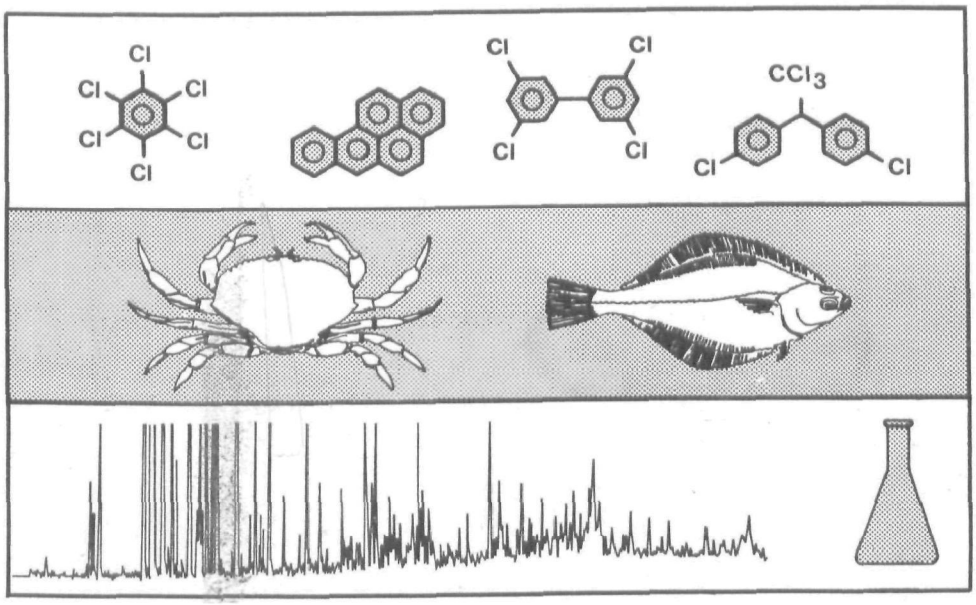
EPA CONTRACT NO. 68-01-6938
TC-3953-03
FINAL REPORT

BIOACCUMULATION MONITORING GUIDANCE:

STRATEGIES FOR SAMPLE REPLICATION AND COMPOSITING



JUNE, 1987

PREPARED FOR:
MARINE OPERATIONS DIVISION
OFFICE OF MARINE AND ESTUARINE PROTECTION
U. S. ENVIRONMENTAL PROTECTION AGENCY

WASHINGTON, DC 20460

TETRA TECH

EPA Contract No. 68-01-6938
TC 3953-03

Final Report

BIOACCUMULATION MONITORING GUIDANCE:

STRATEGIES FOR SAMPLE REPLICATION AND COMPOSITING

for

U.S. Environmental Protection Agency
Office of Marine and Estuarine Protection
Washington, DC 20460

June 1987

by

Tetra Tech, Inc.
11820 Northup Way, Suite 100
Bellevue, Washington 98005

PREFACE

This manual has been prepared by the U.S. Environmental Protection Agency (EPA) Marine Operations Division, Office of Marine and Estuarine Protection in response to requests for guidance from U.S. EPA regional offices and coastal municipalities planning 301(h) monitoring programs for municipal discharges into the marine environment. The members of the 301(h) Task Force of EPA, which includes representatives for the U.S. EPA Regions I, II, III, IV, IX, and X, the Office of Research and Development, and the Office of Water, are to be commended for their vital role in the development of this guidance by the technical support contractor, Tetra Tech, Inc.

This report provides guidance on the selection of appropriate replication and compositing strategies for bioaccumulation monitoring studies. This report is one element of the Bioaccumulation Monitoring Guidance Series. The purpose of this series is to provide guidance for monitoring of priority pollutant residues in tissues of resident marine organisms. These guidance documents were prepared for the 301(h) sewage discharge permit program under the U.S. EPA Office of Marine and Estuarine Protection, Marine Operations Division. Other documents in this series include:

- Estimating the Potential for Bioaccumulation of Priority Pollutants and 301(h) Pesticides (Tetra Tech 1985a)

- Selection of Target Species and Review of Available Bioaccumulation Data, Volumes I and II (Tetra Tech 1987a,b)

- Recommended Analytical Detection Limits (Tetra Tech 1985b).

The statistical analyses conducted in this document are based on the Ocean Data Evaluation System (ODES) Tool No. 14 for Statistical Power Analysis. The Technical Support Document for ODES Statistical Power

Analysis (Tetra Tech 1987c) describes the basis for, and application of, these analytical procedures.

The information provided herein will be useful to U.S. EPA monitoring program reviewers, permit writers, permittees, and other organizations involved in performing nearshore monitoring studies. Bioaccumulation monitoring has become increasingly important in assessing pollution effects; therefore this guidance should have broad applicability in the design and interpretation of marine and estuarine monitoring programs.

CONTENTS

	Page
PREFACE	ii
LIST OF FIGURES	v
LIST OF TABLES	vi
ACKNOWLEDGMENTS	vii
INTRODUCTION	1
MONITORING PROGRAM PERFORMANCE	2
METHODS OF ANALYSIS	2
HYPOTHESIS TESTING	3
POWER ANALYSES FOR INDIVIDUAL TISSUE SAMPLES	5
Analytical Methods	6
Preliminary Analyses	10
Analytical Results	14
Summary	28
COMPOSITE SAMPLING STRATEGIES	29
POWER ANALYSES FOR COMPOSITE SAMPLES	33
Analytical Methods	33
Simulation Analyses	33
Power Analyses	37
Summary	41
SUMMARY AND RECOMMENDATIONS	44
REFERENCES	49

FIGURES

Number		Page
1	Hypothesis testing: possible circumstances and test outcomes	4
2	Frequency distribution for calculated values of the coefficient of variation for 23 historical data sets	15
3	Minimum detectable difference among sampling stations as a function of the coefficient of variation	21
4	Minimum dectectable difference vs. number of replicates at selected levels of unexplained variance for 4 and 6 stations	23
5	Minimum detectable difference vs. number of replicates at selected levels of unexplained variance for 8 and 16 stations	24
6	Minimum detectable difference in the tissue concentration of selected contaminants vs. number of replicates	27
7	Effects of increasing composite sample size on estimate of the mean	35
8	Power of statistical tests vs. number of samples in composite replicate samples	39

TABLES

Number		Page
1	Analysis of variance table for one-way layout	7
2	Summary of data used in power analysis	11
3	Summary of one-way analysis of variance results for historical data	13
4	Results of power analyses showing the minimum detectable difference in the concentration of selected contaminants	16
5	Results of simulation analyses demonstrating the effect of composite sampling on the estimate of the population mean	36
6	Probability of detecting specified levels of minimum detectable differences for selected grab-sampling and composite-sampling strategies	42

ACKNOWLEDGMENTS

This document has been reviewed by the 301(h) Task Force of the Environmental Protection Agency, which includes representatives from the Water Management Divisions of U.S. EPA Regions I, II, III, IV, IX, and X; the Office of Research and Development - Environmental Research Laboratory-Narragansett (located in Narragansett, RI and Newport, OR); and the Marine Operations Division in the Office of Marine and Estuarine Protection, Office of Water.

This technical guidance document was produced for the U.S. Environmental Protection Agency under the 301(h) post-decision technical support contract No. 68-01-6938, Allison J. Duryee, Project Officer. This report was prepared by Tetra Tech, Inc., under the direction of Dr. Thomas C. Ginn. The primary author was Mr. Thomas M. Grieb. Ms. Marcy B. Brooks-McAuliffe performed technical editing and supervised report production.

INTRODUCTION

Monitoring of toxic pollutants in body tissues of marine organisms is an important assessment technique for evaluating effects of coastal sewage discharges and other sources of pollution. A key consideration in the design of bioaccumulation studies is related to the type and number of samples to be analyzed. Measured concentrations of chemicals in organism tissue samples commonly display high levels of variability, resulting from natural biological factors as well as from analytical procedures. Assessment of this variability is an important step in developing an optimal sampling design. Chemical analyses of tissue samples also represent a relatively expensive component of a monitoring program. Without an a priori evaluation of alternative sampling strategies, there is the possibility of analyzing an excessive number of samples (with associated high costs) or of analyzing too few samples where the high variability results in equivocal results.

The objective of this report is to evaluate the applicability of alternative sampling strategies for bioaccumulation monitoring programs. A statistical approach is presented for determining the levels of difference in bioaccumulation that can be reliably detected with varying levels of sampling effort. Example analyses are presented to demonstrate the effects of alternative sampling designs. These example analyses are based on historical data from bioaccumulation monitoring programs that used tissues from individual target species recommended in an earlier report in this series (Tetra Tech 1987a). The results of additional analyses employing simulation methods are used to provide a comparision of grab- and composite-sampling strategies.

MONITORING PROGRAM PERFORMANCE

METHODS OF ANALYSIS

An evaluation of the accumulation of toxic pollutants and pesticides in marine organisms is an important part of 301(h) monitoring programs. The objective of the bioaccumulation component of 301(h) monitoring programs is to determine whether the discharge causes an elevation in the body burden of toxic chemicals in organisms living nearby. This objective is generally addressed by comparing tissue contaminant levels in organisms near the discharge and at a reference area. Measured tissue contaminant levels used for such analyses commonly exhibit a large degree of variability resulting from factors such as measurement errors and natural variability. This variability may be great enough to severely limit the ability to detect statistical changes. However, statistical techniques can be applied to deal with these sources of uncertainty and to make statistically valid comparisons of bioaccumulation levels among monitoring stations.

The 301(h) bioaccumulation monitoring studies are generally designed based on the hypothesis that discharge effects are indicated by measurable differences in bioaccumulation levels among monitoring stations or monitoring events. Given this assumption, statistical techniques can be used to distinguish discharge-related effects from natural variability. This can be accomplished by partitioning field observations into several components. Analysis of variance (ANOVA) techniques, which are commonly used to analyze 301(h) monitoring data, relate observations of interest (e.g., bioaccumulation levels) explicitly to various environmental factors and random errors. This partitioning of field observations can be demonstrated with the ANOVA experimental model shown in Equation (1), which decomposes a single observation (Y_{ij}) into several components:

$$Y_{ij} = \mu + \xi_i + \epsilon_{ij} \quad (1)$$

where:

Y_{ij} = Observations at station i and replicate j of, for example, the tissue concentration of a selected metal

μ = Mean of all Y_{ij} observations

ξ_i = Effect of the i^{th} level of an environmental factor (e.g., station location)

ϵ_{ij} = Random errors not accounted for by either μ or ξ_i.

Under the example model formulation, the effects of environmental factors (e.g., station location) on individual observations can be tested for statistical significance. The null hypothesis tested is that the station location has no effect on observed contaminant concentrations, or stated formally: $\xi_1 = ... = \xi_n = 0$. Similarly, more complex models can be formulated to test for the effect of more than one environmental factor (including time) as well as the statistical significance of the interaction among factors.

