Characterizing Chlorophyll Distributions in Colorado Lakes

The variability of chlorophyll concentrations within and among lakes poses challenges for the establishment of chlorophyll standards and assessment of their attainment. There is some apprehensiveness that patterns of variation are lake-specific to the extent that development of statewide standards may be impractical. Moreover, the nature of distributions for individual lakes, especially if non-normal, has important implications for the way that standards should be defined. These issues can be addressed by examining large sets of lakes and seeking statistical patterns.

Chlorophyll has been measured in many lakes throughout the state, although most lakes have been sampled only a few times. For the purpose of characterizing the statistical distributions of chlorophyll, some constraints are needed. There must be at least 20 observations (not necessarily from the same year) from the averaging period of interest – summer (Jul-Sep). The required number of observations is largely arbitrary. The averaging period represents the Division's current preference based on precedence and the likelihood that those months will be equally weighted in most available data sets. The second point is particularly relevant to averaging periods, like annual or stratification season, which probably do not have equal representation of all months.

Twenty lakes met the screening criteria, and some of them had many more than the minimum number of observations (Table 1). Based on previous examination of many data sets, it was expected that chlorophyll concentrations in most lakes would fit a lognormal[1] distribution. This assumption was tested by means of probability plots using Minitab statistical software. The Anderson-Darling statistic is used to compare goodness of fit between normal and lognormal distributions (Table 1); smaller numbers mean a better fit.

With the exception of Arvada Reservoir, the lognormal distribution fit the data better than the normal distribution; chlorophyll concentrations in Arvada Reservoir fit a normal distribution (Figure 1). Goodness-of-fit values range from 0.4 to 1.3 for lakes other than Arvada Reservoir. Green Mountain Reservoir (Figure 2) and Standley Lake (Figure 3) thus bracket the range of "fits" observed.

Lake	POR	N	A-D normal	A-D lognormal	Data Source
Green Mountain	1984-2001	87	7.3	1.3	SWQC
Dillon	1981-2005	142	3.7	0.9	SWQC
Chatfield	1987-2005	109	7.7	0.4	Chatfield WA
Cherry Creek	1997-2006	93	2.7	0.7	Cherry Creek BWQA
Bear Creek	1991-2006	95	7.2	1.2	Bear Creek WA
Boulder	1993-2006	41	1.1	0.8	Boulder
Barr	2002-2006	29	2.1	0.8	MWRD

[1] All references in this document to logs, log-tranforms, and lognormal assume the use of natural (base e) logarithms.

Lake	POR	N	A-D normal	A-D lognormal	Data Source
Milton	2002-2006	30	3.3	1.1	MWRD
Shadow Mountain	1989-2006	51	5.3	0.5	USGS/USBR
Seaman	2000-2005	25	1.8	0.7	Greeley
Boyd	1999-2005	20	2.5	1.0	Greeley
Loveland	1999-2005	20	0.9	0.7	Greeley
Arvada	1994-2006	73	0.9	2.7	Arvada
Aurora	1997-2006	114	1.6	0.8	Aurora
Quincy	1998-2006	96	2.3	0.7	Aurora
Standley	1995-2006	83	2.9	0.4	Westminster
Granby	1989-2006	42	2.4	0.8	USGS/USBR
Wolford	1995-2005	23	2.2	0.8	USGS
Horsetooth	2000-2006	26	1.0	0.8	USGS/USBR
Carter	1989-2006	27	1.3	0.7	USGS

Table 1. Summary of data sets available for assessment of chlorophyll distributions. Anderson-Darling (A-D) goodness-of-fit values indicate fit to the normal and lognormal distributions; lower values mean better fit. With the exception of Arvada Reservoir, the data conform better to lognormal than to a normal distribution.

