range is the difference between the third quartile (75th percentile) and the first quartile (25th percentile). (p. A3-10)

Kurtosis* is the measure of the degree of peakedness and flatness of a curve. The higher the kurtosis, the closer the points of the curve lie to the mode of the curve. A normal distribution curve has a kurtosis of 3. (p. A3-7)

Lognormal Distribution is the distribution of a variable whose logarithm is normally distributed. (p. A3-15)

Mean is the arithmetic average of a set of numerical observations: the sum of the observations divided by the number of observations (p. A3-7).

Measurement Error is error introduced through imperfections in measurement techniques or equipment.

Median is the value midway (in terms of order) between the smallest possible value and the largest possible value. It is that value above which and below which half the population lies (p. A3-7).

Mode* is that value which, if it exists, occurs most often in a data set. (p. A3-7)

Monte Carlo Analysis (Monte Carlo Simulation) is a computer-based method of analysis developed in the 1940's that uses statistical sampling techniques in obtaining a probabilistic approximation to the solution of a mathematical equation or model. It is a method of calculating the probability of an event using values, randomly selected from sets of data repeating the process many times, and deriving the probability from the distributions of the aggregated data.

Non-parametric Approach is one that does not depend for its validity upon the data being drawn from a specific distribution, such as the normal or lognormal. A distribution-free technique.

Normal Distribution is a probability distribution for a set of variable data represented by a bell shaped curve symmetrical about the mean. (p. A3-14)

Parameter. Two distinct, but often confusing, definitions for parameter are used. In the first usage (preferred), parameters refers to the constants characterizing the probability density function or cumulative distribution function of a random variable. For example, if the random variable W is known to be normally distributed with mean μ and standard deviation σ, the characterizing constants μ and σ are called parameters. In the second usage, parameters are defined as the constants and independent variables which define a mathematical equation or model. For example, in the equation $Z=\alpha X+\beta Y$, the independent variables (X,Y) and the constants (α,β) are all parameters.

Parametric Approach is a method of probabilistic analysis in which defined analytic probability distributions are used to represent the random variables, and mathematical techniques (e.g., calculus) are used to get the resultant distribution for a function of these random variables.

Percentile is the value that exceeds X percent of the observations.

Population is the total collection of observations that is of interest.

Probability (Classical Theory) is the likelihood of an event.

Probabilistic Approach is an approach which uses a group of possible values for each variable to estimate risk.

Probabilistic Density Function (PDF)

Probabilistic Model is a system whose output is a distribution of possible values.

Probability Density Function (PDF) is a distribution of values for a random variable, each value having a specific probability of occurrence. It is alternatively referred to in the literature as the probability function or the frequency function. For continuous random variables, that is, the random variables which can assume any value within some defined range (either finite or infinite), the probability density function expresses the probability that the random variable falls within some very small interval. For discrete random variables, that is, random variables which can only assume certain isolated or fixed values, the term probability mass function (PMF) is preferred over the term probability density function. PMF expresses the probability that the random variable takes on a specific value.

Quantile-Quantile (Q-Q) Plot portrays the quantiles (percentiles divided by 100) of the sample data against the quantiles of another data set or theoretical distribution (e.g., normal distribution). By comparing the data to a theoretical distribution with a straight line, departures from the distribution are more easily perceived. (p. A3-24)

Random Error is error caused by making inferences from a limited database.

Random Number Generator* is a method implemented in a computer program that is capable of producing a series of independent, random numbers.

Random Variable is a quantity which can take on any number of values but whose exact value cannot be known before a direct observation is made. For example, the outcome of the toss of a pair of dice is a random variable, as is the height or weight of a person selected at random from the New York City phone book.

Range* is the difference between the largest and smallest values in a data set.

Regression Analysis (Simple) is the derivation of an equation which can be used to estimate the unknown value of one variable on the basis of the known value of the other variable.

Sampling. One of two sampling schemes are generally employed: simple random sampling or Latin Hypercube sampling. Latin hypercube sampling may be viewed as a stratified sampling scheme designed to ensure that the upper or lower ends of the distributions used in the analysis are well represented. Latin hypercube sampling is considered to be more efficient than simple random sampling, that is, it requires fewer simulations to produce the same level of precision. Latin hypercube sampling is generally recommended over simple random sampling when the model is complex or when time and resource constraints are an issue.

Sensitivity Analysis is an analysis that attempts to provide a ranking of the model's input parameters with respect to their contribution to model output variability or uncertainty. In broader sense, sensitivity can refer to how conclusions may change if models, data, or assessment assumptions are changed.

The difficulty of a sensitivity analysis increases when the underlying model is nonlinear, nonmonotonic or when the input parameters range over several orders of magnitude.

Simple Random Sampling (SRS) is a sampling procedure by which each possible member of the population is equally likely to be the one selected.

Simulation, in the context of Monte Carlo analysis, is the process of approximating the output of a model through repetitive random application of a model.

Skewness is the measure of the degree of deviation of a curve from the norm of a symmetric distribution. The greater the degree of skewness, the more points of a curve lie to one side of the peak of the curve. a normal distribution curve having no skewness is symmetrical, that is to say that there exists a central value a such that f(x-a)=f(a-x), f(x) being the frequency function. (p. A3-7)

Standard Deviation is a measure of dispersion which is expressed in the same units as the measurements. It is a measurement of the variability of a distribution, i.e., the dispersion of values around the mean. Standard deviation is the square root of the variance for a distribution (p. A3-7).

Standard Error of the Mean is the standard deviation of the distribution of possible sample means. This statistic gives one indication of how precise the simulation is.

Stochastic is a term referring to a process involving a random variable.

Triangular Distribution is a distribution with a triangular shape. It is characterized by its minimum, maximum and mode (most likely) values. It is often used to represent a truncated log-normal or normal distribution if there is little information available on the parameter being modeled. (p. A3-14)

Variability refers to observed differences attributable to true heterogeneity or diversity in a population or exposure parameter which cannot be reduced by additional data collection.

Sources of variability are the result of natural random processes and stem from environmental, lifestyle , and genetic differences among humans. Examples include human physiological variation (e.g., natural variation in bodyweight, height, breathing rates, drinking water intake rates), weather variability, variation in soil types and differences in contaminant concentrations in the environment. Variability is usually not reducible by further measurement or study (but can be better characterized).

Variance is a measure of the dispersion, or spread, of a set of values about a mean. Variance is the square of the standard deviation, i.e., the average of the squares of the deviations of a number of observations from their mean value. When values are close to the mean, the variance is small. When values are widely scattered about the mean, the variance is larger.

Bibliography

(1997) Air Force Technical Report on Methods to Quantify Uncertainty in Human Health Risk Assessment - DRAFT, Armstrong Laboratory Occupational and Environmental Health Directorate, Brooks Air Force Base, Texas.

Decisioneering, Inc. (1996) Crystal Ball Version 4.0 User Manual, pages 269-275.

Marriott, F.H.C. (1990), a Dictionary of Statistical Terms -Fifth Edition, Longman Scientific and Technical copublished with John Wiley & Sons: 605 Third Avenue, New York, New York 10158, page 9.

U.S. EPA,(October 4,1996) Guiding Principles for Monte Carlo Analysis -DRAFT, U.S.EPA, 401 M Street SW, Washington, DC 20460.