RRB-87-38

INFERENCES ABOUT THE MEAN FROM CENSORED WATER QUALITY DATA

by
A.H. El-Shaarawi

National Water Research Institute
Canada Centre for Inland Waters
Burlington, Ontario, L7R 4A6

August 1987
NWRI Contribution #87-141

MANAGEMENT PERSPECTIVE

Recently, scientists as well as managers have shown an increasing interest in estimating the levels of race contaminants in different parts of the environment (i.e. water, sediment, food, air). A major difficulty encountered in achieving this is that a substantial portion of sample concentrations of many toxic contaminants is below the limits of detection established by analytical laboratories. Several ad hoc methods were used and reported in the literature for dealing with this problem. It is shown in this paper that all the reported methods are technically inadmissible for estimating the unobserved water quality censored data. Furthermore, the paper presents the correct and natural approach for dealing with this problem. The new approach does not only provide an estimate for the contaminant level but also allows the construction of confidence bounds for the level.

PERSPECTIVE-GESTION

Depuis quelque temps, les chercheurs aussi bien que les gestionnaires s'intéressent de plus en plus à l'estimation des niveaux de contaminants à l'état de trace dans différents milieux (l'eau, les sédiments, la nourriture, l'air, etc.). L'un des principaux obstacles auxquels ils sont confrontés est lié au fait qu'une grande partie des concentrations de nombreux contaminants toxiques dans les échantillons sont inférieures aux seuils de détection établis par les laboratoires d'analyse. Plusieurs méthodes spéciales, ayant fait l'objet de rapports, ont été utilisées pour tenter de résoudre ce problème. Le présent document démontre que toutes les méthodes signalées sont inadmissibles sur le plan technique pour l'estimation de données tronquées non observées ayant trait à la qualité de l'eau. Le document expose également la bonne façon d'aborder le problème. La nouvelle méthode présentée permet non seulement d'estimer le niveau de contaminants, mais également de définir les bornes d'un intervalle de confiance pour ce niveau.

RÉSUMÉ

Plusieurs méthodes couramment utilisées pour calculer par déduction les niveaux de nombreux métaux et contaminants organiques en milieu aquatique à partir de données tronquées de type 1 font l'objet d'analyses critiques et leurs applications sont illustrées au moyen des concentrations de BPC dans des échantillons d'eau provenant de la rivière Niagara. Toutes ces méthodes, à l'exception de la méthode du maximum de vraisemblance, soulèvent différents problèmes, dont la présence d'estimations inadmissibles pour les valeurs tronquées et l'absence d'erreurs-types pour ces estimations. En supposant que la distribution des données était lognormale, on a modifié la méthode de régression logarithmique pour obtenir des estimations admissibles des valeurs tronquées et calculer les propriétés des estimations du maximum de vraisemblance pour la moyenne lognormale. On a également obtenu un intervalle approximatif de confiance pour la moyenne lognormale et démontré comment il pouvait être utilisé pour estimer la charge qui entre dans la rivière Niagara et qui en sort.

INFERENCES ABOUT THE MEAN FROM CENSORED WATER QUALITY DATA

A.H. El-Shaarawi

National Water Research Institute

Canada Centre for Inland Waters

Burlington, Ontario, L7R 4A6

ABSTRACT

Several methods which are commonly used for making inferences about the levels of many metals and organic contaminants in ambient waters from type I censored data are critically evaluated and their applications are illustrated using the concentrations of PCB in water samples from the Niagara River. Difficulties encountered in the application of all the methods, except the method of maximum likelihood, include the occurrence of inadmissible estimates for the censored values and the unavailability of the standard errors for these estimates. Under the assumption that the distribution of the data is lognormal, the log regression method is modified to produce admissible estimates of the censored values and the properties of the maximum likelihood estimates for the lognormal mean are derived. Furthermore, an approximate confidence interval for the lognormal mean is given and its use for estimating the load to and from the Niagara River is illustrated.

INTRODUCTION

For more than a decade the Governments of Canada and United States have been concerned about the occurrences and the levels of toxic pollutants in the Niagara River and Lake Ontario Basins. These pollutants originate from hazardous waste disposal sites as well as industrial and municipal discharges. This led to establishing in 1981 the Niagara River Toxics Committee (NRTC) with representatives from the two countries which resulted in the publication of the 1984 NRTC report. The Binational Data Interpretation Group for the Niagara River produced another report in 1986. Those two reports and that of Allan et al. (1983) and El-Shaarawi et al. (1985) summarize most of the available information on toxic pollutants in the river.

A recurring difficulty encountered during the course of these reports is that a substantial portion of water sample concentrations of many toxic pollutants is below the limits of detection established by analytical laboratories. The Data Interpretation Group dealt with this difficulty by using the log-probability regression (LR) method. Extensive simulation by Gilliam and Helsel (1986) and Helsel (1986) indicate that the LR method is superior to most commonly used methods in estimating the distributional parameters from censored data.

It is shown that the LR method is technically inadmissible for estimating the unobserved water quality censored data which are of type I censoring. The LR method is used as a short-cut method (David, 1981) to estimate the parameters of the normal distribution from type II censored data.

