156

Craig,
Many thanks for
all the support and
encouragement
BH

The Use of Random-Model Tolerance Intervals in Environmental Monitoring and Regulation

Robert W. SMITH

When appropriate data from regional reference locations are available, tolerance-interval bounds can be computed to provide criteria or limits distinguishing reference from nonreference conditions. If the limits are to be to applied to locations and times beyond the original data, the data should include temporal and spatial variation and the tolerance interval calculations should utilize a random crossed or nested ANOVA statistical design. Two computational methods for such designs are discussed and evaluated with simulations. Both methods are shown to perform well, and the adverse effect of using an improper design model is demonstrated. Three real-world applications are shown, where tolerance intervals are used to (1) establish a reference threshold for a benthic community pollution index, (2) set criteria for chemicals in sediments, and (3) establish background thresholds for survival rates in sediment bioassay tests. Some practical considerations in the use of the tolerance intervals are discussed.

Key Words: Bioassay; Bootstrap; Calibration; Pollution index; Sediment criteria; Simulation.

1. INTRODUCTION

Environmental monitoring and regulatory activities focused on determining the presence of impacts or recovery often involve comparisons with reference or background conditions. Reference conditions are usually characterized by indicators and other relevant variables measured at regional locations assumed to represent reference conditions (Bloom 1980; Hughes, Whittier, Rohm, and Larson 1990; Hughes 1995). The resulting data can be used to establish criteria or distinguishing reference from nonreference conditions. It is important to consider and intelligently choose among the analytical options available for establishing these limits.

In an effort to determine impacts, it is a common practice to use ANOVA statistics to compare indicator means for potentially impacted locations with the indicator mean for

Robert W. Smith is an Environmental Statistician working as an independent consultant, P.O. Box 1537, Ojai, CA 93024-1537 (E-mail: rs@robertsmith.net).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 7, Number 1, Pages 74–94

the reference locations. Given natural and random variability, the mean indicator values at different locations within regional reference areas will differ from the overall reference mean (Hurlbert 1984; Wiens and Parker 1995), so deviation from the overall reference mean is not sufficient to distinguish reference from nonreference. Thus, traditional ANOVA methods will often be inappropriate when comparing potentially impacted locations with reference.

The reference data will cover a range of indicator values, and some type of comparison with this range is more appropriate since it incorporates the expected differences in indicator values among the reference locations. Using the actual upper or lower limit of the data range as a standard of comparison for reference conditions would be risky due to uncertainty associated with sampling error. Also, a single unusual value in the sample data could greatly affect the value of the limit used for comparison. A limit incorporating sampling error and corresponding to a relevant quantile toward a tail of the reference data distribution would be more useful (Splitstone 1991; Kilgour and Somers 1998). Such a limit would be a tolerance-interval bound (Hahn and Meeker 1991; Vardeman 1992). Applications using tolerance intervals for comparison with reference conditions have involved monitoring of contaminants in groundwater (Gibbons 1994), soil, vegetation, and snow (Allen and Jones 1998) and benthic infaunal community parameters (Smith 1995, 1998; Smith and Bernstein 1996).

A tolerance-interval bound is simply the upper or lower confidence-interval bound of a quantile of the underlying data distribution. When increasing indicator values are positively correlated with impact, one could choose a quantile (say .90) toward the upper end of the distribution to define a limit. The tolerance-interval bound would then be the upper bound of a $1 - \alpha$ confidence interval for the .90th quantile. Here α is the selected probability or risk that the defined bound (computed from the sample data) does not cover the actual .90th quantile of the underlying data distribution.

Standard parametric and nonparametric tolerance-interval computations (Woodward and Frawley 1980; Gilbert 1987; Hahn and Meeker 1991; Portugal 1992; Vangel 1994; Allen and Jones 1998) are appropriate when the sample data values are independent. Relatively independent observations would be expected in cases where single observations are taken from multiple randomly selected locations during a single time period. Alternately, the sample could involve single observations at randomly chosen times at a single location.

When using tolerance intervals to define limits to reference conditions and these limits will be compared to future measurements, the data sample should contain both temporal and spatial random variability. This requires a sample including multiple times and locations. If the data sample involves only a single sampling event covering locations in space, only spatial variability will be included in the data distribution. For future observations, this increases the risk that natural changes over time will be confused with impact.

When the data include both spatial and temporal random variability, the data values within the different levels of time or location will tend to be positively correlated. In this case, the standard tolerance-interval computations assuming data independence will be inappropriate. Computational methods for computing tolerance intervals from data containing both spatial and temporal variability are not presently well developed, and I am unaware

of any publications where such methodology has been applied to real data. In this article, I describe two computational methods for dealing with data containing both temporal and spatial random variability. The first method is that of Bagui, Bhaumik, and Parnes (1996). The second method, which involves a parametric bootstrap approach (described in reports by Smith and Riege 1998; Hunt et al. 1998), is new.

The objectives of this article are to

(1) describe the tolerance-interval calculations,
(2) use simulations to evaluate the computational methods,
(3) demonstrate the penalty paid when standard tolerance intervals are inappropriately used,
(4) provide examples of real-world applications using the methodology, and
(5) make some practical suggestions for applying tolerance intervals with these methods.

2. STATISTICAL MODEL

In this section, the statistical model for the tolerance interval calculations is discussed. When the data contain both spatial and temporal variability, a useful statistical model will often be a two-way crossed ANOVA model with time and space as random factors (Davis 1994). The model is completely random since, in the target applications, we are interested in generalizing our results to other locations and times not in the sample from which the bounds are computed (Jackson and Brashers 1994). Beckman and Tietjen (1989) derive two-sided tolerance intervals for a random balanced crossed ANOVA design. However, one-sided limits that can be applied to unbalanced data sets will be most useful for the intended applications. One-sided intervals are appropriate for parameters where impact is associated with either an increase or decrease in parameter values (as is most often the case). Also, unbalanced data are common with environmental data sets. The Bagui et al. (1996) and the parametric bootstrap methods are the focus of this article because they are suitable for computing one-sided tolerance-interval bounds with unbalanced data.

In cases where completely different locations are sampled at each sampling time, a random nested model with locations nested within times will be appropriate. However, the simulation results in this article show that the tolerance-interval bounds for the nested model can be computed as a highly unbalanced crossed design, so the nested design is not emphasized.

The crossed statistical model is now more formally described. For simplicity, the statistical model is described in terms of rows, columns, and cells of a table describing the design. The different levels of factor 1 are represented by rows, levels of factor 2 are represented by columns, and combinations of the factor levels are represented by the cells. An observation with a two-way crossed random model can be decomposed as

$$y_{ijk} = \mu + \delta_i + \beta_j + \gamma_{ij} + e_{ijk},$$

where y_{ijk} is the kth observation in the cell defined by the ith row and the jth column, μ is the general mean, δ_i is the effect of the observation being in row i, β_j is the effect of the observation being in column j, γ_{ij} is the interaction effect of row i with column j, and e_{ijk} is the error, or the deviation of the observation from its expectation (the sum of the other terms in the model). The e_{ijk}, δ_i, β_j, and γ_{ij} are assumed to be independent random factors following $\mathrm{N}(0, \sigma_e^2)$, $\mathrm{N}(0, \sigma_\delta^2)$, $\mathrm{N}(0, \sigma_\beta^2)$, and $\mathrm{N}(0, \sigma_\gamma^2)$, respectively.