HYPOTHESIS TESTING

The testing circumstances and outcomes associated with testing the null hypothesis are shown in Figure 1. Four possible outcomes exist:

1. The hypothesis is true and it is not rejected.

2. The hypothesis is true and it is rejected.

3. The hypothesis is false and it is not rejected.

4. The hypothesis is false and it is rejected.

The shaded areas shown in Figure 1 represent incorrect decisions. The incorrect rejection of the null hypothesis is referred to as a Type I error. The probability of a Type I error, designated α, represents the significance level of the statistical test. The incorrect acceptance of the

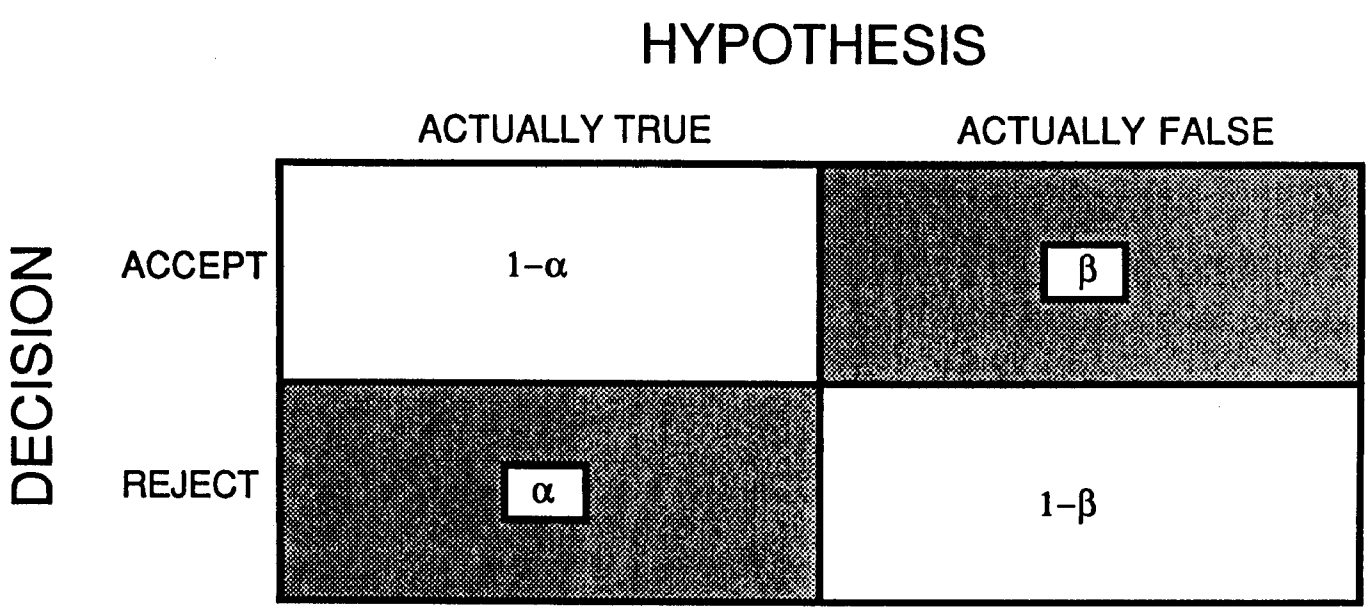


Figure 1. Hypothesis testing: possible circumstances and test outcomes.

null hypothesis is referred to as the β error, where β represents the probability of this incorrect decision. The β error is also known as the Type II error. The probabilities of the correct acceptance and rejection of the null hypothesis are represented by the complements of the Type I and Type II errors, respectively.

The probability of correctly rejecting the false null hypothesis (i.e., of detecting an effect when one exists) is referred to as the power of the statistical test. Because the objective of the bioaccumulation monitoring program is to correctly detect the effects of station location or time of sampling, the power of a statistical test serves as a basis for evaluating the performance of the monitoring program. When existing data are available for the selected monitoring variables, power calculations can be made to provide a quantitative comparison of alternative sampling layouts. For example, the probability of correctly detecting the effects of station location can be determined for a specified level of sampling effort. These methods can also be used to evaluate and interpret statistical analyses in which the null hypothesis has been accepted. In this case, the probability of detecting specific levels of differences between stations or effects associated with different treatments can be determined for the fixed parameters of the experimental design.

POWER ANALYSES FOR INDIVIDUAL TISSUE SAMPLES

The power of the statistical test is determined by the following five study design parameters:

- Significance level of the test
- Number of sampling stations
- Number of replicates

- Minimum detectable difference specified for the monitoring variable

- Residual error variance (i.e., natural variability within the system).

This relationship between the power of a statistical test and the design parameters makes several types of power analyses possible. For example, the power of the test can be determined as a function of the five design parameters. Alternatively, the value for any individual design parameter required to obtain a specified power of the statistical test can be determined as a function of the other four parameters.

For this report, power analyses were conducted using historical data to determine the minimum detectable difference in the tissue concentration of specified contaminants as a function of the number of sample replicates. The purpose of this type of analysis was to determine the level of difference in tissue concentration of contaminants that can be identified in a test of statistical significance. In each individual analysis, the power of the statistical test as well as the other design parameters (i.e., number of stations, significance level, and residual error variance) were held constant. However, a series of these analyses was also conducted for different numbers of stations and values of residual error variance to demonstrate the effect of these design parameters on the ability to detect changes among sampling locations.

Analytical Methods

The results of a one-way ANOVA are usually summarized in a manner similar to that shown in Table 1. The test statistic is the F ratio, which is the ratio of the between-groups mean square (BMS) to the within-groups mean square (WMS). As indicated in Table 1, the WMS is an unbiased estimate of the population variance (σ^2), while the expected value of the BMS is represented by the sum of the population variance and another term representing the actual fixed effects. This added quantity is:

TABLE 1. ANALYSIS OF VARIANCE TABLE FOR ONE-WAY LAYOUT

Source	Sum of Squares	d.f.	Mean Square	E(MS)
Between groups	$\sum_i J_i(\bar{y}_i-\bar{y})^2$	I-1	$SS_B/(I-1)$	$\sigma^2 + (I-1)^{-1}\sum_i J_i(\xi_i-\bar{\xi})^2$
Within groups	$\sum\sum_{ij}(y_{ij}-\bar{y}_i)^2$	n-I	$SS_W/(n-I)$	σ^2
Total	$\sum\sum_{ij}(y_{ij}-\bar{y})^2$	n-1		

where:

y_{ij} = Observation at group (station) i and replicate j

$\bar{y}_i$ = i^{th} group mean

$\bar{y}$ = Overall mean of all i, j observations

I = Number of sampling stations

n = Total number of observations

SS_B = Between groups sum of squares

SS_W = Within groups sum of squares

E(MS) = Expected mean square

J_i = Number of replicates at the i^{th} station

ξ_i = True value of the i^{th} effect

$\bar{\xi}$ = Mean of the treatment effects

σ^2 = Population variance.

$$(I-1)^{-1} \sum J_i (\xi_i - \bar{\xi})^2$$

where:

I = The number of sampling stations
J_i = The number of replicates at the i^{th} station
ξ_i = The true value of the i^{th} effect
$\bar{\xi}$ = The mean of the treatment effects.

Under the null hypothesis, the value of the actual fixed effects term is 0, and the F ratio is equal to 1. When fixed effects are observed in the monitoring program, the value of this term increases and results in an increase in the value of the numerator of the F ratio. Large effects will result in an increase in the power of the test (i.e., the probability of rejecting a false null hypothesis).

In performing power analyses, a set of effects is assumed and the performance of the sampling design is evaluated as if these assumed effects actually occurred. However, when a sample design involves several station locations, many sets of effects can be assumed. For example, alternative hypotheses can be constructed under which actual station effects of a certain magnitude occur at one, two, three, or more of the total number of sampling locations. The magnitude of the effects could also be varied among stations. It can be seen that an infinite number of alternative hypotheses can be constructed for evaluation in power analyses.

The power analyses presented in this report were conducted to provide a conservative evaluation of monitoring program performance. The testable hypothesis used in these analyses was constructed such that the effects occur in the combination that is most difficult to detect. Scheffe (1959) showed that this conservative set of effects is defined by:

$$|\xi_i - \xi_j| = \Delta; \qquad \xi_k = \frac{|\xi_i + \xi_j|}{2}, \quad \text{for all } k \neq i \text{ or } j \tag{2}$$

where:

Δ = The maximum difference in actual effects
ξ_k = The true value of the k^{th} effect.

Equation (2) states that the two effects associated with the hypothesis of interest differ by Δ while all other effects are equal to the mean of these two. For the maximum difference in effects equal to Δ, this arrangement gives the lowest test power.

The power analyses presented in this report were conducted on the Ocean Data Evaluation System (ODES). The power analysis tool available on ODES is described in a user-guidance document (Tetra Tech and American Management Systems 1986). Statistical power analyses and methods of calculation are described by Scheffe (1959) and Cohen (1977).

Recent evidence concerning the robustness of the ANOVA model to deviations from assumptions of normality and equal variances indicates the appropriateness of these statistical methods in environmental monitoring applications (Grieb 1985). However, nonparametric statistical methods such as the Kruskal-Wallis one-way analysis of variance by ranks (Kruskal and Wallis 1952) could also be used for the analysis of these bioaccumulation data. While the statistical analysis results in this report apply to the parametric ANOVA model, these results can also be used to evaluate the corresponding performance of alternative, nonparametric statistical methods by computing the power-efficiency of the nonparametric analog. The power-efficiency of the nonparametric test provides a comparison of the sample size required to achieve the same level of power associated with the corresponding parametric tests. For example, the power-efficiency of statistical Test B relative to Test A is given by:

$$(100)\ \frac{N_A}{N_B}\ \text{percent}$$

where:

N_B = The number of samples required in Test B to achieve the same level of power obtained in Test A with a sample size of N_A.

Calculation of the power-efficiency ratio for the Kruskal-Wallis test is described in Andrews (1954) and Lehmann (1975).

Preliminary Analyses

To conduct the power analysis, it is necessary to obtain an estimate of the residual error variance (i.e., the natural variability not accounted for by the statistical model). This estimate can be obtained by conducting a site-specific preliminary study or by using existing sampling data. For the purposes of demonstrating the power analysis techniques in this report, historical data were compiled and analyzed in a one-way ANOVA to estimate the residual error variance. These estimates were then used as study design parameters in individual power analyses.

A summary of the historical data is provided in Table 2. Data were obtained for five taxa: three fish species (Dover sole, English sole, and winter flounder) and two invertebrate taxa (American lobster and Cancer spp.). Replicate measurements of tissue concentrations of selected contaminants were obtained from various numbers of sample locations. Tissue concentrations of these pollutants were obtained for both muscle and liver tissues. These data were compiled by Tetra Tech (1987a) as part of a review of bioaccumulation data on target species recommended for 301(h) discharge monitoring and were derived from analyses of tissue samples from individual organisms (i.e., no composite samples). The raw data are presented in Tetra Tech (1987b). In general, replicated data for tissue body burdens of priority pollutants are limited. However, while the number of contaminants included in these data is limited, two important chemical groups of concern

TABLE 2. SUMMARY OF DATA USED IN POWER ANALYSIS

Taxon	Type of Tissue	Contaminant	Number of Stations	Number of Replicates	Location	Reference
American lobster (*Homarus americanus*)	Muscle	PCBs, Hg, Cd	4	10	Long Island Sound NY Bight Apex	Roberts et al. (1982)
Dover sole (*Microstomus pacificus*)	Muscle	PCBs, DDT	3	12	Southern California Bight	Sherwood et al. (1980)
	Muscle	Cu	2	6		
	Liver	PCBs, DDT	3	12		
	Liver	Ag, Cd	2	6		
Winter flounder (*Pseudopleuronectes americanus*)	Muscle	PCBs, DDT	4	12	NY Bight Apex	Sherwood et al. (1980)
	Muscle	Cu	3	6		
	Liver	PCBs, DDT	4	12		
	Liver	Cd, Zn	3	6		
English Sole (*Parophrys vetulus*)	Muscle	As, Pb, Hg	6	5	Commencement Bay, WA	Tetra Tech (1985c)
Crab (*Cancer* spp.)	Muscle	PCBs, Pb, Hg	4	5	Commencement Bay, WA	Tetra Tech (1985c)

in terms of bioaccumulation, metals and chlorinated organic compounds, are represented.

The residual error variance design parameter can be viewed as an estimate of the denominator in the F ratio, which is used to evaluate the significance of the ANOVA statistical tests. This quantity is shown in the one-way ANOVA table (Table 1) as the within-groups mean square and represents the average variance within groups. Where sample data are available, this design parameter can be estimated in one of two ways. First, a preliminary ANOVA can be conducted and the value of the within-groups mean square used. Second, the sample variance can be computed from all available data ignoring sample location. The first value provides an estimate of the variance that is unexplained by the statistical model. Therefore, if the effects of sample locations are found to be significant in the F test conducted with the ANOVA, the within-groups mean square will have a value that is less than the overall sample variance.

Because the variance design parameter is an estimate of the denominator in the F ratio, it can be seen that the overall sample variance obtained from existing data provides a larger and, therefore, more conservative estimate for the purposes of conducting power analyses. In this case, the estimated value of the difference that can be detected between stations will be larger than if the power analyses were conducted using the within-groups mean square as an estimate of denominator in the F ratio. However, where available data can be fit to the ANOVA model, the estimate of the within-groups mean square provides a more realistic estimate of the expected value of the denominator in the F ratio.

