# The Application of Power Analysis to Monitoring Data: An Example from the Columbia River

**ESSA** Environmental and Social Systems Analysts Ltd.

B. RAYMOND

# The Application of Power Analysis to Monitoring Data: An Example from the Columbia River

Prepared for

Inland Waters
Environment Canada

by

Trent Berry and David R. Marmorek

ESSA Environmental and Social Systems Analysts Ltd.
3rd Floor, 1765 West 8th Avenue
Vancouver, B.C. V6J 5C6

December 23, 1992

150

*Citation:* **Berry, T. and D.R. Marmorek** 1992. The application of power analysis to monitoring data: an example from the Columbia River. Prepared by ESSA Ltd., Vancouver, B.C. for Inland Waters Directorate, Environment Canada, North Vancouver. ?? pp.

# Table of Contents

# List of Tables

**Table**



# List of Figures

**Figure**

## 1.0 Introduction

An earlier series of workshops for Environment Canada on the fate and effects of pulp mill effluent stressed the importance of proper experimental design for monitoring programs (Marmorek et al. 1992). In particular, the workshops stressed the importance of power analysis to ensure detection of effects of an ecologically significant magnitude. Following the last of these workshops, it was suggested that ESSA conduct a brief analysis of some existing data to help lay the groundwork for proper experimental design in the Fraser River Basin Assessment Program.

This report provides an overview of power analysis and its application to data from monitoring programs. Environment Canada supplied data on metal concentrations in fish and various types of sediments in the Columbia River. The question to be addressed with these data was: **What is the power of detecting spatial and temporal differences or trends?** In order to select appropriate statistical tests and to determine the power of these tests, one needs to know something about the **variability** of the underlying population being sampled. This includes variability at a site and between sites for tests regarding spatial differences; and seasonally and between years for tests of trends or changes in time.

The data provided by Environment Canada for metal concentrations in various types of sediment were insufficient in most cases to determine the level of these types of variability. The sample sizes were too small to provide an unbiased estimate of spatial variability and in many cases no estimate of temporal variability could be obtained due to a lack of time series data. Power can also be determined using a simulated data set if *a priori* estimates of these variances are available. Simulation analysis can be used to evaluate the probability that current sampling and measurement techniques can correctly detect each of several alternative levels of trend or spatial differences. Unfortunately, *a priori* variance estimates were not available for sediment metal concentrations.

For these reasons, this report focuses on the data for tissue metal levels in fish on the Columbia River (Norecol Environmental Consultants 1989). No significant trends were found in these data. On the basis of these results, Norecol recommended reducing the frequency of monitoring. The original analysis was redone for this report in order to assess the level of change which **could** have been detected from this data set given the level of intensity of sampling and an acceptable level of Type II error (i.e. the acceptance of the null hypothesis of no trend when in fact a non-zero trend existed). We also explored the level of replication required to detect more subtle levels of change.

## 2.0 Background

This section of the report establishes the conceptual background necessary for understanding the data analyses presented in section 3.

## 2.1 Experimental Design

The quality of an investigation depends in part on good experimental design. "Experimental design" refers to the logical structure of the experiment. The experimental design specifies a number of components of the study such as the manner in which treatments are assigned to the available experimental units, the number of experimental units receiving each treatment (replicates), the physical arrangement of units, and possibly the temporal sequence of treatment (Hurlbert 1984).

## 2.2 Replication

Unfortunately, the disciplines of statistics and experimental design suffer from an impoverished vocabulary. Many terms are used in different ways in different circumstances. For example, "replication" occurs at several levels such as blocks, experimental units, samples and subsamples. The actual importance of replication at each of these levels varies depending on the goal of the study and the statistical tests which will be preformed with the data. Replication is usually only obligatory at one level, the experimental unit, for statistical inferences. Similarly, "error" can refer to many different quantities or concepts simultaneously, including Type I and Type II errors, random and systematic (bias) errors introduced by the experimenter, variation among replicates (at different levels), or the discrepancy between the true and estimated population parameters.