Furthermore, the standard errors of the estimated parameters using the LR method are not available. A modification of the LR method is given which results in producing admissible results. The method of maximum likelihood (ML) for dealing with normal data is given. Prior to the use of the likelihood method, the normality of the data should be investigated to determine if a transformation is required. Probability plotting such as Q-Q plot (David, 1981) will be useful in this regard. Because the lognormal distribution is commonly used as a model for environmental data, the estimate of the lognormal mean and its standard error are derived. Furthermore, a method (Land, 1972) due to Cox for constructing a confidence interval for the mean of the lognormal distribution is extended to the case of type I censored data. Data from the Niagara River are used to illustrate the methods of this paper.

ESTIMATION METHODS

1. Likelihood Inferences

a. The Normal Case

Let n_0 be the number of water samples with concentrations below the detection limit, X_0 and let $X_1, \ldots, X_n$ be the measured concentrations. Under the assumption that the observations are independent and have a common normal distribution with mean μ and variance σ^2. The likelihood function L for μ and σ is

$$L = \frac{N!}{n_0!\, n!} F^{n_0}(\xi_0)\, f(X_1), \ldots, f(X_n) \qquad (1)$$

where $N = n_0 + n$, $\xi_0 = (X_0 - \mu) / \sigma$,

$$f(X_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(X_i - \mu)^2/2\sigma^2} \qquad (i = 1, 2, \ldots, n)$$

and $$F(\xi_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi_0} e^{-t^2/2} dt$$

The maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ satisfy the equations

$$\bar{X} - \mu = \sigma Y \qquad (2)$$

and $$S^2 + (\bar{X} - \mu)^2 = \sigma^2 [1 + \xi_0 Y] \ , \qquad (3)$$

where $\bar{X} = \sum_1^n X_i/n$, $S^2 = \sum_1^n (X_i - \bar{X})^2/n$

$$Y = \frac{h}{1-h} Z(-\xi_0) \ , \qquad Z(\xi_0) = \frac{e^{-\xi_0^2/2}}{\sqrt{2\pi}\,(1 - F(\xi_0))} \ ,$$

and $h = n_0 / (n_0 + n)$

The large sample variance covariance matrix for $\hat{\mu}$ and $\hat{\sigma}$ is given by

$$V = \frac{\sigma^2}{(n_0+n)(1-F(\xi_0))(\zeta_{11}\zeta_{22}-\zeta_{12}^2)} \begin{bmatrix} \zeta_{22} & -\zeta_{12} \\ -\zeta_{12} & \zeta_{11} \end{bmatrix}$$

$$= \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \qquad (4)$$

where $\zeta_{11} = 1 + Z(\xi_0)\{Z(-\xi_0) + \xi_0\}$,

$\zeta_{12} = Z(\xi_0)\{1 + \xi_0 (Z(-\xi_0) + \xi_0)\}$

and $\zeta_{22} = 2 + \xi \zeta_{12}$.

The above expressions which were derived by Cohen (1959) can be used to construct confidence intervals for μ and σ. The relative likelihood function $R = L / \max_{\mu,\sigma} L$. Contains all the information in the data about μ and σ and its shape justify the use of Cohen's method for a particular data set. One way of examining its shape is to construct the relative likelihood contours. A contour consists of the values of μ and σ which satisfy the equation: $R = C$. Also joint confidence intervals for μ and σ can be shown on the graph using the fact that $-2\ln R$ has approximately a chi-square distribution with two degrees of freedom.

b. The Lognormal Case

The lognormal distribution is used as a model in many environmental applications. Examples can be found in Aitchison and Brown (1981), Helsel (1987), Gilliom and Helsel (1986), El-Shaarawi

and Murthy (1976) and Esterby and El-Shaarawi (1984). Let X_i be a normal distribution with mean μ and variance σ^2, then $Y_i = \exp X_i$ has a lognormal distribution with mean α and variance β^2, where

$$\alpha = \text{Exp}\ (\mu + \sigma^2/2) \quad ,$$

$$\text{and} \quad \beta = \alpha^2\ (\text{Exp}\ \sigma^2 - 1) \quad .$$

To estimate α and β^2 and their standard errors the following are required. Let $\hat{Q}(a,b) = a\hat{\mu} + b\hat{\sigma}^2$ and $Q(a,b) = a\mu + b\sigma^2$ then it is easy to show that the mean and the variance of $\text{Exp}\ \hat{Q}(a,b)$ are

$$\eta(a,b) = \text{Exp}\ \{Q(a,b) + (a^2V_{11} + 4ab\sigma V_{12} + 4b^2\sigma^2V_{22})\ /2\ \}$$

$$D^2(a,b) = \eta(2a,2b) - \eta^2(a,b) \ . \qquad (5)$$

The first of the above expressions shows that $\text{Exp}(a\hat{\mu} + b\hat{\sigma}^2)$ overestimates $\text{Exp}(a\mu + b\sigma^2)$. The degree of bias can be measured by the ratio.

$$B = \eta(a,b) \ / \ \text{Exp}\ \{Q(a,b)\}$$

$$= \text{Exp}\ (a^2\ V_{11} + 4ab\ \sigma\ V_{12} + 4\ b^2\ \sigma^2\ V_{22})\ /2 \qquad (6)$$

