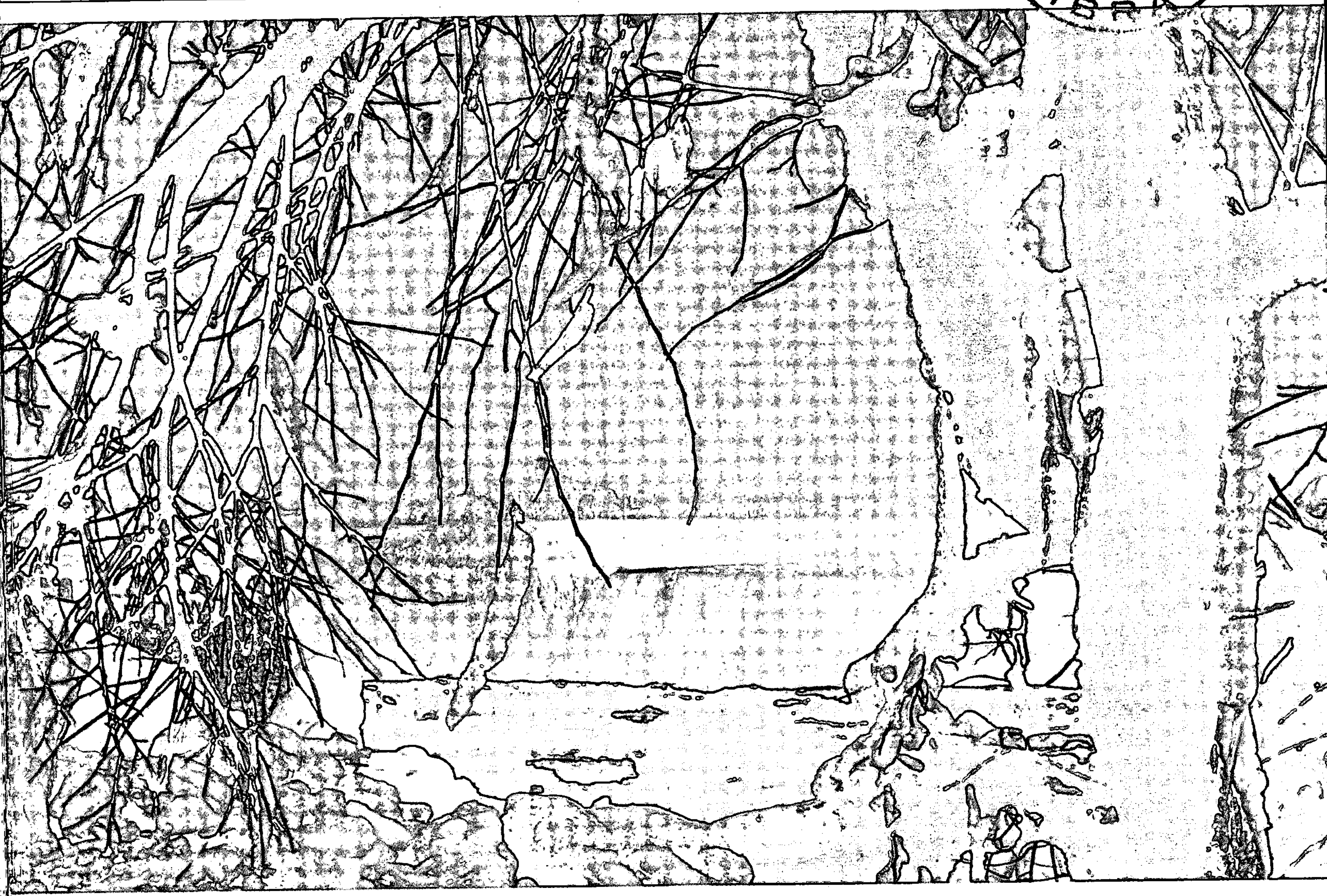
CANADA. INLAND WATERS DIRECTORATE
SCIENTIFIC SERIES
85

Environment Canada — Environnement Canada

The Estimation of Sample Size Required in Chemical Limnology and Autecology of Shelled Invertebrates

