Ecology, 73(4), 1992, pp. 1396–1404

ASSESSING EFFECTS OF UNREPLICATED PERTURBATIONS: NO SIMPLE SOLUTIONS[1]

Allan Stewart-Oaten
Department of Biological Sciences, University of California, Santa Barbara, California 93106 USA

James R. Bence
National Marine Fisheries Service, Southwest Fisheries Science Center, Tiburon Laboratory, 3150 Paradise Drive, Tiburon, California 94920[2] and
Marine Science Institute, University of California, Santa Barbara, California 93106 USA

Craig W. Osenberg
Marine Science Institute, University of California, Santa Barbara, California 93106 USA

Abstract. We address the task of determining the effects, on mean population density or other parameters, of an unreplicated perturbation, such as arises in environmental assessments and some ecosystem-level experiments. Our context is the Before-After-Control-Impact-Pairs design (BACIP): on several dates Before and After the perturbation, samples are collected simultaneously at both the Impact site and a nearby "Control."

One approach is to test whether the mean of the Impact–Control difference has changed from Before to After the perturbation. If a conventional test is used, checks of its assumptions are an important and messy part of the analysis, since BACIP data do not necessarily satisfy them. It has been suggested that these checks are not needed for randomization tests, because they are insensitive to some of these assumptions and can be adjusted to allow for others. A major aim of this paper is to refute this suggestion: there is no panacea for the difficult and messy technical problems in the analysis of data from assessments or unreplicated experiments.

We compare the randomization *t* test with the standard *t* test and the modified (Welch-Satterthwaite-Aspin) *t* test, which allows for unequal variances. We conclude that the randomization *t* test is less likely to yield valid inferences than is the Welch *t* test, because it requires identical distributions for small sample sizes and either equal variances or equal sample sizes for larger ones. The formal requirement of Normality is not crucial to the Welch *t* test.

Both parametric and randomization tests require that time and location effects be additive and that Impact-Control differences on different dates be independent. These assumptions should be tested; if they are seriously wrong, alternative analyses are needed. This will often require a long time series of data.

Finally, for assessing the importance of a perturbation, the *P* value of a hypothesis test is rarely as useful as an estimate of the size of the effect. Especially if effect size varies with time and conditions, flexible estimation methods with approximate answers are preferable to formally exact *P* values.

Key words: environmental assessment; intervention analysis; pseudoreplication; randomization tests.

Introduction

A common problem in basic ecological studies and applied environmental work is to determine whether a particular population, community, or other object of interest has changed after a perturbation to the environment. The answer is often obtained by conducting an experiment, consisting of a number of replicates, each randomly assigned to one of several treatments, and then applying standard statistical analyses.

However, replication with randomly assigned treatments is not always possible. In assessing the effects of a particular power plant, we cannot randomly assign the location of the plant, or build more than one of them. Even in basic ecological work, although we can often randomly assign the perturbation to one or several of the experimental units, costs or the unavailability of replicates may make replication infeasible, particularly in whole ecosystem manipulations (Carpenter 1989, 1990, Carpenter et al. 1989).

In general, the major goal of a study of an unreplicated perturbation is to determine whether the state of the perturbed system differs significantly from what it would have been in the absence of the perturbation. Usually the "state" of the system is the mean value of some univariate or multivariate quantity, such as the population size, average size, or life history parameters

[1] Manuscript received 11 January 1991; revised 26 August 1991; accepted 28 August 1991.

[2] Present address.

of one or more species. We will assume the quantity of interest is univariate, e.g., the population abundance of a single species in a fixed area, although many of the general points we make also apply in the multivariate case.

Because the state of the system in the absence of the effect cannot be observed after the disturbance, we need to estimate what it would have been and compare the estimate statistically with the observed (perturbed) condition. The Before-After-Control-Impact-Pairs (BACIP) design (Stewart-Oaten et al. 1986) accomplishes this by collecting samples at both the Impact site and a nearby "Control" site. These samples are paired, in the sense that the Control and Impact sites are sampled simultaneously (as nearly as possible). Replication comes from collecting such paired samples at a number of times (dates) both Before and After the perturbation.

Each observed difference (e.g., in estimated population density) between the Impact and Control sites during the Before period is taken as an estimate of the mean difference that would have existed in the After period without the perturbation. The observed Impact–Control differences, one for each sample date, constitute a time series; we compare the differences from the Before period to those from the After period; a change in the mean difference indicates that the system at the Impact site has undergone a change relative to the Control site. The general process of estimating a change in a parameter, following a perturbation, has been termed "intervention analysis" (Box and Tiao 1975).

