ICES TECHNIQUES IN MARINE ENVIRONMENTAL SCIENCES

No. 18

Contaminants in marine organisms: Pooling strategies for monitoring mean concentrations

M.D. NICHOLSON
MAFF Fisheries Laboratory
Pakefield Road
Lowestoft, Suffolk, NR33 0HT
United Kingdom

R.J. FRYER
SOAEFD Marine Laboratory
P.O. Box 101, Victoria Road
Aberdeen, AB9 8DB
United Kingdom

INTERNATIONAL COUNCIL FOR THE EXPLORATION OF THE SEA
CONSEIL INTERNATIONAL POUR L'EXPLORATION DE LA MER

Palægade 2–4, DK-1261 Copenhagen K, Denmark

March 1996

ISSN 0903-2606

TABLE OF CONTENTS

Contaminants in marine organisms: Pooling strategies for monitoring mean concentrations

1 INTRODUCTION

Samples of marine organisms collected for contaminant monitoring are often pooled before being chemically analysed. The main reasons for pooling samples are:

1) to obtain a sufficient quantity of tissue to make the chemical analysis possible;

2) to reduce the overall cost of chemical analyses;

3) to improve the precision of the estimated mean contaminant concentration in a population by increasing the sample size without increasing the number of chemical analyses.

However, there are several questions associated with pooling, including:

- what is an appropriate pooling strategy?
- how should data from pools be statistically analysed?
- how should results derived from pooled data be interpreted?

This document is an introduction to the statistical aspects of pooling. Unfortunately, it is not possible to consider all the situations in which pooling might arise, nor to describe the many types of statistical analysis that might be appropriate. The scope is too large, and we do not know all the answers. Here consideration is given to the relatively simple case of estimating the mean concentration of a contaminant in a population; it shows the typical problems encountered in devising an appropriate pooling strategy and statistically analysing data from pools. In particular, it shows how the choice of the number of pools and the number of individuals in each pool allows a balance to be made of the precision of the estimated mean concentration against the sampling and analytical costs incurred in obtaining that estimate.

Although the level of statistical sophistication increases through the text (notably between Sections 3 and 4), it is hoped that all readers will understand the basic ideas and be able to use this document to develop **sensible** pooling strategies. Many readers will be able to develop the theory for their own particular monitoring problems; others will no doubt correct our mistakes and direct us to the literature we have missed.

Much of the following material was developed by the ICES Working Group on Statistical Aspects of Trend Monitoring (ICES, 1987, 1988, 1989, 1990, 1991a, 1992), where pooling questions were addressed as part of the analysis of data collected in the Cooperative ICES Monitoring Studies Programme (CMP) for contaminants in fish and shellfish.

The contents of this document are as follows:

- Section 2 discusses some practical and logistical problems associated with sample collection and pre-treatment.

- Section 3 develops the basic statistical theory of pooling for estimating the mean concentration of a contaminant in a population. For simplicity, the individual sample weights are assumed to be the same. It is shown how to formulate objective strategies for choosing the number of pools and the number of individuals in each pool and apply the results to a real example.

- Section 4 generalizes the theory to allow for varying individual sample weights.

- Section 5 provides a simple overview and summarizes important points that make the statistical analysis of data from pools straightforward.

- Annexes 1 to 3 briefly describe extensions to the statistical theory.

2 PRACTICAL ASPECTS OF POOLING

Uthe and Chou (1987) discussed various practical problems that must be considered when pooling organisms. Some problems arise during sample collection, for example, the difficulty of filling sample quotas. Other problems arise after the samples have been collected, for example, sample pre-treatment, tissue quantity, homogenization and chemical analysis. The main points are:

1) Appropriate guidelines for the collection and storage of each monitoring organism and/or its tissue(s) should be specified and followed (cf. ICES, 1990, Annex 4).

2) Preparing pools in the field is not recommended due to the high risk of contamination by external sources and the time required for autopsy and homogenization. Large samples may cause problems if holding and freezer space are limited.

3) Pooling may increase overall costs. Dissection and preparation must generally be carried out under conditions designed to reduce the risk of contamination. Although the number of chemical analyses may be decreased when individuals are pooled, the number of dissections may increase; this can greatly increase the preparation time, for example, when dissecting muscle from fish.

4) Preparation time is also increased if individual tissues must be homogenized before taking a portion for pool preparation. This procedure is generally necessary for tissues, such as fish livers, where tissue weights may vary considerably between individuals. Further problems arise if the total mass of tissue exceeds the capacity of the homogenizer.

5) Homogenizing different tissues from a single individual or the same tissue from different individuals involves dealing with a wider range of concentrations than would occur when homogenizing a single tissue from a single individual. Harris (1978) describes the difficulties of achieving adequate mixing with materials showing large differences in concentrations. In general, the extra component of mixing variance should be shown to be acceptably small.

6) As the time required for sample preparation increases, so does the risk of contamination by external sources, for example, by atmospheric fall-out. There is also the risk of contamination from the homogenizer itself.

3 BASIC STATISTICAL THEORY FOR POOLING

3.1 Introduction

This section develops the statistical framework for discussing pooling. Throughout, the assumed objective is to estimate the mean concentration of a contaminant in some population of interest and the precision of this estimate. To do this, a number of individuals are sampled at random from the population. The individuals may consist of the whole organism, a particular tissue such as the muscle, or a particular organ such as the liver.

We shall make two assumptions to simplify the theory:

1) The weights of all the individuals are the same. The case in which individual weights vary is discussed in Section 4.

2) There is no analytical error, i.e., all concentrations are measured exactly. This assumption is clearly unrealistic and is relaxed in Section 3.8.

3.2 Individuals Analysed Separately

We begin by considering the simple case in which I individuals are sampled at random and are each analysed separately. Let x_i be the concentration of a contaminant in the ith individual. If μ is the mean concentration in the population, then we can write

$$x_i = \mu + \epsilon_i, \tag{1}$$

where ε_i is the deviation of the ith concentration from μ. The ε_i have zero mean and variance σ^2; σ^2 is known as the population variance.

Since the I individuals are analysed separately, we obtain I observations $x_1 \ldots x_I$. The mean concentration of the contaminant in the population, μ, is estimated by the average of these observations

$$\bar{x} = \frac{1}{I}\sum_{i=1}^{I} x_i. \tag{2}$$

The mean and variance of $\bar{x}$ are found as follows. From equations (1) and (2),

$$\begin{aligned}\bar{x} &= \frac{1}{I}\sum_{i=1}^{I} x_i \\ &= \frac{1}{I}\sum_{i=1}^{I} (\mu + \epsilon_i) \\ &= \mu + \frac{1}{I}\sum_{i=1}^{I} \epsilon_i.\end{aligned}$$

The ε_i have zero mean, so the mean of $\bar{x}$ is μ. Further, the I individuals are sampled at random, so the deviations ε_i are independent and the variance of their sum is given by the sum of their variances, that is,

$$\text{Var}\left[\sum_{i=1}^{I}\epsilon_i\right] = \sum_{i=1}^{I}\text{Var}[\epsilon_i] = I\sigma^2.$$

Hence, since

$$\text{Var}[k\epsilon_i] = k^2\text{Var}[\epsilon_i] = k^2\sigma^2,$$

where k is any constant, the variance of $\bar{x}$ is

$$\begin{aligned}\text{Var}[\bar{x}] &= \text{Var}\left[\frac{1}{I}\sum_{i=1}^{I}\epsilon_i\right] \\ &= \frac{1}{I^2}\text{Var}\left[\sum_{i=1}^{I}\epsilon_i\right] \\ &= \frac{1}{I^2}I\sigma^2 \\ &= \frac{\sigma^2}{I}.\end{aligned}$$

3.3 Individuals Homogenized into One Pool

Now suppose that the I individuals are homogenized into one pool, so we obtain one observation—the pool concentration—denoted by X. If w is the weight of the individuals (assumed constant), then

$$X = \frac{\sum_{i=1}^{I} wx_i}{\sum_{i=1}^{I} w} = \frac{1}{I}\sum_{i=1}^{I}x_i = \bar{x}.$$

By the same arguments as before, the mean of X is μ and

$$\text{Var}[X] = \frac{\sigma^2}{I}.$$

The important thing to note is that $X = \bar{x}$ and that $\text{Var}[X] = \text{Var}[\bar{x}]$. Thus, we get the same estimate of the mean concentration by analysing the individuals separately or by analysing them in one big pool. Further, the variances of the two estimators are the same, even though the estimators have been obtained in different ways.

