STATISTICS FOR HISTORICAL TREND ANALYSIS

For those lakes and ponds that have participated in VLAP for at least 10 consecutive years, we are now analyzing the in-lake data with a simple statistical test. The test is used to determine if there has been a *significant change* in the annual mean value for the three major sampling parameters during the period that the lake or pond has been sampled in VLAP. Specifically, we are using a **linear regression line and regression statistics** to determine if there has been an increase or decrease of the annual mean for chlorophyll-a, Secchi-disk transparency, and total phosphorus. Since this is a new addition to the VLAP annual report, we would greatly appreciate any feedback or comments that you can provide. Statistical analysis can be confusing, so please let us know if the analysis or explanation is difficult to understand.

WHAT ARE STATISTICS?

A statistical test provides a mechanism for making an objective, not subjective, decision about a process. The intent of a statistical test is to determine whether there is enough evidence to "reject" a hypothesis about the process. In the past, we have evaluated the data by "eyeing" (looking at) the trendline to determine if an overall increase or decrease in water quality for these parameters has occurred. This statistical test will allow mathematic equations to determine if there has been a change over time.

HOW WILL STATISTICS BE USED IN VLAP?

For VLAP, we are using a simple linear regression statistical test to determine if there is enough evidence to "reject" the null hypothesis that the annual mean value for the water quality parameter of interest, such as chlorophyll-a concentration, Secchi Disk transparency, or total phosphorus concentration, **has not changed** during the time that that lake or pond has been sampled in VLAP. If there is enough evidence to "reject" the null hypothesis, then we will accept the alternative hypothesis, which says that the mean value **has changed** (either increased or decreased) during the time that the lake or pond has been sampled in VLAP.

> **Ho (the null hypothesis):** The annual mean value for the water quality parameter of interest (either chlorophyll-a concentration, Secchi Disk transparency, or total phosphorus concentration) **has not changed** during the time that the lake or pond has been sampled in VLAP.
>
> **Ha (the alternative hypothesis):** The annual mean value for the water quality parameter of interest (either chlorophyll-a concentration, Secchi Disk transparency, or total phosphorus concentration) **has changed** (either increased or decreased) during the time that the lake or pond has been sampled in VLAP.

SIGNIFICANCE LEVELS

We want to know if the null hypothesis, the annual mean value for the water quality parameter of interest **has not changed** over time, is "true" or "false", which is why we are "testing" it. The alternative hypothesis, the annual mean value for the water quality parameter of interest **has changed** over time, might be true. The procedure to "test" the null hypothesis is constructed so that the risk of "rejecting" the null hypothesis, when it is in fact "true", is relatively small. The risk is referred to as the **significance level** of the "test". By having a significance level for the "test", we feel that we have actually "proved" something when we reject the null hypothesis.

For VLAP we are using a significance level of 0.05, which implies that the null hypothesis is only "rejected" 5 percent of the time, when it is in fact "true". Specifically, this means that only 5 percent of the time, we will be claiming that the annual mean value of the water quality parameter of interest **has changed** over time, when, in reality, it **has not** changed over time. Or, stated in another way, this means that we are 95 percent confident when we "reject" the null hypothesis that the annual mean value of the water quality parameter of interest **has changed** over time.

HOW DO WE DETERMINE IF THE NULL HYPOTHESIS IS "TRUE" OR "FALSE"?

To determine if the null hypothesis is "true" or "false" we look at a probability value. The **probability value (p-value)** of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme or more extreme than that observed by chance alone, if the null hypothesis, is true. The p-value is compared with the significance level, and, if it is smaller, the result of the "test" is significant. Small p-values suggest that the null hypothesis is unlikely to be true. The smaller the p-value is, the more convincing the "rejection" of the null hypothesis is.

Specifically, for VLAP, since we are using a significance level of 0.05, this means that if the p-value is less than 0.05, we will "reject" the null hypothesis, the annual mean value of the water quality parameter of interest **has not changed** over time. If we "reject" the null hypothesis, then we will accept the alternative hypothesis, the annual mean value of the water quality parameter of interest **has changed** over time. Again, the smaller the p-value is, the more convincing the "rejection" of the null hypothesis is.

p-value	**Action**
greater than 0.05	"fail to reject" the null hypothesis
less than 0.05	"reject" the null hypothesis and "accept" the alternative hypothesis

HOW DO WE KNOW IF THE CHANGE OVER TIME IS "INCREASING" OR "DECREASING"?

If we "reject" the null hypothesis and conclude that the annual mean value for the water quality parameter of interest **has changed** since VLAP sampling began, then we will need to determine if the change over time has been an "**increasing trend**" or a "**decreasing trend**". To determine this, we look at the regression coefficient that is assigned to the "x" variable, which is the slope of the regression line. If the "x" variable coefficient is negative, meaning less than 0, then this indicates that the change in the annual mean value of the water quality parameter of interest with respect to time is "decreasing". If the "x" variable, the slope of the regression line, is positive, then this indicates that the change in the annual mean value of the water quality parameter of interest with respect to time is "increasing".