It should be emphasized that the model and associated computations discussed in the next section apply to the situation where single future observations are to be compared with the computed tolerance-interval bounds. The statistical model would need to be modified to apply to comparison with averaged values. Using a model for single observations is the most flexible because it can apply to all future observations regardless of sample size at any one location and time.

3. COMPUTATION OF THE TOLERANCE INTERVAL BOUNDS

The general formula for a parametric one-sided upper tolerance-interval bound is

$$b_{p,\alpha} = \bar{x} + k_{p,\alpha} s, \tag{3.1}$$

and for a one-sided lower tolerance-interval bound

$$b_{p,\alpha} = \bar{x} - k_{p,\alpha} s, \tag{3.2}$$

where $\bar{x}$ is the estimate of the overall mean and s is an estimated standard deviation. The $\bar{x}$ is computed as the mean of all data values in the sample, and the computation of s varies, depending on the method being used. The methods described in the following sections differ only in the manner in which the $k_{p,\alpha}$ and s values are estimated.

The $k_{p,\alpha}$ is computed so that the resulting $b_{p,\alpha}$ estimates will fail to cover the underlying pth population quantile α proportion of the time, thus

$$P[b_{p,\alpha} \leq q_p] = \alpha$$

for an upper bound and

$$P[b_{p,\alpha} \geq q_p] = \alpha$$

for a lower bound, where q_p is the underlying pth quantile of the population distribution (usually unknown). Thus, α is the rate of noncoverage of q_p by the $b_{p,\alpha}$ estimates. The tolerance interval bound is equivalent to a one-sided upper or lower confidence interval bound on q_p (Hahn and Meeker 1991).

3.1 THE STANDARD METHOD

When the sample consists of independent measurements and the data are from a normal distribution, the standard method can apply. In Equations (3.1) and (3.2), s is the computed

standard deviation of the sample data values and

$$k_{p,\alpha} = t_{1-\alpha,n-1,\lambda}/\sqrt{n},$$

where $t_{1-\alpha,n-1,\lambda}$ is the $1-\alpha$ quantile of the noncentral t-distribution with $n-1$ degrees of freedom and a noncentrality parameter of λ. Here n is the sample size and $\lambda = z_p n^{1/2}$, with z_p being the absolute value of the pth quantile of the standard normal distribution.

3.2 The Bagui et al. (1996) Computational Method

This method can be applied to any random model, but the computations specific to the random crossed model are shown here. In Equations (3.1) and (3.2), s is the estimated standard deviation of the mean, computed as

$$s = s_{\bar{x}} = \sqrt{\frac{n_1\hat{\sigma}_\delta^2 + n_2\hat{\sigma}_\beta^2 + n_3\hat{\sigma}_\gamma^2 + n\hat{\sigma}_e^2}{n^2}}, \tag{3.3}$$

where $\hat{\sigma}_\delta^2$ is the estimated variance component for factor 1 (rows), $\hat{\sigma}_\beta^2$ is the estimated variance component for factor 2 (columns), $\hat{\sigma}_\gamma^2$ is the estimated variance component for the interaction between factors 1 and 2, and $\hat{\sigma}_e^2$ is the estimated error variance. In (3.3),

$$n_1 = \sum_{i=1}^{r} n_{i\cdot}^2, \qquad n_2 = \sum_{j=1}^{c} n_{\cdot j}^2, \qquad n_3 = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2, \quad \text{and} \quad n = \sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij},$$

where n_{ij} is the number of observations in the row i, column j cell, $n_{i\cdot}$ is the total number of observations in row i, $n_{\cdot j}$ is the total number of observations in column j, and r and c are the numbers of rows and columns, respectively.

The $k_{p,\alpha}$ is computed as

$$k_{p,\alpha} = t_{1-\alpha,df,\lambda}, \tag{3.4}$$

which is the $1-\alpha$ quantile of the noncentral t-distribution with df degrees of freedom and a noncentrality parameter of λ. Here, df is the approximate degrees of freedom associated with the estimation of $s_{\bar{x}}^2$ (Satterthwaite 1946; Bagui et al. 1996), which is

$$df = \frac{s_{\bar{x}}^4}{[h(MSD)]^2/df_{\text{MSD}} + [h(MSB)]^2/df_{\text{MSB}} + [h(MSG)]^2/df_{\text{MSG}}}. \tag{3.5}$$

MSD is the ANOVA mean square for the rows, MSB is the ANOVA mean square for the columns, and MSG is the ANOVA interaction mean square. The df_{MSD}, df_{MSB}, df_{MSG} are the degrees of freedom associated with the estimation of the respective mean squares and

$$h = \frac{s_{\bar{x}}^2}{MSD + MSB + MSG}.$$

The noncentrality parameter λ is estimated as

$$\hat{\lambda} = z_p\sqrt{\left(\hat{\sigma}_\delta^2 + \hat{\sigma}_\beta^2 + \hat{\sigma}_\gamma^2 + \hat{\sigma}_e^2\right)/s_{\bar{x}}^2}, \tag{3.6}$$

where z_p is the absolute value of pth quantile of the standard normal distribution.

3.3 THE PARAMETRIC BOOTSTRAP METHOD

With this method, s in Equations (3.1) and (3.2) is the estimated standard deviation of the data values, computed as

$$s = \sqrt{\hat{\sigma}_\delta^2 + \hat{\sigma}_\beta^2 + \hat{\sigma}_\gamma^2 + \hat{\sigma}_e^2}. \tag{3.7}$$

The value for $k_{p,\alpha}$ is computed with the following procedure:

(1) The initial values for $\mu, \sigma_\delta^2, \sigma_\beta^2, \sigma_\gamma^2$, and σ_e^2 are set equal to $\bar{x}, \hat{\sigma}_\delta^2, \hat{\sigma}_\beta^2, \hat{\sigma}_\gamma^2$, and $\hat{\sigma}_e^2$, respectively.

(2) We assume that the values for $\mu, \sigma_\delta^2, \sigma_\beta^2, \sigma_\gamma^2$, and σ_e^2 are known exactly, and the pth quantile (for an upper bound) of the underlying data distribution is computed as $q_p = \mu + z_p(\sigma_\delta^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2)^{1/2}$, where z_p is the pth quantile of the standard normal distribution.

(3) N_2 sets of simulated data are created using $\mu, \sigma_\delta^2, \sigma_\beta^2, \sigma_\gamma^2$, and σ_e^2. The simulation method is detailed in the Appendix.

(4) For each of the N_2 data sets, $\bar{x}_i, \hat{\sigma}_{\delta i}^2, \hat{\sigma}_{\beta i}^2, \hat{\sigma}_{\gamma i}^2$, and $\hat{\sigma}_{ei}^2$ ($i = 1$ to N_2) are computed. From the computed variance components, s_i is computed using Equation (3.7).

(5) A chosen $k_{p,\alpha}$ value is used along with all $\bar{x}_i$ and s_i to compute a set of N_2 bounds ($b_{p,\alpha,i}$) using Equation (3.1).

(6) For the bounds computed in step 5, the rate of noncoverage of q_p by all the bounds in the set is computed as $\alpha_0 = m/N_0$, where m is the number of times that $b_{p,\alpha,i} \leq q_p$ (for an upper bound).

(7) Steps 5 and 6 are repeated for a series of $k_{p,\alpha}$ values. The $k_{p,\alpha}$ value producing the minimal value of $|\alpha_2 - \alpha|$ is the final $k_{p,\alpha}$ value used in Equation (3.1).