The historical data described in Table 2 were analyzed using a one-way ANOVA design to obtain values of the within-groups mean square for subsequent power analyses. The results of 23 analyses are shown in Table 3. For each analysis, the estimated mean tissue concentration of the pollutant, the within-groups mean square, and a coefficient of variation are presented. The occurrence of a significant F test is also indicated in this table.

TABLE 3. SUMMARY OF ONE-WAY ANALYSIS OF VARIANCE RESULTS FOR HISTORICAL DATA

Data Set	Taxon	Type of Tissue	Contaminant	Estimated Mean Concentration, $\bar{x}$ (mg/kg)	Within-groups Mean Square ($\hat{\sigma}^2$)	Coefficient of Variation ($\frac{\hat{\sigma}}{\bar{x}} \times 100$)	Significance of Test
1	American Lobster	Muscle	Total PCBs	.152	.0061	51.4	*a
2			Hg	.215	.0076	40.6	*
3			CD	.018	<.0001	28.3	*
4	Dover Sole	Muscle	Total PCBs	.766	.3324	75.3	*
5			Total DDT	1.279	6.7761	203.5	*
6			Cu	.075	.0003	23.1	
7		Liver	Total PCBs	.925	6.1616	268.4	
8			Total DDT	.382	.0615	64.9	
9			Ag	.1162	.0023	41.3	
10			Cd	.7251	.1194	47.7	
11	Winter Flounder	Muscle	Total PCBs	.0906	.0007	29.2	*
12			Total DDT	.0126	<.0001	41.3	
13			Cu	4.389	6.0750	56.2	
14		Liver	Total PCBs	4.015	4.2335	51.2	*
15			Total DDT	.607	.0996	52.0	*
16			Cd	.093	.0023	51.6	*
17			Zn	29.594	17.2450	14.0	
18	English Sole	Muscle	As	5.067	39.2918	123.7	
19			Pb	.247	.0118	44.0	*
20			Hg	.0572	.0006	43.0	*
21	Crab	Muscle	Total PCBs	.0918	.0087	101.6	*
			Pb	.316	.1227	110.8	
			Hg	.094	.0033	61.1	*

[a]F Test significant, $P<0.05$.

Coefficients of variation presented in Table 3 were calculated as the ratio of the square root of the within-groups mean square to the estimated overall mean tissue concentration. This ratio was multiplied by 100 so that the coefficient of variation is expressed as a percentage. These values thus represent a normalized measure of the unexplained variability (uncertainty) within the data set and, as demonstrated below, an important indicator of the level at which statistically significant differences can be detected. A frequency distribution of the observed values of the coefficients of variation is presented in Figure 2. Values ranged from 14.0 to 268.4, but the majority occurred between 40 and 60.

Analytical Results

Results of the power analyses conducted for each historical data set are summarized in Table 4. Results of all analyses are expressed as a percentage of the mean contaminant concentration observed in the particular data set. The presentation of the minimum detectable difference as a percent of the observed mean value, rather than as a concentration of the contaminant, provides a basis for comparing the results obtained for the different data sets. For example, this makes it possible to readily evaluate the effect of the increased sample variability, expressed as an increase in the coefficient of variation, on the ability to detect statistically significant differences among sampling locations. This presentation format also confers a general applicability to these analyses, as the results can be applied to any data set exhibiting the same or similar coefficient of variation. However, as discussed below, it is important to evaluate individual monitoring programs in terms of the value of the contaminant concentration that can be detected among sampling locations.

The analyses presented in Table 4 were conducted with the number of sampling stations fixed at four or eight. This number of stations was selected to represent an expected range in many 301(h) bioaccumulation monitoring programs. The selection of two levels of sampling effort also provided the opportunity to demonstrate the relative effect of an increase in the number of sampling locations on the ability to detect differences in tissue concentrations among stations.

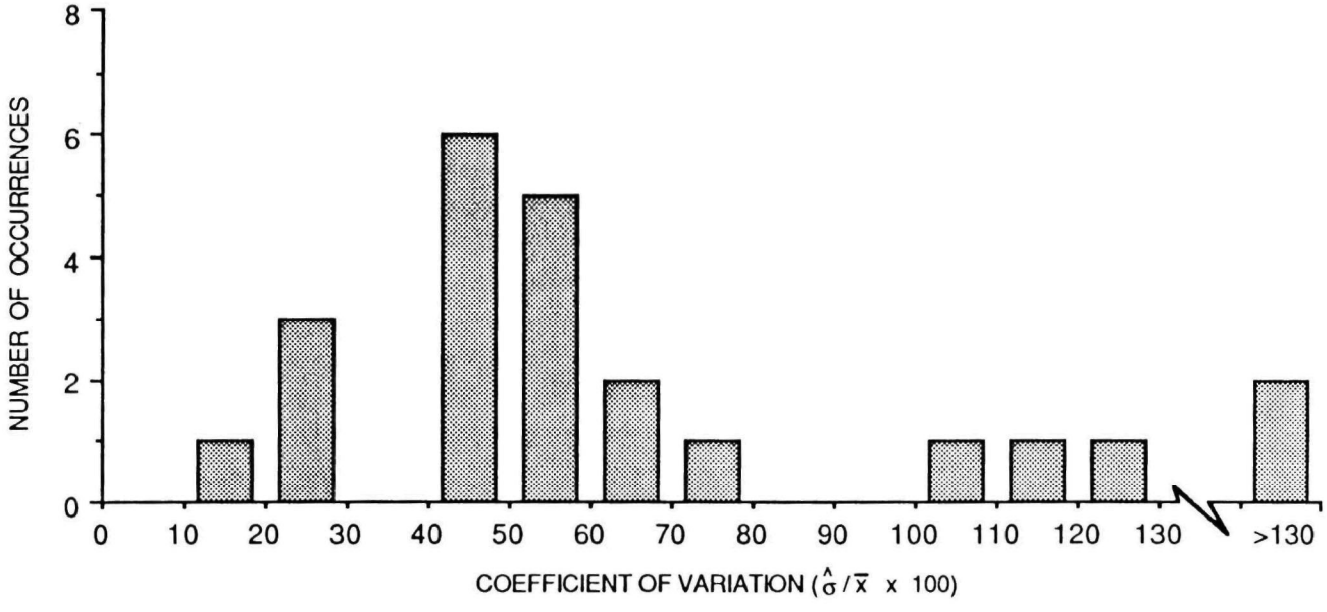


Figure 2. Frequency distribution for calculated values of the coefficient of variation for 23 historical data sets.

TABLE 4. RESULTS OF POWER ANALYSES SHOWING THE MINIMUM DETECTABLE DIFFERENCE IN THE CONCENTRATION OF SELECTED CONTAMINANTS

Data Set	Taxon	Contaminant	Tissue Type[a]	Mean Concentration ($\bar{x}$) mg/kg	Coefficient of Variation ($\frac{\hat{\sigma}}{\bar{x}} \times 100$)	Number of Replicates	Minimum Detectable Difference[b] (Percent of Mean) 4 Stations	8 Stations
1	American lobster	Total PCBs	M	0.152	51.4	2	283	286
						3	178	195
						4	141	158
						5	121	137
						6	108	123
						8	91	104
						10	80	91
						12	72	83
						14	67	76
2	American lobster	Hg	M	0.215	40.6	2	223	226
						3	140	154
						4	112	125
						5	96	108
						6	85	97
						8	72	82
						10	63	72
						12	57	65
						14	53	60
3	American lobster	Cd	M	0.018	28.3	2	156	158
						3	98	107
						4	78	87
						5	67	76
						6	60	68
						8	50	57
						10	44	50
						12	40	46
						14	37	42
4	Dover sole	Total PCBs	M	0.766	75.3	2	414	419
						3	260	285
						4	207	232
						5	178	201
						6	158	179
						8	133	152
						10	117	134
						12	106	121
						14	98	112
5	Dover sole	Total DDT	M	1.279	203.5	2	1,120	1,134
						3	704	772
						4	560	626
						5	481	542
						6	428	485
						8	361	410
						10	317	362
						12	287	328
						14	264	302
6	Dover sole	Cu	M	0.075	23.1	2	127	129
						3	80	87
						4	64	71
						5	55	61
						6	49	55
						8	41	47
						10	36	41
						12	32	37
						14	30	34

TABLE 4. (Continued)

Data Set	Taxon	Contaminant	Tissue Type[a]	Mean Concentration ($\bar{x}$) mg/kg	Coefficient of Variation ($\frac{\partial}{\bar{x}} \times 100$)	Number of Replicates	Minimum Detectable Difference[b] (Percent of Mean) 4 Stations	8 Stations
7	Dover sole	Total PCBs	L	0.925	268.4	2	1,477	1,495
						3	929	1,018
						4	739	826
						5	634	715
						6	565	640
						8	476	541
						10	418	478
						12	378	433
						14	348	398
8	Dover sole	Total DDT	L	0.382	64.9	2	357	362
						3	225	246
						4	179	200
						5	153	173
						6	137	155
						8	115	131
						10	101	116
						12	91	105
						14	84	96
9	Dover sole	Ag	L	0.1162	41.3	2	228	231
						3	143	157
						4	114	127
						5	98	110
						6	87	99
						8	73	83
						10	65	74
						12	58	67
						14	54	61
10	Dover sole	Cd	L	0.7251	47.7	2	262	266
						3	165	181
						4	131	147
						5	113	127
						6	100	114
						8	84	96
						10	74	85
						12	67	77
						14	62	71
11	Winter flounder	Total PCBs	M	.0906	29.2	2	160	162
						3	101	110
						4	80	90
						5	69	78
						6	61	69
						8	52	59
						10	45	52
						12	41	47
						14	38	43
12	Winter flounder	Total DDT	M	.0126	41.3	2	229	231
						3	144	158
						4	114	128
						5	98	111
						6	87	99
						8	74	84
						10	65	74
						12	59	67
						14	54	62
13	Winter flounder	Cu	M	4.389	56.2	2	309	313
						3	194	213
						4	155	173
						5	133	150
						6	118	134
						8	100	113
						10	88	100
						12	79	91
						14	73	83

TABLE 4. (Continued)

Data Set	Taxon	Contaminant	Tissue Type[a]	Mean Concentration ($\bar{x}$) mg/kg	Coefficient of Variation ($\frac{\partial}{\bar{x}} \times 100$)	Number of Replicates	Minimum Detectable Difference[b] (Percent of Mean) 4 Stations	8 Stations
14	Winter flounder	Total PCBs	L	4.015	51.2	2	282	286
						3	177	194
						4	141	158
						5	121	137
						6	108	122
						8	91	103
						10	80	91
						12	72	83
						14	66	76
15	Winter flounder	Total DDT	L	0.607	52.0	2	286	290
						3	180	197
						4	143	160
						5	123	139
						6	109	124
						8	92	105
						10	81	93
						12	73	84
						14	67	77
16	Winter flounder	Cd	L	0.093	51.6	2	284	288
						3	179	196
						4	142	159
						5	122	138
						6	109	123
						8	92	104
						10	81	92
						12	73	83
						14	67	77
17	Winter flounder	Zn	L	17.2450	14.0	2	77	78
						3	49	53
						4	39	43
						5	33	37
						6	30	33
						8	29	28
						10	22	25
						12	20	23
						14	18	21
18	English sole	As	M	5.067	123.7	2	681	689
						3	428	469
						4	341	381
						5	292	330
						6	260	295
						8	219	249
						10	193	220
						12	174	199
						14	160	184
19	English sole	Pb	M	0.247	44.0	2	242	245
						3	152	167
						4	121	135
						5	104	117
						6	93	105
						8	78	89
						10	69	78
						12	62	71
						14	57	65
20	English sole	Hg	M	0.0572	43.0	2	237	239
						3	149	163
						4	118	132
						5	102	115
						6	90	102
						8	76	87
						10	67	76
						12	61	69
						14	56	64

TABLE 4. (Continued)

Data Set	Taxon	Contaminant	Tissue Type[a]	Mean Concentration ($\bar{x}$) mg/kg	Coefficient of Variation ($\frac{\partial}{\bar{x}} \times 100$)	Number of Replicates	Minimum Detectable Difference[b] (Percent of Mean) 4 Stations	8 Stations
21	Cancer crabs (Cancer spp.)	Total PCBs	M	.0918	101.6	2	558	565
						3	351	385
						4	279	312
						5	240	270
						6	213	242
						8	180	205
						10	158	180
						12	143	163
						14	132	150
22	Cancer crabs (Cancer spp.)	Pb	M	.316	110.8	2	610	618
						3	384	420
						4	305	341
						5	262	295
						6	233	264
						8	197	224
						10	173	197
						12	156	179
						14	144	164
23	Cancer crabs (Cancer spp.)	Hg	M	.094	61.1	2	336	340
						3	211	232
						4	168	188
						5	144	163
						6	128	146
						8	108	123
						10	95	109
						12	86	98
						14	79	91

[a]Tissue type: M = muscle, L = liver.