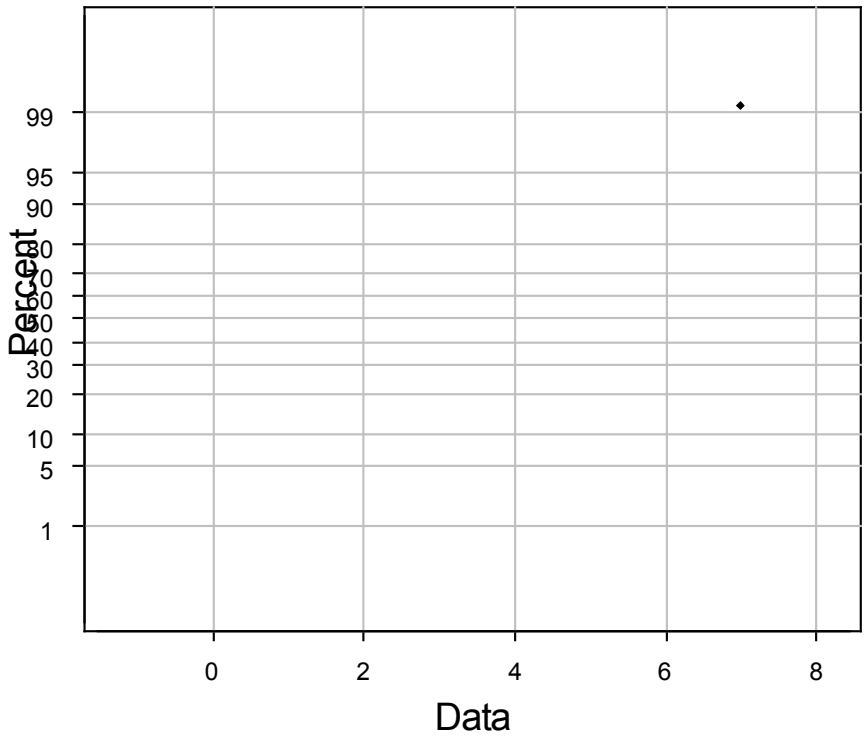

Figure 1. Probability plot of chlorophyll data from Arvada Reservoir.

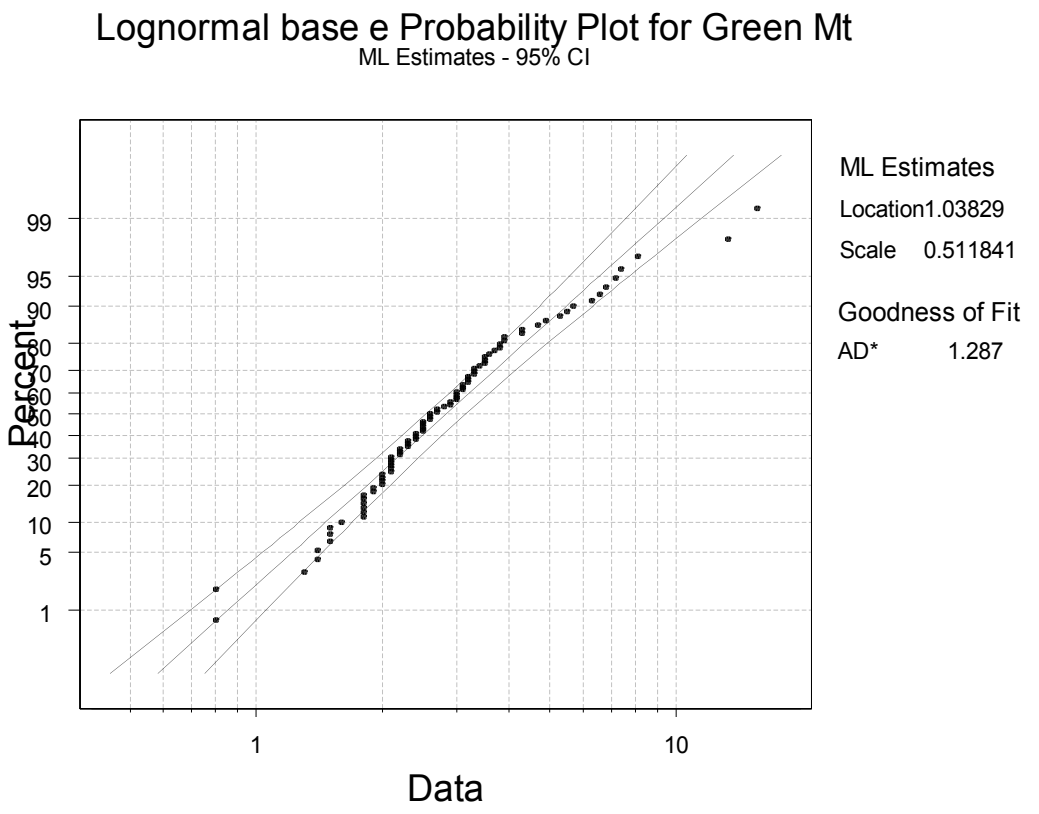

Figure 2. Probability plot of chlorophyll data from Green Mountain Reservoir.