The data on metal levels in fish and various types of sediment may be considered part of a "mensurative experiment" which involves measurements only at one or more points in space or time; thus, space or time are the only "experimental" units or "treatments" (Hurlbert 1984). Essentially, we are interested in spatial and temporal differences in metal levels. This is also sometimes referred to as a "comparative" mensurative experiment.

For example, to demonstrate a difference in time (i.e., temporal trend), it is necessary to have replicates within each time period. Replication is required to test whether there is evidence of a treatment (i.e., time) effect, as opposed to simply an artifact of spatial variability at a site whose characteristics are stable over time. Replicates within a time period will permit an estimate of spatial variability and other random errors (noise) and permit more rigorous conclusions about trends (signals) to be drawn.

Pseudoreplication is "the testing for treatment effects [in this case a temporal trend or spatial difference] with an error term inappropriate to the hypothesis being considered....In mensurative experiments generally, pseudoreplication is often a consequence of the actual *physical space* over which samples are taken or measurements made being smaller or more restricted than the *inference space* implicit in the hypothesis being tested [in this case a portion of the river]." (Hurlbert 1984, 190). Unlike manipulative experiments where treatments must be interspersed with each other in space and time, the "treatments" in mensurative experiments, by their very nature, are usually isolated from each other in either space and/or time (Hurlbert 1984).

Most mensurative experiments of this type are also complicated by the presence of spatial and temporal autocorrelation. Spatial autocorrelation arises when sample sites are within the same region and therefore experience 'treatments' such as weather patterns or precipitation chemistry. Data from rivers is typically spatially autocorrelated since flow will connect upstream sites to downstream ones. Temporal autocorrelation can occur within a site because of such influences as large climatic events which take time to attenuate. Thus, adjacent years tend to be more similar than non-adjacent years.

## 2.3 Statistical Power

One of the key considerations in the design of a monitoring program (i.e., mensurative experiment) is the statistical power of the program to detect spatial differences or changes in time. That is, if we fail to detect a trend we want to make sure that the design was adequate and that this failure was not due to an inability to discriminate signal from noise (natural variation). Ideally, designers of a monitoring program would use historical data or data from pilot studies to choose the appropriate statistical tests to apply and the power of these tests to detect a spatial difference or temporal trend.

Tests of statistical hypotheses generally have two types of error associated with them: Type I error ($\alpha$) or the probability of falsely rejecting the null hypothesis when it is indeed true, and Type II error ($\beta$) or the probability of failing to reject the null hypothesis when it is indeed false (Table 1).

**Table 1: Types of error associated with different decisions given alternative ("true") states of the world.**

| | Decision | |
|---|---|---|
| **"Truth"** | Accept $H_0$ | Reject $H_0$ |
| $H_0$ True | No Error<br>$1-\alpha$ | Type I Error<br>$\alpha$ |
| $H_0$ False | Type II Error<br>$\beta$ | No Error<br>$1-\beta$ |

Power may be thought of as the ability of a statistical test to determine if a null hypothesis is false (i.e. to detect an effect), that is ($1-\beta$) (Toft and Shea 1983). Power largely depends on three factors: the **critical values of $\alpha$** used, the **sample size**, and the **effect size**. The more stringent one's standards for avoiding Type I errors (i.e. the lower the $\alpha$), the greater the probability of committing a Type II error. The size of both errors is reduced by a larger sample size, which increases the reliability of the sample estimate. Effect size refers to the magnitude of an effect that the investigator wants to be able to detect. The larger the desired effect size (relative to the standard deviation), the more easily it can be detected.

These three pieces of information are really all that is required to determine power. Given these three parameters, the actual calculation of power will depend on the statistical test being used to test the significance of results. Thus, the calculation of power only has meaning in the context of a predetermined statistical test. Different tests are used under different assumptions about the data set (and underlying population) being analyzed. Just as each of these tests varies in their ability to detect significant results and in their robustness if assumptions are violated, so does the power of these tests vary. For example, non-parametric methods are used to reduce the sensitivity of statistical tests to violations of normality. These tests, however, are also generally less powerful in their ability to detect a given level of trend.