L.D. Delorme and A.H. El-Shaarawi



SCIENTIFIC SERIES NO. 85
(Résumé en français)

*INLAND WATERS DIRECTORATE,
NATIONAL WATER RESEARCH INSTITUTE,
CANADA CENTRE FOR INLAND WATERS,
BURLINGTON, ONTARIO, 1978.*

Cat. No. En 36-502/85

ISBN 0-662-10080-8

Contract No. KL229-7-1040

THORN PRESS LIMITED

Environment Canada   Environnement Canada

The Estimation of Sample Size Required in Chemical Limnology and Autecology of Shelled Invertebrates

L.D. Delorme and A.H. El-Shaarawi

SCIENTIFIC SERIES NO. 85
(Résumé en français)

INLAND WATERS DIRECTORATE,
NATIONAL WATER RESEARCH INSTITUTE,
CANADA CENTRE FOR INLAND WATERS,
BURLINGTON,ONTARIO, 1978.

Contents

Tables

Illustrations

Abstract

Calculation of sample size is based on the central limit theorem for a normal distribution with mean zero, variance unity and the tail value equal to α. The confidence coefficient is 1 - 2α. In the field of chemical limnology and autecology of shelled invertebrates, most chemical parameters must be transformed to obtain a normal distribution. The exception is pH, which is already transformed. Some parameters, such as carbonate and carbon dioxide, have an exponential distribution because of the carbonate-bicarbonate fence at pH 8.3. The sample size is a function of the parameter measured, primarily variance, mean and range.

Résumé

Le calcul de la taille des échantillons est fondé sur le théorème central limite pour une distribution normale de moyenne 0, de variance 1 et de valeur de la queue égale à α. Le coefficient de confiance est 1 - 2α. Dans le domaine de la limnologie chimique et de l'autécologie des invertébrés à coquilles, la plupart des paramètres chimiques doivent être transformés afin d'obtenir une distribution normale, sauf le pH qui l'est déjà. Certains paramètres, comme la teneur en carbonates et en dioxyde de carbone, ont une distribution exponentielle, car à pH de 8.3, il y a séparation entre les formes carbonate et bicarbonate. La taille des échantillons est fonction du paramètre mesuré, surtout de la variance, de la moyenne et de la dispersion.

The Estimation of Sample Size Required in Chemical Limnology and Autecology of Shelled Invertebrates

L.D. Delorme and A.H. El-Shaarawi

INTRODUCTION

In the planning of a sample survey, a stage is always reached at which a decision must be made about the size of the sample. The decision cannot be made without determining what is expected of the sample. For example, the sample size "n" may be required for estimating the mean value "μ" of a population with probability density function, pdf, f(x) with a predetermined precision. It is well known (when n is large) from the central limit theory that

$$z = \sqrt{n}\,(\bar{x} - \mu) / \sigma \tag{1}$$

has a normal distribution with mean 0 and variance unity (if the variance "σ^2" of f(x) is finite). The confidence interval for μ based on the normal distribution "z" is

$$\text{Prob}\{z^2 \le z_\alpha^2\} = \text{Prob}\{n \le z_\alpha^2\,\sigma^2/d^2\} = 1 - 2\alpha \tag{2}$$

where z_α is the value on the x axis for which the tail value of the normal distribution is equal to α. Confidence coefficient is $1 - 2\alpha$ and the required precision $d = (\bar{x} - \mu)$. Hence, the estimated sample size is

$$n \le z_\alpha^2\,\sigma^2/d^2 \tag{3}$$

To determine n, knowledge of d and σ^2 are necessary. The scientist can usually supply the value of d and sometimes has some knowledge about σ^2 or has historical data which can be used to estimate σ^2. The calculation of σ^2 depends on the form of f(x). Consider the following transformation (Box and Cox, 1964)

$$\begin{aligned} y &= (x^\lambda - 1)/\lambda \qquad \lambda \ne 0 \\ y &= \ln x \qquad \lambda = 0 \end{aligned} \tag{4}$$

where $x > 0$. It is assumed that there exists a value λ for which the distribution of y is normal with mean μ and variance σ^2. El-Shaarawi and Murthy (1976) have shown how to estimate λ and derived the following expressions for the mean and variance of x

$$E(x) = (\lambda\mu + 1)^{\lambda^{-1}} \exp[\sigma^2/2\,(\lambda\mu + 1)^2] \tag{5}$$

$$\text{var}(x) = (\lambda\mu + 1)^{2\lambda^{-1}} \exp(\sigma^2/(\lambda\mu + 1)^2 \{\exp[\sigma^2/(\lambda\mu + 1)^2] - 1\}). \tag{6}$$