Steps 5–7 are similar to the algorithm used by Davies and Gather (1993) to compute a constant (similar to $k_{p,\alpha}$) for a robust outlier-detection technique that is, in principle, very similar to a tolerance interval. In step 7, a successive interpolation procedure was used to zero in on a $k_{p,\alpha}$ that produces a small value for $|\alpha_2 - \alpha|$. This procedure involved repeatedly bracketing the minimum $|\alpha_2 - \alpha|$ for successive series of potential $k_{p,\alpha}$ values. For each successive series, the range of $k_{p,\alpha}$ values decreased and the interval between each $k_{p,\alpha}$ value in the series was decreased by one half. The process was terminated when $|\alpha_2 - \alpha|$ reached a tolerance value ($.5/N_2$) or a maximum number of iterations were exceeded (99).

3.4 ESTIMATION OF VARIANCE COMPONENTS

For the present application, the Henderson method I (Searle, Casella, and McCulloch 1992) was used to compute the variance component estimates because it is appropriate for unbalanced data, the mean squares used in Equation (3.5) are available from the computations, and the variance components can be computed rapidly (as required for simulation and calibration). This method can produce negative variance components, which are treated as zeros in the computations.

4. EVALUATION OF THE METHODS WITH SIMULATIONS

Simulations are used to evaluate the performance of the methods described above. For the tolerance-interval bounds computed from simulated data, the noncoverages of known population percentiles are compared with the nominal α levels. If the methods are working well, the noncoverage should approximate the nominal α level. The following steps describe the simulation procedure:

(1) For a simulation experiment, values are chosen for $\mu, \sigma_\delta^2, \sigma_\beta^2, \sigma_\gamma^2$, and σ_e^2, α, and p. The desired sampling design is defined by specifying the numbers of rows, columns, and cell replicates in the model.

(2) Since the values for $\mu, \sigma_\delta^2, \sigma_\beta^2, \sigma_\gamma^2$, and σ_e^2 are known exactly, the actual pth quantile (for an upper bound) of the underlying data distribution can be computed as $q_p = \mu + z_p(\sigma_\delta^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2)^{1/2}$, where z_p is the pth quantile of the standard normal distribution.

(3) N_0 sets of simulated data are created using chosen values for $\mu, \sigma_\delta^2, \sigma_\beta^2, \sigma_\gamma^2$, and σ_e^2. The simulation method is detailed in the Appendix.

(4) For each of the N_0 data sets, $\bar{x}_i, \hat{\sigma}_{\delta i}^2, \hat{\sigma}_{\beta i}^2, \hat{\sigma}_{\gamma i}^2$, and $\hat{\sigma}_{ei}^2$ ($i = 1$ to N_0) are computed and used to compute a $b_{p,\alpha,i}$ value. This step simply involves computing the tolerance-interval bounds ($b_{p,\alpha,i}$), using the method being evaluated, to each of the N_0 data sets.

(5) The rate of noncoverage of q_p by the $b_{p,\alpha,i}$ values is computed as $\alpha_0 = m/N_0$, where m is the number of times that $b_{p,\alpha,i} \leq q_p$ over all N_0 simulated data sets. Since the rate of noncoverage of q_p is supposed to equal α, α_0 should be approximately equal to α if the method of computing $b_{p,\alpha}$ is accurate and N_0 is a sufficiently large value.

5. CALIBRATION

Both the computational and bootstrap tolerance-interval methods described above are approximations. With the Bagui et al. (1996) computational method, the variance components used in (3.3) and (3.6) are estimates, so the coverage of the pth quantile by $b_{p,\alpha}$ is not assured to be $1 - \alpha$. In addition, the formula for degrees of freedom in (3.5) is only approximate, and with an unbalanced design, the mean squares used in (3.5) are not independent as assumed (Milliken and Johnson 1984). The bootstrap method also treats variance component estimates as known values, and the bounds should reflect the corresponding uncertainty.

Using the evaluation technique described in Section 4, preliminary simulation experiments indicated that the computational method almost always produced noncoverage of the pth quantile at a rate less than α and the parametric bootstrap tended to produce noncoverage at a rate greater than α. To provide $b_{p,\alpha}$ values where the noncoverage of the bounds is closer to α, a bootstrap calibration method using the approach of Efron and Tibshirani (1993, chap. 18) was developed as follows:

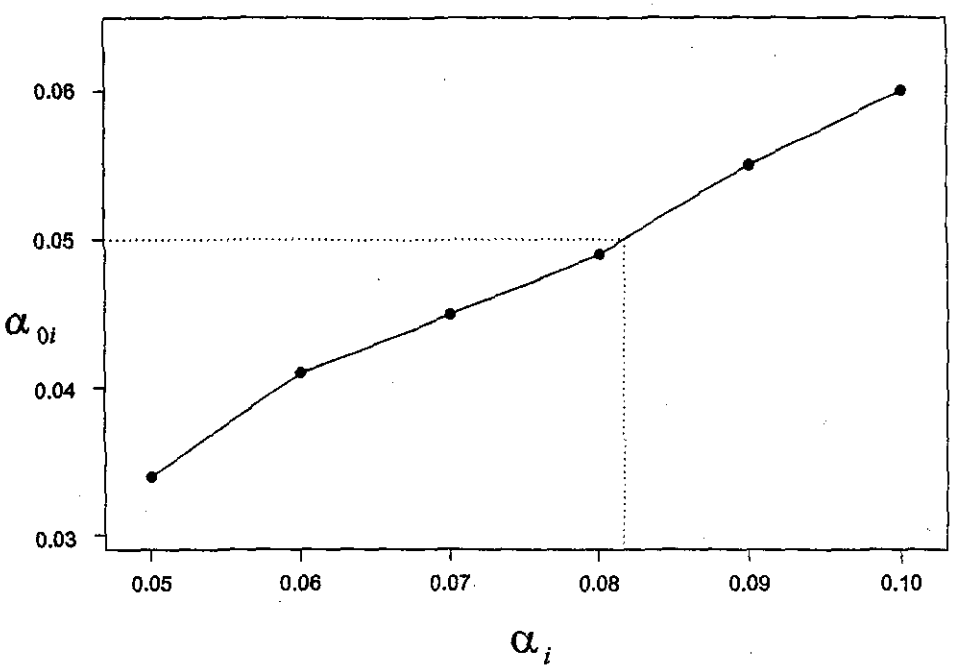

Figure 1. An Example of the Calibration Interpolation Procedure Where the Nominal $\alpha = .05$. Here the final bound is computed with $\alpha' = .082$ to obtain the desired nominal α of .05 for the computational method.

(1) For each of a series of chosen α_i values, the procedure of Section 4 is applied, with $\alpha = \alpha_i$, $N_0 = N_1$, and using the computed $\bar{x}, \hat{\sigma}^2_\delta, \hat{\sigma}^2_\beta, \hat{\sigma}^2_\gamma$, and $\hat{\sigma}^2_e$ values as the initial mean and variance components. The procedure will produce an α_{0i} $(= \alpha_0)$ value for each α_i.

(2) The α_i value corresponding to $\alpha_{0i} = \alpha$ is chosen as the new nominal α, which will be called α'. With the computational method, α' is computed by interpolation (Figure 1). With the bootstrap method, the α_i value corresponding to the α_{0i} closest to α is used as α'.

(3) With the computational method, the variance components and the mean in step 1 are used to compute $k_{p,\alpha'}$, $s_{\bar{x}}$, and the final bound $b_{p,\alpha'}$. With the bootstrap method, a final $k_{p,\alpha'}$ is first computed as the mean of the $k_{p,\alpha',j}$ $(j = 1$ to $N_1)$ values generated in the calibration procedure. Then $b_{p,\alpha'}$ is computed from s (using the variance components in step 1), $\bar{x}$, and the final $k_{p,\alpha'}$.

