

APPENDIX 11

The Use and Interpretation of Frequency Distributions, Cumulative Probability Functions, and Normal Probability Charts Presented in Chapter 4.5 (Estuarine Probabilistic Monitoring Chapter)

When evaluating the results of environmental samples it is often informative to arrange the observed values in various different ways and to evaluate their conformity with known theoretical frequency distributions. There are a number of characteristic shapes of frequency distributions, but the most commonly observed is the bell-shaped normal frequency distribution illustrated in the first line (A) of figures below. In the paired series of examples, the figures in the left hand column (L) are based on the results of an observed sample frequency distribution (estuarine near-bottom dissolved oxygen values from the current report), and those on the right (R) are based on the theoretical normal frequency distribution with the same defining characteristics. The theoretical normal distribution is completely defined by two characteristics - called parameters of the theoretical distribution: (1) the arithmetic average or mean, represented by the lower case Greek letter mu – “ μ ” and (2) the standard deviation, represented by the lower case Greek letter sigma – “ δ ”. When these characteristics are estimated based on a sample they are called sample statistics, and are represented by different symbols – the sample mean by “ \bar{x} ” (called x-bar), and the standard deviation by the letter “ S_x ” (standard deviation of the variable “x”). In the descriptive summary of sample statistics it is also important to identify the number of observations (N) upon which the sample¹ is based.

The visual evaluation of a sample frequency distribution is very useful, in that it reveals the location (or central tendency) and the dispersion (variability) of the observations along a gradient of potential values, and whether the distribution is relatively symmetrical (balanced) or asymmetrical, having more extreme values in one direction or another. Sample observations may be clustered in a single region (unimodal distribution) or in two (bimodal) or more (multimodal) regions of the gradient.

The left figure in the first line below (Figure A-L) is a histogram illustrating the sample distribution of 274 bottom dissolved oxygen (DO) values measured at probabilistic estuarine sites during the 2007-2012 assessment period (N = 274). The arithmetic average of the sample was approximately $\bar{x} = 6.12 \mu\text{g/L}$ DO and the sample standard deviation was $S_x = 1.40 \mu\text{g/L}$ DO. Both of these statistics are subject to measurement and sampling errors (see discussion below). Visual examination of the figure reveals that the sample distribution is unimodal, with the vast majority of the individual observations clustered near to the sample mean of 6.12 $\mu\text{g/L}$, and is relatively symmetrical (*i.e.*, the two tails are of approximately the same length). The theoretical normal distribution with the same mean and standard deviation is illustrated in Figure A-R, to the right, and is also overlaid on the sample frequency distribution of Figure A-L. A sample frequency distribution can never conform exactly to the corresponding normal distribution, because it consists of counts of observations grouped into a series of discrete classes, while the theoretical distribution is based on a continuous function. The fact that the sample distribution approximates the characteristic bell shape of the normal distribution suggests that the arithmetic mean and the standard deviation of the sample (plus the sample size N = 274) are appropriate statistics to describe the sample distribution, because they (parameters μ and δ) are sufficient to completely describe the corresponding (and very similar) normal distribution.

The two figures of line B illustrate the same two distributions in a different form. They are called cumulative distributions², because they order the observations from smallest to largest and represent the proportion of the total number of values (observations) accumulated up to each successively higher value. For example, in the cumulative normal distribution on the right (Figure B-R) the concentration of 5.0 $\mu\text{g/L}$ corresponds to a proportion of approximately 0.220 (22%), which indicates that approximately 22% of the DO values in the distribution are of 5.0 $\mu\text{g/L}$ or less. Similarly, the concentration of 6.12 $\mu\text{g/L}$ (the arithmetic mean) corresponds to a proportion of 0.500 or 50% of the distribution. Half of the values in the distribution are equal to or less than the mean. (This is only true for symmetrical distributions, like the normal distribution.)

