VCU

Monroe Park Campus

Virginia Commonwealth University

Statistical Sciences &
Operations Research

Harris Hall, 4th Floor
1015 Floyd Avenue
P.O. Box 843083
Richmond, Virginia 23284-3083

804 828-0001
Fax: 804 828-8785
TDD: 1-800-828-1120
www.stat.vcu.edu

July 9, 2010

To Whom It May Concern:

The Virginia Department of Health (VDH) collected data from three proprietary systems (Puraflo, AdvanTex, and Ecoflo) in order to establish an interim "end-of-pipe" standard of effluent (BOD, TSS, Fecal Coliform) leaving a treatment unit. The attached document provides a statistical assessment of their methods and findings.

Conclusions from the assessment can be summarized as follows:

a) The raw VDH data obtained is right (or positively skewed) and thus a log-transformation of the data is utilized in order to stabilize its variance and correct for non-normality. The assumption of normality for the log-transformed data is indeed reasonable for the upper tails. This is verified through the use of normal probability plots. The information provided in the upper tail of the distribution of data is most relevant for establishing upper limits on effluent.

b) The use of the standard error of the mean and confidence intervals for the mean are not appropriate when interest lies in where *individual* treatment units will fall. Instead, tolerance intervals should be used for conformance monitoring.

c) Tolerance intervals computed for the VDH data set indicate that the pass/fail criterion for effluent is too low as a larger than likely acceptable percentage (from a manufacturing standpoint) of treatment units will fail the pass/fail criterion. Further investigation and refinement is recommended.

Please do not hesitate to contact me at dedwards7@vcu.edu or (804) 828-2936 with comments/questions.

Sincerely,

David J. Edwards

David J. Edwards, Ph.D.
Assistant Professor of Statistics

Assessment of Virginia Department of Health's Analysis of Alternative Wastewater Systems Data

The Virginia Department of Health (VDH) collected data from three proprietary systems (Puraflo, AdvanTex, and Ecoflo) in order to establish an interim "end-of-pipe" standard of effluent (BOD, TSS, Fecal Coliform) leaving a treatment unit. This assessment is based on the "cleaned-up" version of the VDH data set.

1. *Use of Log Transformation.*

Given the large degree of skewness in the data set, an initial step in the VDH analysis after data "clean-up" was to transform each variable using a natural logarithm transformation in order to help correct for non-normality. While the log-transformed data is not perfectly normal, some departure from normality is expected. Furthermore, as interest lies in establishing an upper bound on effluent (rather than a lower bound), one may direct their focus to the normality of the upper tail of the log-transformed data. That is, statistical techniques that assume normality may be valid even if the entire data set is non-normal so as long as the subset of data of interest is approximately normal.

A normal probability plot is a useful graphical tool for assessing the normality of any given data set. In particular, data is plotted against the theoretical quantiles of a normal distribution in such a way that the points should form an approximate straight line if the data is from a normal distribution. Departures from this line provide an indication of non-normality. Figures 1-9 provide normal probability plots, histograms, and boxplots for each log-transformed variable.