With the computational method, it is known that the noncoverage of the bounds is almost always less than α before calibration, therefore all $a_i \geq \alpha$ only calibrate upward. In the analyses shown in the Results section, $\alpha_i = \alpha, \alpha + d, \alpha + 2d, \ldots, \alpha + 5d$. For $\alpha < .031$, $d = .005$; for $\alpha > .19$, $d = .02$; otherwise, $d = .01$. This provided a greater range of α_i values for larger α.

The rates of noncoverage for the uncalibrated bootstrap method are not as predictable,

Design A - Small, Unbalanced

Levels of δ \ Levels of β	1	2	3
1	3	1	2
2	2	3	1
3	3	2	3

Design B - Balanced w. Replication

Levels of δ \ Levels of β	1	2	3	4	5	6	7
1	3	3	3	3	3	3	3
2	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
4	3	3	3	3	3	3	3

Design C - Balanced w/out replication

Levels of δ \ Levels of β	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1

Design D - Unbalanced

Levels of δ \ Levels of β	1	2	3	4	5	6	7
1	0	3	0	4	2	0	3
2	2	2	2	1	3	0	2
3	3	3	1	3	4	2	1
4	1	2	3	1	3	3	1
5	4	1	4	2	1	3	4

Design E - Very unbalanced

Levels of δ \ Levels of β	1	2	3	4	5	6	7
1	1	3	0	0	0	0	0
2	0	4	2	1	0	0	0
3	0	0	0	3	2	0	0
4	0	0	0	0	2	2	0
5	0	0	0	0	0	1	3

Design F - Nested

Levels of δ \ Levels of β	1	2	3	4	5	6	7
1	1	3	0	0	0	0	0
2	0	0	2	1	0	0	0
3	0	0	0	0	3	2	0
4	0	0	0	0	0	0	2

Figure 2. The Numbers of Cell Replicates for the Designs Used to Evaluate the Methods.

so the α_i include values both above and below α. The potential values used were

$$\alpha_i = .001, .0025, .005, .0075, .01, .02, .03, \ldots, .10, .15, .20, \ldots, .50.$$

The maximum α_i used in this series was 3α.

6. DESIGNS FOR METHODS EVALUATION

To demonstrate the performance of the methods over a range of sampling designs, the tolerance-interval methods are evaluated (see Section 4) using the multiple sampling designs shown in Figure 2. Design A is evaluated over a range of variance components, p, and α values. The other designs assess the effect of design and for the most part are evaluated for a single set of variance components, p, and α values. The logic for the choice of designs is discussed in the Results section.

7. RESULTS

The simulation results evaluating the methods (Section 4) are in Tables 1–4. The designs used are described in Figure 2. The tables show α_0 values, which should approach the

nominal α values used in the simulation. Where variance components are indicated, four 0/1/2 values are shown. These are the actual four variance components used for the simulations. The order of variance components is $\sigma_\delta^2, \sigma_\beta^2, \sigma_\gamma^2$, and σ_e^2, respectively. For example, 1101 indicates that $\sigma_\delta^2 = 1$, $\sigma_\beta^2 = 1$, $\sigma_\gamma^2 = 0$, and $\sigma_e^2 = 1$.

For the computational method, $N_0 = 3{,}000$ (number of simulations, Section 4) and $N_1 = 1{,}000$ (number of calibration simulations, Section 5). For the bootstrap method, $N_0 = 2{,}000$, $N_1 = 300$, and $N_2 = 300$ (number of simulations to compute each $k_{p,\alpha}$, Section 3.3). To speed up the simulations, the values set for N_1 and N_2 were the minimum numbers that produced a reasonably stable result. In practice, higher values for N_1 and N_2 can be used since the process will not be repeated N_0 times, as was the case when evaluating methods.

In the model A scenario, the numbers of rows and columns, and therefore the degrees of freedom, are low. Near worst-case performance from the methods would be expected in this kind of situation. Table 1 shows that the bootstrap and calibrated computational methods

Table 1. Simulation α_0 Values for Design A Over a Range of Variance Component Scenarios. In all simulations, $\alpha = .05$ and $p = .90$. The rows are ordered by the computational with calibration results.

Variance components	*Bootstrap*	*Computational with calibration*	*Computational without calibration*	*Standard*
0001	0.022	0.004	0.001	0.051
0011	0.030	0.012	0.002	0.098
0010	0.021	0.028	0.011	0.205
0111	0.060	0.034	0.013	0.214
1112	0.053	0.035	0.014	0.209
1011	0.048	0.036	0.015	0.206
1222	0.055	0.036	0.013	0.225
1122	0.055	0.038	0.014	0.192
0101	0.085	0.041	0.017	0.230
1001	0.089	0.041	0.018	0.235
2112	0.067	0.043	0.019	0.228
1212	0.075	0.045	0.018	0.237
1121	0.062	0.046	0.021	0.237
1111	0.060	0.047	0.019	0.256
0100	0.040	0.047	0.047	0.465
1000	0.036	0.048	0.048	0.464
2122	0.058	0.049	0.018	0.242
1100	0.031	0.050	0.039	0.389
2121	0.048	0.053	0.026	0.265
1101	0.068	0.054	0.023	0.270
1010	0.052	0.056	0.026	0.313
1221	0.067	0.056	0.025	0.264
2212	0.055	0.056	0.021	0.267
2221	0.050	0.058	0.031	0.274
1110	0.055	0.061	0.034	0.330
0110	0.067	0.065	0.033	0.316
2111	0.077	0.066	0.032	0.283
1211	0.062	0.068	0.035	0.278
2211	0.060	0.070	0.035	0.292
Mean	0.055	0.046	0.023	0.259

Table 2. Simulation α_0 Values for Design A Over a Range of Nominal α Values. In all simulations, $p = .90$ and the variance components are 1111.

Nominal α	Bootstrap	Computational with Calibration	Computational without Calibration	Standard
0.01	0.019	0.011	0.001	0.126
0.03	0.048	0.021	0.008	0.196
0.05	0.060	0.047	0.019	0.256
0.07	0.086	0.063	0.029	0.279
0.10	0.098	0.085	0.050	0.322
0.20	0.221	0.197	0.118	0.422
0.30	0.285	0.307	0.218	0.491
0.40	0.385	0.398	0.317	0.552

Table 3. Simulation α_0 Values for Design A Over a Range of p Values. In all simulations, $\alpha = .05$ and the variance components are 1111.

p	Bootstrap	Computational with Calibration	Computational without Calibration	Standard
0.60	0.056	0.059	0.034	0.240
0.70	0.061	0.061	0.033	0.250
0.80	0.063	0.055	0.027	0.258
0.90	0.060	0.047	0.019	0.256
0.95	0.055	0.041	0.016	0.247
0.99	0.066	0.035	0.017	0.243
Mean	0.060	0.050	0.024	0.249

Table 4. Simulation α_0 Values for Designs B–F. In all simulations, $p = .90$ and $\alpha = .05$.