3.4 Degrees of Freedom

If the variances of $\bar{x}$ and X are the same, and fewer chemical analyses are needed to obtain X, then why not routinely pool all individuals? The reason is that the population variance σ^2 is unknown

and this too must be estimated from the data. If the individuals are homogenized into one pool, then although it is possible to estimate the mean concentration, there is no information about the scatter of concentrations around the mean, so no estimate of σ^2 is possible. Statistically, there are no degrees of freedom in the data to estimate σ^2. Consequently, the variance of X cannot be estimated and the confidence interval for μ, the mean concentration of the population, is infinite.

However, if the individuals are analysed separately, there **is** information on the scatter of concentrations. Specifically, there are I - 1 degrees of freedom in the data to estimate σ^2. In fact, σ^2 is estimated to be

$$s^2 = \frac{1}{I-1}\sum_{i=1}^{I}(x_i - \bar{x})^2$$

and the variance of $\bar{x}$ is estimated to be

$$\frac{s^2}{I}.$$

Confidence intervals for μ can now be constructed. For example, assuming the concentrations x_i are normally distributed, a 95% confidence interval for μ is given by

$$\bar{x} \pm t_{\alpha, I-1}\sqrt{\frac{s^2}{I}},$$

where $t_{\alpha, I-1}$ is the α percentile of a t-distribution on I - 1 degrees of freedom and α is given by

$$\alpha = 1 - \frac{1 - 0.95}{2} = 0.975,$$

(see, for example, Meyer, 1965).

3.5 A More General Pooling Strategy

Of course, the choice is not simply between individuals and one complete pool. We now consider the more general situation in which the individuals are processed as P pools each consisting of I homogenized individuals. Throughout, it is assumed that I is the same in each pool; this is a sensible strategy, since the variance of each pool should be identical, allowing simple statistical analyses. The situation in which there are unequal numbers of individuals in each pool is discussed in Annex 2.

Let x_{pi} be the contaminant concentration of the ith individual in the pth pool and write

$$x_{pi} = \mu + \epsilon_{pi}.$$

Then X_p, the contaminant concentration of the pth pool, is given by

$$X_p = \frac{1}{I}\sum_{i=1}^{I} x_{pi}.$$

As before,

$$X_p = \mu + \frac{1}{I}\sum_{i=1}^{I} \epsilon_{pi},$$

so the mean of X_p is μ and (assuming that individuals have been allocated randomly to pools) the variance of X_p is

$$\mathrm{Var}[X_p] = \mathrm{Var}\left[\frac{1}{I}\sum_{i=1}^{I} \epsilon_{pi}\right] = \frac{\sigma^2}{I}.$$

The mean contaminant concentration of the population is now estimated by the average concentration of the pools

$$\bar{X} = \frac{1}{P}\sum_{p=1}^{P} X_p.$$

The mean value of $\bar{X}$ is μ and

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{PI}.$$

Here, there are P observations, so the population variance σ^2 can be estimated by

$$s^2 = \frac{I}{P-1}\sum_{p=1}^{P}(X_p - \bar{X})^2$$

on $P - 1$ degrees of freedom. The variance of $\bar{X}$ is then estimated to be s^2 / PI and a 95% confidence interval for μ is given by

$$\bar{X} \pm t_{0.975,\,P-1}\sqrt{\frac{s^2}{PI}}.$$

3.6 Choice of Number of Individuals in Each Pool and Number of Pools

An important problem is how to choose suitable values of P and I. Clearly, an important objective is to make the confidence interval for μ suitably small.

In general, for a given total number of individuals, the confidence interval for μ gets wider as the number of pools decreases. For example, the following table shows how the 95% confidence interval for μ changes as 50 individuals are pooled with increasing severity (where the widths of the confidence intervals are expressed relative to that when there is only one individual in each pool). In particular, this table shows the sensitivity of confidence intervals when there are few degrees of freedom. Reducing the number of pools from 50 to 10 produces only a 13% increase in the width of the confidence interval. A decrease by the same factor from 10 to 2 pools produces a $100 \times (6.32 - 1.13) / 1.13 = 460\%$ increase.

Number of pools P	Number in each pool I	Relative confidence interval
50	1	1.00
25	2	1.03
10	5	1.13
5	10	1.38
2	25	6.32
1	50	∞

This table suggests that the best strategy is to have no pooling, or only a limited amount of pooling. However, so far, no account has been taken of the relative costs of collecting samples and conducting chemical analyses. For example, suppose that the cost of collection is 1 unit per individual and of chemical analysis is 5 units per pool. Then the table above can be extended to give

P	I	C.I.	Cost
50	1	1.00	300
25	2	1.03	175
10	5	1.13	100
5	10	1.38	75
2	25	6.32	60
1	50	∞	55

There is clearly a trade-off between small confidence intervals and low costs. The choice of P and I therefore involves finding suitable values that balance these two criteria.

One solution is to set a target for the total costs, and then find the combination of I and P that provides the narrowest confidence interval. Formally, the objective is to find values of I and P that minimize

$$t_{\alpha,P-1}\sqrt{\frac{\sigma^2}{PI}}$$

subject to the constraint

$$PIc_1 + Pc_2 \leq c$$

where c_1 and c_2 are the costs of collecting (and dissecting) an individual and analysing a pool, respectively, and c is the target total cost.

For example, the tables below are based on costs $c_1 = 1$, $c_2 = 5$ and a population variance $\sigma^2 = 1$. The left-hand table shows the total cost of sampling and analysis for different combinations of I and P; the right-hand table shows the corresponding 95% confidence interval.

	Costs						Confidence Interval					
I	P						P					
	2	3	4	5	6	7	2	3	4	5	6	7
1	**12**	**18**	**24**	**30**	**36**	42	**8.98**	**2.48**	**1.59**	**1.24**	**1.05**	0.92
2	**14**	**21**	**28**	**35**	42	49	**6.35**	**1.76**	**1.12**	**0.88**	0.74	0.65
3	**16**	**24**	**32**	**40**	48	56	**5.19**	**1.43**	**0.92**	**0.72**	0.61	0.53
4	**18**	**27**	**36**	45	54	63	**4.49**	**1.24**	**0.80**	0.62	0.52	0.46
5	**20**	**30**	**40**	50	60	70	**4.02**	**1.11**	**0.71**	0.56	0.47	0.41
6	**22**	**33**	44	55	66	77	**3.67**	**1.01**	0.65	0.51	0.43	0.38
7	**24**	**36**	48	60	72	84	**3.40**	**0.94**	0.60	0.47	0.40	0.35
8	**26**	**39**	52	65	78	91	**3.18**	**0.88**	0.56	0.44	0.37	0.33

Although this information would usually be presented graphically, it is presented here in tabular form to make the choice of I and P easier to follow.