To estimate the sample size for distributions belonging to transformation 4, one needs to know σ^2, μ and λ. If one is interested in the case where $\lambda = 0$, which corresponds to the log normal distribution, and suppose one takes

$$d = \delta E(x) = \delta\mu, \quad 0 < \delta < 1 \tag{7}$$

then the sample size corresponding to the log normal distribution is given by

$$n \le z_\alpha^2\,(e^{\sigma^2 - 1})/\delta^2. \tag{8}$$

In the cases when x can take zero values, the transformation 4 can easily be adjusted to include these cases. This adjustment will add an extra unknown parameter, which can be estimated from the data. In particular, if a log transformation is appropriate for transforming the data to normality, a positive constant λ_1 can be added to x such that $x + \lambda_1 > 0$. One way of estimating λ_1 is to minimize the residual sum of squares after estimating the mean value, i.e. to take an estimate for λ_1, the value λ_1, which minimizes the expression

$$S^{(\lambda)} = \Sigma[\log(x + \lambda_1)]^2 - [\Sigma \log(x + \lambda_1)/n]^2. \tag{9}$$

Another interesting situation arises when the data can be fitted to the exponential distribution $\theta e^{-\theta x}$, where θ is an unknown parameter. In this case, the mean and variance of this distribution are given by

$$\begin{aligned} E(x) &= \theta^{-1} \\ \text{var}(x) &= \theta^{-2} \end{aligned} \tag{10}$$