¹ It is worth pointing out here that a common misunderstanding results from different uses of the word “sample.” A single “physical sample” may be directly measured in the field or be collected, transported, and analyzed in a laboratory. In statistics, a single result (value) from this direct measurement or analysis is referred to as an “observation.” In the arrangement of a frequency distribution, or in the comparison of two or more frequency distributions, the distribution of observations is referred to as a “statistical sample” of size “N”. In the example of a sample frequency distribution portrayed here, the distribution is described using three characteristics: (1) the sample arithmetic mean or average - \bar{x} , (2) the sample standard deviation - S_x , and (3) the sample size (number of observations - N - from which the two statistics are calculated). “N” is important because it is an indication of the potential error in the resulting estimates of the true mean and standard deviation (μ and δ) of the population as a whole and the corresponding theoretical frequency distribution. Large samples provide more accurate estimates of theoretical distribution parameters (μ and δ in this case) than do small samples.

² They are often called cumulative distribution functions (CDFs) or cumulative probability functions (CPFs) in the literature.

Comparison of Observed (or Sample) and Theoretical Normal Frequency Distributions

**Observed (Sample) Distribution
of Near-Bottom Dissolved Oxygen (DO in mg/L)**
($\bar{x} = 6.1233$, $S_x = 1.4003$, $N = 274$)

**Theoretical Normal Distribution
of Near-Bottom Dissolved Oxygen (DO in mg/L)**
($\mu = 6.1233$, $\delta = 1.4003$)

Column L

Column R

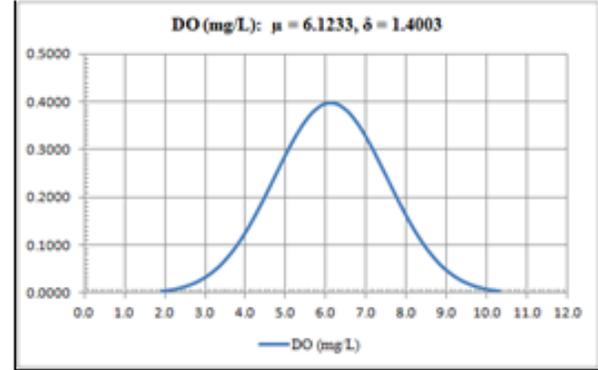
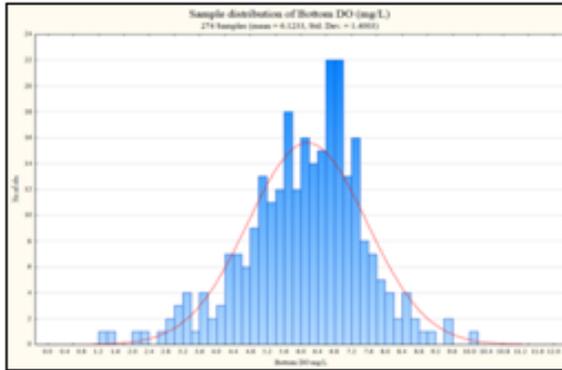


Figure A.