The BACIP design allows for natural differences between the Control and Impact locations, and for changes from the Before to the After period that influence both sites the same way (e.g., resulting from a large-scale change coincident with the putative local impact). Hypothetical examples are shown in Fig. 1.

But the design does not ensure that the assumptions of standard 2-sample tests, for comparing the "Before" set of differences to the "After" set, are satisfied. For the two-sample t test, the assumptions are:

1) Additivity: Time and location (site) effects are additive (i.e., in the absence of the perturbation, the expected Impact–Control difference is the same for all dates).

2) Independence: Observed differences from different dates are independent.

3) Identical Normal Distributions: The distribution of the deviation (observed difference–mean difference) is (a) the same for each time within a period; (b) the same in the After period as in the Before period; (c) Normal.

An adequate analysis must deal with these assumptions, either by supporting them (by arguing for their *a priori* plausibility and/or carrying out tests or other diagnostic procedures) or by showing that the analysis is not sensitive to their violation. This is a messy and complicated part of the analysis, which rarely can dispel doubt altogether. Thus tests needing fewer or more plausible assumptions could be valuable.

Recently, Carpenter et al. (1989) proposed "randomized intervention analysis" (RIA), which employs a BACIP design but uses a randomization test instead of a t test to decide whether there has been a change in the difference between the impact and control sites. They argue that a "distinct advantage" of RIA is that non-Normality does not affect the test results, and imply that this solves problems of temporal trends and time lags. They add that RIA is not affected by heterogeneous variances "unlike . . . the t test," and that the effects of serial correlation will often not lead to equivocal results.

We discuss assumptions (1), (2), and (3) in reverse order, with special reference to RIA, the standard t test, and the Welch (or Welch-Satterthwaite-Aspin) modification of the t test for unequal variances (Snedecor and Cochran 1980:97). We argue: (a) RIA's robustness to non-Normality offers little advantage: the two parametric t tests are also little affected by non-Normality unless sample sizes are very small; (b) the Welch t test is approximately valid when the Before and After distributions have different variances; the other two tests are not, unless sample sizes are nearly equal; (c) the Welch t test is approximately valid when the distributions vary within a period; the others are not, although they are approximately valid if the average Before variance is nearly the same as the average After variance; (d) if the successive differences are not independent, none of the tests is valid; they may be approximately valid if the dependence is weak (and the other assumptions hold); (e) if time and location effects are not additive, none of the tests is valid; they may be approximately valid if the effects are approximately additive.

We also discuss the general application of BACIP. We argue (1) that hypothesis testing, either classical or Bayesian, is less important than estimation of the effect's size and ecological assessment of its importance, and (2) that the appropriate statistical methods will often be unavoidably messy: effects may vary with environmental conditions that can be delineated only roughly, and estimates will depend on models, which are based partly on intuition, guesswork, and mathematical convenience, and must be supported by biological arguments and formal and informal diagnostic checks.

In what follows, we assume there are n_B Before dates and n_A After dates; on the i^{th} Before date, the estimated densities were I_{Bi} at the Impact site and C_{Bi} at the Control, for a difference of D_{Bi}. Similarly we have I_{Aj}, C_{Aj}, and D_{Aj} on the j^{th} After date. The average differences are $D_{B\cdot}$ and $D_{A\cdot}$. The randomization test takes the $(n_B + n_A)$ *values* (the D_{Bi}'s and D_{Aj}'s) as given but, under the null hypothesis, their *assignment* to "Before" or "After" is assumed to be random. The P value for