[b]Power analyses conducted at fixed levels of statistical significance (0.05) and power (0.80).

As indicated in Table 4, these analyses were conducted at fixed levels of statistical significance (α) and power ($1-\beta$) (see Figure 1). The value of the statistical power was set at 0.80. Thus, the probability of detecting the minimum differences shown in Table 4 in a one-way ANOVA statistical test is 0.80. The significance level selected for these power analyses was 0.05, which corresponds to the value most commonly selected in statistical tests.

Values of the coefficient of variation are presented in Table 4 to facilitate comparison of the power analysis results. These data show that for an increase in the value of this measure of unexplained variability, there is a corresponding increase in the minimum detectable difference among sampling stations. This relationship can be seen by examining the results for the smallest and largest values of the coefficient. The smallest value of this measure of variability was obtained for Data Set 17 (Table 4). For the zinc concentration observed in the liver tissue of winter flounder, the computed value of the coefficient of variation is 14.0. Results of the power analyses obtained for this data set indicate that the minimum difference in the zinc concentration that can be detected with four replicate samples at four and eight sampling stations is 39 and 43 percent of the overall mean value, respectively. The largest value for the coefficient of variation (268.4) was obtained for the concentration of total PCBs in the liver tissue of Dover sole (Data Set 7, Table 4). Results of the power analyses obtained for this data set indicate that with four replicate samples at four sampling stations the minimum difference in values of tissue concentration that can be detected among stations is approximately 7 times as great as the overall mean concentration. With four replicate samples at eight stations, this minimum detectable value is over 8 times as great as the observed mean tissue concentration.

Analytical results from all 23 data sets (Table 4) are summarized in Figure 3. For each value of the coefficient of variation, the corresponding minimum difference, expressed as a percentage of the mean, that can be detected with five replicate samples at eight stations is plotted. The data in Figure 3 show that the increase in the minimum detectable difference among sampling stations is a linear function of the coefficient of variation.

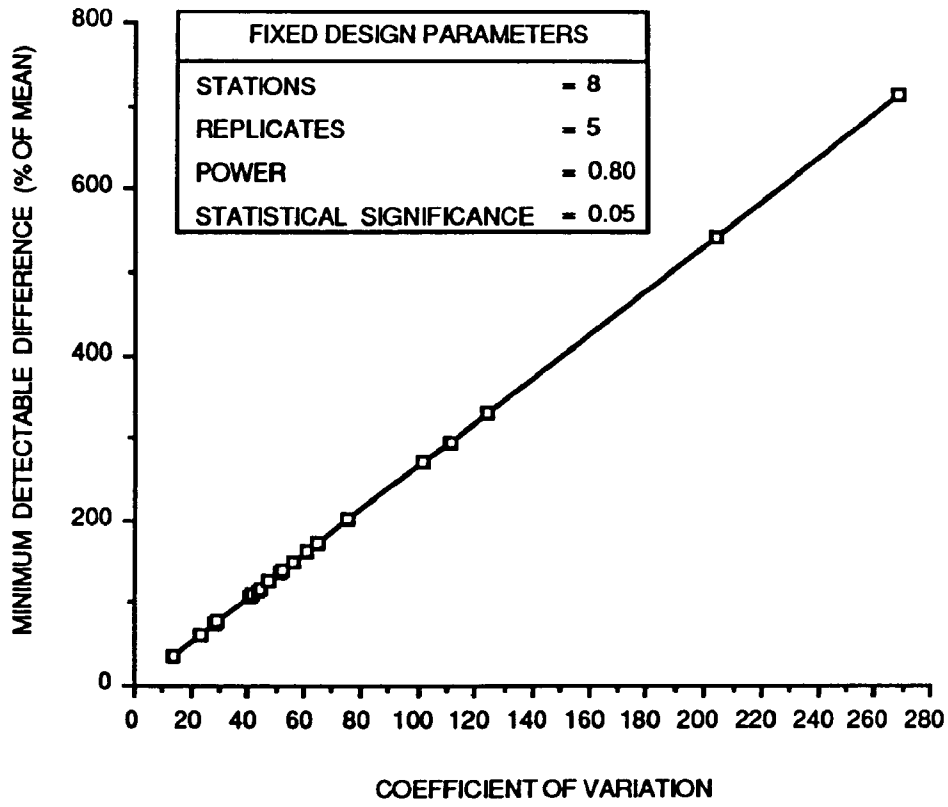


Figure 3. Minimum detectable difference among sampling stations as a function of the coefficient of variation.

From Figures 2 and 3, it can be seen that the greatest proportion of the values of the coefficient of variation for the historical data sets fall within the range of 40-60. Therefore, examination of the power analysis results for data sets with coefficients of variation between 40 and 60 provides an estimate of the expected performance of bioaccumulation monitoring programs. For example, the calculated value of the coefficient of variation for Data Set 2 (mercury, lobster) shown in Table 4 is 40.6. Results of the power analysis for these data indicate that with five replicate samples at either four or eight stations, the difference in the concentration of mercury that could be detected between stations is between 0.206 and 0.232 mg/kg. This minimum detectable difference is approximately equal to the overall observed mean concentration (0.215 mg/kg) of mercury among sampling stations. Data Set 23 (mercury, crab), on the other hand, represents conditions at the other end of this range. In this case, the minimum detectable difference in the concentration of mercury in the muscle tissue of Cancer spp. with the collection of five replicate samples is 144 percent and 163 percent of the mean value for four and eight stations, respectively. Thus, in the majority of the data sets evaluated, the observed coefficient of variation is between 40 and 60, and the collection of five replicates at eight or fewer stations resulted in the ability to detect differences in tissue concentrations of contaminants less than or equal to 163 percent of the overall mean contaminant concentration.

Additional power analyses presented in Figures 4 and 5 were conducted to summarize the results in Table 4 and to demonstrate the importance of the level of unexplained variability, represented by the residual error variance design parameter, in determining the expected performance of a monitoring program. Specifically, these analyses demonstrate the effect of increased levels of unexplained variance on 1) the ability to detect a specified difference between stations and 2) the relative effect of increased numbers of stations on the minimum detectable difference. These analyses were conducted for four levels of unexplained variability. Coefficients of variation were set at 30, 50, 70, and 90. The number of sampling stations was set at 4, 6, 8, and 16. As with the previous analyses presented in Table 4, all calculations were conducted for fixed levels of power (0.8) and statistical significance (0.05).

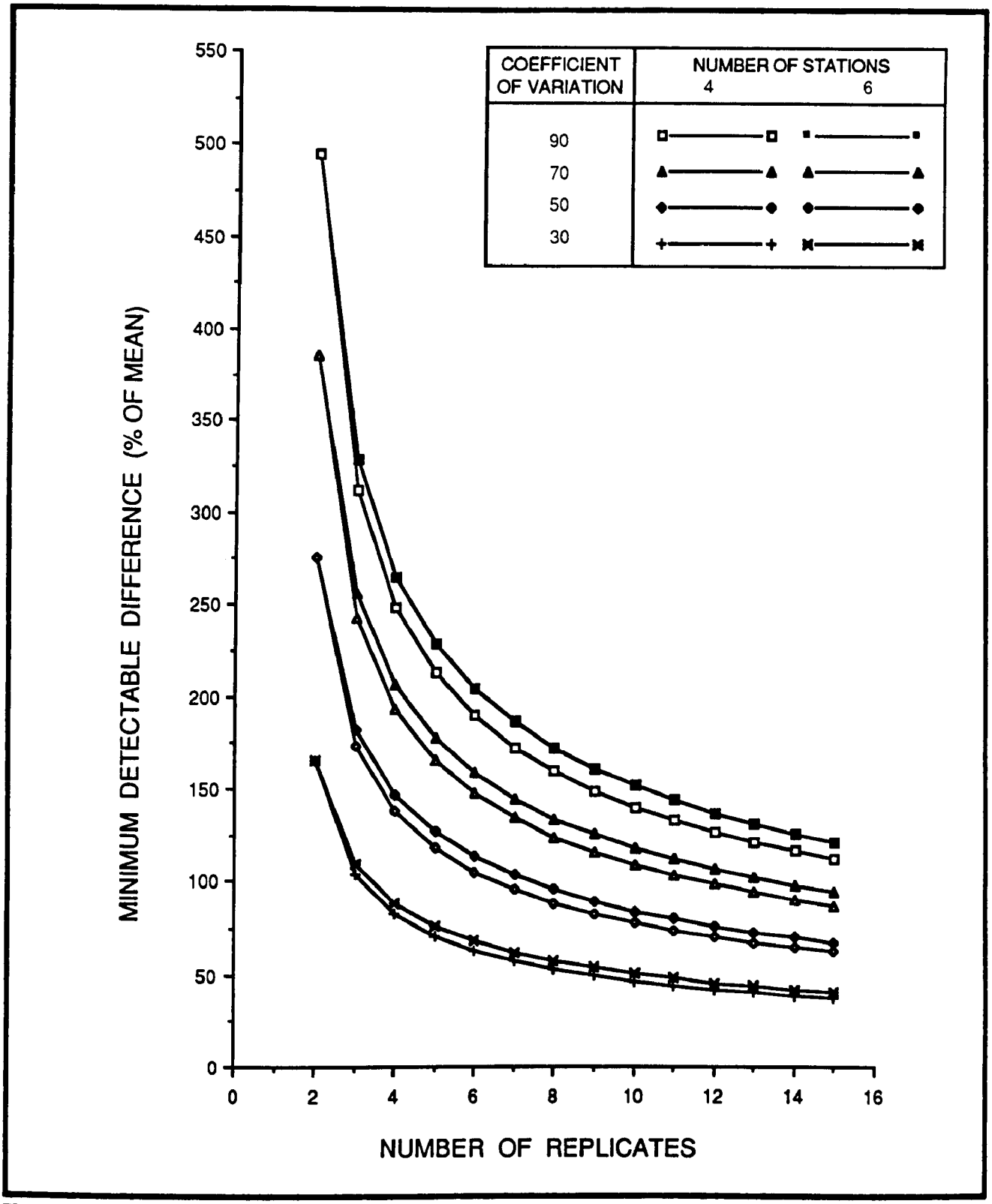


Figure 4. Minimum detectable difference vs. number of replicates at selected levels of unexplained variance for 4 and 6 stations. Power of test = 0.80, significance level = 0.05.