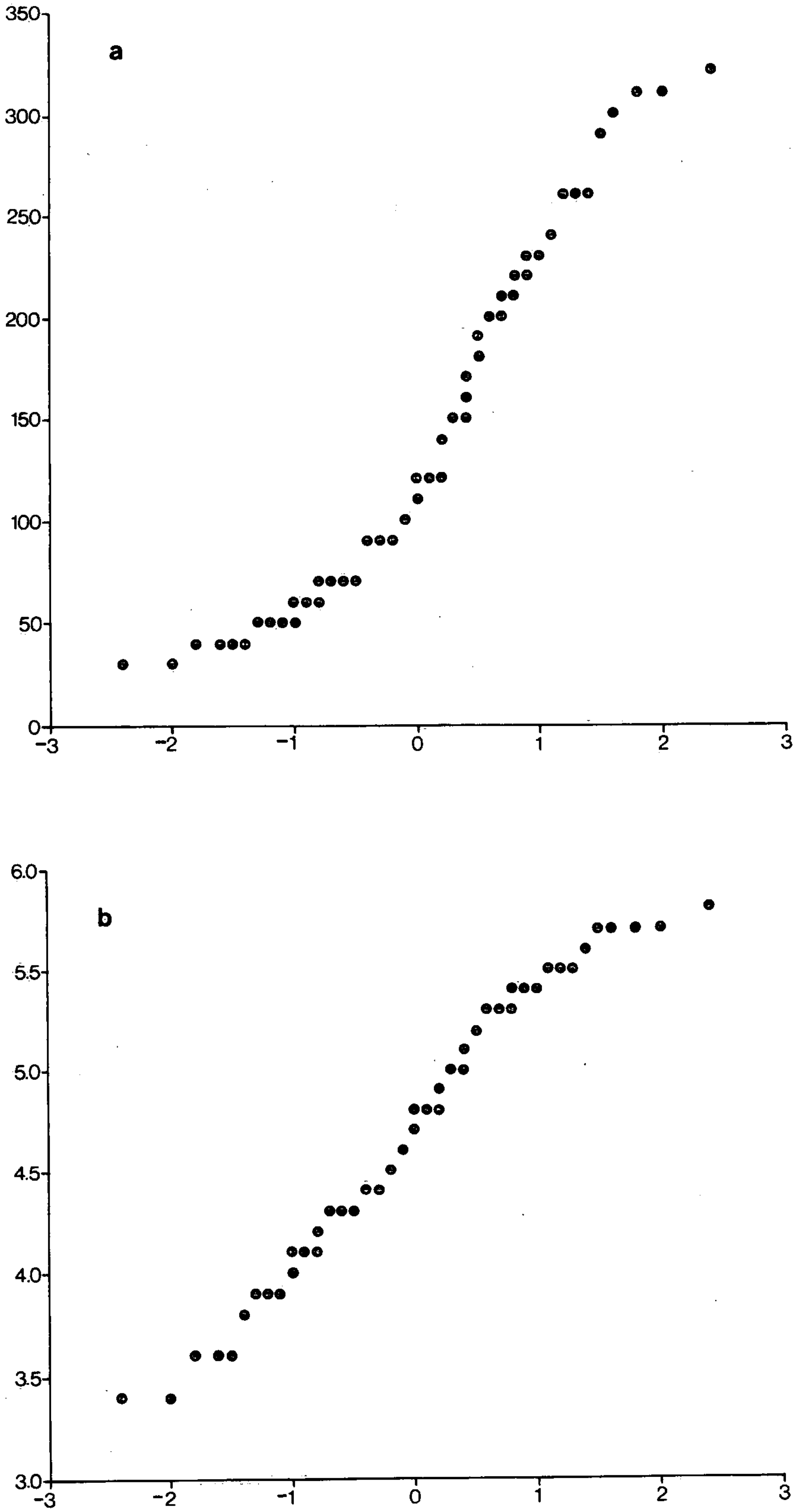


Figure 1. (a) Q-Q plot of bicarbonate values of inland lakes and ponds of Canada. (b) Q-Q plot of transformed bicarbonate values of inland lakes and ponds of Canada.

To estimate the sample size for exponential distributions, one needs to know $\bar{x}$ and δ or d. If one were to take

$$\delta = (\bar{x} - \mu)/\bar{x} \qquad 0 < \delta < 1 \tag{11}$$

then the sample size of the exponential distribution is given by

$$n < z_\alpha^2/2 \quad \text{(relative calculation)} \tag{12}$$

or

$$n < \bar{x}^2 z_\alpha^2/d^2 \quad \text{(absolute calculation)} \tag{13}$$

Using (12), sample size is 384 for $\alpha = 0.1$ and $\alpha = 0.25$.

APPLICATIONS

Chemical Limnology of Inland Lakes and Ponds—National Average

A number of parameters were selected from various lakes and ponds across Canada. These parameters were obtained from a data base containing 6720 samples that was established for use in the study of the autecology of freshwater shelled invertebrates. It is to be emphasized that the records selected are not from a single body of water, but from many. As such, this example shows only the broad variability of some parameters in the natural waters of Canada. The samples were systematically selected at intervals of 50.

It was found that the Q-Q plot (Wilk and Gnandisikan, 1968) for bicarbonate data (Fig. 1a) deviates from linearity. This indicates that the raw data are not normally distributed. On the other hand, the Q-Q plot for the log transformed bicarbonate data (Fig. 1b) indicates linearity, which means that when a log transformation is used, normality is achieved and formula 8 can be used for determining the sample size.

Table 1. Summary of Estimate of Variance and Sample Size Required for a Precision of 0.1 and $\alpha = 0.025$ for the National Average (pH untransformed; CO_3 has exponential distribution)

Parameter	Variance	Sample size $\delta = 0.1$
pH	2.12	14
Conductance	0.10142	41
Calcium	0.10661	44
Bicarbonate	0.13881	58
Chloride	0.27004	120
TDS	0.27748	124
Potassium	0.35045	162
Magnesium	0.38387	180
Sodium	0.82970	497
Sulphate	0.66972	367
Carbonate	34.3	384

Table 1 lists the parameters used, the estimate of σ^2 and the sample size required for $\delta = 0.1$ (within 10% precision) and $\alpha = 0.025$. It is found from the Q-Q plots that all the parameters can be described by a log normal model with the exception of pH, and hence formula 8 is used for the parameters showing log normality, and formula 3 for the pH.