Design	Variance components	Bootstrap	Computational with Calibration	Computational without calibration	Standard
B	1111	0.046	0.051	0.026	0.337
C	1110	0.054	0.053	0.030	0.224
D	1111	0.061	0.061	0.028	0.295
E	1111	0.046	0.048	0.022	0.205
F	1111	0.047	0.044	0.024	0.157
Mean		0.051	0.051	0.026	0.244

work very well on the average, with mean α_0 values very near the desired value of .05. The generally low α_0 values for the uncalibrated computational method demonstrate the value of the calibration procedure. The generally very high α_0 values for the standard method show the risk of applying the improper statistical model when computing tolerance intervals. When the variance components equal 0001, the standard method performs the best. This result is not surprising since, when only the error variance component is greater than zero, the observations will be independent, and the standard method is appropriate.

Results in Table 2 with design A show both the bootstrap and calibrated computational methods again performing well for a series of nominal α values. Table 3 shows a similar result for a series of p values. Table 4 summarizes the results for some larger models. Again, both the bootstrap and calibrated computational methods work very well. Design C

represents a common situation, where at most a single replicate is taken at each location and time. In this case, the error variance component cannot be estimated and will equal zero. Results for designs E and F show that the degree of unbalance does not adversely affect the methods. In fact, design E is essentially nested since each level of δ contains unique levels of β. This suggests that it is not necessary to formulate a separate computational model for the nested design.

8. APPLICATIONS

Three applications are presented below. In the cases shown, the null hypothesis of normality was accepted ($p > .25$) when a Shapiro–Wilk statistic was applied to the pertinent data (Shapiro and Wilk 1965), so a parametric approach seemed justified. The sampling designs included observations taken over time and space, so the two-way random crossed model was appropriate. All computations were performed using the parametric bootstrap method.

8.1 BENTHIC RESPONSE INDEX

A benthic response index (BRI) has recently been developed as an indicator of pollution in the Southern California Bight (Smith et al., 2001). Index development was based on a large calibration data set of benthic infaunal observations covering a wide area over several years between 1973 and 1994. Using ordination analysis of these data (Smith and Bernstein 1985; Bernstein and Smith 1986), 519 taxonomic categories (mostly species) were given pollution tolerance scores. The index value for an observation is computed as the abundance weighted average pollution tolerance score for all taxonomic categories found in that observation. Higher index values indicate increasing pollution effects. Once the index was developed, threshold index values were established to give ecological meaning to different levels of the index. The first threshold of interest was that between reference and minimally affected conditions. To define this threshold, 147 observations from the calibration data set were chosen as reflecting reference conditions. The resulting design involved 4 years by 117 stations. Each cell of the design contained either zero or one replicate. The distribution of index values is shown in Figure 3. The reference threshold was set at an index value of 25, which was tolerance-interval bound for $p = .90$ and $\alpha = .05$.

Once the index reference threshold of 25 was established, it was of interest to examine the index values around Southern California sewage outfalls to see the possible spatial extent and severity of outfall effects. Figure 4 shows a contour map of index values at one outfall. The effect at the terminus of the outfall appears to be minimal and spatially limited. Slightly elevated index values also appear in shallow water. Possible sources of impact in the shallow area are the nearby Santa Ana River and Newport Bay.

8.2 SEDIMENT CRITERIA FOR CHEMICALS IN SAN FRANCISCO BAY

This project involved using chemical measurements from San Francisco Bay sediments to characterize background or ambient conditions (Smith and Riege 1998). This information

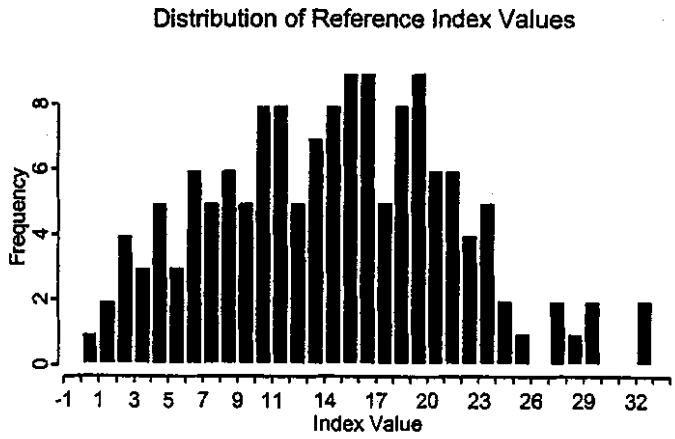

Figure 3. Distribution of BRI Index Values in Reference Areas of the Southern California Bight. The $p = .90$, $\alpha = .05$ upper tolerance-interval bound is 25, which was chosen as the upper threshold for reference conditions.

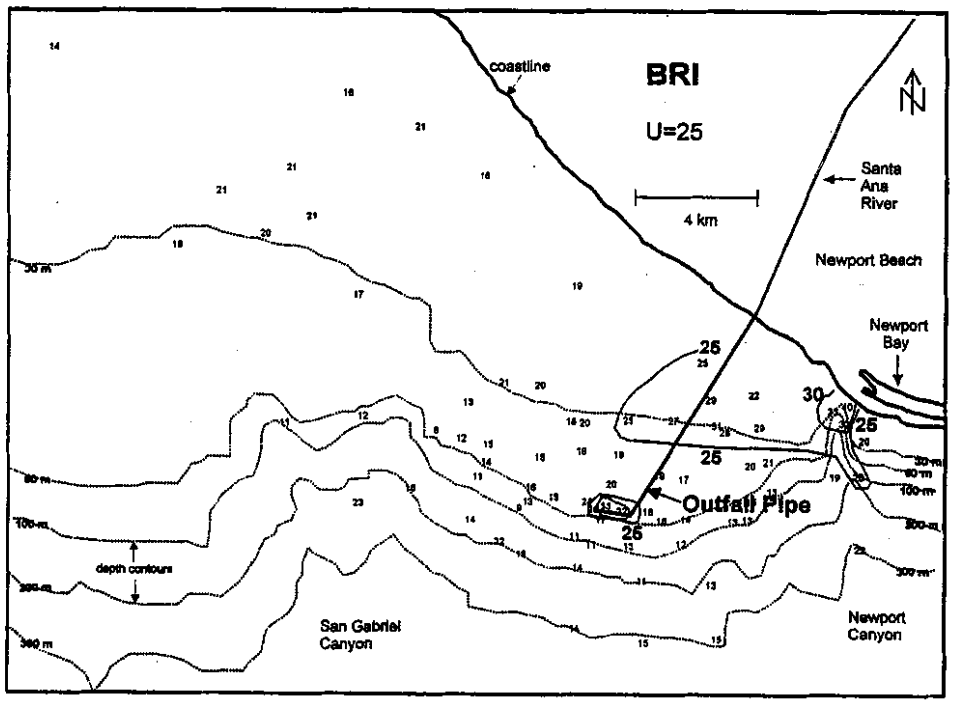

Figure 4. Contour Map of BRI Values in Summer 1994 in the Vicinity of the County Sanitation Districts of Orange County Outfall (From Smith 1998). The numbers on the map are BRI values at the respective locations. Contour lines for 25 and 30 are shown.

would be useful in establishing future sediment criteria for protecting the Bay biological resources. A data set with sediment chemical measurements from 36 locations and 13 sampling times was used to define the reference distribution. Each cell of the design contained 0–3 replicates. Upper tolerance-interval bounds for a series of p values were estimated.