In general, the data support the assumption that a lognormal distribution is reasonable for chlorophyll values from individual lakes. It is almost always a better choice than the normal distribution. This becomes important when defining the basis for assessment. Parametric procedures, when applied appropriately, yield smaller confidence intervals than non-parametric procedures, and the log-normal procedures are better than normal procedures for asymmetrical distributions like chlorophyll.

Knowing that chlorophyll concentrations in any lake are likely to conform to a lognormal distribution leads to predictions about statistical properties. For example, it is generally true of variables with lognormal distributions that the variance increases in proportion to the mean (Sokal and Rohlf 1995). Walker (1985) made use of this tendency to develop a statistical basis for chlorophyll standards in other regions. The relationship is quite strong for Colorado lakes (Figure 4). A power function based on the arithmetic mean[2] (average of untransformed chlorophyll data) explains nearly 95% of the variation in the standard deviation (calculated from untransformed data). The strength of the relationship does much to enable a statewide approach to a chlorophyll standard.

[2] Unless stated to the contrary, mean refers to the arithmetic mean, which is used interchangeably with average throughout this document.

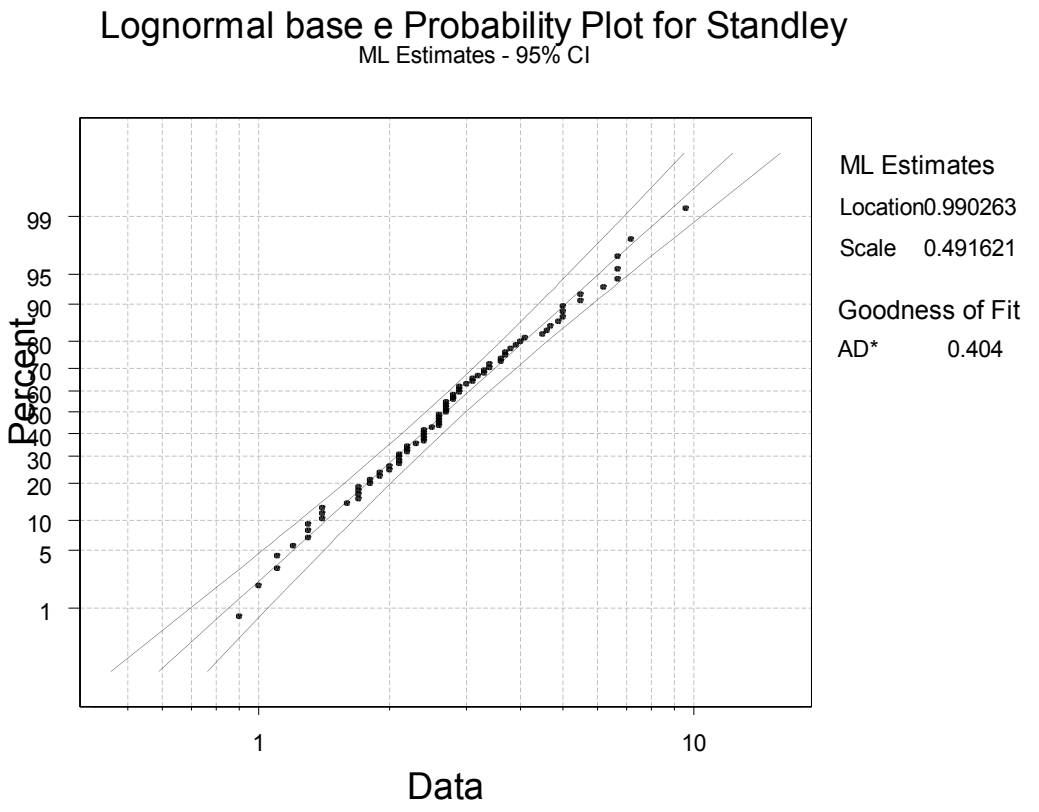

Figure 3. Probability plot of chlorophyll data from Standley Lake.