Suppose the total cost of sampling and analysis must be no more than 40 units. In the tables, the values that satisfy this constraint are shown in boldface. The narrowest confidence interval lies at the bottom of one of the bold columns. Here, given the 40 unit cost constraint, the narrowest confidence interval is $\pm$ 0.71 and is achieved when $I = 5$ and $P = 4$.

A second way of choosing I and P is to set a target for the width of the confidence interval and to meet this target as inexpensively as possible. Suppose, in the example above, we want the confidence interval to be smaller than $\pm$ 1.0. Now we wish to minimize the costs

$$PIc_1 + Pc_2$$

subject to

$$t_{0.975,P-1}\sqrt{\frac{\sigma^2}{PI}} \leq 1.0.$$

The tables of costs and confidence intervals are repeated below. In the confidence interval table, each value less than 1.0 is shown in boldface; the corresponding costs are also shown in bold. Thus, the least expensive way (32 units) of achieving a confidence interval less than $\pm$ 1.0 is to take $I = 3$ and $P = 4$.

	Costs						Confidence Interval					
I	P						P					
	2	3	4	5	6	7	2	3	4	5	6	7
1	12	18	24	30	36	**42**	8.98	2.48	1.59	1.24	1.05	**0.92**
2	14	21	28	**35**	**42**	**49**	6.35	1.76	1.12	**0.88**	**0.74**	**0.65**
3	16	24	**32**	**40**	**48**	**56**	5.19	1.43	**0.92**	**0.72**	**0.61**	**0.53**
4	18	27	**36**	**45**	**54**	**63**	4.49	1.24	**0.80**	**0.62**	**0.52**	**0.46**
5	20	30	**40**	**50**	**60**	**70**	4.02	1.11	**0.71**	**0.56**	**0.47**	**0.41**
6	22	33	**44**	**55**	**66**	**77**	3.67	1.01	**0.65**	**0.51**	**0.43**	**0.38**
7	24	**36**	**48**	**60**	**72**	**84**	3.40	**0.94**	**0.60**	**0.47**	**0.40**	**0.35**
8	26	**39**	**52**	**65**	**78**	**91**	3.18	**0.88**	**0.56**	**0.44**	**0.37**	**0.33**

In summary, the two basic strategies for choosing the number of pools and the number of individuals in each pool are:

- to obtain the narrowest confidence interval for a given cost, or
- to minimize the cost of obtaining a given width confidence interval.

However, this is not the end of the story. It is important to consider whether the narrowest confidence interval that can be obtained for a given cost is narrow enough; if not, then either more money is required, the objectives of the project should be changed, or the work should not be done at all. Similar arguments apply if the cost of obtaining a given width confidence interval is outside the budget of the project.

3.7 More Complicated Cost Functions

So far, it has been assumed that the costs of sampling and analysis are directly proportional to the number of individuals collected and the number of analyses, respectively. In practice, this simple rule may not always apply and more complicated cost functions might be required.

For example, collecting many individuals might require more than one field trip and analysing many pools might require more laboratory facilities or analysts. Suppose that N_{max} is the maximum number of individuals that can be collected on a single field trip and P_{max} is the maximum number of analyses that can be made without having to increase the resources in a laboratory. Then, if c_3 is the cost of a field trip and c_4 is the cost of increasing resources, the cost function could be extended to

$$PIc_1 + Pc_2 + \text{Integer}\left[1 + \frac{PI - 1}{N_{\max}}\right] c_3 + \text{Integer}\left[\frac{P - 1}{P_{\max}}\right] c_4.$$

3.8 Analytical Error

In practice, contaminant concentrations are measured with analytical error and this affects both the choice of suitable values of I and P and the estimation of the mean concentration of the population.

Inevitably, the statistical theory is slightly more complicated.

The **measured** concentration of the pth pool can be written

$$X_p = \frac{1}{I}\sum_{i=1}^{I} x_{pi} + \delta_p,$$

where

$$\frac{1}{I}\sum_{i=1}^{I} x_{pi}$$

is the true concentration of the pth pool and δ_p is the error contributed by the chemical analysis. We shall assume that the analytical errors δ_p are normally distributed with zero mean (i.e., the analytical method is unbiased) and variance σ_a^2, known as the analytical variance.

Since

$$x_{pi} = \mu + \epsilon_{pi},$$

it follows that

$$X_p = \mu + \frac{1}{I}\sum_{i=1}^{I} \epsilon_{pi} + \delta_p.$$

Thus, the mean of X_p is μ and the variance of X_p is

$$\begin{aligned} \mathrm{Var}[X_p] &= \mathrm{Var}\left[\mu + \frac{1}{I}\sum_{i=1}^{I} \epsilon_{pi} + \delta_p\right] \\ &= \mathrm{Var}\left[\frac{1}{I}\sum_{i=1}^{I} \epsilon_{pi}\right] + \mathrm{Var}[\delta_p] \\ &= \frac{\sigma^2}{I} + \sigma_a^2. \end{aligned}$$

The mean concentration of the population is again estimated by the average concentration of the pools

$$\bar{X} = \frac{1}{P}\sum_{p=1}^{P} X_p.$$

The mean value of $\bar{X}$ is μ and

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{PI} + \frac{\sigma_a^2}{P}.$$

In this case, neither the population variance σ^2 nor the analytical variance σ_a^2 can be estimated separately, because it is not possible to determine how much of the scatter in the observations is due to population variability and how much to analytical variability. However, the variance of $\bar{X}$ can still be estimated to be

$$\frac{1}{P(P-1)}\sum_{p=1}^{P}(X_p - \bar{X})^2,$$

on P - 1 degrees of freedom, and a 95% confidence interval for μ is

$$\bar{X} \pm t_{0.975,P-1}\left[\frac{1}{P(P-1)}\sum_{p=1}^{P}(X_p - \bar{X})^2\right]^{1/2}.$$

Again, it is necessary to choose values of I and P that are optimum in some sense. Both the methods of Section 3.6 can be generalized. For example, values of I and P that minimize the width of the confidence interval around μ subject to some total cost constraint are found by minimizing

$$t_{\alpha,P-1}\sqrt{\frac{\sigma^2}{PI} + \frac{\sigma_a^2}{P}}$$

subject to

$$PIc_1 + Pc_2 \leq c.$$

3.9 An Application

The above results will now be used to develop a pooling strategy for estimating the mean concentration of zinc in mussels (*Mytilus edulis*) at a Swedish laboratory. The example and the sources of the data are described in detail by van der Meer (1990).

The first thing to note is that estimates of the population and analytical variances are required to choose optimum values of I and P. The best way of estimating these variances is by conducting a pilot study, such as the one described in Section 3.10, below. However, variance estimates obtained in previous studies are used here.

First, the analytical variance σ_a^2 was estimated using data from three Swedish laboratories, some of which were reported in an ICES intercalibration exercise (Berman and Boyko, 1986). Based on six replicate analyses, the following estimates of σ_a^2 were obtained:

Laboratory	Sample	s_a^2
1	a	26.7
1	b	25.5
2	a	21.4
2	b	80.0
3	a	110.0
	Average	52.7

The analytical variance was estimated to be the average of these values

$$s_a^2 = 52.7.$$

Typically, analytical variances might be expected to vary between laboratories, so averaging variances across laboratories (as above) must be done with some caution. For example, if we were trying to establish a pooling strategy for laboratory 1, then it would be sensible to use only variance estimates from that laboratory. However, sometimes there are not adequate data to do this, so averaging variances across laboratories is a reasonable way of obtaining a rough estimate of the likely analytical variance.

The combined population and analytical variance $\sigma^2 + \sigma_a^2$ was estimated from observed concentrations in individual mussels collected by Sweden as part of the ICES CMP. The following table shows estimates from several areas and years. Zinc concentrations were expressed on a dry weight basis in mg kg^{-1}. The number of observations on which each estimate is based is denoted by n.