"x" variable coefficient (slope of the regression line)	**Trend Interpretation**
negative (less than 0)	decreasing trend
positive (greater than 0)	increasing trend

HOW DO WE KNOW THE STRENGTH OF THE TREND?

If we have "rejected" the null hypothesis and have concluded that the annual mean value for the water quality parameter of interest **has changed** since VLAP sampling began, and we have also determined if the change over time has been an "**increasing trend**" or "**decreasing trend**", then we will want to know how **strong** of an increase or decrease this trend is. The strength of the trend can be reported as a **percent change over time**. To calculate the percent change in time for the water quality parameter of interest, we divide the slope of the regression line, the regression coefficient that is assigned to the "x" variable, by the mean value of the water quality parameter of interest over time. To calculate the mean value over time, we simply add together the annual mean value for the water quality parameter of interest for each sampling season, and then divide this total by the number of years the lake or pond has been sampled through VLAP. This number represents the percent change in the water quality parameter interest over time. The **larger** the percent change over time for the water quality parameter of interest indicates the **greater** the strength of the trend.

As an example, let's discuss the historical chlorophyll-a data from Kezar Lake in North Sutton:

1. We inputted the historical data from 1988 to 2001 for the chlorophyll-a concentration into the computer software program, and the results of the regression gave a **p-value of 0.007**, which is less than 0.05, so we **"rejected"** the null hypothesis and **"accepted"** the alternative hypothesis, which says that the chlorophyll-a concentration has changed over time.
2. Since the coefficient of the **"x" variable (slope of the regression line) is "– 0.392",** we know that the change in the annual mean chlorophyll-a concentration since the lake has been sampled is a **decrease**.

3. Now we want to know how **strong** the decrease is, so we calculate the **percent change in the annual mean chlorophyll-a concentration over time** (as shown below):

Sampling Season	**Mean Annual Chlorophyll-a concentration (mg/m^3)**
1988	12.20
1989	7.55
1990	9.93
1991	7.85
1992	5.47
1993	11.31
1994	8.76
1995	6.73
1996	6.08
1997	3.84
1998	6.22
1999	7.13
2000	5.31
2001	5.14
Total (sum of annual means) =	103.51
Overall Mean (Total/number of sampling seasons) =	6.90
"x"-variable coefficient (slope of regression line) =	-0.392
Percent Change over Time ("x" variable coefficient/Overall mean) x 100 =	-5.65%

4. This calculation shows the average percent change over time is –5.65%. Specifically, this means that the annual mean chlorophyll-a concentration in Kezar Lake has **decreased on average by 5.65% per year** during the sampling period 1988 – 2001. We know that this is a decrease because there is a negative sign.

HOW DO WE KNOW HOW MUCH OF THE CHANGE IN THE WATER QUALITY PARAMETER OF INTEREST IS CORRELATED WITH TIME?

To determine how much or how little of the change in the water quality parameter of interest is correlated with time, or stated another way, to determine the percentage of the variability in the water quality parameter of interest that is explained by the variability in time, we look at the **R-squared value**. The R-square value is a measure of the degree of relationship between two variables "x" and "y". Again, the "x" variable is time, meaning sampling season, and the "y" variable is the annual mean value of the water quality parameter of interest. The R-squared value can have any value

between "0" and "+1". An **R-squared value of "0"** indicates that there is no correlation, meaning that there is no amount of variability in the "y" variable (the water quality parameter of interest) that is explained by the variability in the "x" variable (time). An **R-squared value of "1"** indicates that there is a perfect correlation, meaning that all the variability in the "y" variable (the water quality parameter of interest) is explained by the variability in the "x" variable (time).

R-squared value	**Relationship between "x" and "y" variables**
0	no correlation, no variability explained
0 to 1	some correlation, some variability explained
1	all variability explained

Let's look at the Kezar Lake data again:

1. We determined the strength of the decrease in the annual mean chlorophyll-a concentration in Kezar Lake from 1988 to 2001 by calculating the percent change in the annual mean chlorophyll-a concentration over time. We determined that the annual mean chlorophyll-a concentration in Kezar Lake has **decreased on average by approximately 5.65 percent** per year during the sampling period 1988 to 2001.
2. Now we want to know how much of the 5.6 percent decrease is explained by the variation in time. To do this, we look at the **correlation coefficient (R-squared value)** that was generated by the regression.
3. The results of the regression gave an **R-squared value of 0.46.**
4. This means that approximately half of the decrease in the annual mean chlorophyll-a concentration is explained by the variation in time. Since the R-square value is not "1", which would indicate that all of the variability in the annual mean chlorophyll-a concentration is explained by the variability in time, this means that there may be other variables that explain the variability in the chlorophyll-a concentration. These other variables could be the total phosphorus concentration in the lake or the amount of precipitation during the summer. We would need to conduct a multiple variable regression to determine what additional variables account for the remainder of the variation in the annual mean chlorophyll-a concentration.