STATISTICAL DETERMINATION OF THE OPTIMAL SAMPLE SIZE OF SECONDARY EFFLUENT BOD_5 AND SS

ERIC C. TSAI

Environment Canada, Environmental Protection Service, 804, 9942–108 Street, Edmonton, Alberta, Canada, T5K 2J5

(Received April, 1984)

Abstract. A statistical technique is presented for determining the optimal sample size required to estimate the true geometric mean with an allowable error at a desired level of confidence. Attention is focused on its application in the monitoring of secondary effluent BOD_5 and SS. It is concluded that continuous monitoring of effluent BOD_5 or SS throughout the year may generate much more data than are required for practical purposes. This statistical method may be used by operators or regulatory agencies to formulate cost-effective monitoring schemes. Records of the sample size data of secondary effluent BOD_5 and SS across Canada are also discussed.

1. Introduction

The question of sample size – i.e., the number of samples – is an important parameter to be considered in the design of an effluent quality monitoring program. A recent survey (Tables I and II) done by EPS (Environmental Protection Service) shows that in Canada, 79.8% and 68.4% of activated sludge treatment plants have a sample size of less than 40 in a year for effluent BOD_5 (biochemical oxygen demand at 5 days) and SS (suspended solids), respectively. Ideally, effluent quality variables should be monitored every day in order to obtain the true performance of a treatment plant and its total discharges of pollutants into a receiving body. However, there are some constraints, for example, manpower and money which limit the number of samples which can be obtained. The problem is, then, obtaining the required accuracy in the estimate of a parameter using a limited number of samples. The purpose of this paper is to present a statistical model for determining the optimum sample size.

2. Optimal Sample Size Model

It is possible to make probability statements concerning the behaviour of an effluent quality random variable if the form of the population distribution is known. It has been shown (Dean and Forsythe, 1976; Niku *et al.*, 1979) that the secondary effluent BOD_5 and SS data can be best described by a log-normal distribution. Standard statistical calculations may be applied to log-distributed concentrations if their values are first log-transformed. Therefore, where the log-normal distribution applies, the sample mean of the logs of the concentrations $\bar{X}_{\ln x}$, will lie between $\mu_{\ln x} - Z_{1-\alpha/2}(\sigma_{\ln x}/n^{1/2})f$ and $\mu_{\ln x} + Z_{1-\alpha/2}(\sigma_{\ln x}/n^{1/2})f$ with a risk of α. Expressing this as a pair of inequalities, one can write:

$$\mu_{\ln x} - Z_{1-\alpha/2}(\sigma_{\ln x}/\sqrt{n})f \le \bar{X}_{\ln x} \le \mu_{\ln x} + Z_{1-\alpha/2}(\sigma_{\ln x}/\sqrt{n})f \tag{1}$$

Environmental Monitoring and Assessment **5** (1985) 177–183. 0167–6369/85.15.

TABLE I

Sample size of secondary effluent BOD_5 (Canada)

Types of plants	Number of plants			
	$0<n<40$	$40\leq n<100$	$100\leq n<200$	$200\leq n\leq 366$
Conventional	54	15	9	5
Contact stabilization	21	0	0	0
Extended aeration	85	6	4	0
High rate activated sludge	18	2	3	1
Overall	79.8%	10.3%	7.2%	2.7%

n = number of daily samples in a year (Source: Mundat, EPS)

TABLE II

Sample size of secondary effluent SS (Canada)

Types of plants	Number of plants			
	$0<n<40$	$40\leq n<135$	$135\leq n<225$	$225\leq n\leq 366$
Conventional	46	14	7	18
Contact stabilization	18	1	0	0
Extended aeration	66	14	4	3
High rate activated sludge	15	3	3	1
Overall	68.4%	15.1%	6.6%	9.9%

n = number of daily samples in a year (Source: Mundat, EPS)

with 100 α% significance,
where: X_i = effluent quality variable (BOD_5 or SS), mg 1^{-1}.

$$\bar{X}_{\ln x} = \frac{1}{n}\sum_{i=1}^{i=n} \ln X_i, \qquad \text{mg } 1^{-1}.$$

$$\mu_{\ln x} = \frac{1}{N}\sum_{i=1}^{i=N} \ln X_i, \qquad \text{true mean (i.e. population mean)}$$

of the logs of the variable, mg 1^{-1}.