The value of B can be estimated by replacing μ and σ in (6) by $\hat{\mu}$ and $\hat{\sigma}$. Taking a=1 and b=0.5 in (5) and (6) yield the mean, the variance and the bias that result from the use of the estimate $\hat{\alpha} = \text{Exp}\ (\hat{\mu}+\hat{\sigma}^2/2)$

for the lognormal mean α. In the absence of censoring we have $V_{11}=\sigma^2/n$, $V_{12}=0$ and $V_{22}=2\,\sigma^4/(n-1)$. Hence

$$\eta(1,1/2) = \text{Exp}\{\mu + \frac{\sigma^2}{2} + \frac{\sigma^2}{2n} + \frac{\sigma^6}{n-1}\}$$

$$D^2(1,1/2) = \eta^2(1,\ 1/2)\ \{\text{Exp}\ \{\frac{\sigma^2}{n} + \frac{2\sigma^6}{n-1}\} - 1$$

and $$B = \text{Exp}\ \{\frac{\sigma^2}{2n} + \frac{\sigma^6}{n-1}\}$$

which shows that $\hat{\alpha}$ is asymptotically unbiased and a consistent estimate for α. A correction for the bias leads to estimating α by $\hat{\alpha}=\hat{\alpha}/\hat{B}$, where $\hat{B}$ is the value of B when μ and σ are replaced by $\hat{\mu}$ and $\hat{\sigma}$.

c. Approximate Confidence Interval for the Lognormal Mean

This is an extension of the confidence interval for the lognormal mean due to Cox and presented in Land (1972) to the case of censored data. It is based on the fact that the asymptotic distribution of $\hat{\mu} + 1/2\ \hat{\sigma}^2$ is normal with mean $\mu + 1/2\ \sigma^2$ and variance $v_{11} + 2V_{12}\sigma + \sigma^2V_{22}$. This leads to the approximate confidence limits for α as

$$\hat{\alpha}\ \text{Exp}\ \{-Z_{1-\alpha/2}\ \sqrt{V_{11}+2V_{12}\hat{\sigma}+V_{22}\hat{\sigma}^2}\}\leq\alpha\leq\hat{\alpha}\ \text{Exp}\{Z_{1-\alpha/2}\ \sqrt{V_{11}+2V_{12}\sigma+V_{22}\sigma^2}\} \tag{10}$$

One basic advantage of this interval is the fact that it always produces positive values for the limits and hence it is admissible.

Conditional on the observed flow data and under the assumption that the flow and the concentration data are independent, the above

confidence interval can be used to obtain confidence limits for the mean loading. This is done by multiplying the limits given in (10) by the mean of the flow rate during the period of data collection.

2. The Log-Probability (LR) Method

This method was first applied to environmental data by Hashimoto and Trussell (1983) and its performance along with several other methods was evaluated by Gilliom and Helsel (1986) and Helsel (1987). The results of this work indicate that the method is robust for estimating the mean and the variance under a wide range of underlying distributions.

The steps followed in estimating μ and σ are as follows: (1) fit the regression line of $X_{(n_o+i)}$ on the normal scores Z_{n_o+i} for $i=1,2,\ldots,n$, where $X_{(n_o+i)}$ is the ith largest observation among $X_1,\ldots,X_n$ and $Z_{n_o+i} = F^{-1}((n_o+i)/(n_o+n+1))$ where F^{-1} is the inverse normal distribution function; (2) use the intercept and the slope of the regression line to provide initial estimates for μ and σ, respectively; (3) estimate the censored values $X_{(1)},\ldots,X_{(n_o)}$ by $\hat{X}_{(1)},\ldots,\hat{X}_{(n_o)}$ from the regression equation; and finally (4) assuming the lognormal model for the data, the LR method estimates the mean and variance as

$$m_1 = \left\{ \sum_{i=1}^{n_o} e^{\hat{X}_{(i)}} + \sum_{i=1}^{n} e^{X_{(n_o+i)}} \right\} / (n_o + n) \qquad (10)$$

$$m_2 = \left[\sum_{i=1}^{n_o} e^{2\hat{X}_{(i)}} + \sum_{i=1}^{n} e^{2X_{(n_o+i)}} - (n_o + n)\, m_1^2 \right] / (n_o+n-1) . \qquad (11)$$

3. Modifications of the LR Method

Information in the detection limit X_0 is ignored in the application of the LR method. As a result some of the estimated censored values $X_{(1)},\ldots,X_{(n_0)}$ may fall between X_0 and $X_{(n_0+1)}$, which is inadmissible. This can be avoided by adding the values X_0 and Z_{n_0} to the data used for fitting the regression line. This is reasonable since n_0/n_0+n estimates $F(\xi_0)$. Another modification can be made by replacing the normal score $Z_{(i)}$ by the exact expected value $W_{(i)}$ of the ith standard ordered normal deviates which are available from the Biometrics Tables for Statisticians (Pearson and Hartley, 1976) for up to a sample of size 100. In the tables the variance-covariance matrix of the order statistics (available for up to a sample of size 20) could be used for performing weighted regression analysis.

APPLICATIONS

The data used to illustrate the methods of this paper represent the concentrations (ng/L) of PCB in water samples from the Niagara River which were collected by Environment Canada, at Fort Erie (FE) and Niagara-on-the-Lake (NOTL) every two weeks between 12/12/1984 and 19/3/1986. Figure 2 shows the time plot of the data on the log scale. There are 34 observations with the numbers of censored values at 9 and 7 for NOTL and FE respectively. Figure 3 shows the

probability plot (Q-Q) for both the raw and the log transformed data. The graphs indicate that the normal distribution is a better model for fitting the logs of the data and show that two of the concentration values from the NOTL water samples are very different from the rest of the data. The Q-Q plot of the logs of 32 observations (i.e. after eliminating the largest two values) from NOTL is much closer to the normality.