Year	Area	n	$s^2 + s_a^2$
1981	43G1	20	510
1981	46G1	20	1156
1982	43G1	20	5357
1982	46G1	13	3844
1983	43G1	25	871
1983	46G1	21	3326
1984	43G1	25	1778
1984	46G1	25	4545
1986	46G1	25	541
1987	43G1	25	442
1987	46G1	25	836
1988	43G1	25	483
1988	46G1	25	7316
		Weighted average	2311.8

The combined population and analytical variances were estimated by the weighted average of these values to be

$$s^2 + s_a^2 = 2311.8.$$

The average zinc concentrations were similar in the two data sets (106 mg kg^{-1} in the CMP data, 152 mg kg^{-1} in the intercalibration data) so it is plausible that the variance estimates from the two studies can be combined. Thus, the population variance is estimated, by subtraction, to be

$$s^2 = 2311.8 - 52.7 = 2259.1.$$

Cost estimates were based on a personal communication from N. Green of the Norwegian Institute for Water Research, Oslo. The handling costs (sampling, dissection, etc.) were estimated to be 25 Norwegian Kroner per individual (i.e., $c_1 = 25$). The chemical costs were estimated to be 1500 Norwegian Kroner per analysis (i.e., $c_2 = 1500$).

The Norwegian pooling scheme usually consists of three pools each containing 50 mussels. The total cost is therefore

$$50 \times 3 \times 25 + 3 \times 1500 = 8250 \text{ Kr.}$$

The Swedish scheme consists of 25 individual mussels, giving a total cost of

$$1 \times 25 \times 25 + 25 \times 1500 = 38125 \text{ Kr.}$$

However, 95% confidence limits are given by

$$\pm 4.303 \sqrt{\frac{2259.1}{150} + \frac{52.7}{3}} = \pm 24.6 \text{ mg kg}^{-1}$$

and

$$\pm 2.064 \sqrt{\frac{2259.1}{25} + \frac{52.7}{25}} = \pm 19.8 \text{ mg kg}^{-1}$$

for Norway and Sweden, respectively. Thus, based on these figures, the Norwegian pooling scheme is less precise than the Swedish scheme, although this precision is achieved at approximately 20% of the cost.

To find out whether there is an alternative to the Norwegian pooling scheme, giving smaller confidence intervals without costing more, we need to minimize

$$t_{0.975, P-1} \sqrt{\frac{\sigma^2}{PI} + \frac{\sigma_a^2}{P}}$$

subject to

$$PIc_1 + Pc_2 \leq c$$

where

$$\sigma^2 = 2259.1, \qquad \sigma_a^2 = 52.7$$

and

$$c_1 = 25, \qquad c_2 = 1500, \qquad c = 8250.$$

Table 1 shows costs and confidence limits for a range of values of P and I. The values for which costs are below 8250 Kr are shown in bold. The smallest confidence limits for which the costs are below 8250 Kr are $\pm$ 19.8 mg kg^{-1}, achieved when $I = 22$ and $P = 4$. Thus, by coincidence, it is possible to recreate the confidence limits of the Swedish pooling scheme at the cost of the Norwegian pooling scheme.

Table 1 Costs and confidence intervals associated with various levels of pooling.

	Costs					**Confidence Intervals**				
	P					*P*				
I	2	3	4	5	6	2	3	4	5	6
1	**3050**	**4575**	**6100**	**7625**	9150	**432.0**	**119.4**	**76.5**	**59.7**	50.5
2	**3100**	**4650**	**6200**	**7750**	9300	**308.9**	**85.4**	**54.7**	**42.7**	36.1
3	**3150**	**4725**	**6300**	**7875**	9450	**255.0**	**70.5**	**45.2**	**35.2**	29.8
4	**3200**	**4800**	**6400**	**8000**	9600	**223.3**	**61.7**	**39.5**	**30.9**	26.1
5	**3250**	**4875**	**6500**	**8125**	9750	**210.8**	**55.8**	**35.7**	**27.9**	23.6
6	**3300**	**4950**	**6600**	**8250**	9900	**186.1**	**51.5**	**33.0**	**25.7**	21.7
7	**3350**	**5025**	**6700**	8375	10050	**174.1**	**48.1**	**30.8**	24.1	20.3
8	**3400**	**5100**	**6800**	8500	10200	**164.5**	**45.5**	**29.1**	22.7	19.2
9	**3450**	**5175**	**6900**	8625	10350	**156.6**	**43.3**	**27.7**	21.6	18.3
10	**3500**	**5250**	**7000**	8750	10500	**150.0**	**41.5**	**26.6**	20.7	17.5
11	**3550**	**5325**	**7100**	8875	10650	**144.3**	**39.9**	**25.6**	19.9	16.9
12	**3600**	**5400**	**7200**	9000	10800	**139.5**	**38.6**	**24.7**	19.3	16.3
13	**3650**	**5475**	**7300**	9125	10950	**135.2**	**37.4**	**23.9**	18.7	15.8
14	**3700**	**5550**	**7400**	9250	11100	**131.5**	**36.3**	**23.3**	18.2	15.4
15	**3750**	**5625**	**7500**	9375	11250	**128.1**	**35.4**	**22.7**	17.7	15.0
16	**3800**	**5700**	**7600**	9500	11400	**125.1**	**34.6**	**22.2**	17.3	14.6
17	**3850**	**5775**	**7700**	9625	11550	**122.4**	**33.8**	**21.7**	16.9	14.3
18	**3900**	**5850**	**7800**	9750	11700	**119.9**	**33.2**	**21.2**	16.6	14.0
19	**3950**	**5925**	**7900**	9875	11850	**117.7**	**32.5**	**20.8**	16.3	13.7
20	**4000**	**6000**	**8000**	10000	12000	**115.6**	**32.0**	**20.5**	16.0	13.5
21	**4050**	**6075**	**8100**	10125	12150	**113.7**	**31.4**	**20.1**	15.7	13.3
22	**4100**	**6150**	**8200**	10250	12300	**112.0**	**31.0**	**19.8**	15.5	13.1
23	**4150**	**6225**	8300	10375	12450	**110.4**	**30.5**	19.5	15.3	12.9
24	**4200**	**6300**	8400	10500	12600	**108.9**	**30.1**	19.3	15.0	12.7
25	**4250**	**6375**	8500	10625	12750	**107.5**	**29.7**	19.0	14.9	12.6
30	**4500**	**6750**	9000	11250	13500	**101.7**	**28.1**	18.0	14.0	11.9
35	**4750**	**7125**	9500	11875	14250	**97.3**	**26.9**	17.2	13.4	11.4
40	**5000**	**7500**	10000	12500	15000	**93.9**	**26.0**	16.6	13.0	11.0
45	**5250**	**7875**	10500	13125	15750	**91.1**	**25.2**	16.1	12.6	10.6
50	**5500**	**8250**	11000	13750	16500	**88.9**	**24.6**	15.7	12.3	10.4
55	**5750**	8625	11500	14375	17250	**87.1**	24.1	15.4	12.0	10.2

3.10 Estimating the Population and Analytical Variances

Estimates of the population and analytical variances are required to find suitable values of I and P. The best way to estimate these variances is to conduct a pilot study. Other methods include using variance estimates published elsewhere or estimates obtained in previous studies (cf. Section 3.9). However, there is no guarantee that these estimates will be appropriate to the current study.