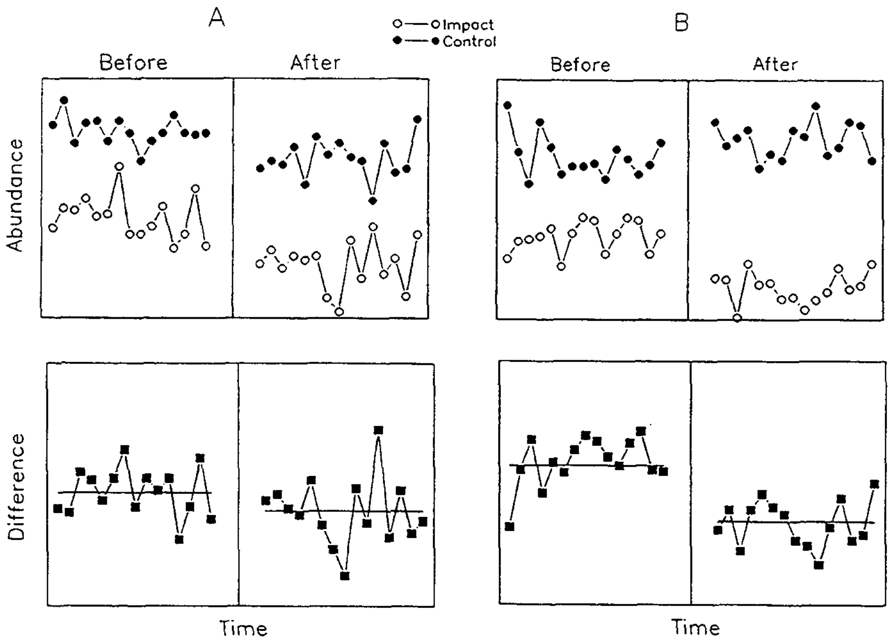

FIG. 1. Hypothetical examples of data collected using the Before-After-Control-Impact-Pairs (BACIP) design. (A) A case where average abundance is greater in the Control area than in the Impact area and where average abundance falls from Before to After. Note that the average difference between Impact and Control does not change significantly from Before to After (bottom panel), indicating that there has been no effect of the perturbation. (B) For comparison, a case where the perturbation has reduced the abundance of the species at the Impact site, leading to a decline in the difference from Before to After (bottom panel).

the test is then the fraction of the $(n_B + n_A)!/(n_B!n_A!)$ possible assignments that give a larger value of the test statistic than was actually observed (Pratt and Gibbons 1981: Chapter 6). The randomization t test uses $D_{B\cdot} - D_{A\cdot}$ (or an equivalent) as the test statistic.

IDENTICAL NORMAL DISTRIBUTIONS

It is likely that one or more of these assumptions will fail. Many biological observations are non-Normal. Even without perturbation effects, distributions may well change between periods (Before and After), e.g., due to long-term weather patterns. They may also change *within* periods, e.g., the variance of an estimate of population density may be greater in summer than in winter.

Parametric t *tests*

Non-normality.—Strong evidence that the standard and Welch t tests are little affected by non-Normality comes from studies of both large and small sample sizes.

For large sample sizes, it is a direct result of the Central Limit Theorem: the usual t statistics are all approximately Normal, provided only that the parent distributions have finite variances.

For small sample sizes, there are a few analytical studies (Efron 1969, Tan 1982) providing evidence of the t test's robustness to non-Normality, but the main evidence comes from simulations, e.g., Yuen and Dixon (1973), Yuen (1974), Murphy (1976), Posten (1978, 1979), Tiku (1980), Gans (1981), Tiku and Singh (1982). Several others are reviewed by Glass et al. (1972).

A serious problem arises only from strong skewness. If $D_{B\cdot}$ and $D_{A\cdot}$ have different skewness, or have the same (non-zero) skewness but different variances, then $D_{B\cdot} - D_{A\cdot}$, the numerator of the t statistics, will have a skewed distribution. But such skewness is unlikely to be strong. Since I_{Bi} and C_{Bi} are estimates of similar things (e.g., population densities) based on similar sampling effort, they are likely to have similar skewness and variance: most of the skewness should cancel in the difference, $D_{Bi} = I_{Bi} - C_{Bi}$. More skewness is lost by averaging to get $D_{B\cdot}$, and still more in the difference, $D_{B\cdot} - D_{A\cdot}$, if these are similarly skewed, as is likely. If histograms of the D_{Bi}'s and D_{Aj}'s show pronounced skewness that is likely to persist through averaging and differencing, a modification of the Welch t test (Cressie and Whitford 1986) seems to solve the problem.

Distributions change between periods.—This creates little problem unless both variances and sample sizes are unequal, in which case the Welch t test is approximately valid, but the standard t test is not. The two

t statistics have the same numerator, $D_{B\cdot} - D_{A\cdot}$, which is approximately Normal by the Central Limit Theorem, except for the problem of skewness just described. Validity depends on the denominator, whose square should approximate the variance of $D_{B\cdot} - D_{A\cdot}$, with a relative error that approaches 0 as sample size increases.