Figure 4 shows the contours of the relative likelihood functions and the joint 95% large sample confidence intervals for μ and σ using the NOTL and FE data. From the shape of these contours it seems that the likelihood surfaces can be reasonably approximated by the bivariate normal distribution and hence the asymptotic theory provides a good approximation for making inferences about the paramaters.

Table 1 presents the results of the analysis in the log-space. Specifically, the estimates of μ and σ using the method of maximum likelihood and the LR method along with their associated 95% confidence intervals are given in the table. The results show that there can be substantial differences in the point estimates between the two methods. The maximum likelihood estimates, the maximum likelihood corrected for bias estimates and the LR method estimates for the lognormal mean are given in Table 2 along with their estimated standard errors for the maximum likelihood estimates. Also the approximate confidence intervals for the median and for the lognormal mean are also given in the Table. From the results it appears that there are substantial differences between the estimates of the mean

using the maximum likelihood method and LR method and the bias in the estimate of the lognormal mean can be substantial.

CONCLUSIONS

This paper recommends the use of the method of maximum likelihood for making inferences about the mean of the lognormal distribution from censored data. The maximum likelihood estimate is biased but the bias can be corrected as shown in the paper. Moreover, the likelihood function method provides a natural way for obtaining an approximate confidence interval for the mean and for the median of the lognormal distribution. On the other hand, the LR method can lead to inadmissible estimates for the censored values of the concentration which are below the level of detection and expressions for the standard errors of these LR estimates are not available.

REFERENCES

Aitchison, J. and J.A.C. Brown. The Lognormal Distribution. 176 pp., Cambridge University Press, New York, 1957.

Allan, R.J., A. Mudroch and M. Munawar (ed). The Niagara River-Lake — Ontario Pollution Problem. J. Great Lakes Res. 9, 232 pp., 1983.

Cohen, A.C., Jr. Simplified Estimators for the Normal Distribution When Samples are Singly Censored or Truncated. Technometrics, 1(3), 217-237, 1959.

Davis, H.A. Order Statistics. 2nd Ed., 360 pp., John Wiley, New York, 1981.

El-Shaarawi, A.H., S.R. Esterby, N.D. Warry and K.W. Kuntz. Evidence of Contaminant Loading to Lake Ontario from the Niagara River. Can. J. Fish. Aquat. Sci., 42, 7, 1985.

El-Shaarawi, A.H. and C.R. Murthy. Probability Distribution of Concentrations Measured in the Wake of a Continuous Point Source in Coastal Currents. J. of Phys. Oceanography, 6, 5, 1976.

Esterby, S.R. and A.H. El-Shaarawi. Coliform Concentrations in Lake Erie - 1966 to 1970. Hydrobiologia, 111, 1984.

Gilliom, R.J. and Dennis R. Helsel. Estimation of Distributional Parameters for Censored Trace Level Water Quality Data. 1. Estimation Techniques. Water Resources Research, 22, 2, 1986.

Hashimoto, L.K. and R.R. Trussell. Evaluating Water Quality Data Near the Detection Limit. Paper presented at the Proceedings of the American Water Works Assoc. Advanced Technology Conference, Am. Water Works Assoc., Las Vegas, Nev., June 5-9, 1983.

Helsel, D.R. Estimation of Distributional Parameter for Censored Water Quality Data. Developments in Water Science (ed. A.H. El-Shaarawi and R.E. Kwiatkowski), V. 27, Elsevier, 1986.

Land, C.E. An Evaluation of Approximate Confidence Interval Estimation Methods for Lognormal Means. Technometrics, 14, 1972.

Pearson, E.S. and H.O. Hartley. 1976. Biometrika tables for statisticians. 3rd Edition. Charles Griffin and Company Ltd., Vol. 1.

Niagara River Toxics Committee. A Joint Publication of New York State Department of Environmental Conservation, Environment Canada, U.S. Environmental Protection Agency and Ontario Ministry of the Environment, October 1984.

Joint Evaluation of Upstream/Downstream Niagara River Monitoring Data 1984-1986. Prepared by Data Interpretation Group. A Joint Publication of New York State Department of Environmental Conservation, Environment Canada, U.S. Environmental Protection Agency and Ontario Minstry of Environment, October 1986.

Table 1. Summary of the statistical analyses in the log space.

Station	n_o	n	Method of Maximum Likelihood								
			Estimates		Variance and Covariances			95% Confidence Intervals for		LR Method	
			μ	σ	$V(\mu)$	$V(\sigma)$	$C(\mu,\sigma)$	μ	σ	μ	σ
FE	7	27	0.2438	1.5603	0.0780	0.0483	-0.0070	(-0.2950, 0.7827)	(1.1564,1.9659)	0.2882	1.6243
NOTL	9	25	0.3307	2.1632	0.1591	0.1030	-0.0207	(-0.4316, 1.0898)	(1.5346,2.7928)	0.4431	2.1357
NOTL*	9	23	0.1121	1.6380	0.0977	0.0600	-0.0106	(-0.4784, 0.6992)	(1.1581,2.1185)	0.3614	1.4055

*The two largest observations are not included in the analysis.

Table 2. Summary of the analysis in the measurement space.

	Method of Maximum Likelihood						LR Method	
Station	$\hat{\alpha}$	$\sigma(\hat{\alpha})$	$\hat{\alpha}_u$	Median	95% Confidence Interval for the Mean	95% Confidence Interval for the Median	$\tilde{\alpha}$	$\tilde{B}$
FE	4.3106	2.2741	3.9725	1.2762	(1.9521, 9.5187)	(0.7445,2.1874)	3.4291	5.0386
NOTL	14.4456	21.9674	11.1241	1.3897	(3.5027,59.5746)	(0.6495,2.9737)	16.2696	51.2991
NOTL*	4.2783	2.7102	3.8500	1.1167	(1.7391,10.5251)	(0.6198,2.0121)	3.1884	5.2342

*The two largest observations are not included in the analysis.