The relationship between the standard deviation and the mean provides a powerful basis for generalizing about the variance associated with a standard of any magnitude in any lake. For example, if the chlorophyll standard were set at 15 ug/L, the expected variance could be estimated with a high degree of certainty from the regression equation in Figure 4. Armed with mean and standard deviation, it is possible to estimate confidence intervals for the mean and prediction intervals for individual observations. In other words, it becomes possible to develop a statistical rationale for assessing attainment of the standard for any given lake.

A simple screening procedure that could be applied to any lake with a chlorophyll standard uses the statistical concept of a "prediction interval" (as explained in Helsel and Hirsch 2002). Each new chlorophyll measurement is compared with the distribution associated with the standard to determine if it is likely that the new observation came from the same distribution underlying the standard or from a different distribution (i.e., one with a larger mean). The prediction interval, which is computed from the distribution associated with the standard, defines the range of chlorophyll values expected for a certain percentage, let's say 90%, of the existing distribution. When dealing with a standard, it makes sense to state this definition such that 90% of the values lie below a particular concentration (i.e., a one-sided prediction limit). A new measurement that exceeds this 90% threshold is unlikely to be a member of the distribution corresponding to the standard (there is still a 10% chance that it is from the same distribution); i.e., the chlorophyll standard in that lake is unlikely to be attained.

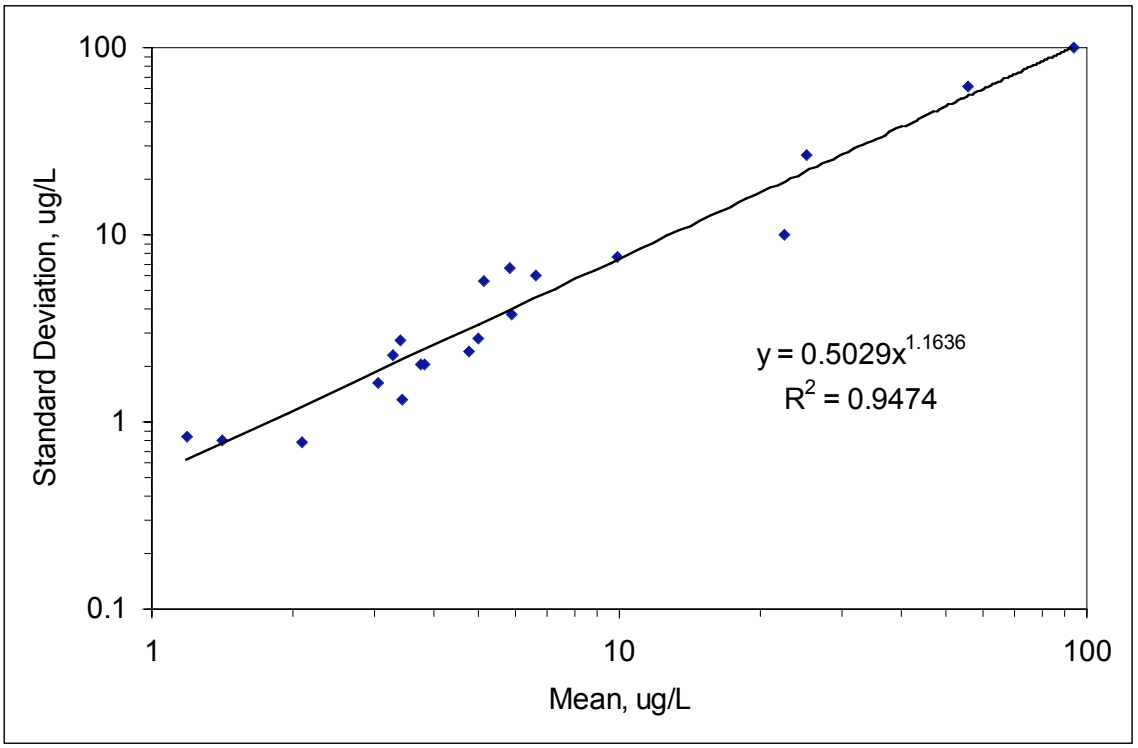

Figure 4. Power function characterizing the relationship between the mean and the standard deviation of Jul-Sep chlorophyll concentrations (untransformed) measured in a set of Colorado lakes.