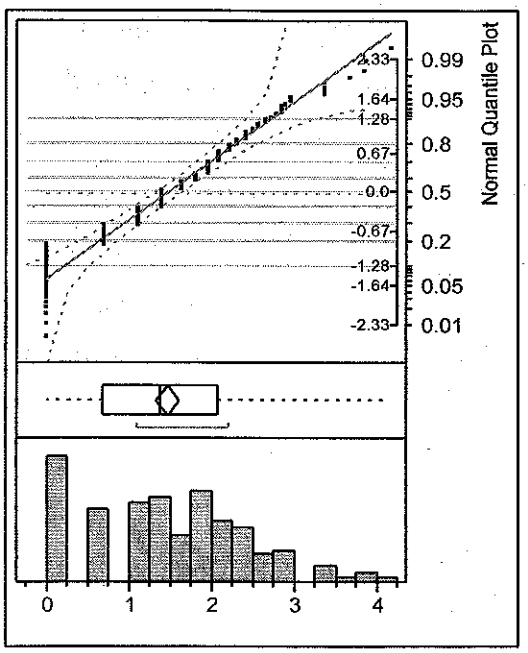

Figure 1. Log Puraflo BOD

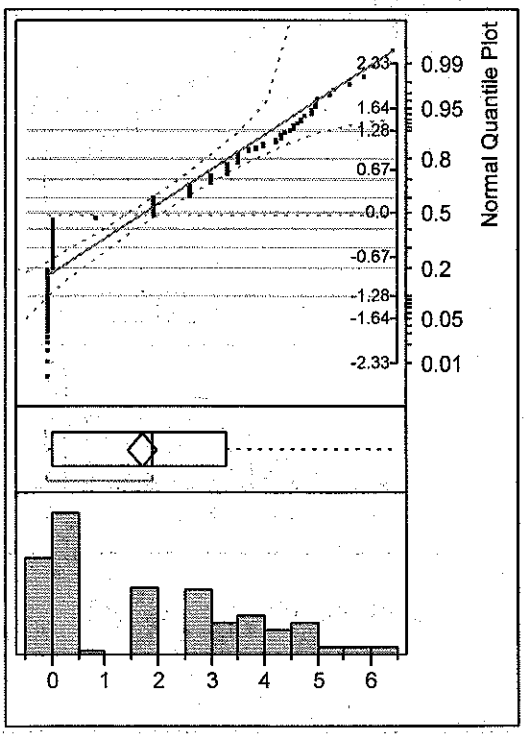

Figure 2. Log Puraflo TSS

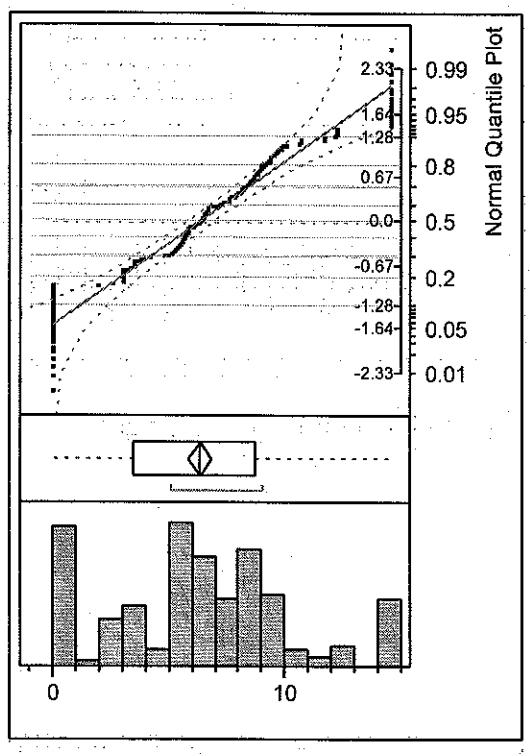

Figure 3. Log Puraflo Fecal Coliform

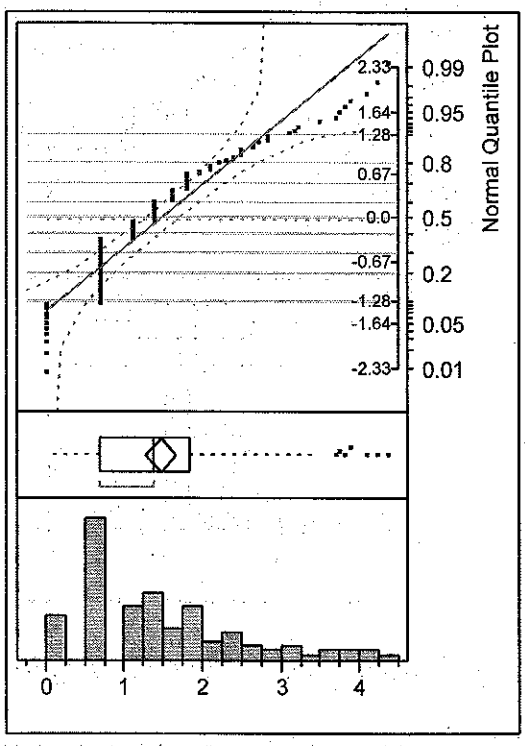

Figure 4. Log AdvanTex BOD

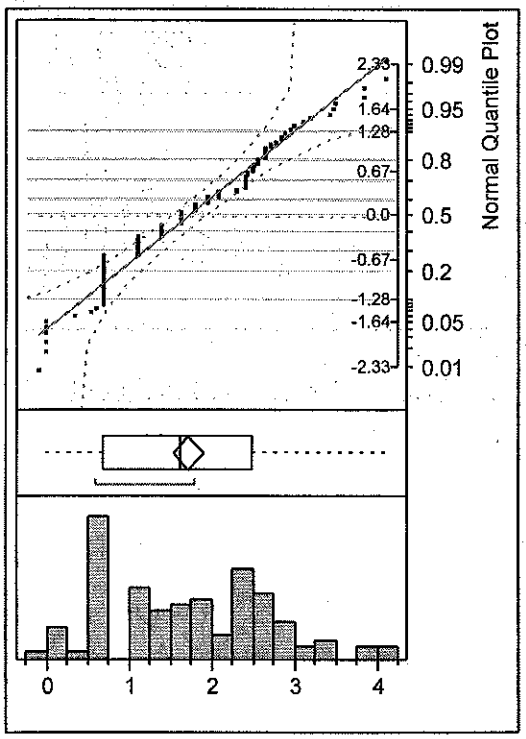

Figure 5. Log AdvanTex TSS

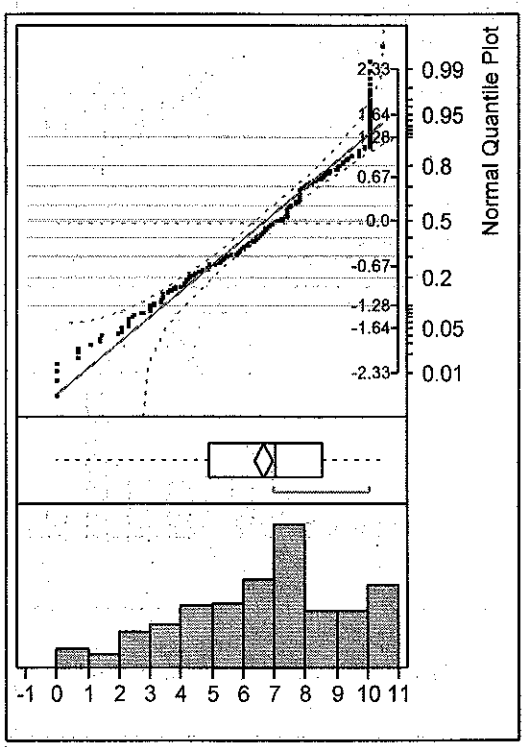

Figure 6. Log AdvanTex Fecal Coliform

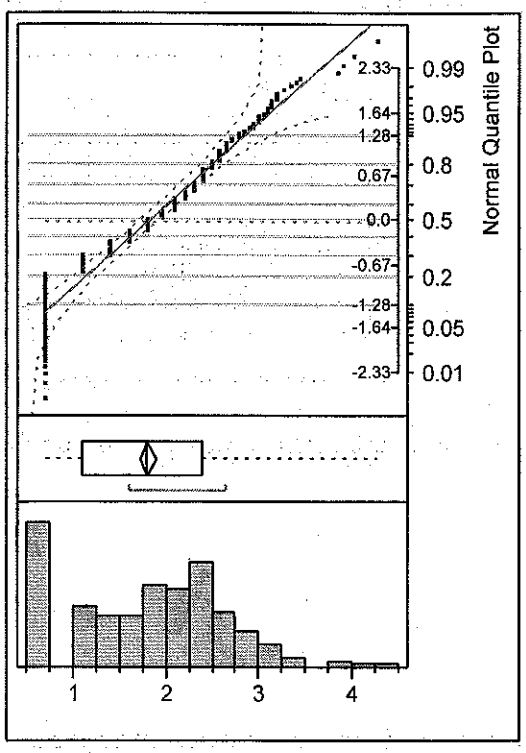

Figure 7. Log Ecoflo BOD

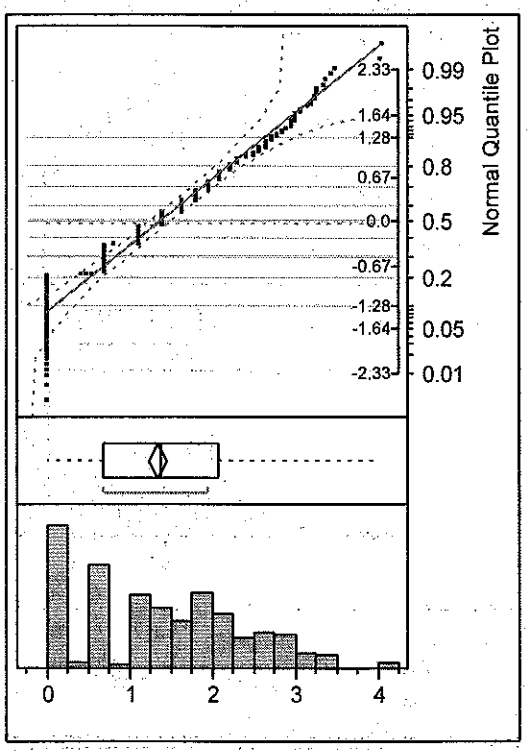

Figure 8. Log Ecoflo TSS

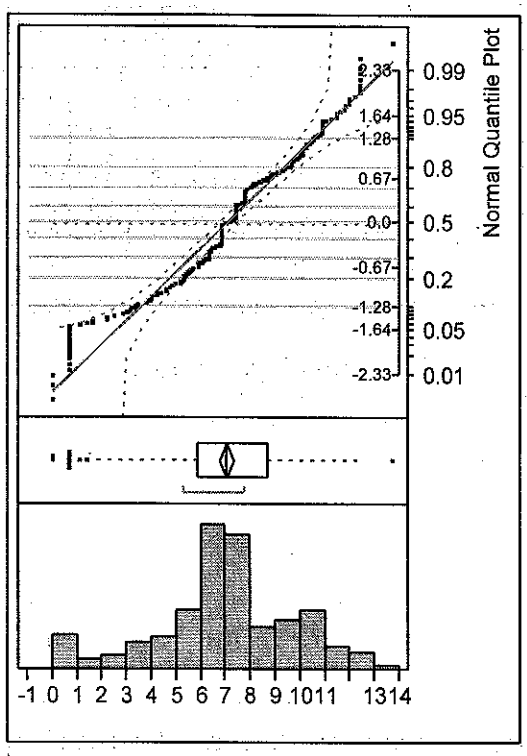

Figure 9. Log Ecoflo Fecal Coliform

With the exception of log-transformed Puraflo Fecal Coliform and log-transformed Advantex Fecal Coliform, the upper tails of each distribution appear to reasonably follow a normal distribution. More severe departure from normality is evident in the lower tails. Although some departure from normality is present in the upper tails of Puraflo and Advantex Fecal Coliform variables, the deviation from normality is not severe. Therefore, it is reasonable to utilize normality-based inference techniques as done in the VDH analysis.

Note that with the log-transformation, back-transforming the mean (by exponentiation) of the transformed data will never be the same as the mean of the untransformed data. In particular, back-transforming log-transformed data yields the geometric mean of the original data rather than