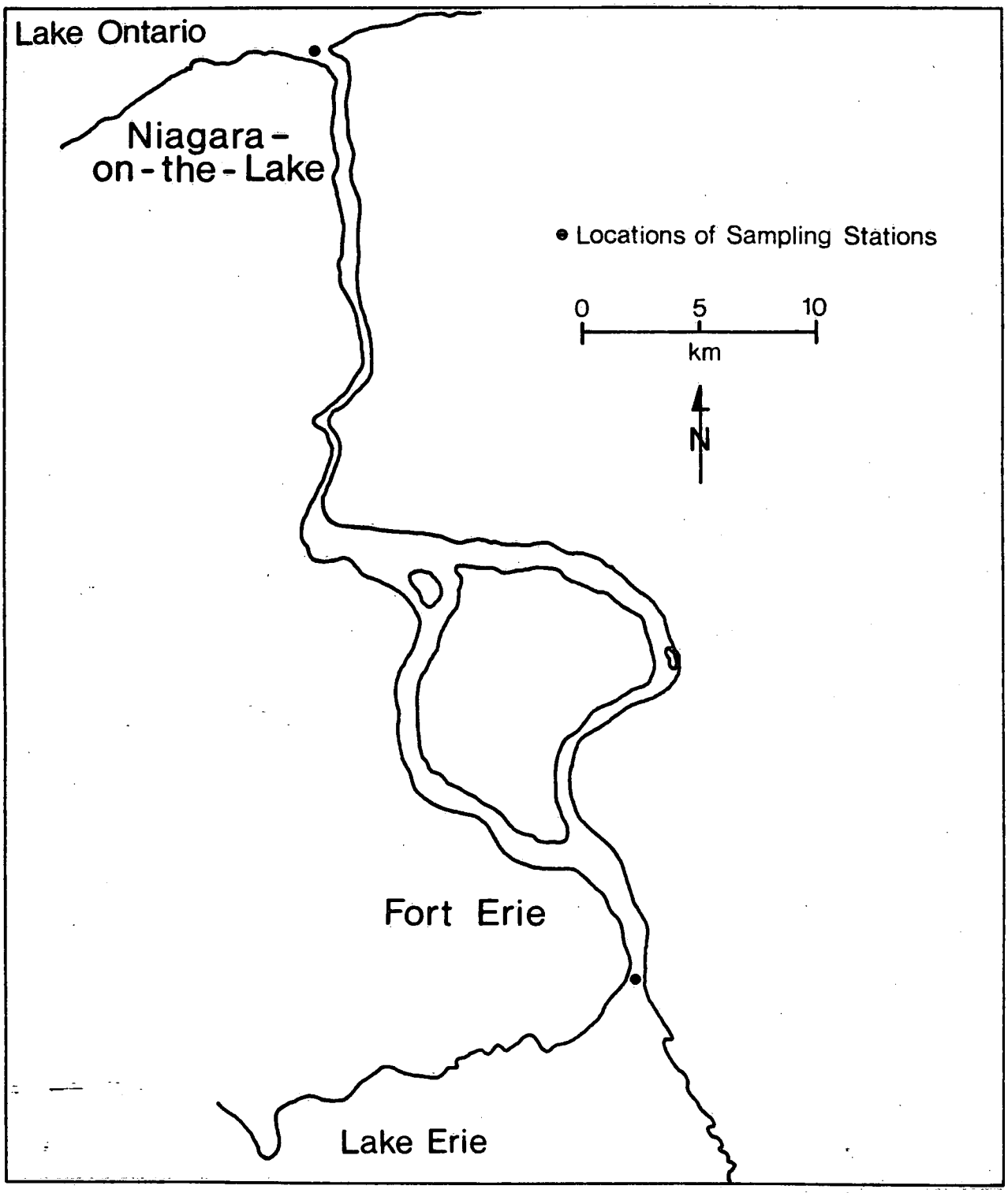


Figure 1 Location of sampling stations on the Niagara River.