The variance of $D_{B\cdot} - D_{A\cdot}$ is $S^2 = \sigma_B^2/n_B + \sigma_A^2/n_A$, where the σ^2's are the true variances. For the Welch t test, the denominator is

$$S_W = \surd[s_B^2/n_B + s_A^2/n_A].$$

For the standard t test, the denominator can be written as

$$S_s = \surd[M_B s_B^2/n_A + M_A s_A^2/n_B].$$

The s^2's are the usual variance estimates and, for $Z = A$ or B,

$$M_Z = (1 - 1/n_Z)/[1 - 2/(n_B + n_A)].$$

As n_B and n_A approach ∞, the s^2's approach the σ^2's and the M_Z's approach 1. Thus $S_W{}^2$ does approach S^2, but

$$S_s{}^2 \text{ approaches } s_B^2/n_A + s_A^2/n_B = R^2S^2,$$

where $R^2 = (r\sigma_B^2 + \sigma_A^2)/(\sigma_B^2 + r\sigma_A^2)$ and $r = n_B/n_A$.

This shows that the Welch t test gives the correct level for large sample sizes; the simulations mentioned above verify this for moderate sample sizes. It also shows that, when its nominal level is α, the standard t test rejects the (true) null hypothesis that the Before mean is less than or equal to the After mean with probability approximately $\Phi(-Rz_\alpha)$, where Φ is the standard $N(0, 1)$ cumulative distribution function and z_α is the point for which $\Phi(-z_\alpha) = \alpha$. If either the sample sizes or the variances are approximately equal, $R \approx 1$ and the test is approximately valid. But if the smaller sample comes from the distribution with the larger variance, $R < 1$ and the test rejects more frequently than advertised. In the reverse case, it rejects less frequently.

One option is to use a test of equality of variances to decide whether to use the standard or the Welch t test. Simulations by Gans (1981) indicate that this is inferior to direct use of the Welch t test. In particular, it rejects too frequently when $R < 1$ and one variance is about half the other.

Distributions change within periods.—For this problem, some general results are given for estimates of location by Stigler (1976), for one-sample t tests by Cressie (1982) and for two-sample t tests by Cressie and Whitford (1986). The main large sample results are similar to those just described. The numerator of both t tests, $D_{B\cdot} - D_{A\cdot}$, is approximately Normal. (There is a condition for this, roughly that the variances not be so dissimilar that most of the variability of $D_{B\cdot}$ or $D_{A\cdot}$ comes from a small subset of the observations; see Feller 1966:491.) Its variance is $\sigma_{B\cdot}{}^2/n_B + \sigma_{A\cdot}{}^2/n_A$, where $\sigma_{B\cdot}{}^2 = \Sigma\, \sigma_{Bi}{}^2/n_B$ and $\sigma_{Bi}{}^2$ is the variance of the i^{th} "Before" difference. The Welch t test is approximately valid in general, because $S_w{}^2$ approaches this variance. $S_s{}^2$ does so only if $\sigma_{B\cdot}{}^2 = \sigma_{A\cdot}{}^2$, i.e., the standard t test is valid for unequal sample sizes only if the average Before and After variances are the same.

For moderate sample sizes, the Welch t test may be "liberal": its true rejection probability may be slightly greater than the nominal value because its degrees of freedom are overestimated. The standard formula divides an estimate of $2\{E[S_w{}^2]\}^2$ by an estimate of $V[S_w{}^2]$. With heterogeneous variances, the latter estimate, $s_B{}^4/n_B{}^2(n_B - 1) + s_A{}^4/n_A{}^2(n_A - 1)$, is biased low: roughly, for Normal variables, $s_B{}^4$ approaches $(\sigma_{B\cdot}{}^2)^2$ instead of the desired $\Sigma\, \sigma_{Bi}{}^4/n_B = (\sigma_{B\cdot}{}^2)^2 + V(\sigma_B{}^2)$, where $V(\sigma_B{}^2)$ is the variance of the set $\sigma_{B1}{}^2, \sigma_{B2}{}^2, \ldots$ (Cressie and Whitford 1986). But, since variances must be positive, $V(\sigma_B{}^2)$ is unlikely to be significantly larger than $(\sigma_{B\cdot}{}^2)^2$, which is overestimated by $s_B{}^4$, so the correct degrees of freedom are likely to be at least half the nominal value. If the nominal value is 30 or more, this error has little effect. Unfortunately, we know of no simulation studies of this case.

Randomization tests

The assumptions for randomization tests (which are sometimes called permutation tests) are usually satisfied in experiments by the investigator's deliberate random assignment of units to treatments. This is not possible in intervention analysis: one cannot randomly assign sampling times to "Before" and "After." Instead it is assumed that "Nature" does the random assigning: under the null hypothesis, the "Before" and "After" observations are assumed to be independent draws from a common distribution.

Thus all of the assumptions listed in the *Introduction* are required, except only assumption 3(c), Normality. The user of RIA must show either that these assumptions hold or that RIA remains valid when they fail.