There are three uses of the calculation of power in the analysis presented here. The first is to provide some indication of our confidence in the conclusion of an existing monitoring program that no temporal trend exists - i.e., the probability of correctly rejecting $H_0$ of no effect (i.e., trend), if a real effect of some of a specific magnitude (i.e., trend) actually does exist (Toft and Shea 1983). It could also be used to indicate the **"detectable effect size"** or magnitude of temporal trend which could be detected with sufficient confidence under current conditions (i.e., monitoring regime and years of data) (Cohen 1988). Finally, power could be employed to determine when a desired effect or trend could be detected with sufficient power given a particular sampling regime and assumptions about the variances within the data. All of these uses provide important inputs into decisions about the conclusions which may be draw from a monitoring program or opportunities for increasing the power of a monitoring program.

The problem in the present analysis is to detect and quantify a trend or change in a time series. Trends are masked by systematic or random fluctuations in variables due to seasonal effects, spatial variations, and sampling or analytical variability (all commonly referred to as 'noise'). There are three steps in dealing with these types of variation: 1) proper experimental design to minimize variation which obscures the signal of interest; 2) data manipulation and transformation to control for systematic variation due to seasonal or spatial variability (e.g., Seasonal Kendall Test[1]), and 3) statistical tests and analyses to test the significance of changes and trends (e.g., Time Series Analysis[2]).

---

[1] The seasonal Kendall test is a technique for detecting monotonic trends in data while accounting for known seasonal variability. Non-parametric correlational statistics are calculated for each month and used to determine a weighted average for the year, thus controlling for seasonality (Hirsch et al. 1982).

[2] Time series analysis demands a relatively large data set to detect serial correlations with different time lags.

# 3.0 Metal Levels in Fish

## 3.1 Background

From 1980 to 1983 Environment Canada and the British Columbia Ministry of Environment conducted a joint study of metal levels in fish in the Columbia River upstream of the International Boundary. The study was intended to address concerns regarding metal discharges from a lead-zinc smelter in Trail and possible mobilization of biologically available metals in several upstream reservoirs. The fish sampled were large-scale suckers (*Catostomus macrocheilus*), mountain whitefish (*Prosopium williamsoni*), walleye (*Stizostedion vitreum*) and sturgeon (*Acipenser transmontanus*). Muscle tissues from sample fish were analyzed for mercury, cadmium, chromium, copper, lead and zinc.

Based on the results of this 3-year program, an ongoing monitoring program was recommended. Annual sampling was undertaken (timed to correspond with the presence of walleye in the study area) for a three-year period from 1986 to 1988. In addition to the above species, rainbow trout (*Oncorhynychus mykiss*, formerly *Salmo gairdneri*) were also sampled. The data were analyzed by Norecol Environmental Consultants in 1989. One of the objectives of their analysis was to determine whether there had been significant changes in fish tissue metal concentrations between 1980 and 1988 in order to determine whether annual monitoring should be continued or whether monitoring frequency should be reduced (Norecol 1989).

Norecol's analysis of the data showed that mercury levels in fish did not differ significantly among years and that there was no evidence that other metal levels in Columbia River fish were either increasing or decreasing. On the basis of these results they recommended that monitoring frequency be reduced to a sampling interval of three to five years.

Essentially, failure to reject the null hypothesis (i.e., no change in metal concentrations) was taken to mean that there was indeed no change (i.e., the null hypothesis is true). However, the authors of this report did not report $\beta$, the probability of making a Type II error (the probability of failing to reject the null hypothesis of no trend when there is indeed a trend in metal levels present. We felt it would be instructive to calculate the level of change in fish tissue metal concentrations which might have been detected with sufficiently high power (i.e., low Type II error). These calculations are essential if policy recommendations or management prescriptions are to be developed based on the results of the analysis.

The choice of an acceptable level of Type II error will vary depending on the context of the problem. Ideally, decision makers would weigh the relative costs of each error in determining an appropriate level for each. In the present example, the cost of incorrectly concluding that there was a change in metal levels (Type I error) might include the cost of continuing to monitor at past frequencies and/or the cost of any abatement measures undertaken. The cost of incorrectly concluding no trend when one does indeed exist (Type II error) may include the cost of failing to undertake abatement measures. This cost may also be amplified if failure to detect a change results in reduced monitoring frequency and therefore even lower power to detect current or future changes in concentrations.