Interpreting the levels of chemicals in sediments is complicated by the fact that the concentrations of most chemicals increase as the sediment particle size decreases. Thus, to be useful, the tolerance-interval bounds need to vary with the sediment size. Metals showed a monotonic increase in concentrations with increasingly finer sediments (% fines). For some metals, the relationship was slightly nonlinear and there was a tendency for more variability in the concentrations of finer sediments. A regression model that accommodates such relationships (Chatterjee and Price 1977) is

$$\log(y_i + c) = \log(a) + \beta x_i + \log(\varepsilon_i),$$

where y_i is the ith chemical measurement, x_i is the % fines in the sediments associated with measurement i, $\log(a)$ is the intercept, β is the slope, $\log(\varepsilon_i)$ is the residual, and c is a constant added to all y_i to prevent indeterminate values when $y_i = 0$, or to provide a better fitting model. Standard least squares linear regression was used to compute $\log(a)$ and β. The tolerance-interval bounds were computed from the residuals, which were adjusted for the effect of sediment size by the βx_i term. The resulting tolerance-interval bound represents a positive distance off the regression line predicting $\log(y_i + c)$. The tolerance-interval upper bounds in the original concentration units were computed as

$$U_i = e^{\log(a)+\beta x_i + u} - c,$$

where U_i is the upper bound for sediment i, u is the tolerance interval bound computed from the residuals, and e is the base of the natural logarithm. This model ignores the uncertainties involved in estimating $\log(a)$ and β, which are small compared with the variability associated with the residuals.

Figure 5 shows the results for nickel. This application contrasts with the alternate approach of defining limits based solely on chemical concentrations where environmental harm is known to appear. Most of the values for nickel were above a level associated with toxicity in test organisms (SFEI 1997). It is very unlikely that a location in the bay can be cleaned up to a completely nontoxic level for a very long period of time since mixing and sediment transport will tend to restore background levels even after the metal is removed from an impacted location. Thus, the tolerance-interval bounds could be used to indicate what sediment criteria limits could be practically enforced.

8.3 IDENTIFICATION OF TOXIC HOT SPOTS IN SAN FRANCISCO BAY

This study was a part of the California Bay Protection and Toxic Cleanup Program (BPTCP). The initial focus of the program has been the identification of toxic hot spots in fine sediments of the Bay (Hunt et al. 1998; Hunt, unpublished manuscript). Hot spots were defined as localized areas where elevated concentrations of toxic pollutants are found in

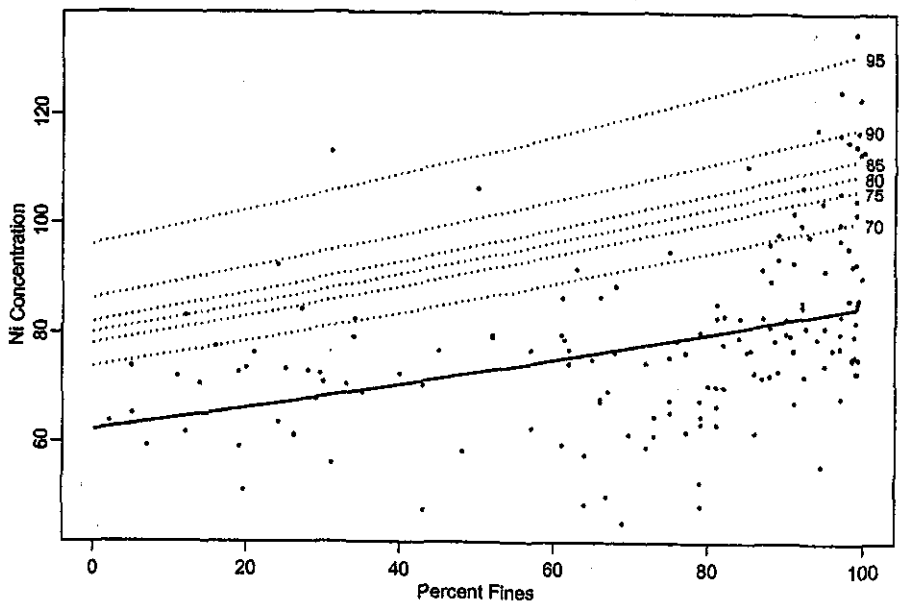

Figure 5. Plot of Nickel Values Versus Percent Fines of the Sediment (From Smith and Riege 1998). The solid line is the regression line, and the dotted lines are the upper tolerance-interval bounds for $p = .70, .75, \ldots, .95$ (labeled on the right), with $\alpha = .05$. Concentration is in mg/kg. Percent fines is percent silt plus clay.

association with adverse biological impacts. Bioassay tests with marine organisms were used to measure biological impacts. However, bioassay tests from sediments in even the cleaner parts of the Bay will usually show some toxicity. A screening tool was needed to distinguish between background toxicity levels and more extreme toxicity. Reference sediment data were obtained from five locations during three survey periods, with three replicates per location-survey. Four cells were empty, giving a total of 33 observations. Several bioassay tests were applied to each sediment sample. Tolerance-interval bounds were computed from the bioassay results to distinguish between background toxicity and more seriously toxic potential hot spots.

The data distribution for the *Ampelisca abdita* bioassay is shown in Figure 6. Here the data variable is a measure of survival, so adverse impacts are associated with lower survival values, and lower tolerance-interval bounds for $p < .50$ are of interest. The computed bounds for $p = .01, .05., .10, .16$, and .20 were computed as 54.7, 65.3, 70.9, 75.1, and 77.5, respectively ($\alpha = .05$). The choice of the value of p to use when screening sediments for toxicity is a regulatory decision that has to balance environmental protection, cost and feasibility of cleanup, and politics.

9. DISCUSSION AND CONCLUSIONS

The simulation results show that both the bootstrap and computational (with calibration) tolerance-interval methods work very well for the two-way crossed random design. They

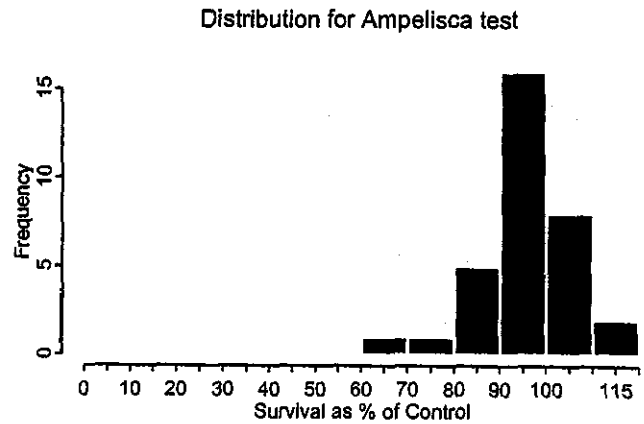

Figure 6. Distribution of Bioassay Results for the Amphipod Ampelisca abdita *(After Hunt et al. 1998).*

also show that using a standard tolerance interval when the crossed model is appropriate can lead to very inflated rates of α.

Often, the assumption of normality of the underlying distribution will be untenable and the parametric computations will produce questionable bounds. If the observations are independent, the nonparametric analogs of the standard method can be used. When the crossed or nested model is appropriate, the following options can be considered:

(1) Transform the data to better approximate normality (Box and Cox 1964) before computing the tolerance intervals. However, for some data, no meaningful transformation will produce sufficient normality. Also, if the data are transformed, the back-transformed bounds should be examined closely. In practice, extreme bounds that may not be useful can sometimes result from this approach.

(2) Presently, there is no nonparametric analog to the tolerance intervals with a crossed or nested random design. One could go ahead and use nonparametric methods for independent observations. To compensate for the fact that the observations are not actually independent, a more extreme p value can be used in place of the planned p value. The relative amount of adjustment needed will depend on the distribution of variance components. For example, in Table 1, it is evident that, as the greater the proportion of variance found in the error and interaction components, the closer the coverage of the bounds is to the nominal α, meaning that less adjustment would be needed.