A simple example of a pilot study is provided here. Suppose we take P pools with I individuals in each. Suppose we then take R subsamples from each pool and measure the contaminant concentration of each subsample. Let X_{pr} be the measured concentration of the rth subsample from the pth pool, let

$$\bar{X}_p = \frac{1}{R}\sum_{r=1}^{R} X_{pr}$$

be the average of these measurements for the pth pool and let

$$\bar{X} = \frac{1}{P}\sum_{p=1}^{P} \bar{X}_p$$

be the average of all the measured concentrations.

The analytical variability σ_a^2 is estimated, from the scatter of replicate measurements made on each pool, to be

$$s_a^2 = \frac{1}{P(R-1)}\sum_{p=1}^{P}\sum_{r=1}^{R}\left(X_{pr} - \bar{X}_p\right)^2.$$

The population variance σ^2 is estimated from the scatter of measurements made on different pools, as follows. The variance of $\bar{X}_p$ can be shown to be $\sigma^2 / I + \sigma_a^2 / R$, which is estimated by

$$\frac{1}{P-1}\sum_{p=1}^{P}\left(\bar{X}_p - \bar{X}\right)^2.$$

Hence, σ^2 is estimated to be

$$s^2 = \frac{I}{P-1}\sum_{p=1}^{P}\left(\bar{X}_p - \bar{X}\right)^2 - \frac{Is_a^2}{R}.$$

Sometimes s^2 is negative, in which case it is usual to replace it with zero.

For example, suppose $P = 20$, $I = 10$, $R = 2$ and the measured concentrations X_{pr} are given in the following table:

	r	
p	1	2
1	136.1	139.0
2	114.9	128.6
3	102.6	101.1
4	102.8	91.5
5	120.6	99.7
6	79.9	81.8
7	106.3	107.4
8	137.0	122.8
9	102.1	104.4
10	100.4	103.3
11	115.2	127.2
12	96.4	110.5
13	91.4	97.1
14	104.6	97.8
15	104.5	100.2
16	71.9	87.3
17	104.1	99.8
18	94.7	80.8
19	113.8	121.4
20	105.0	105.1

Then, the variance estimates are $s_a^2 = 48.2$ and $s^2 = 1.95 \times 10^3$.

It is not possible to recommend values of P, I and R appropriate for all pilot studies, although generally, the larger the better. However, some rough guidelines are as follows. The values of I and R are the least important. A suitable value of I should strike a balance between having to collect too many individuals and not having enough material for the chemical analyses. R should be at least 2, to ensure some replicate analyses. The value of P is the most important, because the degrees of freedom for estimating σ^2 and σ_a^2 are P - 1 and $P(R - 1)$, respectively. Thus, with a 'large' P, there will be sufficient degrees of freedom for estimating both variances whatever the values of R and I. As a rough guide, $P = 20$ will estimate the quantities σ_a^2 and $\sigma^2 / I + \sigma_a^2 / R$ to within approximately 0.5 and 1.75 times their true values. Usually, this will give the order of magnitude of σ^2 and σ_a^2, which should be sufficient for the costing analysis. If the estimates of σ^2 and σ_a^2 are found to be unsatisfactory, the pilot study could be repeated with a larger P.

A disadvantage of the pilot study described above is that it implicitly assumes that the mixing of the pools and the subsampling are perfect, i.e., that the pooled individuals have really been homogenized. This is probably not a realistic assumption. Mixing variances are discussed in more detail in Annex 1.

4 DIFFERENT WEIGHTS OF TISSUES IN POOLS

4.1 Introduction

So far, it has been assumed that the weights of all the individuals in a pool are identical, presumably because a fixed weight of tissue has been taken from each individual for analysis. This section develops the more complicated theory required if the individual weights vary. For convenience, the weight of tissue taken from an individual is called the 'weight of the individual'.

Assume that individuals are allocated to pools at random and let w_{pi} be the weight of the ith individual in the pth pool. The true contaminant concentration of the pth pool is then

$$\frac{\sum_{i=1}^{I} w_{pi} x_{pi}}{\sum_{i=1}^{I} w_{pi}}$$

and the measured concentration is

$$X_p = \frac{\sum_{i=1}^{I} w_{pi} x_{pi}}{\sum_{i=1}^{I} w_{pi}} + \delta_p . \qquad (3)$$

The important thing to note is that the pool concentration depends on the individual weights. Consequently, the choice of pooling strategy and the statistical analysis of data from pools must take account of these weights.

The statistical properties of the concentration of a contaminant in a pool when individual weights vary are discussed in Section 4.2. The theory is quite difficult, and those daunted by long statistical formulae can jump to Sections 4.3 to 4.5, below, that deal with the special (and simple) case when concentration and weight are independent. Section 4.6 briefly considers whether concentration and weight are independent in practice.

4.2 Statistical Properties

The mean and variance of X_p can be thought of in two ways:

1) The (unconditional) mean and variance apply before sampling when the individual weights are unknown. They are generally used when choosing the number of pools and the number of individuals per pool.

2) The conditional mean and variance apply after sampling when the individual weights are known. The mean and variance are conditional on having observed these particular weights and are generally used when estimating the mean concentration of the population.

The conditional mean and variance of X_p

First, consider the mean and variance of X_p conditional on a particular set of weights w_{pi}. To make the conditioning explicit, these are denoted by $\mathrm{E}[X_p \mid w_{pi}]$ and $\mathrm{Var}[X_p \mid w_{pi}]$, respectively. By

similar arguments to those in Section 3.2, above, from equation (3)

$$\mathrm{E}[X_p \mid w_{pi}] = \mu + \frac{\sum_{i=1}^{I} w_{pi}\mathrm{E}[\epsilon_{pi} \mid w_{pi}]}{\sum_{i=1}^{I} w_{pi}}$$

$$\mathrm{Var}[X_p \mid w_{pi}] = \frac{\sum_{i=1}^{I} w_{pi}^2 \mathrm{Var}[\epsilon_{pi} \mid w_{pi}]}{\left[\sum_{i=1}^{I} w_{pi}\right]^2} + \sigma_a^2.$$

Simplification of these expressions requires that specific assumptions be made about the conditional distribution of concentration given weight. In particular, if concentration and weight are independent, then

$$\mathrm{E}[\epsilon_{pi} \mid w_{pi}] = 0$$

$$\mathrm{Var}[\epsilon_{pi} \mid w_{pi}] = \sigma^2$$

so that

$$\mathrm{E}[X_p \mid w_{pi}] = \mu \tag{4}$$

$$\mathrm{Var}[X_p \mid w_{pi}] = \frac{\sigma^2 \sum_{i=1}^{I} w_{pi}^2}{\left[\sum_{i=1}^{I} w_{pi}\right]^2} + \sigma_a^2. \tag{5}$$

Thus, if concentration and weight are independent, the (conditional) mean of X_p is μ.

The unconditional mean and variance of X_p

Now consider the unconditional mean and variance of X_p. Here, all possible sets of weights w_{pi} are considered, so write

$$w_{pi} = \mu_w + \eta_{pi},$$

where μ_w is the mean weight in the population and η_{pi} is the deviation of w_{pi} about μ_w, assumed to have zero mean and variance σ_w^2.