In Norecol's analysis, the data set for metal levels in fish tissues was initially evaluated using the F-test from the analysis of variance (ANOVA) and analysis of covariance (ANCOVA) with fish length as the covariate. Which of these tests was more appropriate was determined by the significance of the pooled regression coefficient from the ANCOVA. The ANCOVA was found to be more appropriate for the data on mercury levels in all species and for zinc levels in mountain whitefish and rainbow trout. The data were also tested for homogeneity of variances using Bartlett's test. If this test showed significant heterogeneity, the analysis was repeated using ln-transformed data. Log transformed data were used for copper concentrations in all species and zinc levels in mountain whitefish and rainbow trout. When the F-test for differences among years was found to be significant, multiple range tests were used to determine if monotonic trends were present.

## 3.2 Results of Power Analyses

Norecol's statistical analyses were repeated for this report in order to obtain the inputs for an *a posteriori* analysis of power. An *a posteriori* power analysis is a test of the power of an experiment (in this case a monitoring program) to detect change (i.e., trends) after the data have been collected. A statistical power analysis program by Bornstein and Cohen (1988) was used to complete the analysis. The inputs required by this program are the group means (i.e., within-year means) and within-group standard deviations. Where ANCOVA was used in the original analysis, power was determined using an F-test with the adjusted group means.

The results of the original F-tests are presented in Table 2. Power was computed for non-significant results. The detectable effect size (i.e., detectable percentage change among any two years) with 90% power are also provided in Table 2 wherever this could be calculated. That is, we report the size of effect which could have been detected with 90% confidence given the number of fish analyzed in Norecol (1989). In the case of an F-test (not a true test for monotonic trends but simply a quick way of screening data for changes) the detectable effect size represents the minimum change between **any two years** that could have been detected in the ANOVA or ANCOVA given the variability in the data set. Thus, the detectable effect size may be averaged over the sampling interval to determine the minimum level of trend which could have been detected.

**Table 2:** **Significance of F-tests from Analyses of Variance/Covariance for among-year differences in metal concentrations in fish of the Columbia River, as reported in Norecol (1989). Wherever it could be calculated, the detectable effect size at a power level of 90% is reported below F-tests which were found to be not significant.**

| Metal | Largescale Suckers | Mountain Whitefish | Walleye | Rainbow Trout |
|---|---|---|---|---|
| Mercury[d] | not significant<br><br>Detectable effect size[c]:<br><br>1.23 s.d. or a 42% deviation from the among group mean | not significant<br><br>Detectable effect size[c]:<br><br>1.26 s.d. or a 60% deviation from the among group mean | not significant<br><br>Detectable effect size[c]:<br><br>2.4 s.d. or a 194% deviation from the among group mean[e] | not significant<br><br>Detectable effect size[c]:<br><br>1.3 s.d. or a 57% deviation from the among group mean |
| Cadmium | significant only with 1980-81 data[b] | not significant<br><br>Detectable effect size[c]:<br><br>1.32 s.d. or a 119% deviation from the among group mean | not tested[a] | not tested[a] |
| Chromium | significant only with 1980-81 data[b] | significant only with 1980-81 data[b] | significant only with 1980-81 data[b] | not tested[a] |
| Copper | significant ($p<0.01$) | significant ($p<0.01$) | significant ($p<0.01$) | significant ($p<0.01$) |
| Lead | significant only with 1980-81 data<br><br>Detectable effect size[c]:<br><br>1.32 s.d. or a 85% deviation from the among group mean | not significant<br><br>Detectable effect size[c]:<br><br>1.32 s.d. or a 14% deviation from the among group mean | not significant[b] | not tested[a] |
| Zinc | significant only with 1980-81 data<br><br>Detectable effect size[c]:<br><br>1.45 s.d. or a 33% deviation from the among group means | significant ($p<0.05$) | significant ($p<0.05$) | significant ($p<0.05$) |

**Table 2: (continued)**

[a] Statistical tests were not performed in original analysis by Norecol because the metal was not detectable in the indicated species. Power analyses were therefore not calculated for these data.

