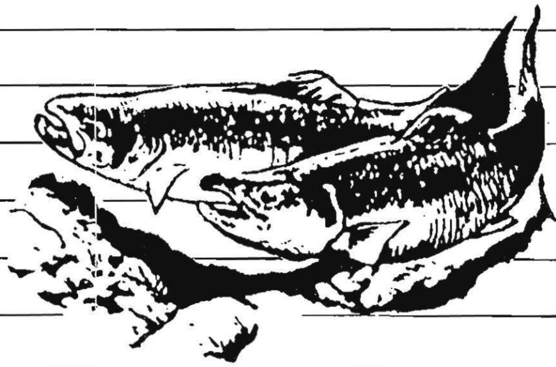
Research Information bulletin

U.S. DEPARTMENT OF THE INTERIOR
NATIONAL BIOLOGICAL SURVEY



NO. 23
DATE 1994

Improper Formation of Composite Samples Jeopardizes Data Interpretation

Our purpose is to alert researchers to potential problems in forming a suitable composite sample and in analyzing the data correctly. One of the practices associated with environmental studies is the combining of samples into a composite that is analyzed to obtain a single average measure. For example, five samples of sediment or five fish are combined (homogenized), and a single analytical measurement of contaminant load is obtained from the composite. The formation of a composite may be justified when the cost of analysis is high in relation to the cost of collecting individual samples. Composite samples may be essential when the analytical technique requires more material than a single sample or individual specimen can provide. We investigated the properties of composite samples using computer simulations based on actual contaminant concentrations measured from individual fish.

Size Variation Among Fish Contributes to the Bias of Incorrectly Formed Composite Samples

Whole fish are often combined in a large mixer or homogenizer when forming a composite sample for fish contaminant analysis. This procedure, which we call the Batch Method, leads to biased estimates of contaminant concentration because larger fish contribute proportionally more mass to the composite. Although fish of the same age class are sometimes used to form the composite, size variation within an age class may be significant. The correct procedure, the Individual Method, is to homogenize each fish separately, obtain an equal-sized subsample from each homogenate, and combine the subsamples to form the composite.

Simulations With Striped Bass Data Illustrate the Bias

To demonstrate the effects of the two procedures to form a composite, we worked with observations on individual contaminant concentrations in 195 striped bass (*Morone saxatilis*) between 55 and 116 cm long. Contaminant concentrations in these fish ranged between 0.1 and 40.7 ppm; the mean contaminant concentration was 3.57 and the variance was 24.10 (Fig. 1). We then produced computer-generated composites by combining groups of five fish randomly selected (with replacement) from the original sample; we simulated the formation of 10,000 composites following the well-described bootstrap procedure.

The Individual Method Yields Unbiased Estimates

Concentration of each composite sample was estimated in two ways, corresponding to the two procedures described previously. The Individual

Method simulated the combining of equal portions of homogenate from each fish; the Batch Method simulated the homogenate formed from whole fish. The composite resulting from the Individual Method has a contaminant concentration equal to a simple average of the concentrations in the five fish. Contaminant concentration in the composite formed by the Batch Method is a weighted average concentration for five fish where fish mass is the weighting factor. The results of these bootstrap simulations demonstrate how the mean and variance are affected by the method used to form the composite.

The mean obtained by the Individual Method (Table) is similar to the true (population) mean, 3.57. The Individual Method gives an unbiased estimate of the population mean. The mean obtained by the Batch Method is significantly greater than the true mean ($t = 2.90$, $P < 0.05$), thus showing the bias in the Batch Method computations. We believe the bias of the Batch Method depends, in part, on the distribution of fish weight in the population and in the sample. In some special instances, the Batch Method may produce an unbiased estimate of the mean concentration; we found this to be true when the range of fish weight was restricted in the original population.

The variance among 10,000 composites is approximately one-fifth the population variance: $4.75 \times 5 = 23.75$ (Individual Method) and $5.65 \times 5 = 28.25$ (Batch Method). That is, if we multiply the variance among composites by the number of fish per composite, we can estimate the true variance. Our observation on the relation of the composite variance and the true variance is grounded in statistical sampling theory, as we demonstrate later. the Individual Method provides an unbiased estimate of the population mean and variance and is therefore recommended for composite formation.

The Sample Variance From a Set of Composites Is Not an Estimate of the Population Variance

Because researchers often form more than one composite to characterize a population, we formed 20 composite samples of five fish each using the Individual Method. We repeated these procedures 1,000 times for each set of 20 composites. The means of each bootstrap trial (20 composite samples) were plotted and the mean of these means and the average variances were calculated (Fig. 2). Note the change in the x-axis scale compared with Fig. 1. Most of the sets of composite samples had means near 3.57, the true mean. However, the means ranged from 2.45 to 5.61. We calculated the 95% confidence intervals around the composite means and found that 9.9% of the confidence intervals did not include the true mean, 3.57. Notice that we selected an alpha level of 0.05, but the actual significance level was 0.099. The reason for this difference in alpha is that the average estimated variance of the composite samples (4.90) underestimated the true population variance (24.10) by about a factor of 5.