Then, from equation (3)

$$X_p = \frac{\sum_{i=1}^{I} (\mu_w + \eta_{pi})(\mu + \varepsilon_{pi})}{\sum_{i=1}^{I} (\mu_w + \eta_{pi})} + \delta_p .$$

For convenience, drop the suffix p and let all summations run from $i = 1$ to I. Then

$$\begin{aligned} X_p &= \frac{\sum \mu\mu_w(1 + \eta_i/\mu_w)(1 + \varepsilon_i/\mu)}{\sum \mu_w(1 + \eta_i/\mu_w)} + \delta_p \\ &= \frac{\mu}{I}\left[I + \frac{\sum \varepsilon_i}{\mu} + \frac{\sum \eta_i}{\mu_w} + \frac{\sum \varepsilon_i\eta_i}{\mu\mu_w} \right]\left[1 + \frac{\sum \eta_i}{I\mu_w} \right]^{-1} + \delta_p . \end{aligned}$$

Ignoring all error terms greater than order two,

$$\begin{aligned} X_p &\simeq \frac{\mu}{I}\left[I + \frac{\sum \varepsilon_i}{\mu} + \frac{\sum \eta_i}{\mu_w} + \frac{\sum \varepsilon_i\eta_i}{\mu\mu_w} \right]\left[1 - \frac{\sum \eta_i}{I\mu_w} + \frac{(\sum \eta_i)^2}{I^2\mu_w^2} \right] + \delta_p \\ &\simeq \frac{\mu}{I}\left[I + \frac{\sum \varepsilon_i}{\mu} + \frac{\sum \eta_i}{\mu_w} + \frac{\sum \varepsilon_i\eta_i}{\mu\mu_w} - \frac{\sum \eta_i}{\mu_w} - \frac{\sum \varepsilon_i \sum \eta_i}{I\mu\mu_w} - \frac{(\sum \eta_i)^2}{I\mu_w^2} + \frac{(\sum \eta_i)^2}{I\mu_w^2} \right] + \delta_p \\ &\simeq \frac{\mu}{I}\left[I + \frac{\sum \varepsilon_i}{\mu} + \frac{\sum \varepsilon_i\eta_i}{\mu\mu_w} - \frac{\sum \varepsilon_i \sum \eta_i}{I\mu\mu_w} \right] + \delta_p . \end{aligned}$$

The mean of X_p is given approximately by

$$\mathrm{E}[X_p] \simeq \mu + \frac{\sigma_{xw}}{\mu_w} - \frac{\sigma_{xw}}{I\mu_w} = \mu + \frac{\sigma_{xw}}{\mu_w}\left[1 - \frac{1}{I} \right],$$

where σ_{xw} is the covariance between concentration and weight. Thus,

- the mean of X_p does not equal μ (i.e., X_p is a biased estimator of μ) unless $\sigma_{xw} = 0$ or $I = 1$, that is, unless the weights are constant, or there is no correlation between concentration and weight, or there is no pooling anyway;

- the bias in X_p increases from 0 to σ_{xw} / μ_w as I increases from 1 to infinity.

A general approximation to the variance of X_p is more complicated and is not given here. However, an approximate formula for the variance can be obtained when concentration and weight are independent. From the expression

$$\mathrm{Var}[X_p] = \mathrm{E}[\mathrm{Var}[X_p \mid w_{pi}]] + \mathrm{Var}[\mathrm{E}[X_p \mid w_{pi}]]$$

and equations (4) and (5),

$$\mathrm{Var}[X_p] = \mathrm{E}\left[\frac{\sigma^2\sum_{i=1}^{I} w_{pi}^2}{\left(\sum_{i=1}^{I} w_{pi}\right)^2} + \sigma_a^2\right] + \mathrm{Var}[\mu]$$

$$= \sigma^2\mathrm{E}\left[\frac{\sum_{i=1}^{I} w_{pi}^2}{\left(\sum_{i=1}^{I} w_{pi}\right)^2}\right] + \sigma_a^2.$$

Dropping the suffix p as before,

$$\left(\sum w_i^2\right)\left(\sum w_i\right)^{-2} = \left(\sum (\mu_w + \eta_i)^2\right)\left(\sum (\mu_w + \eta_i)\right)^{-2}$$

$$= \left[I + 2\frac{\sum \eta_i}{\mu_w} + \frac{\sum \eta_i^2}{\mu_w^2}\right]\left[I + \frac{\sum \eta_i}{\mu_w}\right]^{-2}$$

$$\simeq \frac{1}{I^2}\left[I + 2\frac{\sum \eta_i}{\mu_w} + \frac{\sum \eta_i^2}{\mu_w^2}\right]\left[1 - 2\frac{\sum \eta_i}{I\mu_w} + 3\frac{(\sum \eta_i)^2}{I^2\mu_w^2}\right]^{-2}$$

$$\simeq \frac{1}{I^2}\left[I + 2\frac{\sum \eta_i}{\mu_w} + \frac{\sum \eta_i^2}{\mu_w^2} - 2\frac{\sum \eta_i}{\mu_w} - 4\frac{(\sum \eta_i)^2}{I\mu_w^2} + 3\frac{(\sum \eta_i)^2}{I\mu_w^2}\right]$$

$$= \frac{1}{I^2}\left[I + \frac{\sum \eta_i^2}{\mu_w^2} - \frac{(\sum \eta_i)^2}{I\mu_w^2}\right].$$

Hence,

$$\mathrm{Var}[X_p] \simeq \frac{\sigma^2}{I}\left[1 + \frac{\sigma_w^2}{\mu_w^2}\left[1 - \frac{1}{I}\right]\right] + \sigma_a^2.$$

Thus, the variance of X_p increases as the weights become more variable, i.e., as σ_w^2 increases. For a fixed number of individuals per pool I, the variance of X_p is minimized when the weights are constant, in which case

$$\mathrm{Var}[X_p] = \frac{\sigma^2}{I} + \sigma_a^2$$

as before.

4.3 Estimating the Mean Concentration

Having taken individuals with weights w_{pi} and measured pool concentrations X_p, the mean concentration of the population is estimated by

$$\bar{X} = \frac{1}{P}\sum_{p=1}^{P} X_p .$$

Provided that concentration and weight are independent, the mean of $\bar{X}$ is μ and

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{P^2}\sum_{p=1}^{P} \frac{\sum_{i=1}^{I} w_{pi}^2}{\left[\sum_{i=1}^{I} w_{pi}\right]^2} + \frac{\sigma_a^2}{P},$$

which is estimated by

$$\frac{1}{P(P-1)}\sum_{p=1}^{P}(X_p - \bar{X})^2$$

on P - 1 degrees of freedom.

4.4 Statistically Weighted Means

If the population and analytical variances are well known, then an improved estimator of the mean concentration is obtained by taking a statistically weighted average of the pool concentrations,

$$\bar{X}^* = \frac{\sum_{p=1}^{P} \lambda_p X_p}{\sum_{p=1}^{P} \lambda_p},$$

where

$$\lambda_p = (\mathrm{Var}[X_p])^{-1} = \left[\frac{\sigma^2 \sum_{i=1}^{I} w_{pi}^2}{\left[\sum_{i=1}^{I} w_{pi}\right]^2} + \sigma_a^2\right]^{-1},$$

(Draper and Smith, 1981). Again, if concentration and weight are independent, the mean of $\bar{X}^*$ is μ and the variance of $\bar{X}^*$ is

$$\mathrm{Var}[\bar{X}^*] = \left[\sum_{p=1}^{P} \lambda_p\right]^{-1},$$

which is estimated to be

$$\frac{\sum_{p=1}^{P} \lambda_p (X_p - \bar{X}^*)^2}{(P-1)\left[\sum_{p=1}^{P} \lambda_p\right]}$$

on P - 1 degrees of freedom.

The estimator $\bar{X}^*$ is better than $\bar{X}$ in the sense that it generally has smaller variance. However, if the population and analytical variances are not well known, the use of $\bar{X}^*$ can cause more problems than it solves.