$\sigma_{\ln x}$ = true standard deviation of the logs of the variable, mg 1^{-1}.
n = number of daily samples collected or number of sampling days.
N = number of population = number of days over the period of interest.
f = finite population correction equal to $[(N - n)/N]^{1/2}$.
α = risk associated with a decision.
Z = standard normal variate.

The value of $Z_{1-\alpha/2}$ may be obtained from the standard normal table. For instance, if $\alpha = 5\%$, one can obtain $Z_{1-\alpha/2} = 1.96$. $\sigma_{\ln x}$ can be shown to be (Benjamin and Cornell, 1970):

$$\sigma_{\ln x} = \sqrt{\ln(V^2 + 1)} \tag{2}$$

in which $V = \mu_x/\sigma_x$ = coefficient of variation; σ_x = true standard deviation of the variable; and μ_x = true mean of the variable.

The antilog of the mean of the logs of the variable is known as the geometric mean. The accuracy of the sample geometric mean may be expressed as:

$$\text{error of sample geo. mean,}$$
$$\% = \frac{|\text{sample geo. mean} - \text{true geo. mean}|}{\text{true geometric mean}} \times 100\% \tag{3}$$

Accordingly, one can express the errors of the upper and lower limits of Equation (1), P_u and P_L, as follows:

$$P_u, \% = \frac{\exp[\mu_{\ln x} + Z_{1-\alpha/2}(\sigma_{\ln x}/\sqrt{n})f] - \exp \mu_{\ln x}}{\exp \mu_{\ln x}} \tag{4}$$

$$P_L, \% = \frac{\exp \mu_{\ln x} - \exp[\mu_{\ln x} - Z_{1-\alpha/2}(\sigma_{\ln x}/\sqrt{n})f]}{\exp \mu_{\ln x}}. \tag{5}$$

Hale (1972) also derived an equation similar to Equation (4). Wastewater treatment facilities do not provide $\sigma_{\ln x}$ in their monthly and annual reports. Therefore, it is necessary to express $\sigma_{\ln x}$ in terms of V. Substituting Equation (2) into Equations (4) and (5) yields:

$$P_u, \% = \exp[Z_{1-\alpha/2}\sqrt{\ln(V^2 + 1)}\sqrt{1/n - 1IN}] - 1 \tag{6}$$

$$P_L, \% = 1 - \exp[- Z_{1-\alpha/2}\sqrt{\ln(V^2 + 1)}\sqrt{1/n - 1/N}] \tag{7}$$

Since the log-normal distribution is skewed to the right, the distance from the upper limit to the true geometric mean is larger than that from the lower limit. Therefore, for a given sample size, P_L is less than P_u. Equation (6) can be rewritten as:

$$n = \frac{Z^2_{1-\alpha/2}\ln(V^2 + 1)}{\ln^2(1 + P_u) + \dfrac{Z^2_{1-\alpha/2}\ln(V^2 + 1)}{N}} \tag{8}$$

3. Discussion

From Equation (8), one can determine the optimum sample size required to obtain sample geometric means within a predetermined accuracy with a confidence of $100(1 - \alpha)\%$. This can be approached by answering the following questions:

(1) How much variability is presented in the population of the effluent variable?

(2) What size of risk (α) are we willing to take of incorrectly choosing n?

(3) How close do we wish to have the sample geometric mean to the true geometric mean?

If numerical values can be estimated for the above questions, then the sample size may be determined. Variability of effluent BOD_5 or SS may be measured by the coefficient of variation. The value of the coefficient of variation may be taken from past experience or estimated from other similar plants (Niku *et al.*, 1979). Lin (1974) performed statistical analyses on secondary effluent BOD_5 and SS daily data of 13 activated sludge treatment plants across Canada and the northern United States; and Niku and Schroeder (1981), 43 activated sludge treatment plants across the United States. Activated sludge process types included in Niku and Schroeder's work were complete mix, conventional, step feed, contact stabilization, plug flow, and extended aeration. It is found from Niku and Schroeder's and Lin's studies that the range of values of the coefficients of variation on a long term basis (i.e., sample sizes ranging from 200–365 in one year) for effluent BOD_5 is 0.32 to 1.27 with an average of 0.67, while for effluent SS values range from 0.32–1.70 and average 0.81.