**L. Histogram of Observed Results
(Sample Frequency Distribution)**

**R. Theoretical Normal Frequency Distribution
based on the mean and standard deviation of the sample**

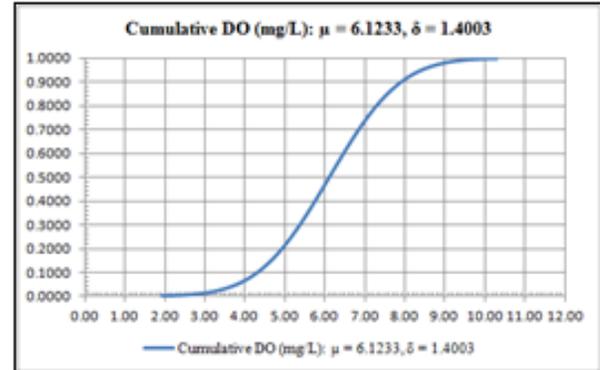
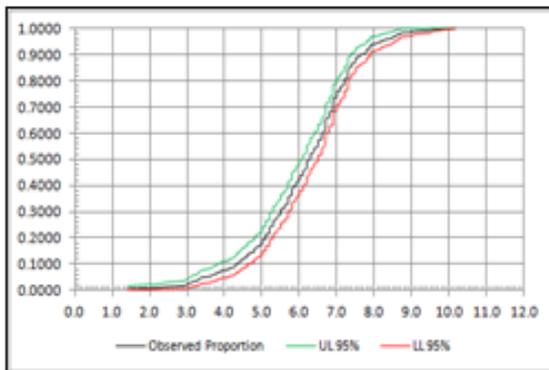


Figure B.

**L. Cumulative Sample Distribution with 95%
Confidence Interval**
($\bar{x} = 6.1233$, $S_x = 1.4003$, $N = 274$)

R. Cumulative Normal Distribution

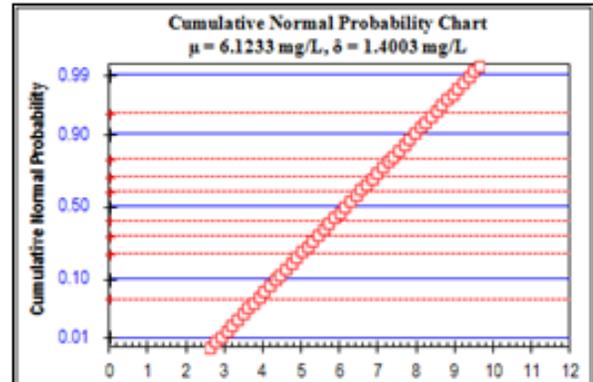
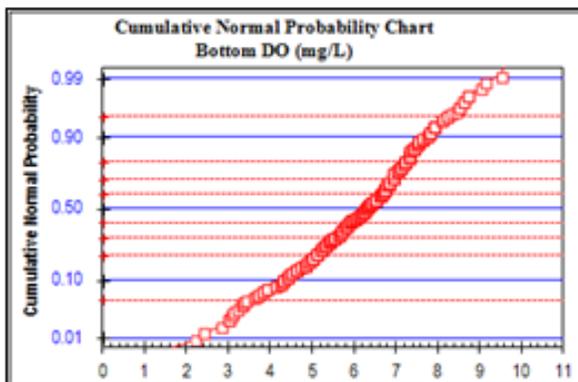


Figure C.

**L. Cumulative Normal Probability Chart of
The Sample Distribution**

**R. Cumulative Normal Probability Chart of
the Normal Distribution**

Note that there are three lines in the figure of the cumulative sample distribution (Figure B-L). The center (black) curve corresponds to the observed cumulative sample distribution but, because it is based on a sample, it is only an estimate of the true cumulative distribution. The estimates calculated for it, as well as those of the sample mean and standard deviation, are subject to sampling errors, which can be reduced but not eliminated by collecting a larger number of physical samples (observations), along with measurement errors, which can be minimized but not eliminated by better instrumentation and frequent calibration. The upper (green) and lower (red) curves in Figure B-L represent the estimated upper and lower 95% confidence limits, respectively, of the cumulative distribution; we have a confidence of 95% that the true value of the cumulative distribution occurs in the interval between these two limits. We don't need these limits for the theoretical cumulative distribution because when the two parameters (μ and δ) are known, they determine the exact distribution (no error). In the cumulative sample distribution, the value of 5.0 $\mu\text{g/L}$ (black line) corresponds to approximately 0.180, indicating that approximately 18% of the observed values are equal to or less than this value. The calculated 95% confidence interval at this concentration extends from 0.1297 (13%) to 0.2407 (24%), and includes the 22% estimate from the corresponding theoretical normal distribution. From the cumulative sample distribution, the calculated 95% confidence interval for the mean of 6.12 $\mu\text{g/L}$ extends from 0.3822 (38%) to 0.5231 (52%), which also includes the 50% estimate from the theoretical normal distribution.