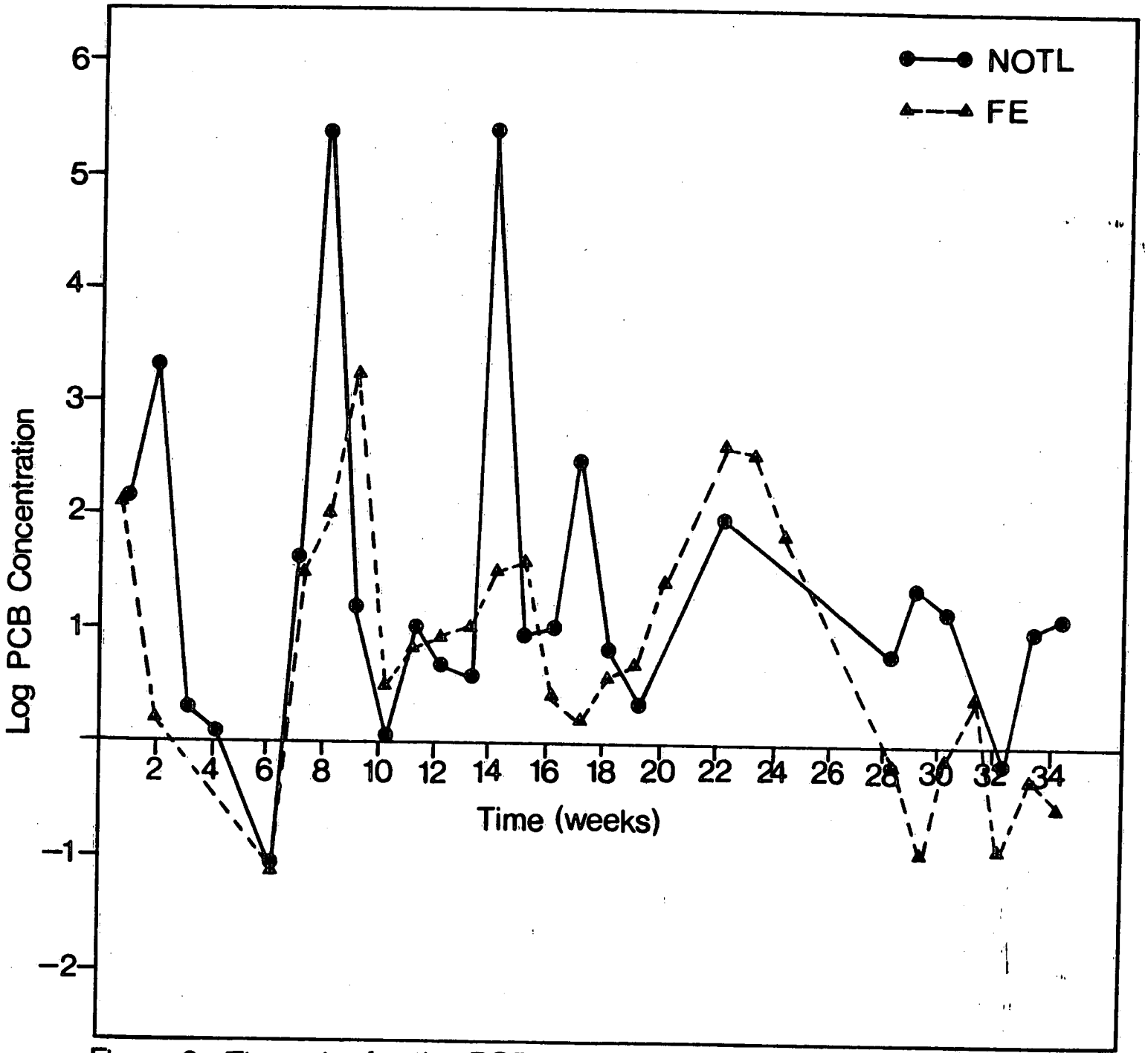
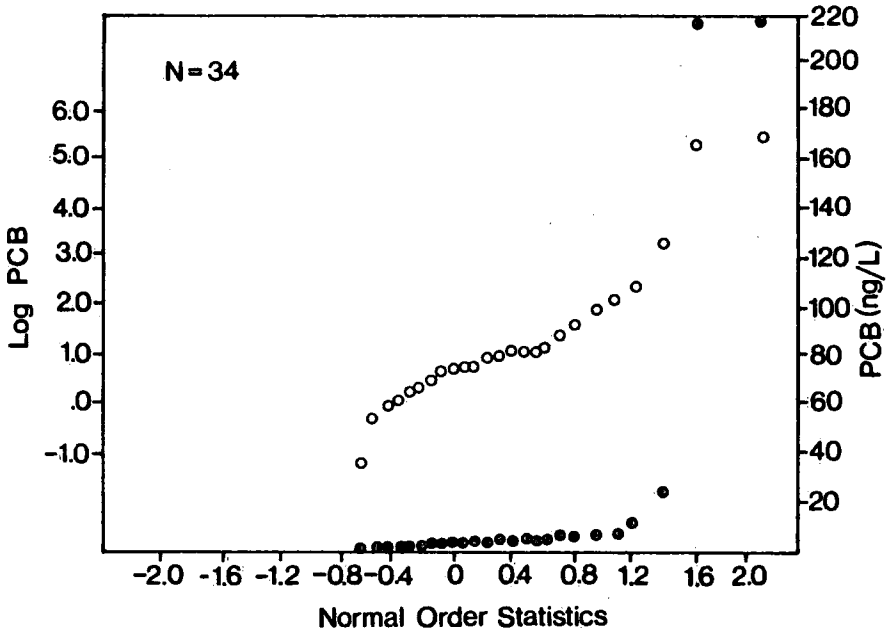
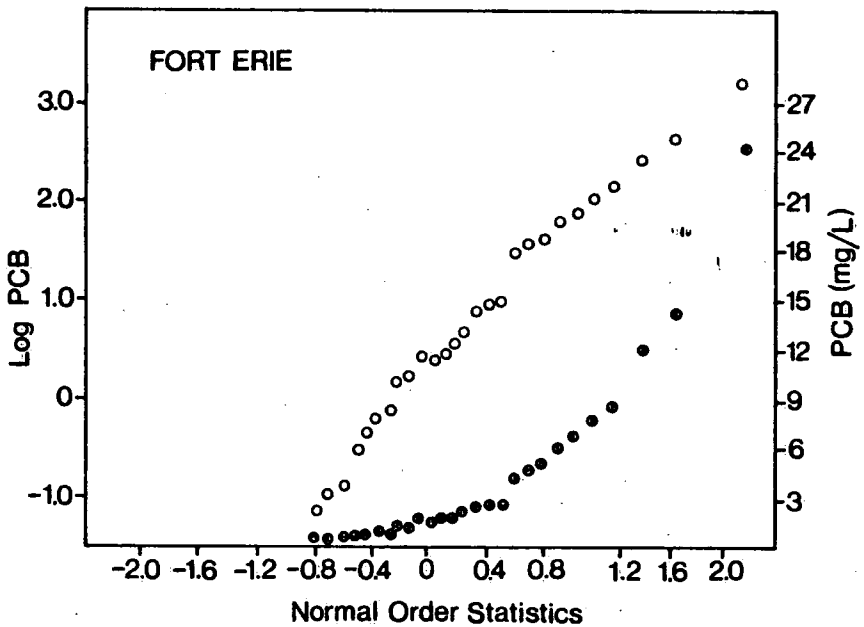