Figure 1 shows curves of the optimal sample size calculated by using Equation (8) for $\alpha = 5\%$ and $N = 365$, and illustrates several important points. It can be seen that

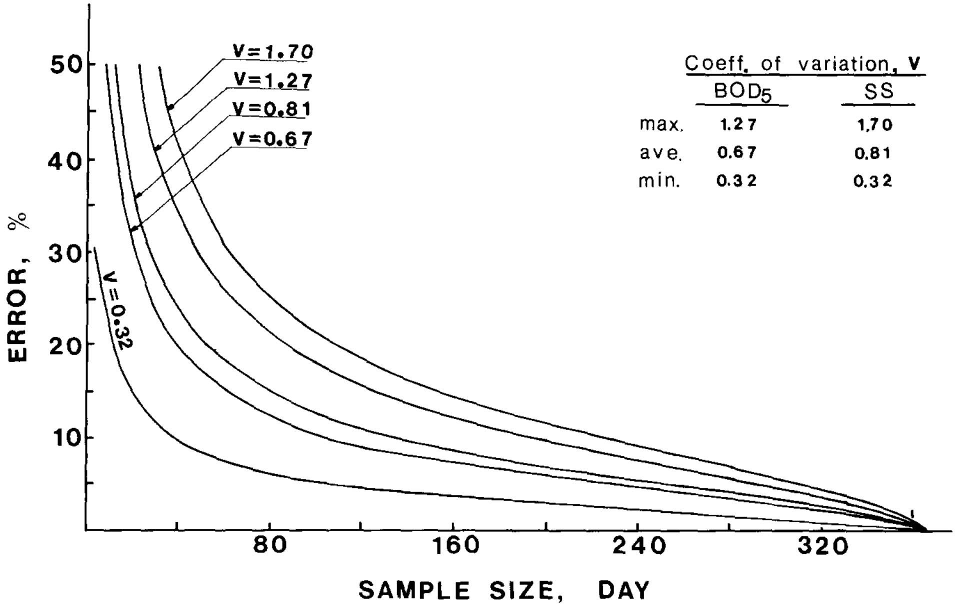

Fig. 1. Error of annual sample geometric mean vs sample size with 95% confidence.

a size too small would not give the desired accuracy, and for a given accuracy with a level of confidence, the optimal sample size increases with the coefficient of variation. For instance, to obtain an error of the annual sample geometric mean within 10 % with 95 % confidence (i.e., 95 % of the time), for $V = 0.32$, the optimal sample size is 40 in a year, and for $V = 1.27$, about 200 in a year. It may be of interest to note from Table I that in Canada, approximately 80 % of activated sludge treatment plants have a sample size of less than 40 for effluent BOD_5, but only about 3 % of the plants have a sample size of greater than 200 in a year. As mentioned earlier, 0.32 and 1.27 are the minimum and maximum values of the coefficients of variation of effluent BOD_5 observed in activated sludge treatment plants, respectively. An implication of this is that any activated sludge treatment plants with a sample size of greater than 200 in a year for effluent BOD_5 observations may achieve an error within 10 % for 95 % of the time, and that for those with a sample size of less than 40 in a year their annual sample geometric means error of the upper limit of the 95 % confidence interval will be greater than 10 %. The maximum optimal sample size of effluent SS is 225 in a year for a plant with $V = 1.70$ (Figure 1). It is clear, therefore, that continuous effluent BOD_5 or SS monitoring throughout the year may, depending on the value of V, provide much more data than are required for practical purposes.

Figure 2 depicts the relationship between the sample size and the error of the monthly sample geometric mean. It was plotted by using 30 days per month and values of the long term coefficient of variation for comparison purposes. One can see that it takes about 18 daily samples for a plant with $V = 0.32$ to achieve a desired accuracy at an error within 10 % with 95 % confidence; with $V = 0.67$, 25 daily samples; with $V = 1.27$, 28 daily samples. This indicates that even for a plant that has most stable performance