(3) At times, nonnormality is caused by one or a few outlier observations. From a regulatory perspective, one would not want to set limits overly influenced by a small number of outliers, which may be due to unknown errors or conditions. Using the outliers in the computations could lead to extreme tolerance-interval limits that expose the environment to unreasonable risk. Thus, removal of outliers can produce more environmentally conservative limits and also allow for use of parametric

tolerance intervals if the outlier removal makes the assumption of normality more tenable.

(4) The parametric bootstrap computations could be modified to be appropriate for different (nonnormal) distributions.

Another assumption of the methods is that the levels of the factors and the replicates are randomly selected. The analyst should carefully consider the possible effects of nonrandom sampling on the resulting bounds. If the data are unrepresentative or some other imbalance exists in the distribution in time or space, then balance may need to be restored by removing or obtaining more data.

Frequently, a systematic form of sampling is used. If the reference locations or times are in random order (Gilbert 1987) or in quasi-random order (Barnett 1991), the systematic data can be used as if it were random without biasing the variance estimates. The degree to which the data meet these criteria will mostly depend on the amount of autocorrelation (Cliff and Ord 1981) among the locations and times. Autocorrelation will prevent the criteria from being met. Often, the autocorrelation will decrease as the temporal or spatial distance between the factor levels increases.

There are some indicators that both increase and decrease in response to pollution. For example, measures of species diversity initially increase in value with moderate increases in organic enrichment and then decrease as the enrichment increases further (Pearson and Rosenberg 1978). In such cases, a two-sided tolerance interval would be useful. Rather than use separate computations for two-sided intervals, it is simpler to compute both lower (with $p < .50$) and upper (with $p > .50$) one-sided bounds. These bounds will be similar to the bounds for a two-sided interval (Hahn and Meeker 1991).

It should be noted that tolerance intervals may not always be the most powerful approach to impact assessment. When sufficient before- and after-impact data are available at reference and impact locations, repeated measures ANOVA models can potentially provide much more powerful tests (Green 1979, 1993; Bernstein and Zalinski 1983; Stewart-Oaten, Murdoch, and Parker 1986; Faith, Humphrey, and Dostine 1991; Underwood 1991, 1993, 1994). However, when an impact is detected with a more powerful approach, the tolerance-interval bounds for reference (if available) can help put the impact in perspective. For example, if the indicator value at an impacted location is still inside or very near the tolerance-interval bound, then the impact could be judged as minimal. Indicator values further outside the reference bound would be associated with more serious impacts.

Time is a random component in the proposed tolerance-interval statistical model. Interestingly, some of these repeated measure ANOVA models mentioned in the previous paragraph consider time as a fixed factor (Green 1993; Underwood 1993, 1994), while others consider time a random factor (Bernstein and Zalinski 1983; Stewart-Oaten et al. 1986). VanLeeuwen, Murray, and Urquhart (1996) include both a fixed temporal trend component and a random time (year) component. The tolerance-interval statistical model could be modified to include time as a fixed factor. However, when time is fixed, the $\bar{x}$ in Equations (3.1) and (3.2) will need to be the mean for the time period of the observation

being compared with the tolerance-interval bound. To estimate this mean, a sample of spatial locations for the time period of each future observation would be required. Thus, the tolerance-interval model with random time is more general in a sense that it can be applied to observations whether or not there is any replication of locations during the time period of the observation.

As with tolerance intervals, prediction intervals (Whitmore 1986; Hahn and Meeker 1991; Vardeman 1992) are also useful for statistically defining boundaries enclosing proportions of the data distribution from which a random sample is taken. Prediction intervals define limits that will apply to a specific chosen number of future observations and are very useful where a monitoring program is well defined with specific planned comparisons to determine impact (Davis and McNichols 1987; Boswell, O'Connor, and Patil 1994; Gibbons 1994, 1996). My focus has been on applications where a limit is set and the number of future comparisons to that limit is unknown. For this situation, tolerance intervals are more appropriate since the tolerance-interval limits defined apply to all future observations regardless of the actual number of observations this might involve.

The program used to compute tolerance intervals and evaluate the methods is available upon request from the author.

ACKNOWLEDGMENTS

Partial financial support for this project was supplied by California State Water Resources Control Board, EcoAnalysis Inc., EPA Region 9, and the San Francisco Bay Regional Water Quality Control Board. Craig Wilson, Tom Gandesbery, Don Stevens, Terry Fleming, John Hunt, Bruce Thompson, and Karen Taberski provided encouragement and valuable discussions. Steve Weisberg, Don Stevens, Michael Kellogg, and two anonymous referees provided suggestions for improving the manuscript.

[Received November 1999. Accepted March 2001.]

REFERENCES

Allen, O. B., and Jones, R. (1998), "Distribution-Free Estimates of Quantiles of the Distribution of a Contaminant in Environmental Media," *International Journal of Environment and Pollution*, 9, 140–151.

Bagui, S. C., Bhaumik, D. K., and Parnes, M. (1996), "One-Sided Tolerance Limits for Unbalanced M-Way Random-Effects ANOVA Models," *Journal of Applied Statistical Science*, 2/3, 135–148.

Barnett, V. (1991), *Sample Survey Principles and Methods*, New York, Oxford University Press.

Beckman, R. J., and Tietjen, G. L. (1989), "Two-Sided Tolerance Limits for Balanced-Effects ANOVA Models," *Technometrics*, 31, 185–197.

Bernstein, B. B., and Smith, R. W. (1986), "Community Approaches to Monitoring," *IEEE Oceans '86 Conference Proceedings*, 934–939.

Bernstein, B. B., and Zalinski, J. (1983), "An Optimum Sampling Design and Power Tests for Environmental Biologists," *Journal of Environmental Management*, 16, 35–43.

Bloom, S. A. (1980), "Multivariate Quantification of Community Recovery," in *The Recovery of Damaged Ecosystems*, ed. J. Cairns, Jr., Ann Arbor, MI: Ann Arbor Science, pp. 141–151.

Boswell, M. T., O'Connor, J. S., and Patil, G. P. (1994), "A Crystal Cube for Coastal and Estuarine Degradation: Selection of Endpoints and Development of Indices for Use in Decision Making," in *Handbook of Statistics*, eds. G. P. Patil and C. R. Rao, Amsterdam: Elsevier Science B. V., pp. 771–790.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211–252.

Chatterjee, S., and Price, B. (1977), *Regression Analysis by Example*, New York: Wiley & Sons.

Cliff, A. D., and Ord, J. K. (1981), *Spatial Processes*, London: Poin Limited.

Davies, L., and Gather, U. (1993), "The Identification of Multiple Outliers," *Journal of the American Statistical Association*, 88, 782–792.

Davis, C. B. (1994), "Environmental Regulatory Statistics," in *Handbook of Statistics*, eds. G. P. Patil and C. R. Rao, Amsterdam: Elsevier Science B. V., pp. 817–865.

Davis, C. B., and McNichols, R. J. (1987), "One-Sided Intervals for at Least p of m Observations From a Normal Population on Each of r Future Occasions," *Technometrics*, 29, 359–370.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Faith, D. P., Humphrey, C. L., and Dostine, P. L. (1991), "Statistical Power and BACI Designs in Biological Monitoring: Comparative Evaluation of Measures of Community Dissimilarity Based on Benthic Macroinvertebrate Communities in Rockhole Mine Creek, Northern Territory, Australia," *Australian Journal of Marine and Freshwater Research*, 42, 589–602.