Figure 2. Time plot for the PCB data





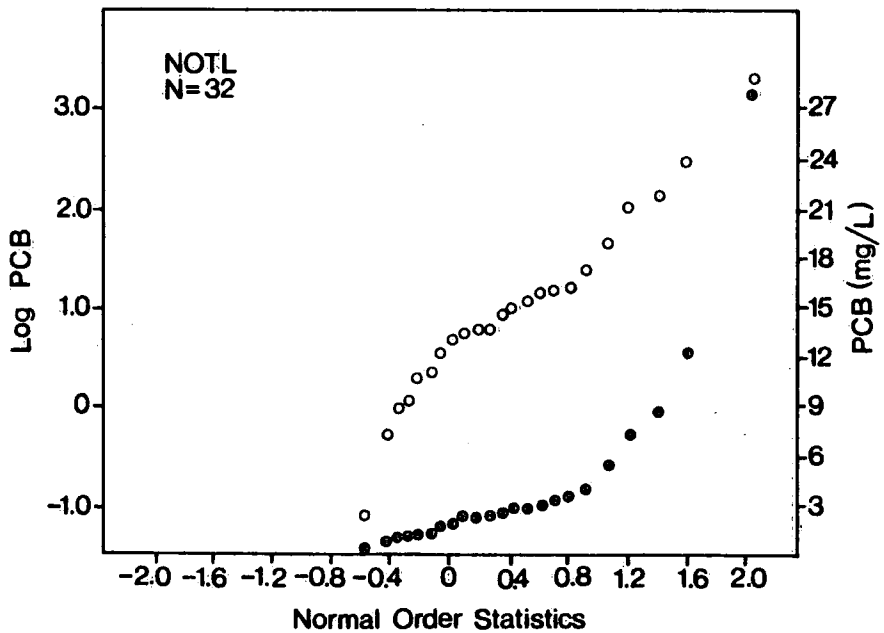


Fig. 3. Q-Q Plots for the PCB data.