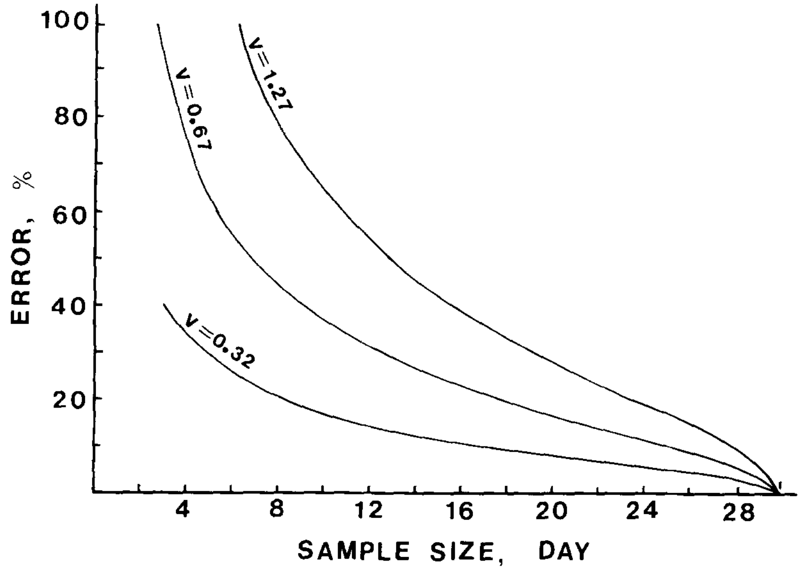

Fig. 2. Error of monthly sample geometric mean vs sample size with 95 % confidence.

(i.e., $V = 0.32$), it almost requires a continuous sampling program during weekdays if the desired accuracy of the monthly geometric mean is set at an error within 10% at a 95% level of confidence. Thus, any attempt to obtain short term averages of effluent BOD_5 or SS with reasonable accuracy may be costly.

The aforementioned discussion is based on the log-normal distribution of the random variable X. If the random variable is not log-normally distributed, practical implications of the central limit theorem provide for continued use of Equations (1) and (8). The central limit theorem states that the distribution of the sample mean approaches the normal distribution as the sample size n increases, even when the parent population is not normal. It would be plausible to assume the normality for the distribution of $\bar{X}_{\ln x}$ if n is greater than 4 (Bendat and Piersol, 1971).

4. Summary and Conclusion

On the basis of the assumption of random and independent observations from a log-normal population, a statistical model for determining the optimal sample size of secondary effluent BOD_5 and SS has been presented. This model is also applicable to a non-lognormally distributed effluent variable if its sample size is large. The model can be used by a plant operator or a regulatory agency to formulate an effluent monitoring plan for an individual municipal or industrial wastewater treatment facility, based on actual historical operation data.

To achieve an annual sample geometric mean error within 10% at a 95% confidence level, the optimal sample sizes for activated sludge effluent BOD_5 have been found to range, depending on the coefficient of variation, from 40–200 in a year; and for effluent SS, from 40–225 in a year. In Canada, most activated sludge treatment plants have a sample size of less than 40 in a year suggesting their annual true geometric means have not been accurately estimated.

Acknowledgement

The cooperation of the Inventory Management Division, EPS, Ottawa, for retrieving sample size data from MUNDAT, is gratefully acknowledged. The opinions expressed in this paper are those of the author alone; they should not be construed as representing the policy of the Environmental Protection Service.

References

Bendat, J. S. and Piersol, A. G.: 1971, *'Random Data: Analysis and Measurement Procedures'*, Wiley-Interscience, New York, pp. 110–111.

Benjamin, J. R. and Cornell, A.: 1970, *'Probability Statistics and Decision for Civil Engineers'*, McGraw-Hill, New York, pp. 266–267.

Dean, R. B. and Forsythe, S. L.: 1976, 'Estimating the Reliability of Advanced Waste Treatment, Part 2', *Water and Sew. Work*, July, 57–60.

Hale, W. E.: 1972, 'Sample Size Determination for the Log-normal Distribution', *Atmospheric Environment* **6**, 419–422.

Lin, K. C.: 1974, 'Significance of Temperature in the Activated Sludge Process', Ph.D. Dissertation, University of Toronto, Toronto, Canada.

Niku, S., Schroeder, E. D ., and Samaniego, F. J.: 1979, 'Performance of Activated Sludge Processes and Reliability-Based Design', *J. Water Pollut. Control Fed.* **53**, 2841–2857.

Niku, S. and Schroeder, E. D.: 1981, 'Stability of Activated Sludge Processes Based on Statistical Measures', *J. Water Pollut. Control Fed.* **53**, 457–470.