The equation for the prediction interval uses the mean and the standard deviation of the log-transformed values, but the power function relationship between mean and standard deviation is based on untransformed data. An algorithm is used to bridge the gap (Aitchison and Brown 1963, as shown in Walker 1985). The standard deviation of the log-transformed values (SL) can be estimated from the standard deviation and the mean of the untransformed values (MA and SA) as follows: $SL^2 = LN[1+(SA/MA)^2]$. Thus, for a chlorophyll standard (assumed to be the "population mean") of 15 ug/L, the corresponding standard deviation for the untransformed data would be estimated as follows: $SA = 0.5029*15^{1.1636} = 11.75$, from which the variance of the log-transformed values would be estimated as $SL^2 = LN[1+(11.75/15)^2]$, or 0.478, yielding a standard deviation of 0.691. The mean of the log-transformed values (ML) can be estimated from the arithmetic mean as follows: $ML = LN(MA) – 0.5SL^2$. For this example, ML = LN(15) – 0.5*0.478, or 2.47.

The prediction interval also requires a sample size estimate and a value from the t-distribution. For the purpose of estimating the prediction interval of the standard, sample size is assumed to be very large; a sample size of 100 is chosen arbitrarily. The justification is based on the origin of the standard deviation for the distribution of the standard, which was calculated from a regression line derived from a set of 20 lakes from which more than 1200 chlorophyll measurements were available. The choice does not

matter greatly insofar as the t-distribution, which matches the normal distribution for $N=\infty$, yields values within a few percent of the normal distribution for $N>30$.

The one-sided upper prediction limit (PL) for a standard of 15 ug/L, with $\alpha=0.10$ and $n=100$, is calculated as shown below (from Helsel and Hirsch 2002). In the equation, the mean and standard deviation are for the log-transformed values.

$PL=\exp[\text{mean}+t_{(\alpha,n)}\sqrt{(s_y^2+s_y^2/n)}]$
$PL=\exp[2.47+1.29\sqrt{(0.478+0.478/100)}]$
$PL=29$
If α were set to 0.05, the upper PL would 37.4.

The prediction interval can be used to assess individual samples from any lake. In this sense, it can be an early warning tool. If the chlorophyll concentration in one sample is greater than 29 ug/L, that sample has only a 10% chance of belonging to a population with a mean of 15 ug/L. Put another way, it is a strong indication that the lake would not attain a chlorophyll standard of 15 ug/L.

Chlorophyll concentrations also can be assessed in terms of the observed seasonal mean. In this case, the observed mean is compared to the 90^{th} or 95^{th} percentile confidence interval for the distribution of the standard. The H-statistic procedure is often used for locating the percentile as an untransformed value (Gilbert 1987), but concerns have been raised about this statistic (e.g., Singh et al. 1997). The Division is still reviewing options for estimating confidence limits.

Citations

Aitchison, J and JAC Brown. 1963. The Lognormal Distribution. Cambridge University Press (not seen).

Gilbert, RO. 1987. Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold, New York.

Helsel, DR and RM Hirsch. 2002. Statistical Methods in Water Resources. Chapter A3, Book 4, Techniques of Water-Resources Investigations of the United States Geological Survey. Publication available at: http://water.usgs.gov/pubs/twri/twri4a3/

Singh, AK, A Singh, and M Engelhardt. 1997. The lognormal distribution in environmental applications. EPA Technology Support Center Issue. EPA/600/R-97/006. http://www.hanford.gov/dqo/training/lognor.pdf

Sokal, RR and FJ Rohlf. 1995. Biometry, 3rd edition. WH Freeman and Co., NY. 887 pages.

Walker, WW. 1985. Statistical bases for mean chlorophyll A criteria. Pages 57-62 in "Lake and Reservoir Management - Practical Applications", Proc. 4th Annual Conference, North American Lake Management Society, McAfee, New Jersey, October 1984.