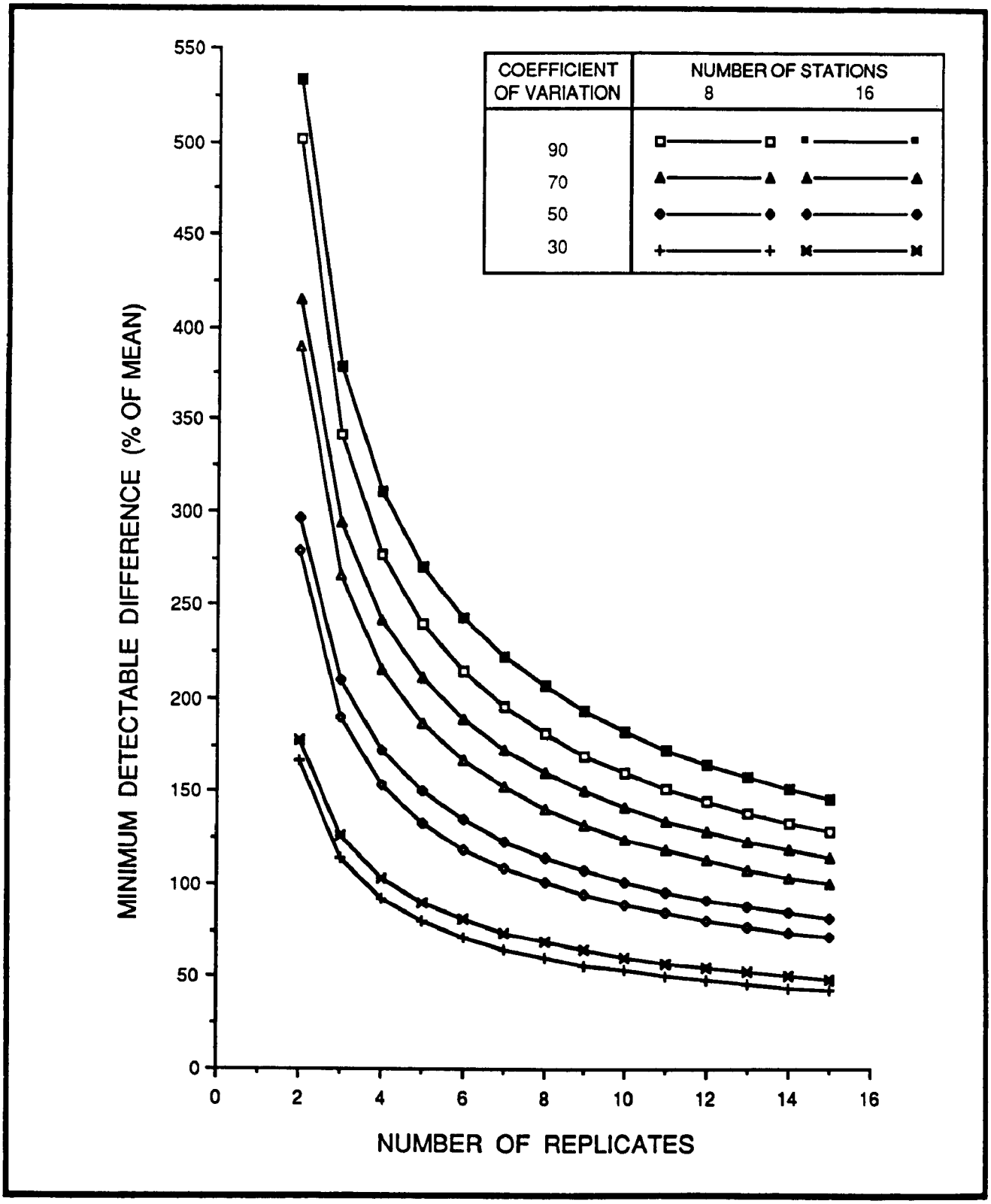


Figure 5. Minimum detectable difference vs. number of replicates at selected levels of unexplained variance for 8 and 16 stations. Power of test = 0.80, significance level = 0.05.

Results of these analyses, like those presented in Table 4, have general applicability, because the minimum detectable difference is expressed as a percentage of the mean. Additionally, the power curves are presented for coefficients of variation representing a wide range of unexplained variability in the sampling environment. These curves can be used to evaluate monitoring program performance for sampling designs using 4-16 stations and for sampling data exhibiting coefficients of variation between 30 and 90. This range includes the majority of historical data sets compiled for this study.

These results show that for an increase in the level of unexplained variance, the minimum detectable difference between sampling stations increases. For example, in Figure 4 it can be seen that with five replicates at four stations the minimum detectable difference between stations ranges from approximately 70 percent of the mean for a coefficient of variation of 30 to 212 percent of the mean for a coefficient of variation of 90. Correspondingly, both figures show that as the level of unexplained variance increases, a greater level of sample replication is required to detect a specified level of difference. For example, in a sample design with four sampling stations (Figure 4), the number of replicate samples required to detect a difference between stations equal to the mean is 3, 7, 12, and about 17, respectively, for coefficients of variation of 30, 50, 70, and 90.

Results of these analyses also demonstrate that for a fixed level of sample variability, the minimum detectable difference between stations increases as the number of stations increases. This increase is small, however, compared to the effect of increased variability in the sampling environment. For example, in Figure 5, for a coefficient of variation equal to 30, the minimum difference detectable with five replicates is approximately 80 and 90 percent of the mean for 8 and 16 stations, respectively. In general, monitoring program performance, measured by the ability to detect specified differences among stations, is increased for a fixed level of sampling effort by the collection of more replicates at fewer stations. However, the effect of number of stations on program performance is small relative to that of the number of replicate samples.

In Table 4 and Figures 3 through 5, the minimum detectable difference is expressed in terms of a percentage of the overall mean. As indicated, this provides a direct basis for the comparison of results among data sets, and the results presented in Figures 4 and 5 allow a quick evaluation of the expected performance of a large number of study designs. However, in many monitoring programs, there may be an interest not only in the relative change in contaminant concentrations among sampling locations, but also in the minimum value of the contaminant concentration that can be detected. In fact, results of power analyses used to evaluate individual monitoring program design are generally expressed in terms of the measured units.

In Figure 6, results of power analyses are shown for selected data sets presented in Table 4. In each of the four plots shown, the minimum detectable difference in the concentration of a selected contaminant is shown as a function of the number of sample replicates. Additionally, the mean concentration of the particular contaminant in each example case is shown to indicate the relative performance of monitoring programs at the different levels of unexplained variability.

Power analyses conducted with Data Set 11 (Table 4) are shown in Figure 6a. The number of replicates required to detect a specified difference in the concentration of total PCBs in the muscle tissue of winter flounder between stations is shown. These data are characterized by a relatively low coefficient of variation (29.1), indicating a low level of unexplained variation. As a result, small differences in the contaminant concentration can be detected with low levels of sample replication. Four or more replicates will provide adequate replication to detect differences of approximately 0.09 mg/kg of the contaminant in muscle tissue.

The results of power analyses conducted with data collected from sampling environments exhibiting increasing levels of unaccountable variation are shown in Figures 6b, c, and d. The calculated coefficients of variation for these data sets are 40.6, 64.9, and 101.6, respectively. As the plots indicate, successive increases in the coefficient of variation are accompanied by a decrease in the ability to detect differences relative to the

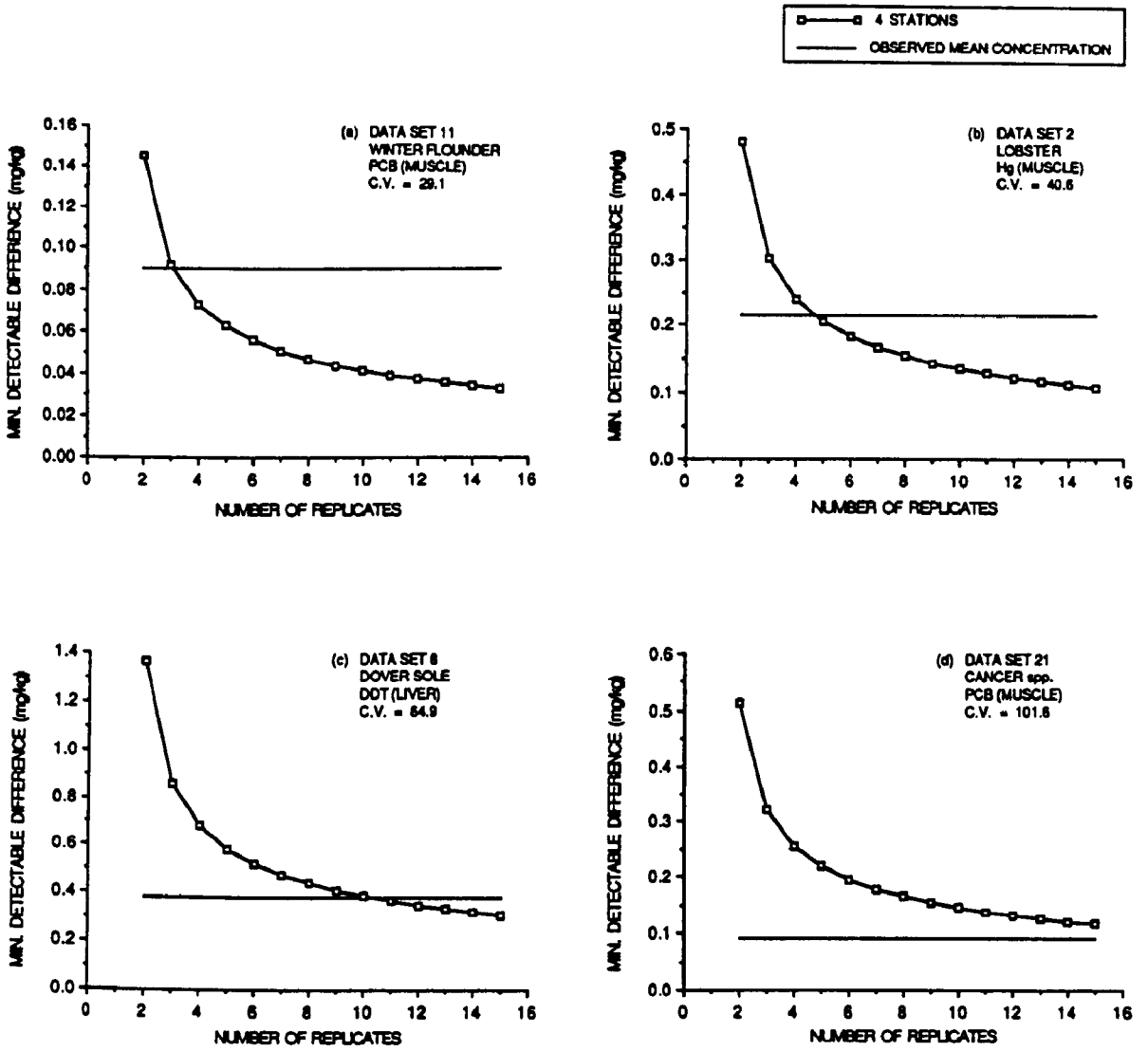


Figure 6. Minimum detectable difference in the tissue concentration of selected contaminants vs. number of replicates.

overall mean concentration among stations. However, in Figures 6b and 6d, the contaminant concentrations that can be detected at comparable levels of sampling effort are similar. Likewise, from Figure 6c, 10 replicates at each station are required to detect a difference approximately equal to the overall mean among stations. In comparison, even with 15 replicate samples at each station, the minimum detectable difference in the contaminant concentration is greater than the overall mean in Figure 6d. However, the values of the minimum detectable differences in terms of the contaminant concentration corresponding to 10 replicates at each station are 0.39 mg/kg (Figure 6c) and 0.16 mg/kg (Figure 6d). Thus, while the minimum detectable difference in terms of a percentage of the overall mean is greater in one analysis (Figure 6d), the minimum detectable contaminant concentration is considerably less than that found in the other analysis (Figure 6c). These results indicate the importance of evaluating the performance of monitoring programs both in terms of the relative change in contaminant concentration that can be detected among sampling locations as well as the minimum contaminant concentration that can be detected.

Summary

1. Analyses of 23 historical data sets on tissue contaminants indicate that, with the collection of individual tissue samples, a difference of $\leq$163 percent of the mean could be detected in the majority of cases (assumes five replicates at eight or fewer sampling stations, $\alpha = 0.05$, power = 0.80).

2. Many important chemicals (e.g., PCBs in Dover sole and crabs) displayed much higher variability, however. In these cases, a similar analytical design could only detect differences of about 200-700 percent of the mean.

COMPOSITE SAMPLING STRATEGIES

The historical data sets compiled for this report (Table 2) were based on similar sampling and analytical approaches. In all cases, tissue samples were obtained from selected organisms and analyzed individually to determine the concentration of particular contaminants. This type of sample is referred to as a grab sample, since the individual tissue samples are used to provide an estimate of the contaminant concentration in the tissue of specified populations. In each data set presented, a fixed number of these individual estimates was obtained and analyzed statistically to estimate distributional parameters of the underlying population and to make statistical comparisons of these parameters among sampling sites.