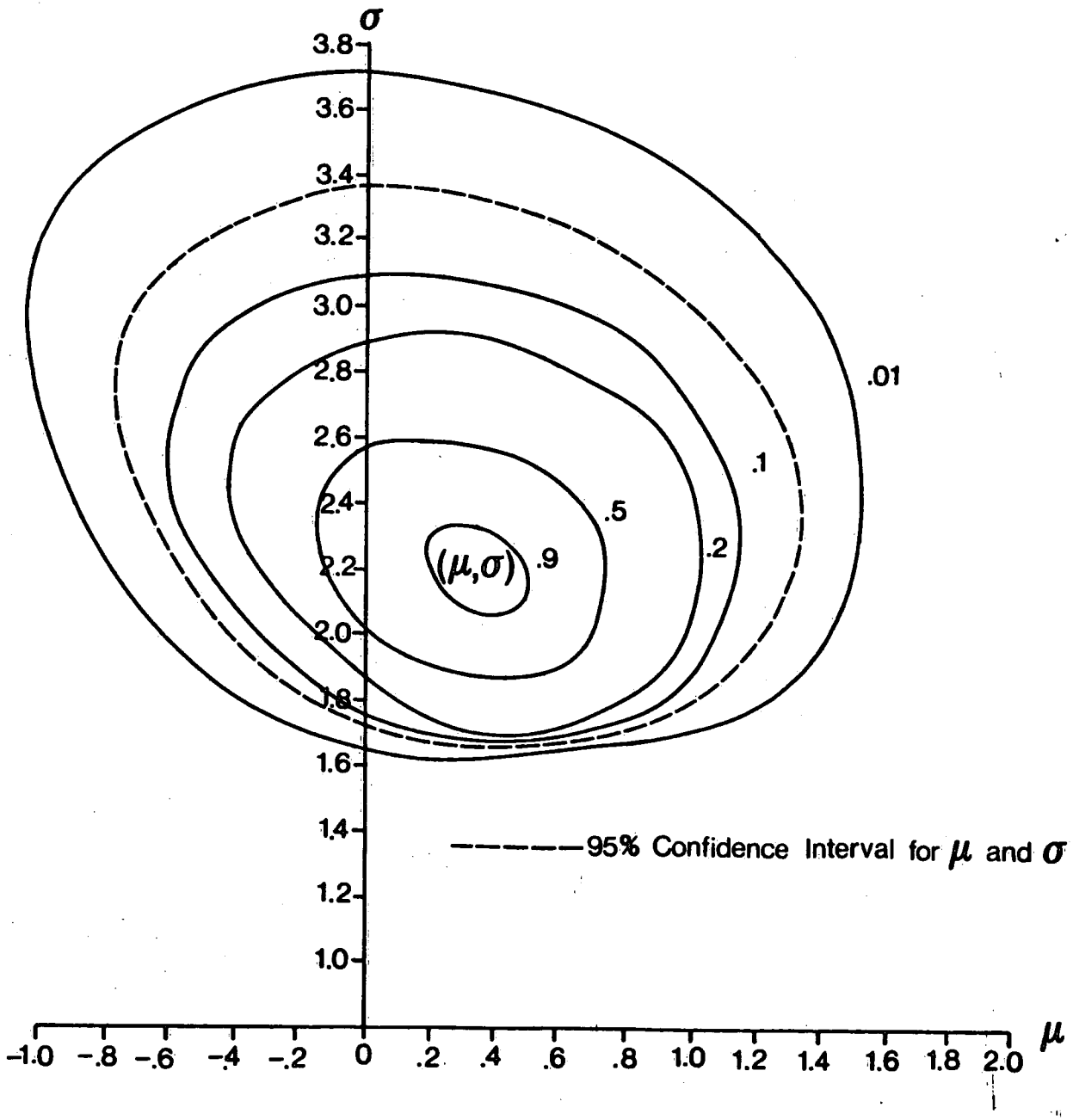


Fig. 4a. Relative Likelihood Contours for μ and σ (Niagara On The Lake)

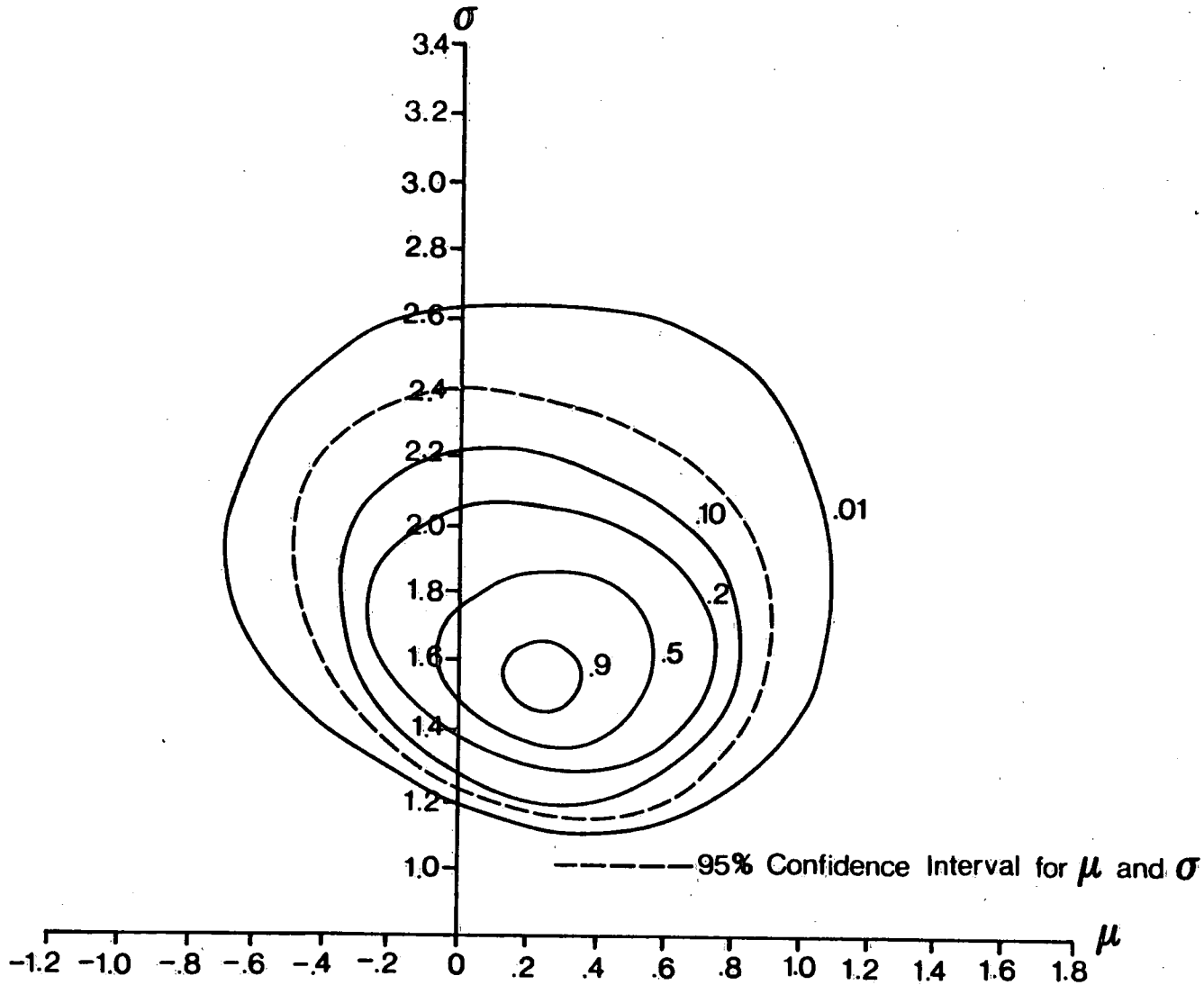


Fig. 4b. Relative Likelihood Contours for μ and σ (Fort Erie)