4.5 Choosing the Number of Pools and the Number of Individuals per Pool

If concentration and weight are independent, the mean concentration of the population will be estimated by

$$\bar{X} = \frac{1}{P}\sum_{p=1}^{P} X_p,$$

(see Section 4.3). To choose an appropriate number of pools P and number of individuals per pool I, it is necessary to know the variance of $\bar{X}$. At this stage, the particular weights of individuals are unknown and so we use the variance of X which considers all possible combinations of weights (see Section 4.2), namely,

$$\text{Var}[\bar{X}] = \frac{\sigma^2}{PI}\left[1 + \frac{\sigma_w^2}{\mu_w^2}\left[1 - \frac{1}{I}\right]\right] + \frac{\sigma_a^2}{P}.$$

Both methods discussed in Section 3.6 can be used to choose suitable values of I and P. For example, values of I and P that minimize the width of the confidence interval for μ subject to some total cost constraint are found by minimizing

$$t_{\alpha,P-1}\left[\frac{\sigma^2}{PI}\left[1 + \frac{\sigma_w^2}{\mu_w^2}\left[1 - \frac{1}{I}\right]\right] + \frac{\sigma_a^2}{P}\right]^{1/2}$$

subject to

$$PIc_1 + Pc_2 \leq c.$$

4.6 What Happens in Practice

In practice, the relationship between concentration and weight must be considered, since $\bar{X}$ will be a biased estimator of μ unless concentration and weight are independent. This will generally involve collecting additional data.

Previous studies have suggested that concentration is related to weight in some instances and not in others. For example, ICES (1991b), investigating contaminant levels in mussels, regressed log contaminant burden on log shell weight for cadmium, copper, lead, mercury and zinc. The regression coefficients tended to be close to unity, so that

$$\log(\text{contaminant burden}) \simeq \text{constant} + \log(\text{shell weight}).$$

Further, for data from twelve areas, shell weight was approximately proportional to tissue weight, implying that

$$\log(\text{contaminant burden}) \simeq \text{constant} + \log(\text{tissue weight})$$

and hence

$$\log(\text{concentration}) \simeq \log\left[\frac{\text{contaminant burden}}{\text{tissue weight}}\right] \simeq \text{constant}.$$

Thus, in this study, there was no evidence of a relationship between concentration and weight.

Boyden (1974) found a tendency for contaminant concentrations in mussels to decrease with tissue weight for copper, iron, lead and zinc, although not for cadmium or nickel. However, it may be that this relationship only occurs at times of gametogenesis (Phillips, 1976), with the greater number of gametes produced by larger individuals leading to greater dilution of the body burden by the increased tissue weight. This problem can be avoided by taking care in the choice of sampling period.

In a study of contaminants in fish liver, Nicholson *et al.* (1991) found that concentrations of cadmium expressed on a tissue weight basis decreased with the weight of fat in the liver, but not the liver weight. However, for PCBs, concentration depended on both fat and liver weight.

If concentration and weight are related, then one approach is to take an equal weight of tissue from each individual, since then $\bar{X}$ will be an unbiased estimator of μ. However, it must be recognised that a relationship between concentration and weight is potentially very informative; weight is a surrogate for age, so changing concentrations with weight might reflect the availability of the contaminant over time.

5 DISCUSSION, OVERVIEW AND SUMMARY

Throughout Sections 3 and 4 it was assumed, for simplicity, that

- there are the same number of individuals in each pool,
- the population and analytical variability are normally distributed.

These assumptions will often be met in practice, at least approximately so in the case of normality. If either is not satisfied, the statistical theory becomes more complicated, and sometimes intractable. Annexes 2 and 3 discuss what happens if there are unequal numbers of individuals in pools and if the population and analytical variability are lognormally distributed, respectively.

Two methods for choosing the number of pools and the number of individuals per pool have been considered:

1) to achieve the narrowest confidence interval for a given cost,

2) to achieve a specified confidence interval as inexpensively as possible.

Whichever method is used, it is important to ensure that the original objectives of the monitoring programme are met. For example, suppose we want to compare the mean contaminant concentration to an environmental standard. The confidence interval should then be sufficiently narrow so that large violations of the standard are likely to be detected. What is meant by a 'large' violation will vary, and should be carefully considered in the design of the programme. The probability that a particular violation will be detected is measured by the power of the programme (Cohen, 1977). This is not discussed further here, but see Nicholson and Fryer (1992) and Fryer and Nicholson (1993) for more details.

Often we have objectives other than the estimation of the mean concentration of a contaminant in a population. A different pooling scheme might then be required. For example, to detect changes in contaminant levels from one time period to the next, the population sampled may be restricted to animals within a certain size range. It will be necessary to recognize this in the pooling strategy. See, for example, Nicholson and Portmann (1985) and ICES (1987, 1988, 1989, 1990, 1991a, 1992).

Complications also arise if we are interested in a number of contaminants and consequently analyse more than one contaminant in the same individuals. These complications occur because

- the population and analytical variances are likely to vary between contaminants,
- the errors for the same individual are likely to be correlated,
- the criteria for choosing the optimum number of pools and number of individuals per pool will involve several confidence intervals.

Nevertheless, in conclusion, DO NOT WORRY! If you make sure that there are the same number of individuals in each pool and there are many pools, most of the time your monitoring will function as desired.

6 ACKNOWLEDGEMENTS

The authors would like to thank the members of the Working Group on Statistical Aspects of Environmental Monitoring for their helpful comments during the preparation of this document: in particular, Jaap van der Meer (Netherlands Institute for Sea Research) and Bill Warren (Department of Fisheries and Oceans, St. Johns, Nfld., Canada) for their constructive suggestions and reviews.

ANNEX 1

MIXING VARIANCE AND REPLICATE ANALYSES OF INDIVIDUAL OR POOLED SAMPLES

In Section 3 of the main text, it is assumed that the whole individual or pool is chemically analysed. In practice, subsamples may be taken, either because there is more tissue than required or to provide replicate chemical analyses.

Replicate analyses may provide useful information on the analytical variance. Furthermore, if the analytical method is subject to intermittent sources of error, replicate analyses may be a suitable means of filtering out potentially erroneous results. However, this assumes that the individual or pool is thoroughly mixed, and that the variability in concentration within the sample is either zero or small relative to the analytical variance. In practice, this might not be the case.

Mixing variability can be introduced as follows. Let X_{pr} be the measured concentration of the rth replicate analysis on the pth pool, let $\bar{X}_p$ be the average of the measurements on the pth pool and let $\bar{X}$ be the average of the measurements on all the pools. Then, the variance of $\bar{X}_p$ becomes

$$\mathrm{Var}[\bar{X}_p] = \frac{\sigma^2}{I} + \frac{\sigma_a^2}{R} + \sigma_m^2,$$

where R is the number of replicate analyses and σ_m^2 is the mixing variance. This formula is a simplification which assumes that the amount of material in the subsamples is small relative to the amount in the pool. The variance of $\bar{X}$ is then given by

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{PI} + \frac{\sigma_a^2}{PR} + \frac{\sigma_m^2}{P},$$

which is estimated by

$$\frac{1}{P(P-1)}\sum_{p=1}^{P}(\bar{X}_p - \bar{X})^2,$$

on P - 1 degrees of freedom.

The problem is now to choose values of P and I that, e.g., minimize the width of the confidence intervals around $\bar{X}$ subject to some total cost constraint.

Unless the variability introduced by subsampling a pool is considered to be a problem which must be incorporated into the pooling strategy, the practical consequences of mixing variability can be avoided. If the subsampling procedure is pre-defined and constant from sample to sample, the contribution from mixing variance will be the same for all samples, and can simply be incorporated into the analytical variance, and thus ignored. If a more complicated analysis is required, an estimate of σ_m^2 will be required. Various methods are described by Kassmaul and Anderson (1967), Brown and Fisher (1972) and Rohde (1976).