An alternative to the analysis of tissue from individual organisms is the analysis of composite samples. Composite tissue sampling consists of mixing tissue samples from two or more individual organisms collected at a particular site and analyzing this mixture as a single sample. The analysis of a composite sample, therefore, provides an estimate of an average tissue concentration for the individual organisms that make up the composite sample. This composite sampling strategy is often used in effluent sampling (Schaeffer and Janardan 1978; Schaeffer et al. 1980) to estimate average concentrations of water quality variables in cases where the individual chemical analyses are expensive but the cost of collecting individual samples is relatively small. Composite sampling is also used in the collection of samples from bioaccumulation monitoring systems containing caged specimens of bivalve molluscs (Risebrough et al. 1980; Gordon et al. 1980). The collection of composite samples is also required in other cases where the tissue mass of an individual organism is insufficient for the analytical protocol. An evaluation of the appropriateness of composite sampling in bioaccumulation monitoring programs is provided below.

Composite sampling of the tissue from selected organisms represents an attempt to prepare a sample that will represent the average concentration.

If $X_1, X_2, \ldots X_n$ represent the contaminant concentration of n tissue samples from individual organisms, these samples can be mixed to obtain a single composite observation:

$$Z = \sum_{i=1}^{n} \omega_i X_i \tag{3}$$

where:

ω_i = The proportion of total sample contributed by the i^{th} component.

Rhode (1976) has shown that the expected value and variance of Z are given by:

$$E(Z) = \mu \tag{4}$$

$$Var(Z) = \frac{\sigma^2}{n} + n\sigma_\omega^2\sigma^2 \tag{5}$$

where:

μ = Population mean
σ^2 = Population variance
σ_ω^2 = Variance of the composite proportions
n = Number of samples in each composite.

If the values of ω_i in Equation (3) are equal for all i, then the numerical value of Z is equivalent to the mean of the n samples, $\bar{X}$, where:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{6}$$

In this case, for the cost of analyzing a single composite sample, an estimate of the mean of n grab samples is obtained. However, a consequence

of selecting the composite sampling strategy in the above example is the loss of information concerning individual sample variability. As shown below, the range of values (minimum and maximum concentrations) contributing to the mean concentration is not known.

The above results apply to single composite samples. However, replicate composite samples can also be used in bioaccumulation monitoring programs. The basic sample design previously described for the historical data sets involved the collection of replicate grab samples from two or more locations and the statistical comparison of the mean values among sampling locations. As an alternative to this design, replicate composite samples, each composed of tissue from several organisms, could be collected at specified sampling locations with the objective of obtaining a more accurate estimate of the true mean at each location and to increase the power of the statistical tests.

The comparison of single composite and replicate grab samples can be extended to replicate composite samples (Rhode 1976, 1979). The mean of m composite samples (Z_1, Z_2, ..., Z_m) is given by:

$$\bar{Z} = \frac{\sum_{i=1}^{m} Z_i}{m} \tag{7}$$

The expected value and variance of $\bar{Z}$ are given by:

$$E(\bar{Z}) = \mu \tag{8}$$

$$Var(\bar{Z}) = \frac{\sigma^2}{mn} + n\sigma_\omega^2\sigma^2 \tag{9}$$

The consequences of Equations 8 and 9 are pertinent to the evaluation of composite sampling strategies for bioaccumulation monitoring programs. For example, when the composites consist of samples of equal mass (i.e., the same mass tissue is taken from each organism) ($\sigma_\omega^2=0$), then:

$$\frac{\text{Var } \bar{X}}{\text{Var } \bar{Z}} = n \tag{10}$$

where:

$$\text{Var } \bar{X} = \frac{\hat{\sigma}^2}{m} \tag{11}$$

$$\text{Var } \bar{Z} = \frac{\hat{\sigma}^2}{mn} \tag{12}$$

m = Number of replicate samples (replicate or composite) used in the estimate of the population variance ($\hat{\sigma}^2$)

n = Number of samples constituting the composite sample.

Thus, from Equation 10, it can be seen that the collection of replicate composite tissue samples at specified sampling locations will result in a more efficient estimate of the mean (i.e., the variance of the mean obtained with replicate composite samples is smaller than that obtained with the collection of replicate grab samples). From Equation 9, it should also be noted that for unequal proportions of composite samples (i.e., tissue mass), the variance of the series of composite samples increases and, in extreme cases, exceeds the variance of grab samples. A table of values for the upper bound of the variance of the proportions (σ_ω^2) that lead to such an increase in composite variance is presented in Schaeffer and Janardan (1978). However, these tabulated values are extremely high when compared with expected values of σ_ω^2 associated with preparing tissue-sample composites. For example, using the Dirichlet model for compositing probabilities, Rhode (1979) gave:

$$\frac{\text{Var } \bar{X}}{\text{Var } \bar{Z}} = \frac{n+1}{2} \tag{13}$$

as the increase in precision that can be achieved at the additional cost of the compositing process. For the analyses presented below, it was assumed

that the composite samples consist of individual samples of equal proportions and therefore that $\sigma_\omega^2=0$.

POWER ANALYSES FOR COMPOSITE SAMPLES

Analytical Methods

Historical data that could be used to evaluate the applicability of composite sampling in bioaccumulation monitoring programs were not available. Instead, simulation methods were used to make a direct comparison of grab-and composite-sampling strategies. Simulation refers to the use of numerical techniques to generate random variables with specified statistical properties. For the analyses described below, computer programs were written 1) to produce individual random samples from populations with statistical properties similar to those of the historical data described in Table 2 and Figure 2 and 2) to construct composite samples.

The algorithms used to generate the individual random samples are described in Rubinstein (1981). All algorithms used required the generation of independent random variables uniformly distributed over the interval 0, 1. The program developed to perform these simulations used the congruential method described by Lewis et al. (1969). Normally distributed variables were generated using the approach developed by Box and Muller (1958).

Two sets of analyses are described below. In the first set, simulation methods were used to show the effect of sample compositing on the estimate of the population mean. Power analyses were used in the second set of analyses to demonstrate the effect of increasing the number of samples in a composite sample on the probability of detecting specified levels of differences among stations.

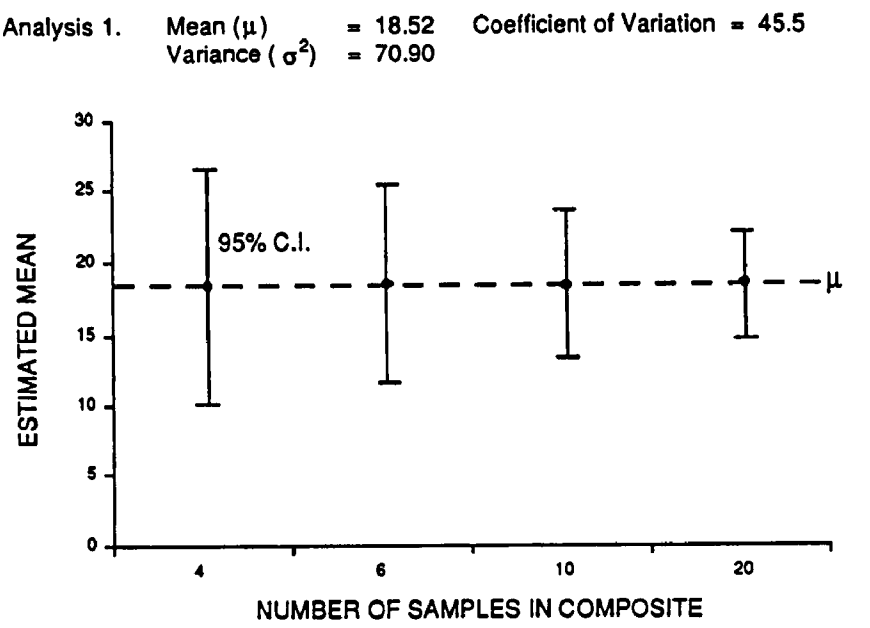
Simulation Analyses

The first set of analyses was conducted to demonstrate the effect of increasing the number of individual samples in the composite on the estimate

of the mean. The simulated sampling consisted of randomly selecting 10,000 composite samples from two populations exhibiting two different levels of variability in the sampling environment. The mean value in both populations was fixed at 18.52, but the population variances were set at 70.90 or 354.19, corresponding to coefficients of variation of 45.5 and 101.6, respectively. These population characteristics were selected as representative of the range of values for the coefficient of variation observed in the historical data sets (Table 2 and Figure 2). Coefficients of variation of 40-50 percent were measured in several historical data sets for metals, including American lobster muscle (mercury), Dover sole liver (silver and cadmium), and English sole muscle (lead and mercury). Coefficients of variation of approximately 100 percent were observed for arsenic in English sole muscle and for lead and PCBs in Cancer spp.

To demonstrate the effect of increasing the number of samples constituting the composite sample, the sample variance and the range of observed values were recorded in each experiment. The results of these experiments are summarized in Figure 7 and Table 5. A graphic display of the increase in the ability to estimate the population mean obtained by increasing the number of individual samples in composite samples is provided in Figure 7. The 95 percent confidence intervals shown in Figure 7 represent the range within which 95 percent of all samples in the simulation experiments fell. As the number of individual samples per composite increased, the observed range of mean values decreased substantially.

The actual values obtained at the boundaries of these confidence intervals shown in Figure 7 (minimum and maximum values) are presented in Table 5. In Analysis 1, sampling was conducted from a normal population with a mean of 18.52 and a variance of 70.90. Therefore, 95 percent of all values in this population ranged from approximately 1.7 to 35.4. This range (33.7) would be expected from randomly selecting a large number of individual samples from the specified population. For the same specified population, composite sampling resulted in a much more precise estimate of the population mean. With four individual samples in each composite, 95 percent of all values obtained from the 10,000 simulated samples were between 10.4 and 26.7. This range (16.3) is approximately 50 percent of the range that would



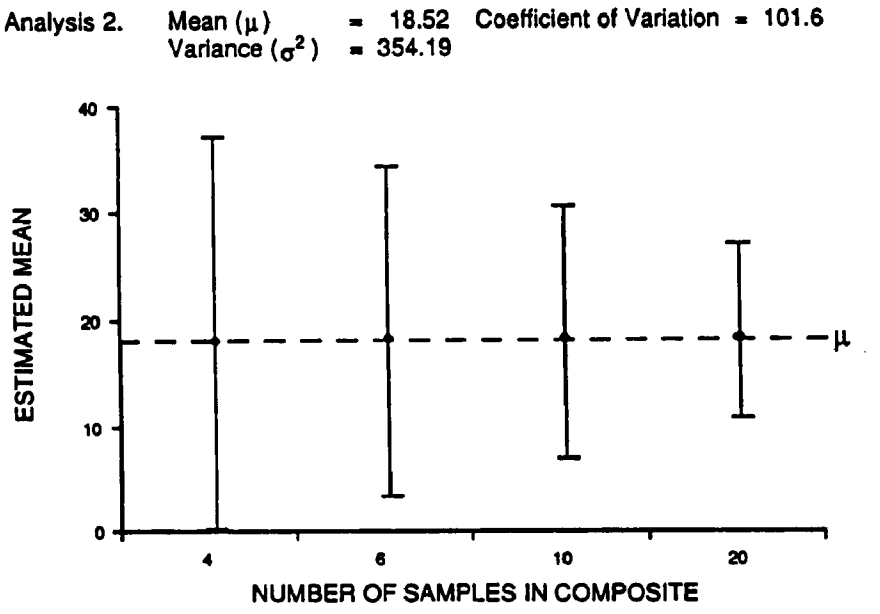


Figure 7. Effects of increasing composite sample size on estimate of the mean.

TABLE 5. RESULTS OF SIMULATION ANALYSES DEMONSTRATING THE EFFECT OF COMPOSITE SAMPLING ON THE ESTIMATE OF THE POPULATION MEAN

Analyses 1. Mean (μ) = 18.52 Coefficient of Variation = 45.5
Variance (σ^2) = 70.90

Number of Samples in Composite	Observed Variance	95 Percent Confidence Interval		
		Minimum Value	Maximum Value	Observed Range
4	17.29	10.4	26.7	16.3
6	12.31	11.6	25.4	13.8
10	7.02	13.3	23.7	10.4
20	3.49	14.9	22.2	7.3

Analyses 2. Mean (μ) = 18.52 Coefficient of Variation = 101.6
Variance (σ^2) = 354.19

Number of Samples in Composite	Observed Variance	95 Percent Confidence Interval		
		Minimum Value	Maximum Value	Observed Range
4	88.29	0.1	36.9	36.8
6	60.34	3.3	33.7	30.4
10	35.57	6.8	30.2	23.4
20	16.73	10.5	26.5	16.0

be obtained with the collection of a similarly large number of individual grab samples from this population. Furthermore, an increase in the number of samples from 4 to 20 in each composite sample decreased the range of values that define this 95 percent confidence interval by approximately another 50 percent.