The third pair of figures (Line C) still represents cumulative frequency distributions, observed sample distribution on the left and theoretical distribution on the right, but the Y-axis (ordinate) scale has been modified in such a way that the theoretical cumulative normal distribution becomes a straight line, rather than an s-shaped curve. Both the lower and the upper ends of the scale have been expanded to straighten out the two tails of the distribution. Note that the scale is truncated at the proportions 0.01 on the lower end and 0.99 on the upper end. This is because the tails of the theoretical normal distribution are extremely long (essentially infinite in each direction) and never reach true zero (0.000) or unity (1.000). The extremities of the scale could be extended to 0.001 and 0.999 but, in reality, we would only see such extreme values in very large samples. They are very rare and, because of the rarity of values at the extremes of the distribution, these are the regions where the sample distribution is most likely to deviate from the expected theoretical distribution because of sampling error (even with very large samples). Some examples of such deviations will be presented below. For the present, it is sufficient to note that the observed distribution of near-bottom dissolved oxygen values (Figure C-L) conforms fairly well to the straight line of the corresponding normal distribution in Figure C-R. It is slightly bowed (concave upward) in the middle, where the frequencies of observations between 6.6 and 7.2 $\mu\text{g/L}$ in the sample distribution exceed the expected normal distribution (Figure A-L), but another sample of equal size might differ in another fashion.

Approximations to the normal frequency distribution, as illustrated and discussed above, are expected and often observed when an environmental characteristic is in equilibrium with its surroundings. Random ambient variations in other related characteristics (such as water temperature, wind speed, and surface mixing in the case of dissolved oxygen) may induce similar random (and symmetrical) variations in the characteristic of interest. Such variations give rise to an approximate normal distribution... "When the outcome is produced by many small effects acting additively and independently, its distribution will be close to normal. The normal approximation will not be valid if the effects interact multiplicatively (instead of additively), or if there is a single external influence that has a considerably larger magnitude than the rest of the effects."³

The following examples illustrate approximate normal distributions observed in environmental samples from Virginia's near-shore oceanic and estuarine waters, along with a few examples where external influences have noticeably altered the shapes of the distributions.

Figure D illustrates the cumulative sample frequency distribution of dissolved inorganic phosphorus (DIP or orthophosphate) from 50 sites (55 physical samples) in Virginia's near-surface oceanic waters, along with the corresponding cumulative normal probability distribution. DIP in these waters is in equilibrium with oceanic sources and receives relatively little influence from anthropomorphic (human generated) sources.

³ Taken from a discussion of the univariate normal distribution in Wikipedia, the free online encyclopedia.

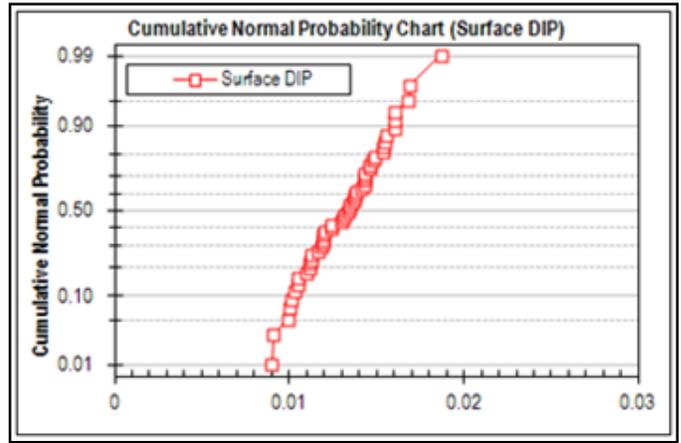
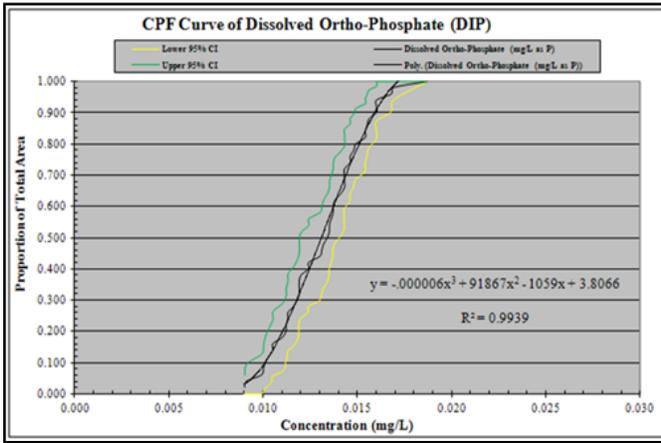