the usual arithmetic mean. With right skewed data (as is the case with the VDH data), the geometric mean is always less than the arithmetic mean. Since the arithmetic mean is highly influenced by extreme values, the log-transformation provides a way of lessening such influence.

2. *Use of Standard Error of the Mean.*

In order to establish an upper bound for "end-of-pipe" standard of effluent leaving a treatment unit, VDH utilized the log transformed upper limit of a 99% confidence interval for the mean effluent transformed back to native units. This has led to the following pass/fail criteria for effluent: ≤10 mg/l of BOD_5, ≤10 mg/l of TSS, and ≤2000 cfu/100ml of Fecal coliform.

Letting $\overline{X}$ represent the mean of a log-transformed variable, VDH computed a 99% confidence interval for each variable as follows:

$$\overline{X} \pm 2.576\left(\frac{s}{\sqrt{n}}\right)$$

where s is the standard deviation of the log-transformed variable and n is the number of observations. To compute the converted upper confidence limit, one needs only to exponentiate the log-transformed upper confidence limit. The quantity, $s/\sqrt{n}$, is known as the standard error of the mean and provides a metric for variability of the sample mean.

One misconception regarding the use of confidence intervals is that a confidence interval covers a particular proportion of the population. Said another way, a common misuse of confidence intervals is to address the spread of individual data values. Rather, confidence intervals provide an interval estimate of a parameter (some unknown characteristic) of a population (such as the mean). However, if interest lies in where individual values should be, as it appears to be in the VDH study, the use of the standard error of the mean is inappropriate.

Instead, a statistical tolerance interval is more applicable for compliance monitoring studies. From Hahn and Meeker (1991), a tolerance interval is defined to be an interval that one can claim to contain at least a specified proportion, p, of the population with a specified degree of confidence. Such an interval would be of interest in setting limits on the process capability for a product manufactured in large quantities. An upper tolerance limit is one that guarantees that at least p percent of population measurements will not exceed this upper limit and has the form:

$$\overline{X} + Ks$$

where K is a constant that depends on p and the level of confidence desired. Tables of values of K can be found Hahn and Meeker (1991). The above formula for an upper tolerance limit is valid when the normality assumption is reasonable (as it is for our purposes). It is worth noting that nonparametric (or distribution free) tolerance limits exist for any probability distribution. However, they have limited practical value as they require substantially larger sample sizes then are available in this study (Montgomery (2009).