Similar results were obtained in the second experiment that involved simulated sampling from a population with the same mean value (18.52), but with the variance increased to 354.2. The simulated collection of composite samples, each consisting of four individual samples from the population, resulted in reduction in the range of values in the 95 percent confidence interval by a factor of 2. A similar reduction in this range was obtained by increasing the number of samples in the composite to 20.

Power Analyses

The results of power analyses conducted with the historical data sets of individual grab samples are presented above in Figures 4 through 6. In these analyses, the minimum detectable difference between sampling stations was shown as a function of the number of replicate grab samples at each station. These results are shown for specified sets of design parameters (i.e., number of stations, significance level of the test, residual error variance, and power of the test). To demonstrate the effect of sample compositing on the power of the statistical test of significance, additional power analyses were conducted. In these analyses, the number of stations (5), number of replicate samples at each station (5), significance level of the test (0.05), residual error variance level, and level of minimum detectable difference were fixed. The power of the test or probability of detecting the specified minimum difference was then calculated as a function of the number of individual samples constituting each replicate composite sample.

Power analyses were conducted for three levels of sample variability. All design parameters except the residual error variance were identical in each set of analyses. Values of the residual error variance were selected to represent the range of values found in the historical data sets (Table 2

and Figure 2). The coefficients of variation selected for these three sets of analyses were 45.5, 101.6, and 203.5. The highest level of variability (coefficient of variation = 203.5) is equal to that measured in Dover sole for DDT concentrations in muscle tissue.

The results of the power analyses conducted at the two lower levels of sample variability are shown in Figure 8a. In each analysis, the probability of statistically detecting a difference equal to the overall sample mean among stations increases with the collection of replicate composite samples at each station and as the number of samples constituting the composite increases. For example, in Analysis 1 (Figure 8a), conducted at the lowest level of sample variability (coefficient of variation = 45.5), the probability of detecting the specified difference among stations with five replicate grab samples (i.e., number of samples = 1) is 0.70. With the collection of five replicate composite samples, each composed of two individual samples at each station, the power of the test increases to 0.96, and with four or more samples per composite, the detection of the specified difference between stations is virtually assured.

The benefits of the composite sampling strategy are more apparent from the analysis conducted at the intermediate level of sample variability (Analysis 2, Figure 8a). The probability of statistically detecting the specified difference among stations with the collection of five individual grab samples (number of samples = 1) is only 0.17. The power of the test increases to 0.59 with the collection of 5 replicate composite samples, with each composite composed of 4 samples in equal proportions, to 0.90 with 8 samples per composite, and to 0.96 with 10 samples per composite.

The results of both sets of analyses shown in Figure 8a also demonstrate the phenomenon of diminishing returns for continued increases in the number of samples per composite. In Analysis Set 1, for example, virtually no increase in the power of the statistical test was achieved with increasing the individual sample size above three. In the second analysis set, substantial increases in statistical power were achieved by increasing the number of samples in each composite from 2 to 10. However, with each successive increase in sample size, the relative benefit was reduced until

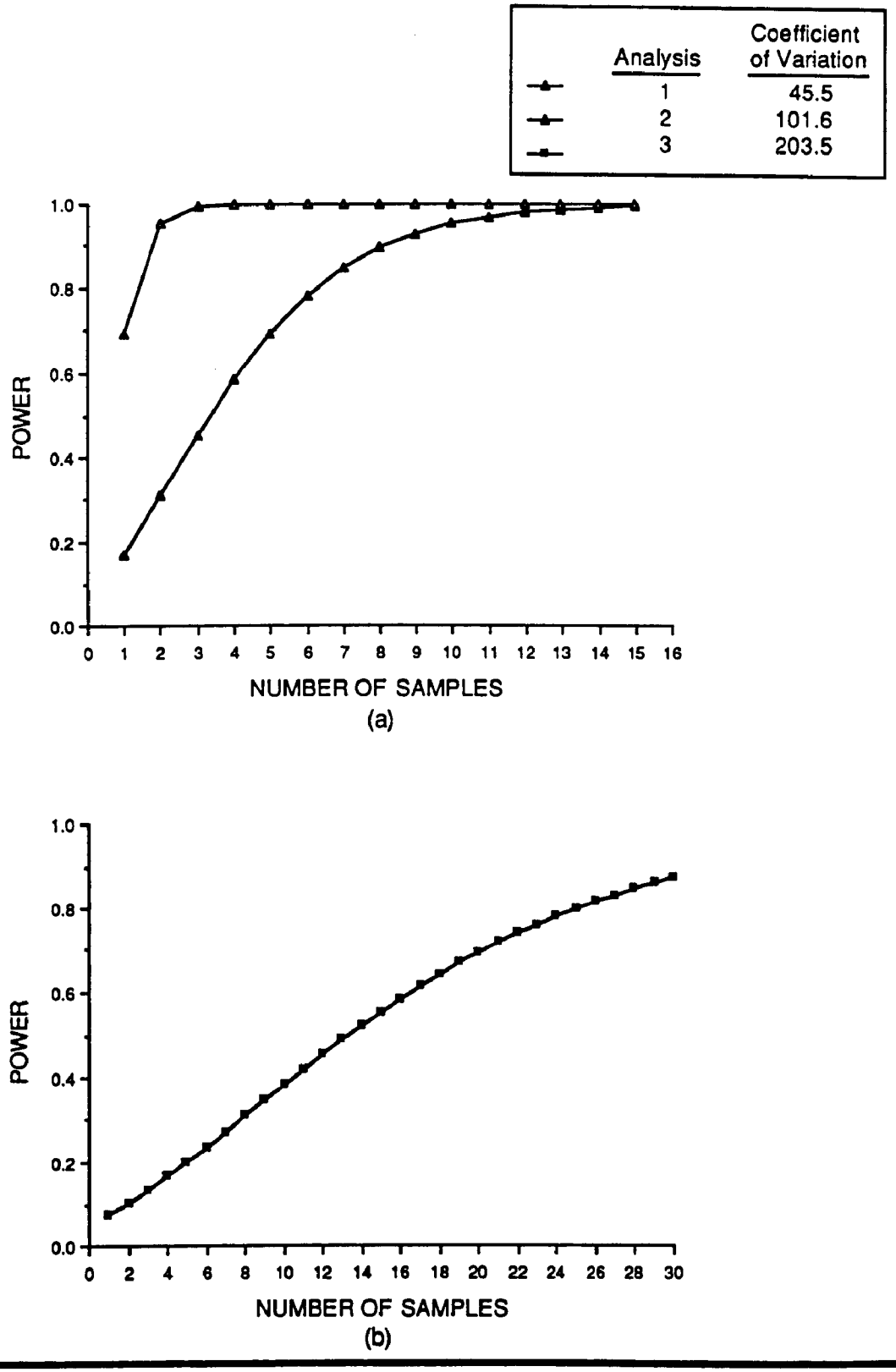


Figure 8. Power of statistical tests vs. number of samples in composite replicate samples. Fixed design parameters: number of stations = 5, number of replicates = 5, significance level = 0.05, minimum detectable difference = 100 percent of overall mean value.

very little was gained by increasing the sample size above 10. This phenomenon is analogous to that observed in the results of the power analyses presented in Figures 4 through 6. In these previous analyses, the benefit achieved in the minimum detectable difference between stations also decreased with the addition of each successive replicate grab sample. However, the main difference is that while the cost of collecting and processing additional replicate samples (grab or composite) is substantial, the cost of collecting additional samples for each composite replicate sample is relatively small.

The results of the power analyses conducted at the highest level of sample variability selected are shown in Figure 8b. These results are shown separately from the first two sets of analyses to include a larger range of values for the number of samples in each composite. These results are directly comparable, however, and provide additional evidence of the increased power obtained by the collection of replicate composite samples. These analyses indicate the relatively low statistical power associated with samples displaying a high level of natural variability. Under these conditions (coefficient of variation = 203.5), the probability of detecting a difference among stations equal to the overall mean with the collection of five individual replicate grab samples is 0.08. The power of the test is doubled with the collection of replicate composite samples composed of four samples (power = 0.17). With the collection of 10 samples per replicate composite sample, the power is increased to 0.38. However, given the high level of background variability, the collection of replicate composite samples composed of 25 individual samples each is required to obtain a testing power of 0.80. Additionally, due to the diminishing returns associated with increasing the number of samples per composite, the collection of replicate composite samples consisting of 32 samples each is required to obtain a statistical power of approximately 0.90.

A final set of power analyses was conducted to provide a direct comparison of grab-sampling and composite-sampling strategies. The number of stations (5), significance level of the test (0.05), and residual error variance level (coefficient of variation = 101.6) were fixed in all analyses. In individual analyses, the probability of detecting specified

minimum differences between stations was determined for selected numbers of replicate grab samples and for a fixed number of replicate composite samples at each station composed of selected numbers of individual samples. The results of these analyses are summarized in Table 6.

Results of the first three analyses shown in Table 6 demonstrate the effect of increasing the number of replicate grab samples on the ability to detect statistically significant differences between sampling stations. For example, the probability of detecting a difference equal to the overall mean among stations is increased from 0.17 to 0.35 by increasing the number of replicate grab samples at each of the five stations from 5 (Analysis 1) to 10 (Analysis 3). These first three results demonstrate from a different perspective what was previously shown in Figures 4 through 6--that an increase in the number of replicate samples is accompanied by an increase in the ability to identify differences among sampling stations. The results of Analyses 4 through 6 presented in Table 6 demonstrate the effect of sample compositing on the ability to detect differences between stations. These analyses were conducted for five replicate composite samples at each station and various numbers of individual samples per composite. These results, therefore, are directly comparable to those provided in Analysis 1 for the collection of five replicate grab samples at each station. Comparison of the probability of detecting a difference between stations equal to the overall mean in Analyses 1, 4, 5, and 6 indicates that a substantial increase in the power of the statistical test can be achieved by the collection of replicate composite samples. These analyses demonstrate that the collection of five replicate composite samples each consisting of four samples will increase the power to 0.59. The power is further increased to 0.96 by increasing the samples in each composite sample to 10.

Summary

1. Based on simulation results for a given number of samples, composite sampling results in a much more precise estimate of the mean than analysis of grab samples.

TABLE 6. PROBABILITY OF DETECTING SPECIFIED LEVELS OF MINIMUM DETECTABLE DIFFERENCES FOR SELECTED GRAB-SAMPLING AND COMPOSITE-SAMPLING STRATEGIES.
FIXED DESIGN PARAMETERS: NUMBER OF STATIONS = 5, SIGNIFICANCE LEVEL = 0.05, COEFFICIENT OF VARIATION = 101.6

Analysis Number	Number of Replicate Samples at Each Station	Number of Samples in Composite	Minimum Detectable Difference (expressed as a proportion of the overall mean)				
			0.25	0.50	1.0	1.5	2.0
			Corresponding Probability of Detection				
1	5	1 (grab sample)	0.06	0.08	0.17	0.35	0.59
2	8	1 (grab sample)	0.06	0.10	0.27	0.58	0.85
3	10	1 (grab sample)	0.07	0.11	0.35	0.71	0.94
4	5	4	0.08	0.17	0.59	0.93	1.00
5	5	8	0.10	0.31	0.90	1.00	1.00
6	5	10	0.11	0.38	0.96	1.00	1.00

2. The precision of the estimated mean increases with increasing numbers of individual samples constituting a composite sample.

3. Because of reduced sample variance, composite sampling results in a considerable increase in statistical power over grab sampling (for a given number of samples analyzed).

4. For most contaminants, the collection of six to eight samples per composite results in adequate statistical power, with little relative gain in power for additional samples.

SUMMARY AND RECOMMENDATIONS

This document describes the use of power analyses in designing 301(h) bioaccumulation monitoring programs and provides evaluations of alternative sampling strategies. These methods can be used to evaluate alternative designs on the basis of the level of sampling effort required to obtain desired levels of precision. For example, existing data can be analyzed to determine the minimum differences in contaminant concentrations that can be detected for selected levels of sample replication. The probability of detecting specific levels of differences in tissue contaminant concentrations for alternative sampling designs can also be determined.