Chemical Limnology of Lake Kipabiskau, Saskatchewan

Up to 14 replicate samples are available for Lake Kipabiskau. The samples were collected at 10-day to 2-week intervals from spring to late fall. Table 2 lists the parameters, the estimate of variance and the sample size required for a precision of within 10%. Four of the 12 samples indicate that n is not large enough to produce the estimated mean with the predetermined precision. However, if the required precision is decreased to within 12% for chloride and iron, to within 13% for sodium, and within 15% for copper, then the available samples are sufficient to produce an estimate of the mean with a 95 per cent confidence coefficient.

Table 2. Summary of Estimate of Variance and Sample Size Required for Precision of 0.1 (except where noted) for Lake Kipabiskau (pH untransformed)

Parameter	Variance	Sample size $\delta = 0.1$
Calcium	0.00044	3
Magnesium	0.00606	5
Carbonate	0.00976	6
pH	0.47	7
TDS	0.00967	7
Bicarbonate	0.02064	11
Oxygen	0.01869	11
Chloride	0.03395 (0.03395)*	17 (12)*
Sodium	0.03709 (0.02515)†	19 (8)†
Iron	0.04424 (0.03235)*	22 (11)*
Copper	0.04269 (0.04269)‡	22 (10)‡

*Precision $\delta = 0.12$ (12%).
†Precision $\delta = 0.13$ (13%).
‡Precision $\delta = 0.15$ (15%).

In comparison with the sample sizes of the national average (see Table 1), the sample size required for estimating the mean with the same precision is much smaller for Lake Kipabiskau. This is undoubtedly caused by the smaller variability within the range of the given parameters.

Sample Size for the Autecology of Shelled Invertebrates

A total of 223 samples are available in which the ostracode species *Candona renoensis* (Delorme, 1970) has been found. Of concern was whether or not a sufficient number of samples had been obtained for estimating the mean value of the various parameters. As was the case for the national average, the samples are from various ponds, lakes and streams within a given zoogeographic province. Table 3 lists the parameters, the estimate of the variance and the sample size for a precision of the mean value within 10 per cent. The samples were randomly selected from a population of 223.

It can be seen from Table 3 that the sample size is a function of the parameter and the requirements of the biological species for that parameter. In other words, to arrive at a suitable mean estimate for the various parameters, the size of the sample required varies with the parameter rather than with the biological species as a function of the variance, mean and range. The "ranking" of the number of samples required for estimating the mean may be a function of the tolerance to the parameter by the biological species and can only be confirmed by the use of other statistical methods.

Table 3. Summary of Estimate of Variance and Sample Size Required for Precision of 0.1 for the Ostracode *Candona renoensis* (pH untransformed)

Parameter	Variance	Sample size at $\delta = 0.1$
pH	0.07	3
TDS	0.05494	24
Conductance	0.06821	30
Chloride	0.06951	30
Bicarbonate	0.08111	33
Potassium	0.09529	38
Magnesium	0.11638	47
Calcium	0.19443	82
Sulphate	0.22177	96
Carbonate	0.22177	99
Sodium	0.31546	144

REFERENCES

Box, G.E.P. and Cox, D.R., 1964. Analysis of transformation (with discussion). J. R. Stat. Soc., Vol. B26, 211-252.

Delorme, L.D., 1970. Freshwater ostracodes of Canada. Part III, Family Candonidae. Can. J. Zool., Vol. 48, No. 5, 1099-1127.

El-Shaarawi, A. and Murthy, C.R., 1976. Probability distribution of concentrations measured in the wake of a continuous point source in coastal currents. J. Phys. Oceanogr., Vol. 6, No. 5, 735-740.

Wilk, M.B. and Gnandisikan, R., 1968. Probability plotting methods for the analysis of data. Biometrika, Vol. 55, 1-17.