ANNEX 2

UNEQUAL NUMBERS OF INDIVIDUALS IN POOLS

This Annex considers what happens when the number of individuals varies between pools. Let I_p be the number of individuals in the pth pool. The measured contaminant concentration of the pth pool is

$$X_p = \mu + \frac{1}{I_p}\sum_{i=1}^{I_p} \varepsilon_{pi} + \delta_p,$$

which has mean μ and variance

$$\mathrm{Var}[X_p] = \frac{\sigma^2}{I_p} + \sigma_a^2.$$

Thus, pools with a large number of individuals have smaller variance than pools with a small number of individuals.

As in Section 4 of the main text, if the population and analytical variances are well known, the population mean contaminant concentration can be estimated by a weighted average of the concentrations X_p, where the statistical weights λ_p are given by

$$\lambda_p = \left[\frac{\sigma^2}{I_p} + \sigma_a^2\right]^{-1}.$$

ANNEX 3

LOGNORMALITY

Sometimes the distribution of concentrations in the population is skewed and is better described by a lognormal distribution than a normal distribution. If x is the contaminant concentration of an individual picked at random from the population, this means that

$$y = \log(x)$$

has a normal distribution with mean υ and variance τ^2, for example. One approach is then to work on a log-scale and estimate the mean log-concentration υ. However, as shown below, the interactions between pooling and data transformation may introduce more problems than they solve. (An alternative approach is to assume that the concentrations have a gamma rather than a lognormal distribution; see, e.g., McCullagh and Nelder, 1989.)

First consider the situation in which there is no analytical variability. If there is no pooling, the estimation of υ is straightforward. Suppose, that I individuals have been picked at random from the population. Let x_i be the concentration of the ith individual and let

$$y_i = \log(x_i).$$

The mean log-concentration is then estimated by

$$\bar{y} = \frac{1}{I}\sum_{i=1}^{I} y_i.$$

The mean of $\bar{y}$ is υ, so it is an unbiased estimator. Further, $\bar{y}$ has variance τ^2 / I, which is estimated by

$$\frac{1}{I(I-1)}\sum_{i=1}^{I}(y_i - \bar{y})^2$$

on I - 1 degrees of freedom.

Now suppose that there are P pools with I individuals in each pool. As before, let x_{pi} be the contaminant concentration of the ith individual in the pth pool and let X_p be the contaminant concentration of the pth pool. Let

$$Y_p = \log(X_p)$$

be the log-concentration of the pth pool. Again υ is estimated by

$$\bar{Y} = \frac{1}{P}\sum_{p=1}^{P} Y_p.$$

However, this estimator is now biased. The mean of $\bar{Y}$ is

$$v + \frac{\tau^2(1 - I^{-1})}{2}$$

(Aitchison and Brown, 1957), so the bias is positive and increases with the degree of pooling. As I varies between 1 and infinity, the mean of $\bar{Y}$ varies between v and $\log(\mu)$, where μ is the mean concentration on the original untransformed scale. The variance of $\bar{Y}$ is unaffected by the pooling and is given by

$$\frac{\tau^2}{PI}$$

which is estimated by

$$\frac{1}{P(P - 1)} \sum_{p=1}^{P} (Y_p - \bar{Y})^2$$

on P - 1 degrees of freedom.

If the analytical variability is also lognormally distributed and proportional to concentration, the variance of $\bar{Y}$ becomes

$$\mathrm{Var}[\bar{Y}] = \frac{\tau^2}{PI} + \frac{\tau_a^2}{P},$$

where τ_a^2 is the analytical variability on a log-scale, and is again estimated to be

$$\frac{1}{P(P - 1)} \sum_{p=1}^{P} (Y_p - \bar{Y})^2$$

on P - 1 degrees of freedom.

Thus, if the concentrations are lognormally distributed, pooling introduces bias. As the degree of pooling increases, so does the bias. In practice, the bias can be estimated given an estimate of the population variance τ^2. Whether the bias is large enough to cause concern will depend on, e.g., the variance of $\bar{Y}$, the magnitude of v, and the reason for estimating v in the first place.

If the numbers of individuals per pool or the weights of the individuals in a pool vary, then the bias will vary from pool to pool and the estimation of the mean log-concentration becomes even more complicated. Inconsistent pooling should be avoided, if possible.

REFERENCES

Aitchison, J., and Brown, J.A.C. 1957. The Lognormal Distribution. University of Cambridge Department of Applied Economics. Monograph 5. Cambridge University Press.

Berman, S.S., and Boyko, V.J. 1986. Report on the results of the seventh intercalibration exercise on trace metals in biota. Part 1. ICES Cooperative Research Report, No. 138.

Boyden, C.R. 1974. Trace element content and body size in molluscs. Nature, London, 251: 311–314.

Brown, G.H., and Fisher, N.I. 1972. Subsampling a mixture of sampled material. Technometrics, 14: 663–668.

Cohen, J. 1977. Statistical Power Analysis for the Behavioural Sciences. Academic Press Inc., New York.

Draper, N.R., and Smith, H. 1981. Applied Regression Analysis, Second Edition. John Wiley & Sons, New York.

Fryer, R.J., and Nicholson, M.D. 1993. The power of a contaminant monitoring programme to detect linear trends and incidents. ICES Journal of Marine Science, 50: 161-168.

Harris, W.E. 1978. Sampling, manipulative, observational, and evaluative errors. American Laboratory, 10: 31–39.

ICES. Report of the 1987 meeting of the Working Group on Statistical Aspects of Trend Monitoring. ICES CM 1987/E:24.

ICES. 1988. Report of the 1988 meeting of the Working Group on Statistical Aspects of Trend Monitoring. ICES CM 1988/E:27.

ICES. 1989. Report of the Working Group on Statistical Aspects of Trend Monitoring. ICES CM 1989/E:13.

ICES. 1990. Report of the Working Group on Statistical Aspects of Trend Monitoring. ICES CM 1990/Poll:6.

ICES. 1991a. Report of the Working Group on Statistical Aspects of Trend Monitoring. ICES CM 1991/Poll:2.

ICES. 1991b. Statistical analysis of the ICES Cooperative Monitoring Programme data on contaminants in fish liver tissue and *Mytilus edulis* (1978–1988) for the determination of temporal trends. ICES Cooperative Research Report, No. 176.

ICES. 1992. Report of the Working Group on Statistical Aspects of Trend Monitoring. ICES CM 1992/Poll:1.

Kassmaul, K., and Anderson, R.L. 1967. Estimation of variance components in two-stage nested design with composite sampling. Technometrics, 9: 373–389.

McCullagh, P., and Nelder, J.A. 1989. Generalized Linear Models, Second Edition. Chapman and Hall, London.

Meyer, P.L. 1965. Introductory Probability and Statistics. Addison Wesley, Reading, Massachusetts.

Nicholson, M.D., and Fryer, R.J. 1992. The statistical power of monitoring programmes. Marine Pollution Bulletin, 24: 146–149.

Nicholson, M.D., Green, N.W., and Wilson, S.J. 1991. Regression models for assessing trends in cadmium and PCBs in cod livers from the Oslofjord. Marine Pollution Bulletin, 22: 77–81.

Nicholson, M.D., and Portmann, J.E. 1985. The precision of estimated mean levels of metals in fish tissues. ICES CM 1985/E:32.

Phillips, D.J.H. 1976. The common mussel *Mytilus edulis* as an indicator of pollution by zinc, cadmium, lead and copper. I. Effects of environmental variables on the uptake of metals. Marine Biology, 38: 59–69.

Rohde, C.A. 1976. Composite sampling. Biometrics, 32: 273–282.

Uthe, J.F., and Chou, C.L. 1987. The use of pooled samples in measuring time trends in contaminant levels in marine biota. Annex 7. ICES CM 1987/E:24.

Van der Meer, J. 1990. Pooling may economize a sampling programme: Part II. Annex 5. ICES CM 1990/Poll:6.