[b] Significance results suspect due to zero variance in some cells (metal undetectable in all samples from one or more years). Power was not calculated for these data since zero variance also undermines the analysis of power.

[c] This refers to the effect size which could have been detected with approximately 90% power (i.e., 10% Type II error). Detectable effect size is reported first in terms of standard deviations from the group mean and then as a percentage deviation from the group mean.

[d] The ANCOVA was found to be the most appropriate test for mercury levels in all fish. Power analyses for the results of ANCOVA were calculated using the ANOVA tables in Cohen (1988) and the adjusted group means. The remaining power analyses for levels of other metals were based on the standard F-test.

[e] This result is highly biased since there was only one sample in each of 1980 and 1981 and very small samples in 1987 and 1988.

It is worth reviewing a specific example from Table 2. In the case of mercury concentrations in Largescale Suckers, the minimum detectable effect size among any two years with 90% power was a 42% deviation from the among group mean (or 0.074 µg/g). In other words there is a 90% chance that the F-test would have detected a change of 0.074 µg/g or larger between 1980 and 1988. This translates into a minimum monotonic change of 0.00925 µg/g/year over the eight years in which largescale suckers were sampled or a 5.3% change per year. This was one of the lowest detectable effect sizes calculated; several other species and metals showed much higher detectable effect sizes (i.e. less ability to detect trends with 90% confidence).

Figure 1 shows the size of trends in [Hg] in Largescale Suckers detectable with 90% power, given the number of fish sampled and analyzed in Norecol (1989). Note that increases of 3.1-5.2% would have raised the [Hg] above the 2 µg/g guideline of the Health Protection Branch of Health and Welfare Canada - the recommended limit for native populations who consume large quantities of fish (Wheatley 1979, cited in Norecol 1989). However, increases of 3.1-5.2% would not have been detectable with 90% confidence or power. The mean [Hg] was 0.16 ug/g in Largescale Suckers in 1980. A 3.1% increase per year translates into 0.04 µg/g over eight years (i.e. 0.2 - 0.16). We estimate that the sample size used by Norecol (1989) would have had only a 41% chance of detecting a change of this magnitude. Again we stress that mercury in walleye showed relatively low detectable effect sizes due to relatively low within group variance.

Figure 2 shows the changes in power (1-β) as a function of detectable effect size and sample size, for [Hg] in Largescale Suckers. The number of fish actually analyzed by Norecol (1989) is at the extreme left hand side of the graph (i.e. 86). Therefore, the power with which the indicated effect sizes could have been detected by Norecol (1989) is the y-intercept of each of the curves. Detection of a change of 0.04 µg/g (the example discussed above) is illustrated by the second highest curve. With 86 fish, one would have only 41% confidence of detecting a trend of this magnitude. More than 200 fish would be required to detect such a trend with 90% confidence.

A number of caveats must be attached to both the original analysis and the power analysis prepared for this report. First, the data collected in 1980 and 1981 employed different detection limits than data collected in subsequent years. This may explain the significance of differences in the levels of cadmium, chromium and lead. Furthermore, there are substantial interannual variations in copper and zinc concentrations. This interannual noise makes detection of monotonic trends (signal) quite difficult to detect. Some explanation of this interannual variability may help refine sampling design or data analysis (much as seasonal data can be de-trended to detect interannual trends). Finally, in some cases, the original analysis could not be repeated. For example, the ANCOVA table for mercury in Largescale Suckers (1980 to 1988) in Appendix 3-2 of the Norecol Report could not be duplicated.