Gibbons, R. D. (1994), *Statistical Methods for Groundwater Monitoring*, New York: John Wiley & Sons.

——— (1996), "Some Conceptual and Statistical Issues in Analysis of Groundwater Monitoring Data," *Environmetrics*, 7, 185–199.

Gilbert, R. O. (1987), *Statistical Methods for Environmental Pollution Monitoring*, New York: Van Nostrand Reinhold.

Green, R. H. (1979), *Sampling Design and Statistical Methods for Environmental Biologists*, New York: Wiley-Interscience—John Wiley & Sons.

——— (1993), "Application of Repeated Measures Designs in Environmental Impact and Monitoring Studies," *Australian Journal of Ecology*, 18, 81–98.

Hahn, G. J., and Meeker, W. Q. (1991), *Statistical Intervals. A Guide for Practitioners*, New York: John Wiley & Sons.

Hughes, R. M. (1995), "Defining Acceptable Biological Status by Comparing With Reference Conditions," in *Biological Assessment and Criteria*, eds. W. S. Davis and T. P. Simon, Boca Raton: Lewis, 31–47.

Hughes, R. M., Whittier, T. R., Rohm, C. M., and Larsen, D. P. (1990), "A Regional Framework for Establishing Recovery Criteria," *Environmental Management*, 14, 673–683.

Hunt, J. W., Anderson, B. S., Phillips, B. M., Newman, J., Tjeerdema, R. S., Puckett, H. M., Stephenson, M., Fairey, R., Smith, R. W., and Taberski, K. M. (1998), "Evaluation and Use of Sediment Reference Sites and Toxicity Tests in San Francisco Bay," Technical Report, The California State Water Resources Control Board.

Hurlbert, S. H. (1984), "Pseudoreplication and the Design of Ecological Field Experiments," *Ecological Monographs*, 54, 187–211.

Jackson, S., and Brashers, D. E. (1994), *Random Factors in ANOVA*, Sage University Paper Series on Quantitative Applications in the Social Sciences 07-098, Newbury Park: Sage.

Johnson, M. E. (1987), *Multivariate Statistical Simulation*, New York: John Wiley & Sons.

Kilgour, B. W., and Somers, K. M. (1998), "The Statistics of Testing Generic Biological Criteria," *SETAC News*, 18, 20–22.

Milliken, G., and Johnson, D. E. (1984), *Analysis of Messy Data* (Vol. I), *Designed Experiments*, New York: Van Nostrand Reinhold.

Pearson, T. H., and Rosenberg, R. (1978), "Macrobenthic Succession in Relation to Organic Enrichment and Pollution of the Marine Environment," *Oceanography and Marine Biology Annual Review*, 16, 229–311.

Portugal, S. (1992), "A SAS Program to Compute Factors for One-Sided Tolerance Limits for a Normal Distribution," *Proceedings of the 17th Annual SAS Users' Group International Conference*, 1231–1234.

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110–114.

Searle, S., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: John Wiley and Sons.

SFEI. (1997), *1996 Annual Report: San Francisco Estuary Regional Monitoring Program for Trace Substances*, Richmond, CA: San Francisco Estuary Institute.

Shapiro, S. S., and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52, 591–611.

Smith, R. W. (1995), *The Reference Envelope Approach to Impact Monitoring*, Technical Report, U.S. EPA, Region IX.

——— (1998), *Analysis of 1994 Orange County Outfall Benthic Data in a Regional Context*, Technical Report, County Sanitation Districts of Orange County, Fountain Valley, CA.

Smith, R. W., Bergen, M., Weisberg, S. B., Cadien, D., Dalkey, A., Montagne, D., Stull, J. K., and Velarde, R. G. (2001), "Benthic Response Index for Assessing Infaunal Communities on the Mainland Shelf of Southern California," *Ecological Applications*, 11, 1073–1087.

Smith, R. W., and Bernstein, B. B. (1985), *Index 5: A Multivariate Index of Benthic Degradation*. Technical Report, Brookhaven National Laboratory for NOAA.

——— (1996), "Quantifying and Testing for Effects of Outfalls on Biological Communities," *Proceedings of the Oceans '96 Monitoring Strategies Symposium*, 2, 619–623.

Smith, R. W., and Riege, L. (1998), "San Francisco Bay Sediment Criteria Project: Ambient Analysis Report," Technical Report, California Regional Water Quality Control Board, San Francisco Region.

Splitstone, D. E. (1991), "How Clean Is Clean, Statistically?" *Pollution Engineering*, March, 90–96.

Stewart-Oaten, A., Murdoch, W. W., and Parker, K. R. (1986), "Environmental Impact Assessment: 'Pseudoreplication' in Time?" *Ecology*, 67, 929–940.

Underwood, A. J. (1991), "Beyond BACI: Experimental Designs for Detecting Human Environmental Impacts on Temporal Variations in Natural Populations," *Australian Journal of Marine and Freshwater Research*, 42, 569–587.

——— (1993), "The Mechanics of Spatially Replicated Sampling Programmes to Detect Environmental Impacts in a Variable World," *Australian Journal of Ecology*, 18, 99–116.

——— (1994), "On Beyond BACI: Sampling Designs That Might Reliably Detect Environmental Disturbances," *Ecological Applications*, 4, 3–15.

Vangel, M. G. (1994), "One-Sided Nonparametric Tolerance Limits," *Communications in Statistics—Simulation*, 23, 1137–1154.

VanLeeuwen, D., Murray, L. W., and Urquhart, N. S. (1996), "A Mixed Model With Both Fixed and Random Trend Components Across Time," *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 435–453.

Vardeman, S. (1992), "What About the Other Intervals?" *The American Statistician*, 46, 193–197.

Whitmore, G. A. (1986), "Prediction Limits for a Univariate Normal Observation," *The American Statistician*, 40, 141–143.

Wiens, J. A., and Parker, K. R. (1995), "Analyzing the Effects of Accidental Environmental Impacts: Approaches and Assumptions," *Ecological Applications*, 5, 1069–1083.

Woodward, W. A., and Frawley, W. H. (1980), "One-Sided Tolerance Limits for a Broad Class of Lifetime Distributions," *Journal of Quality Technology*, 12, 130–137.

APPENDIX: DATA SIMULATION

This appendix briefly describes how simulated data are generated from a set of variance components and a mean. First, a multivariate random normal generator (Johnson 1987) is used to produce cell means in the sampling design. For a simulation, let M be an $r \times c$ matrix of cell means for the crossed design, where r = number of rows and c = number of columns in the design. If m_i is the ith column of M, then

$$m_i = AY_i + X,$$

where Y_i is an $r \times 1$ column vector of $\mathrm{N}(0, 1)$ standard random normal deviates, X is an $r \times 1$ column vector of $\mathrm{N}(\mu, \sigma_\delta^2)$ random normal deviates, and A is an $r \times r$ matrix such that $AA' = \Sigma$. Here Σ is an $r \times r$ variance–covariance matrix with $\sigma_\beta^2 + \sigma_\gamma^2$ in the diagonal

and σ_β^2 in the off-diagonal. Matrix A is computed by Choleski factorization of Σ. Once the cell means in M are computed, the replicate values in each cell of the design are simulated. A data value is simulated as an $N(m_{ij}, \sigma_e^2)$ random normal deviate, where m_{ij} is the cell mean of the ith row and the jth column of M. The number of replicates simulated in each cell equals the number of replicates in the sampling design.