The following results detail log-transformed upper 99% tolerance limits converted back to natural units based on Hahn and Meeker (1991):

		Log Transformed		Log 99% Upper Tolerance Limit						
	n	Mean	Std. Dev.	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9	p=0.95	p=0.99
Puraflo BOD	184	1.46	0.97	1.63	1.88	2.16	2.48	2.95	3.33	4.07
Puraflo TSS	184	1.71	1.85	2.03	2.51	3.04	3.66	4.55	5.29	6.70
Puraflo Fecal Coliform	234	6.31	3.97	6.92	7.95	9.07	10.41	12.29	13.86	16.84
AdvanTex BOD	114	1.46	1.02	1.69	1.95	2.25	2.60	3.10	3.52	4.32
AdvanTex TSS	115	1.71	0.98	1.93	2.19	2.47	2.81	3.29	3.70	4.47
AdvanTex Fecal Coliform	280	6.67	2.51	7.02	7.67	8.38	9.21	10.39	11.37	13.24
Ecoflo BOD	333	1.8	0.79	1.90	2.11	2.33	2.59	2.96	3.27	3.86
Ecoflo TSS	337	1.34	0.98	1.47	1.72	2.00	2.32	2.78	3.16	3.89
Ecoflo Fecal Coliform	337	7.05	2.71	7.40	8.10	8.85	9.76	11.02	12.08	14.08

	Converted 99% Upper Tolerance Limit						
	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9	p=0.95	p=0.99
Puraflo BOD	5.10	6.56	8.63	11.99	19.05	28.07	58.54
Puraflo TSS	7.61	12.32	20.84	39.04	94.74	198.90	811.13
Puraflo Fecal Coliform	1012.66	2834.46	8700.27	33067.35	217164.40	1046680.00	20554580.00
AdvanTex BOD	5.40	7.06	9.47	13.47	22.20	33.76	74.89
AdvanTex TSS	6.88	8.91	11.83	16.63	26.94	40.40	87.23
AdvanTex Fecal Coliform	1123.21	2145.62	4338.78	10021.80	32567.57	87114.53	564729.84
Ecoflo BOD	6.69	8.22	10.26	13.37	19.38	26.43	47.57
Ecoflo TSS	4.34	5.59	7.36	10.20	16.13	23.66	48.90
Ecoflo Fecal Coliform	1631.93	3281.17	7007.40	17250.74	61224.11	176105.58	1306686.03

As an example for interpretation, consider Puraflo BOD. Then, we are 99% confident that at least 50% of manufactured treatment units have BOD levels less than 5.1 mg/L. Likewise, we are 99% confident that at least 80% of treatment units have BOD levels less than 11.99 mg/L and finally, we are 99% confident that at least 95% of treatment units have BOD levels less than 28.07 mg/L. Therefore, a pass/fail criterion of 10 mg/L of BOD likely underestimates the likelihood of failure since 20-30% of treatment units will indicate a BOD level greater than 10mg/L. Further refinement is recommended based on these tolerance limits. Similar conclusions can be made for AdvanTex and Ecoflo regarding BOD. (Interpretations for TSS and Fecal Coliform also follow in a similar manner.)

3. *Summary*

a) The assumption of normality is reasonable for the upper tails of the log-transformed data. This can be verified through the use of normal probability plots as shown.

b) The use of the standard error of the mean and thus confidence intervals for the mean are not appropriate when interest lies in where individual treatment units will fall. An upper 99% confidence limit of say 5.19 for effluent only states that we are 99% confident that the mean effluent level for all treatment units of a certain brand is less than 5.19 mg/L. Rather, tolerance intervals should be used for conformance monitoring.

c) Tolerance intervals computed for the VDH data set indicate that the pass/fail criterion for effluent is too low as a larger than likely acceptable percentage (from a manufacturing standpoint) of treatment units will fail the pass/fail criteria. Further investigation and refinement is recommended.

4. *References*

Hahn, G. and Meeker, W.Q. (1991). *Statistical Intervals: A Guide for Practitioners,* Wiley, New York, NY.

Montgomery, D.C. (2009). *Introduction to Statistical Quality Control,* Wiley, New York, NY.