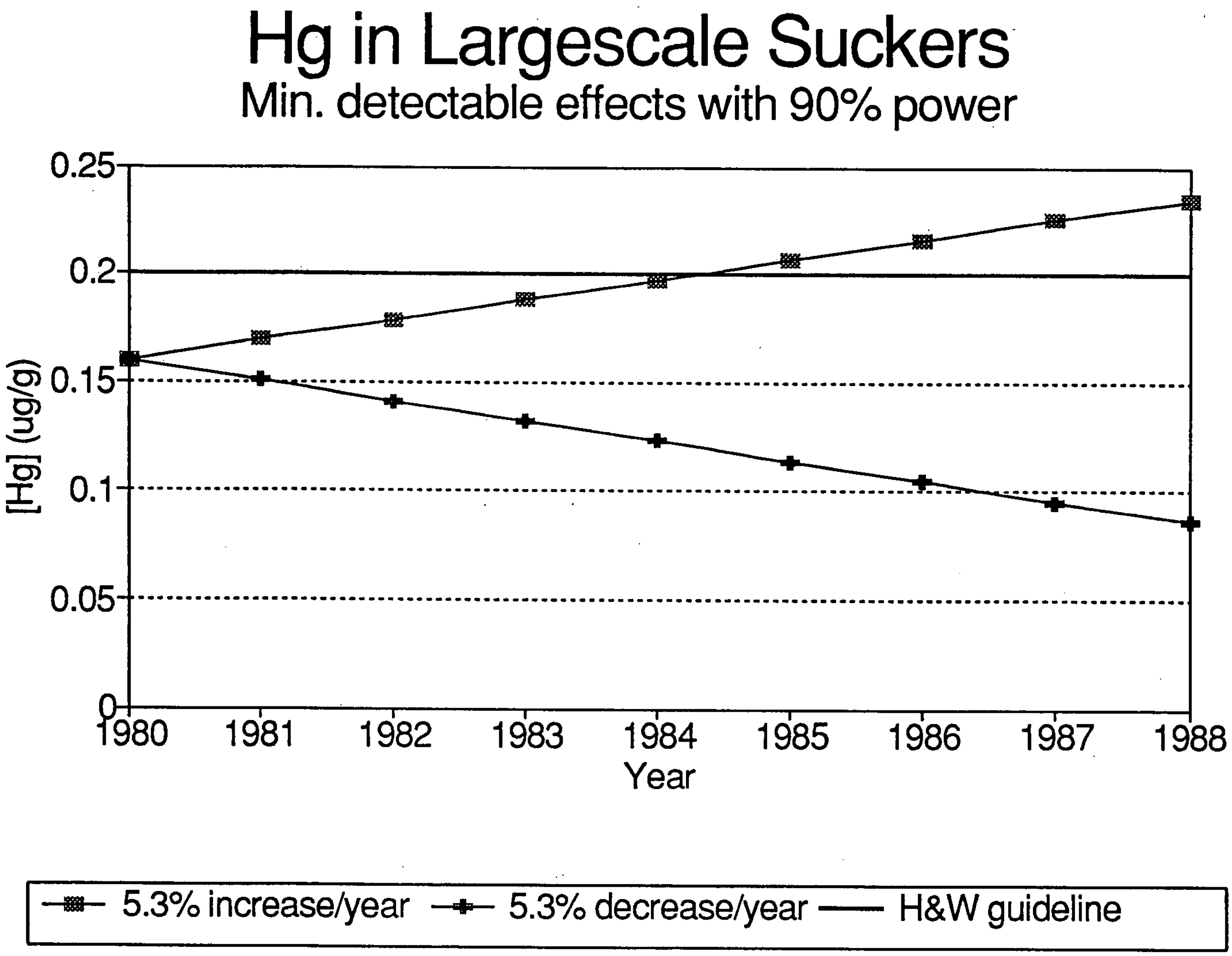


Figure 1. Minimum detectable trends in $[Hg]_{largescale\ suckers}$ with 90% confidence (power), given the number of fish sampled in Norecol (1989). The trend lines shown begin at the 1990 mean concentration and proceed up or down at the minimum detectable rate. Trends less than this would not be detectable with 90% confidence. The recommended limit for native populations who consume large quantities of fish (0.2 µg/g) is also shown.