From elementary statistics, we know that the magnitude of the variance is inversely proportional to sample size. Although sample size for the set of composites seems to be 20, it is actually 100 fish because the composites are made of 5 fish each. According to sampling theory, the variance among fish is σ^2. The variance among composite samples is $\sigma^2/5$, and the variance among the set of 20 composites is $(\sigma^2/5)/20$. The raw variance estimated from the set of 20 composite samples (4.90) is actually an estimate of $\sigma^2/5$, not σ^2. Thus, the apparent or raw variance estimated from a set of composites underestimates the variance among individuals from the population under study. A confidence interval constructed without regard to the number of fish sampled will be incorrect in that the significance level will be too optimistic.

Using sets of composite samples and the Individual Method to prepare composites, we can estimate the true population mean. The accuracy of the estimate of the population variance from a set of composites depends on the number of composite samples and other factors that we have not examined here. An approximation of the population variance can be obtained from the statistical literature.

The Correct Formation of Composites Must Be Considered Before Sample Collection and Processing

These examples illustrate that statistical analyses are an important part of the experimental design process and should be considered at the onset of a project. For composite samples, statistical analyses include the proper formation of composites. We recognize that the formation of composites is sometimes essential, so our goal in this study was to highlight the proper procedure of composite formation and the correct statistical interpretation of the results.

For further information contact:

Mary C. Fabrizio, Anthony M.Frank, or
Jacqueline F. Savino
Great Lakes Center
1451 Green Road
Ann Arbor, MI 48105
(313) 994-3331

Table. Contaminant concentrations (ppm) in simulated five-fish composites of striped bass (*Morone saxatilis*), 55–116 cm long.[a]

Method	Mean	Variance
Individual	3.55	4.7455
Batch	3.64	5.6486

[a]Results are based on 10,000 bootstrap trials.

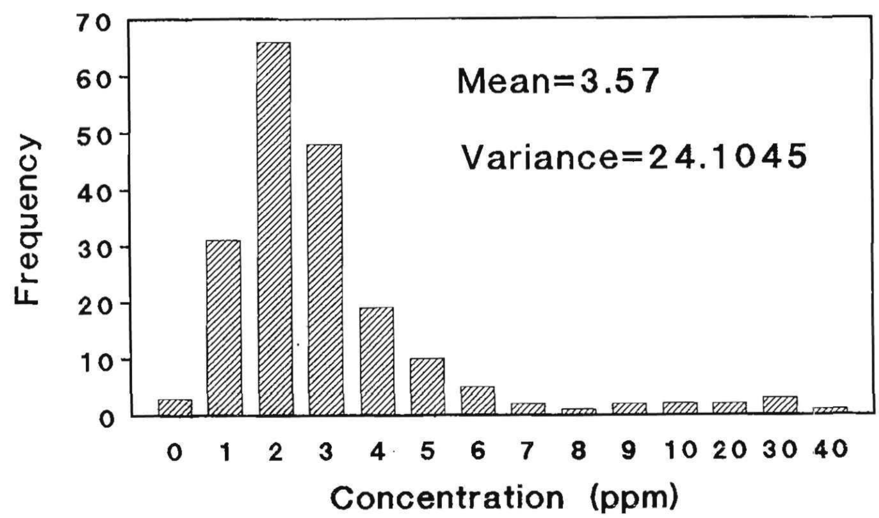


Fig. 1. Contaminant concentrations in striped bass (*Morone saxatilis*).

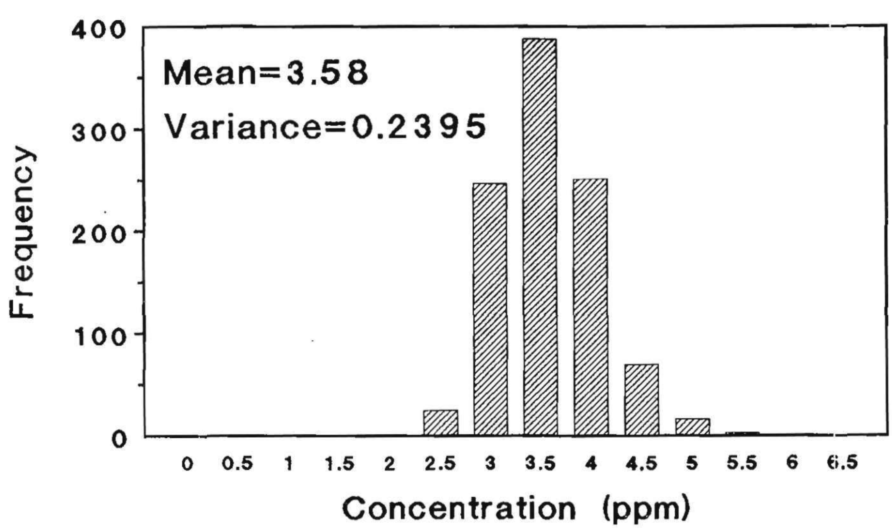


Fig. 2. Mean concentration for 1,000 simulated 20-composite tests.