Figure D - Offshore Near-Surface Dissolved Inorganic Phosphorus (DIP - mg/L) from the 2010 Oceanic Survey. The sample distribution (N = 50) exhibits an approximately normally distributed variation with a minimum of human disturbance (relatively straight line in the Normal Probability Chart). (Taken from Chapter 4.8 - 2010 NEAR-SHORE OCEANIC SURVEY, of DEQ's 2012 Water Quality Assessment 305(b) / 303(d) Integrated Report – DEQ-WQA 2012.)

Figure E summarizes the cumulative sample frequency distribution of dissolved inorganic phosphorus from 273 probabilistic estuarine sites sampled between June 2007 and October 2012. The frequency distribution is highly skewed to the right (assymetric), having many measurements with elevated concentrations, extending the right tail of the distribution well beyond what would be expected under undisturbed conditions. Anthropomorphic additions of phosphorus, primarily from the use of chemical fertilizers and/or manure on agricultural fields, golf courses, and residential lawns, have skewed the distribution toward higher concentrations of DIP.

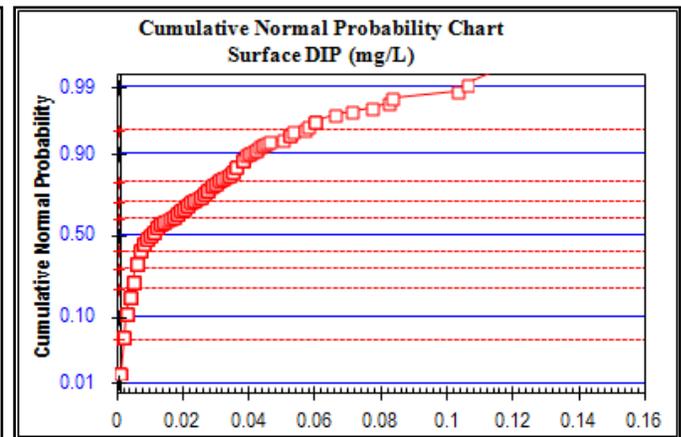
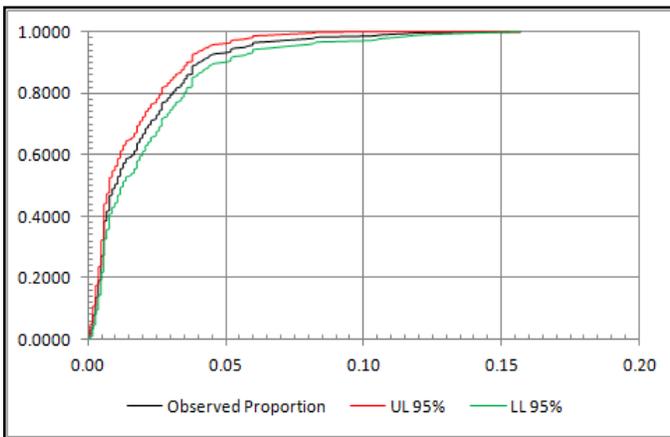
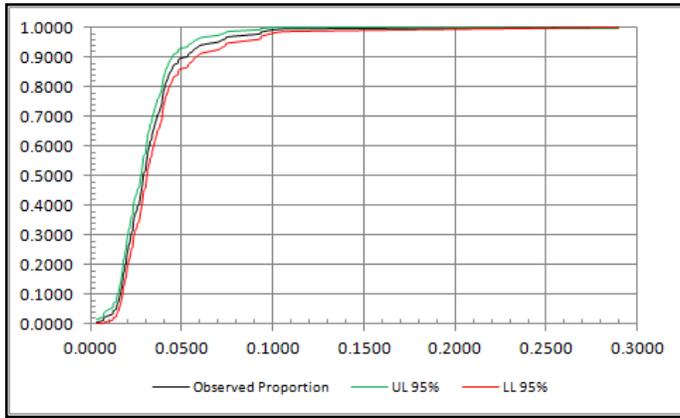