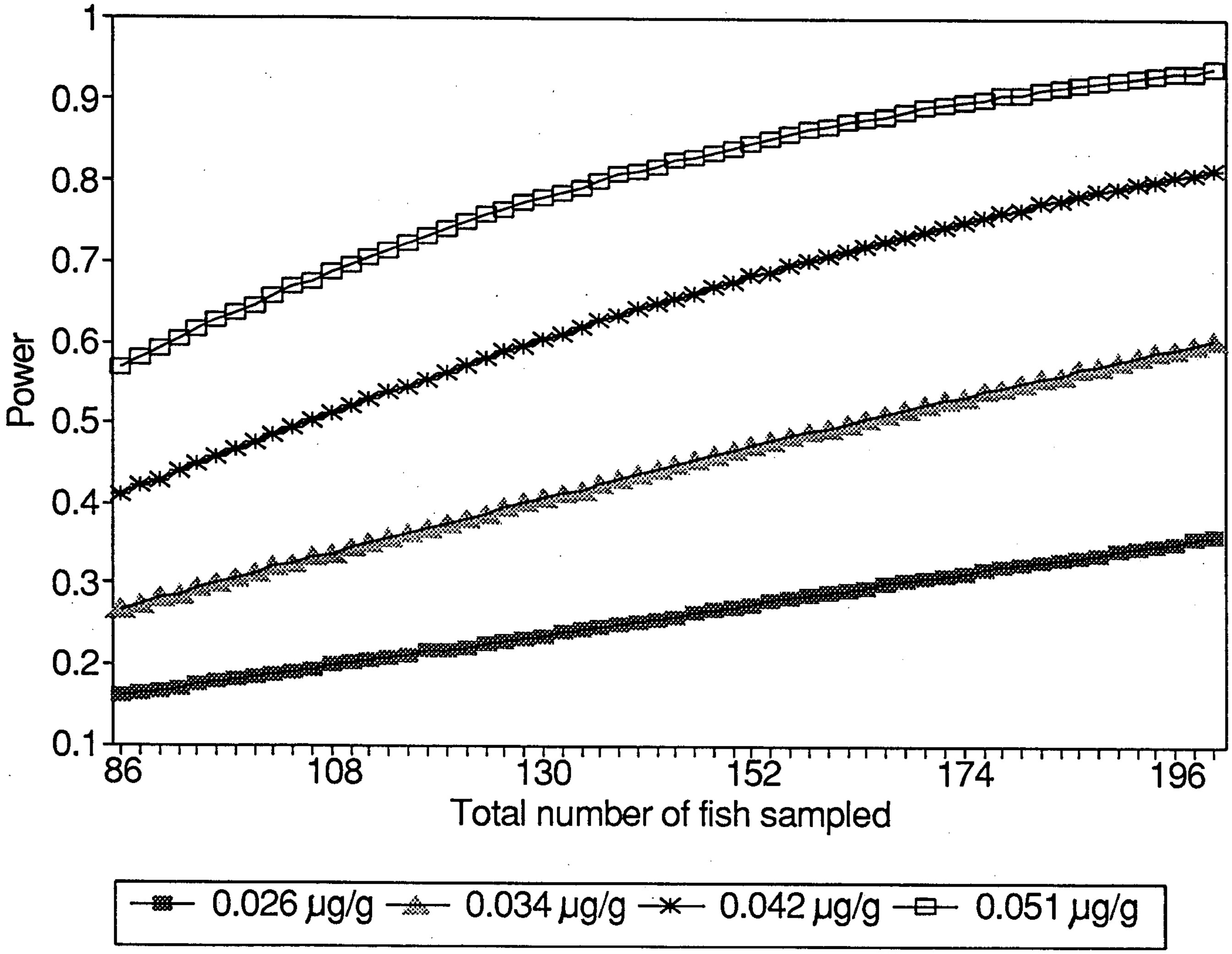


Figure 2. *A priori* power analysis for mercury concentrations in Largescale Suckers. Lines show the changes in power (1-β) as a function of detectable effect size and sample size (total number of fish sampled between 1980 and 1988, assumed to be equally distributed across years). The minimum detectable effect size represents the change in mercury concentrations among any two years. The effect sizes shown correspond to 0.43sd, 0.57sd, 0.71sd, and 0.85sd (respectively).