The example analyses presented in this report were conducted on the Ocean Data Evaluation System (ODES) using the statistical power analysis tool. The ODES Power Analysis Tool can be used to assess bioaccumulation monitoring programs from two perspectives. In the monitoring program design phase, these techniques can be used in a prospective manner to evaluate alternative design parameters such as numbers of samples and sampling stations. The techniques can also be used retrospectively when monitoring data are available to evaluate overall monitoring program performance. For example, if a greater statistical power was desired for future data, the relative benefits in power of increasing numbers of replicate samples could be evaluated relative to increased program costs.

The use of power analyses in designing bioaccumulation studies was demonstrated with both historical and simulated data. Twenty-three historical data sets were compiled from published reports and analyzed. Data were obtained for a total of five common marine species, three body tissues, and measured values of nine contaminants. These data encompassed a wide range of sample variability, and the results of the analyses conducted provide an indication of the approximate levels of statistical power that can be achieved with the collection of replicate grab samples at selected sampling locations. Simulation techniques, sometimes referred to as Monte

Carlo techniques, were used to produce data from specified sampling distributions with fixed parameters. These data were essential for the evaluation of grab- vs. composite-sampling strategies because equivalent historical data for composite samples were not available.

In addition to the description and demonstration of power analysis techniques, a primary objective of this document was to evaluate composite- vs. grab-sampling strategies. Based on the results presented in this report, the collection of replicate composite samples is recommended for most bioaccumulation monitoring programs. The results of the analyses using simulated data demonstrated that the collection of replicate composite tissue samples at selected sampling locations provides a better estimate of the population mean. The results of power analyses using these simulated data also demonstrate that the corresponding decrease in the sample variance that is achieved with composite sampling leads to an increase in the power of statistical tests. For example, with an overall coefficient of variation of approximately 100, the analyses demonstrated that the probability of detecting a difference in tissue contaminant concentrations equal to the overall mean among 5 stations was 0.17 with the collection of 5 replicate grab samples at each station and 0.59, 0.90, and 0.96 with the collection of 5 replicate composite samples consisting of 4, 8, and 10 samples, respectively.

Based on the levels of variability in the measurements of tissue contaminant concentrations that were identified in the historical data sets and the results of analyses presented in this report, it was concluded that the collection of replicate composite samples makes it feasible to distinguish elevated tissue concentrations of contaminants between sampling locations. However, the selection of the appropriate numbers of replicate composite samples and numbers of samples per replicate will depend on site-specific levels of sample variability in the tissues and contaminants of concern. When these kinds of historical data are available for a particular study site, the tools demonstrated in this document can be used to make quantitative comparisons of alternative sampling designs and to select the appropriate level of sampling effort. Where historical data are not available, pilot studies may be conducted to estimate the level of

expected variability in contaminant concentrations for selected species and tissues. Alternatively, the observed level of variability in tissue concentrations for selected contaminants and species at other locations could be used to estimate sample variability. Where these data cannot be obtained, the collection of five replicate composite samples, each consisting of equal amounts of tissue from six individual organisms, is recommended. The selection of these design parameters assumes a coefficient of variation calculated among stations of approximately 100 percent and will result in the ability to detect a difference between stations equal to the overall mean concentration among stations with an estimated probability of detection (power of the statistical test) equal to 0.80. Note that most data sets reviewed herein had coefficients of variation less than 100. Therefore, the recommended design will generally result in either a lower detectable difference or higher power than stated above. If this general design specification is used, power analyses should be conducted after site-specific data are available to evaluate the exact probability of detecting specific levels of differences between stations.

The objectives of a bioaccumulation monitoring program should be evaluated prior to selecting a sampling strategy. Composite sampling methods are appropriate for monitoring programs that have as a primary objective the determination of differences in contaminant tissue concentrations among sampling stations. However, as shown in Equations 9 and 11, the variance of composite samples is substantially less than the population variance. As a consequence, the range of values obtained from composite samples is not representative of the true range of tissue concentrations in individual organisms of the sampled population. These results demonstrate that composite sampling will not detect extreme values. Therefore, bioaccumulation monitoring programs using a composite sampling strategy may not detect the existence of tissue concentrations that exceed legal limits or action levels for contaminants in fish tissues. For example, from the historical data compiled for this study, the mean concentration of total PCBs in Dover sole muscle tissue (Data Set 4, Table 3) was 0.766 ppm. However, 2 of the 36 values in this data set exceed the U.S. Food and Drug Administration legal limit of 2.0 ppm (U.S. Food and Drug Administration 1984). These values would not have been detected in a monitoring program based on the

collection of composite samples. Therefore, if the objective of a bioaccumulation study is to determine compliance with specified tissue concentration limits, the program should include the collection of tissue samples from individual organisms. This objective could be accomplished by two different monitoring strategies. In the first, the entire study could be designed to collect replicate grab samples. In this design there would be a much lower statistical power to detect among-station differences than could be accomplished with a composite sampling strategy using the same number of samples. Alternatively, the program could be designed to collect replicate composite samples at all sampling stations with the collection of supplemental individual tissue samples at areas of expected high tissue concentrations. This program design would enable a more efficient assessment of among-station differences in addition to providing an assessment of regulatory compliance in specified areas of concern.

A primary objective of 301(h) bioaccumulation monitoring programs is to determine whether the discharge causes an increase in the body burden of toxic chemicals in indigenous organisms. Monitoring programs may also use caged molluscs as sentinel organisms to evaluate uptake of toxic pollutants. For these kinds of studies on indigenous or transplanted organisms, a composite sampling strategy is recommended. Evaluation of effects on recreational and commercial fisheries is another important component of some 301(h) monitoring programs. Where such fisheries are included in the assessment, it is important to document whether tissue contaminant levels exceed applicable criteria or standards. In these cases, the 301(h) monitoring program may contain the dual objectives discussed previously. Based on the statistical evaluations conducted herein, it is recommended that such dual-objective programs be designed to collect composite tissue samples at all sampling stations, and that they also include the collection of individual tissue samples for commercial and recreational species in selected areas.

An additional concern relative to selecting between analysis of individual organisms and composite samples involves an evaluation of the numbers of organisms required for collection. As stated previously, the analytical costs associated with processing grab samples and composite

samples are essentially equal. Overall cost differences between the two strategies are associated with the additional time needed to collect organisms and process tissues for composite sampling. For the general composite sampling design recommended above, 30 organisms would be required at each sampling station. For a comparable design involving analysis of individual organisms, only five organisms would be required at each sampling station. Therefore, the decision on appropriate sampling strategy should also involve an assessment of the feasibility of collecting the required numbers of organisms.

REFERENCES

Andrews, F.C. 1954. Asymptotic behavior of some rank tests of analysis of variance. Ann. Math. Stat. 25:724-736.

Box, G.E.P., and M.E. Muller. 1958. A note on the generation of random normal deviates. Ann. Math. Stat. 29:610-611.

Cohen, J. 1977. Statistical power analysis for the behavioral sciences. Academic Press, New York, NY.

Gordon, M., G.A. Knauer, and J.H. Martin. 1980. Mytilus californianus as a bioindicator of trace metal pollution: variability and statistical considerations. Mar. Pollut. Bull. 11:195-198.

Grieb, T.M. 1985. Robustness of the analysis of variance in environmental monitoring applications. Report EA 4015. Electric Power Research Institute, Palo Alto, CA. 72 pp.

Kruskal, W.H., and W.A. Wallis. 1952. Use of ranks in one-criterion variance analysis. J. Am. Statist. Assoc. 47:583-612.

Lehmann, E.L. 1975. Nonparametrics: statistical methods based on ranks. Holden-Day, Inc., San Francisco, CA. 457 pp.

Lewis, P.A.W., A.S. Goodman, and J.M. Miller. 1969. A pseudo-random number generator for the System/360. IBM Syst. J. 8:199-200.

Rhode, C.A. 1976. Composite sampling. Biometrics 32:278-282.

Rhode, C.A. 1979. Batch, bulk, and composite sampling. pp. 365-377. In: Sampling Biological Populations. R.M. Cormack et al. (eds). International Co-operative Publishing House, Fainand, MD.

Risebrough, R.W., B.W. deLappe, E.F. Letterman, J.L. Lane, M. Firestone-Gillis, A.M. Springer, and W. Walker II. 1980. California mussel watch: 1977-1978. Volume III - Organic Pollutants in Mussels, Mytilus californianus and M. edulis along the California Coast. Water Quality Monitoring Report No. 79-22. Prepared by Bodega Marine Laboratory, Bodega Bay, CA, for California State Water Resources Control Board, Sacramento, CA. 109 pp. + appendices.

Roberts, A.E., D.R. Hill, and E.C. Tifft. 1982. Evaluation of New York Bight lobsters for PCBs, DDT, petroleum hydrocarbons, mercury, and cadmium. Bull. Environ. Contam. Toxicol. 29:711-718.

Rubinstein, R.Y. 1981. Simulation and the Monte Carlo method. John Wiley and Sons, New York, NY. 278 pp.

Schaeffer, D.J., and K.G. Janardan. 1978. Theoretical comparison of grab and composite sampling programs. Biometrical J. 20:215-227.

Schaeffer, D.J., H.W. Kerster, and K.G. Janardan. 1980. Grab versus composite sampling: a primer for the manager and engineer. Environ. Manage. 4:157-163.

Scheffe, H. 1959. The analysis of variance. John Wiley and Sons, New York, NY. 477 pp.

Sherwood, M.J., A.J. Mearns, D.R. Young, B.B. McCain, R.A. Murchelano, G. Alexander, T.C. Heeson, and T.-K. Jan. 1980. A comparison of trace contaminants in diseased fishes from three areas. Southern California Coastal Water Research Project, Long Beach, CA. 131 pp.

Tetra Tech. 1985a. Bioaccumulation monitoring guidance: 1. estimating the potential for bioaccumulation of priority pollutants and 301(h) pesticides discharged into marine and estuarine waters. Final Report prepared for Marine Operations Division, Office of Marine and Estuarine Protection, U.S. Environmental Protection Agency. EPA Contract No. 68-01-6938. Tetra Tech, Inc., Bellevue, WA. 56 pp. + appendices.

Tetra Tech. 1985b. Bioaccumulation monitoring guidance: 3. recommended analytical detection limits. Final Report prepared for Marine Operations Division, Office of Marine and Estuarine Protection, U.S. Environmental Protection Agency. EPA Contract No. 68-01-6938. Tetra Tech, Inc., Bellevue, WA. 23 pp.

Tetra Tech. 1985c. Commencement Bay nearshore/tideflats remedial investigation. Final Report. Volumes 1 and 2. Prepared for Washington State Department of Ecology under Contract No. C84031. Tetra Tech, Inc., Bellevue, WA.

Tetra Tech. 1987a. Bioaccumulation monitoring guidance: 2. selection of target species and review of available bioaccumulation data, volume I. EPA 430/9-86-005. U.S. Environmental Protection Agency, Marine Operations Division, Office of Marine and Estuarine Protection, Washington, DC. 52 pp.

Tetra Tech. 1987b. Bioaccumulation monitoring guidance: 2. selection of target species and review of available bioaccumulation data, volume II: appendices. EPA 430/9-86-006. U.S. Environmental Protection Agency, Marine Operations Division, Office of Marine and Estuarine Protection, Washington, DC.

Tetra Tech. 1987c. Technical support document for ODES statistical power analysis. Final Report prepared for Marine Operations Division, Office of Marine and Estuarine Protection, U.S. Environmental Protection Agency. EPA Contract No. 68-01-6938. Tetra Tech, Inc., Bellevue, WA. 34 pp. + appendix.

Tetra Tech and American Management Systems, Inc. 1986. ODES user's guide: supplement A - description and use of Ocean Data Evaluation System (ODES) tools. Prepared for U.S. Environmental Protection Agency. Tetra Tech, Inc., Bellevue, WA.

U.S. Food and Drug Administration. 1984. Polychlorinated biphenyls (PCBs) in fish and shellfish; reduction of tolerances; final decision. U.S. FDA, Rockville, MD. Federal Register, Vol. 49, No. 100. pp. 21514-21520.