Figure E - Estuarine Near-Surface DIP from 2007-2012 Estuarine Surveys. Initially normally distributed at lower concentrations (relatively straight line in extreme lower left portion of the Normal Probability Chart), but showing strong influence of anthropomorphic addition of phosphorus as deviation from the normal distribution (curvilinear plot in upper right portion of the Normal Probability Chart).

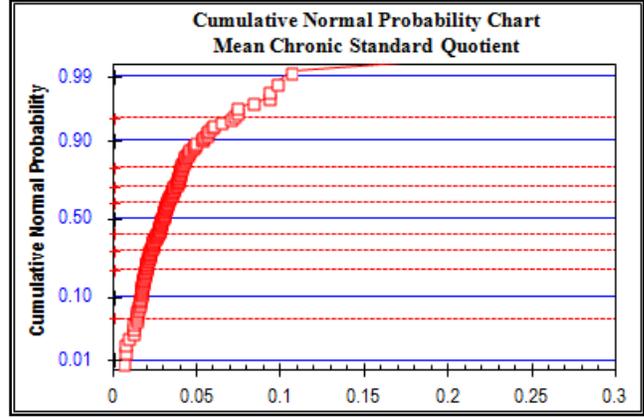
Figure F summarizes the cumulative sample frequency distribution of dissolved metals concentrations, expressed as the arithmetic average (mean) quotient of observed dissolved metals concentrations divided by their respective saltwater chronic standards for eight dissolved metals (As, Cd, Cu, Pb, Hg, Ni, Se, and Zn).

Chronic Standard quotient (CSq) = Observed dissolved concentration / chronic saltwater standard
 e.g., for Copper: $CSq_{Cu} = \mu\text{g/L dissolved Cu observed} / \text{Chronic saltwater standard for Cu (6.0 } \mu\text{g/L)}$

Mean quotients below a value of approximately 0.04 (~80% of the sites) appear to be approximately normally distributed – relatively straight line in the lower left portion of the cumulative normal probability chart – reflecting typical variability among undisturbed sites. The 20% of sites with mean quotients above 0.04 are skewed toward



Mean CSq – Saltwater chronic standard quotient



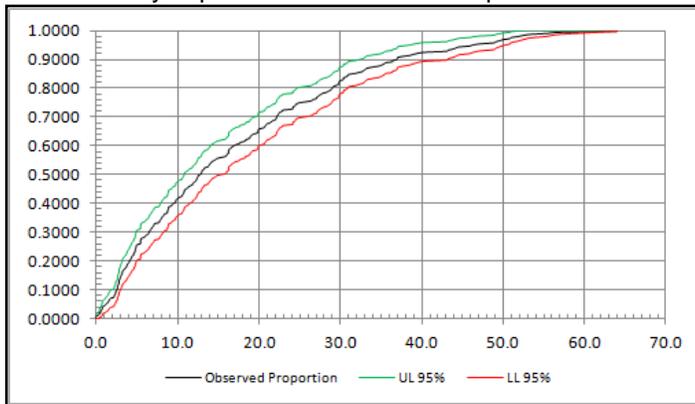
Mean CSq – Saltwater chronic standard quotient