## 4.0 Discussion

The purpose of the above analysis was to illustrate a common problem with making recommendations on analyses of monitoring data without simultaneous analyses of power. Recommendations to reduce monitoring frequency should be assessed against the risk of not having detected a change of a given size when one actually existed. Given the very low power of the preceding data set, these costs may be quite high. Power analyses should become standard procedures in data analysis from these sorts of monitoring programs. They give important additional information to decision makers. It is also much easier to undertake such analyses at the time of the original analysis.

Such power analyses can also be used to generate *a priori* assessments of how an existing monitoring program could be changed to increase its effectiveness. For example, nomograms such as those generated for the fish mercury data in Figure 2 could be used in the decision to reduce or in this case, possibly increase monitoring frequency. Ideally, such *a priori* power analyses would be conducted during the design phase of a monitoring program in order to maximize benefits while minimizing costs (i.e., to choose and optimal sampling regime). These analyses can be based on simulated data sets, existing data sets or as part of an initial pilot program to assess the adequacy of a particular design. Pilot studies are likely to be particularly important in the first year of the Fraser River Basin Assessment Program, so that the monitoring activities will be sufficiently intensive to capture effect sizes of interest.

## 5.0 References

Bornstein, M. and J. Cohen. 1988. Statistical power analysis: a computer program. New Jersey: Lawrence Erlbaum Associates, Inc.

Cohen, J. 1988. Statistical power analysis for the behavioral science. Second edition. L. Erlbaum Associates, Hillside, New Jersey, 567 pp.

Cohen, J. Statistical power analysis for the behavioral sciences. Second Edition. New Jersey: Lawrence Erlbaum Associates, Publishers.

Gilbert, R. O. 1987. Statistical methods for environmental pollution monitoring. New York, NY: Van Nostrand Reinhold Company.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. New York, NY: John Wiley & Sons.

Green, R. H. 1989. Power analysis and practical strategies for environmental monitoring. Environmental Research 50: 195-205.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54: 187-211.

Loftis, J. C., R. C. Ward, R. D. Phillips, and C. H. Taylor. 1989. An evaluation of trend detection techniques for use in water quality monitoring programs. Corvalis, OR: Environmental Research Laboratory, Office of research and Development, E.S. Environmental Protection Agency.

Marmorek, D.R., J. Korman, D.P. Bernard, and T. Berry. 1992. Ecosystem fate and effects of pulp mil effluents in the Fraser River: Identification of research and monitoring priorities. Prepared by ESSA Ltd., Vancouver, B.C. for Inland Waters Directorate, Environment Canada, North Vancouver, 151 pp.

Marmorek, D.R. and T. M. Webb. 1986. Review of data analysis techniques for the watershed manipulation project. Prepared for Northrop Services, Inc., Corvalis, Oregon by ESSA Environmental and Social Systems Analysts Ltd. Vancouver, B.C..

McAllister, M. K. and R. M. Peterman. 1992. Experimental design in the management of fisheries: a review. North American Journal of Fisheries Management 12: 1-18.

Norecol Environmental Consultants. 1989. Statistical analyses of metal levels in fish of the Columbia River near international boundary, 1980 to 1988. Prepared for Environment Canada, Water Quality Branch, Inland Waters, Pacific and Yukon Region, Vancouver, B.C..

Peterman, R. M. and M. J. Bradford. 1987. Statistical power of trends in fish abundance. Canadian Journal of Fisheries and Aquatic Sciences 44: 1879-1889.

Porter, P. S. 1986. Statistical analysis of water quality data affected by limits of detection. Fort Collins, CO: Department of Agricultural and Chemical Engineering, Colorado State University.

Toft, C. A. and P. J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. The American Naturalist 122: 618-625.

Wheatley, B. 1979. Methylmercury in Canada, exposure of Indian and Inuit residents to methylmercury in the Canadian environment, Volume 1. Ministry of National Health and Welfare, Medical Services Branch, Ottawa, Ontario.