Figure F – Estuarine Near-Surface Dissolved Metals Concentrations from 2008 – 2011. Expressed as the arithmetic average (mean) of the quotients of observed metal concentrations divided by their respective chronic saltwater standards for the metals As, Cd, Cu, Pb, Hg, Ni, Se, and Zn. Ten of the 182 sites were in tidal freshwaters and their quotients were based on chronic freshwater standards for the same metals. The maximum value observed (0.2897) represents the 99.9th percentile and cannot be plotted on the chart.

higher concentrations, indicating slight to moderate stress from elevated concentrations of one or more dissolved metals. Only five sites had mean standard quotients above 0.0900 (97.6th percentile, upper 2.4%), indicating a more severe stress, and in only one case was a chronic standard exceeded.

Water clarity is expressed as percent of photosynthetically active radiation (PAR) that is available at a depth of 1.0 meter, relative to what is available at the water's surface. It is one of the few stressors (along with dissolved Oxygen and pH) in which lower values indicate increased stress. In the cumulative normal probability chart of Figure G available PAR values above 40% are all from sites within the Chesapeake Bay mainstem or in its embayments or the lower, deeper reaches of minor tidal tributaries to the Bay. Available PAR values from 40% to approximately 15% fall in a relatively straight line in the chart, indicating that they are approximately normally distributed with a typical variation about some mean value (approximately 27.5%). Available PAR values lower than 15% (152 sites – 55.7%) deviate sharply (downward) from the straight line and indicate moderately to severely limited PAR availability.

Water Clarity expressed as %PAR @ depth of 1.0 meter



Water Clarity expressed as %PAR @ depth of 1.0 meter

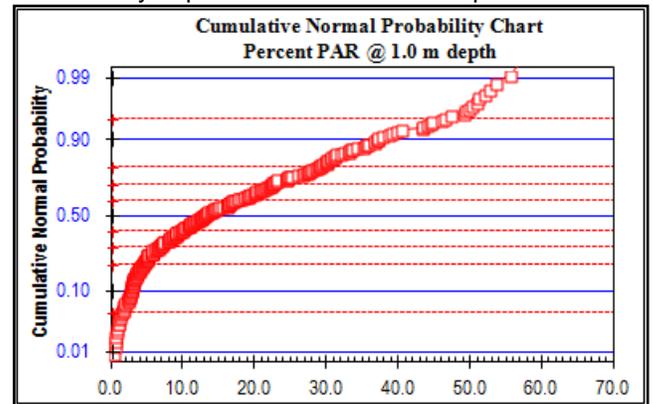


Figure G – Water Clarity expressed as Percent of Available Photosynthetically Active Radiation (PAR) at a Depth of 1.0 Meter. One of the few environmental stressors where a decrease in value indicates a greater stress (dissolved Oxygen and pH are the others). All sites with %PAR values of 40% or more were in the Chesapeake Bay mainstem or in its embayments and minor tributaries (21 sites = 7.7%).

The compound distributions discussed below illustrate cases where an irregular cumulative normal probability distribution may result from typical variations of natural conditions rather than from contamination or intensification of a specific stressor, or where distinct distributions merge as opposed to a gradient of stressor intensity modifying the shape of a more natural distribution.

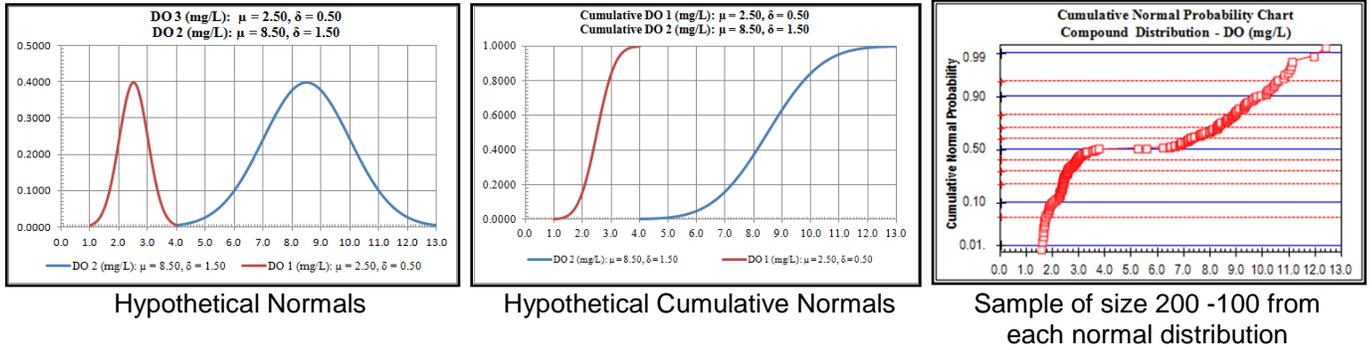


Figure H – Hypothetical Case where Two Normal Distribution Overlap. The two hypothetical normal distributions were defined to describe two populations of dissolved Oxygen concentrations with different means and standard deviations, and barely overlapping value ranges. The cumulative normal probability chart was constructed from a composite random sample of size $N = 100$ observations from each of the two normal distributions. Note that the distribution with the largest standard deviation (DO 2 - broader) has a less steep slope in its cumulative forms.

Figure I – Compound Bottom Temperature Distribution Observed During the 2010 Near-shore Oceanic Survey. (DEQ-WQA, 2012) The relatively straight line in the upper right portion of the chart represents bottom temperatures at sites with well-mixed waters from surface to bottom, while the lower left portion of the chart represents cooler temperatures at the bottom of an unmixed (more stratified) water column. In this case, the two different distributions have very similar standard deviations (*i.e.*, the two distributions are of approximately the same width), and consequently have similar slopes. ($N = 50$)

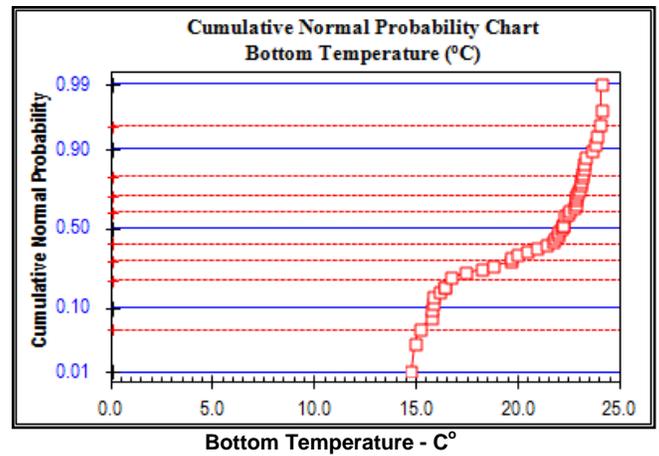


Figure J – Compound Distribution of Bottom Dissolved Oxygen Concentrations Observed During the 2010 Near-shore Oceanic Survey. (DEQ WQA, 2012). The relatively straight line in the upper right portion of the plot represents bottom DO concentrations in well-mixed oceanic waters, isolated from continental waters by broad areas of estuarine marshes and/or barrier islands, beaches or dunes. The cluster of values in the lower left portion of the figure represents a localized region of coastal Delmarva, between Chincoteague channel and Wachapreague, where nutrient-laden (DO depressed) continental waters are better able to mix with oceanic waters. ($N = 50$)