The randomization test is *not* valid for unequal variances. For large sample sizes, it is invalid in the same way, and to the same extent, as the standard t test discussed above. The limiting level and power of the randomization test are the same as those of the standard t test. For equal variances, this result was proved by Hoeffding (1952), with the restriction that the original distributions have finite third absolute moments, in our notation, $E\,|D_{Bi}|^3 < \infty$ and $E\,|D_{Ai}|^3 < \infty$, which is satisfied in almost all realistic cases. Romano (1990) proves it without requiring either equal variances or the third moment restriction. Our moderate sample (20 and 40) simulations with Normal variables agreed closely with these asymptotic results. One of us has also extended Hoeffding's proof to the case where variances change within periods (A. Stewart-Oaten, *unpublished manuscript*): Romano's work suggests the third moment restriction is unnecessary here, too. Romano also shows that the randomization test based on medians is invalid for non-identical distributions, even for equal sample sizes, unless the Before and After probability densities at their medians are equal or satisfy an unlikely condition.

For small sample sizes, it is easy to construct examples for which the randomization t test is invalid for non-identical distributions, even when the variances are the same.

INDEPENDENCE

Standard two-sample tests, including t tests and randomization tests (when these are based on randomization by "Nature" rather than by an experimenter), assume that the D_{Bi}'s and D_{Ai}'s are independent.

In the assessment problem, the most likely violation is positive serial correlation: observations (D_{Bi}'s and/or D_{Aj}'s) close in time may tend to be close in value. In this case, the variance of the average of the differences, e.g., $V(D_{B\cdot})$, is no longer the variance of a single observation divided by the sample size, e.g., $V(D_{Bi})/n_B$, but is larger. If this is not allowed for, all these tests will reject true null hypotheses more frequently than advertised, because observed averages will be less precise than they are assumed to be.

The observed D_{Bi}'s and D_{Aj}'s vary for two reasons. One is sampling error: the estimated Impact–Control difference at a given sampling time will not exactly equal the true difference at that time. But our concern is not with this "true difference," which itself varies naturally over time: any particular Before and After values are almost certain to be different even if there is no perturbation effect. Our concern is with the mean of the "true difference," i.e., the mean of the stochastic process of which the entire set of true differences over a period is a single realization (see Stewart-Oaten et al. 1986).

Correlation can arise from the second source of variation: the deviation between the true difference and its mean. This potential problem has been termed "pseudoreplication in time" (Hurlbert 1984). Two deviations will be correlated if the time between them is short enough that the same random events (births, deaths, movements, etc.) play significant roles in both. The variation in the true difference would then be underrepresented in the sample, leading to underestimation of the variance of $D_{B\cdot}$ or $D_{A\cdot}$.

Whether serial correlation in the observed differences is sufficient to invalidate the test for an effect must be assessed by formal tests and by *a priori* arguments and models based on knowledge of the populations under study. Stewart-Oaten et al. (1986) present arguments and a simple (though easily extended) model suggesting that, provided the additivity assumption holds, only large, local events (occurring at one site but not the other) should introduce serial correlation. Non-local events (e.g., storms) should have similar population consequences at both Impact and Control, and thus cancel (at least approximately) when we take differences. Small events (e.g., individual births and deaths) should not affect the populations for far into the future, and are likely to be swamped by the sampling errors (which are independent).

Correlation should be insignificant if sampling times are sufficiently separated so that a single event is unlikely to have a large local effect for more than one time. Arguments and models indicating how large a separation is needed should depend on the organism. For some populations, e.g., those which are short lived, highly mobile or strongly density dependent, local changes, even if large, will have only a brief effect.

For others, observations a year or more apart may be significantly correlated. A sedentary species whose larvae or seeds disperse unevenly in space over a short annual recruitment/settlement period is likely to have much the same local population within a year (between one recruitment period and the next) but quite different populations between years: variation in recruitment might be a long-lasting large local effect. Another case occurs when dispersion between Impact and Control sites is rare, as for lakes. For example, Osenberg et al. (1988) analyzed size-specific growth rates of sunfish in eight lakes over a 10-yr period and found that half of the interpretable variation arose from lake X year interactions. Since the growth of these fishes is closely tied to the availability of their resources (Mittelbach 1988, Mittelbach et al. 1988, Osenberg et al. 1988, Osenberg and Mittelbach 1989), these data suggest that the abundances of the invertebrate prey also exhibit lake X year effects. Some fish populations are also known to exhibit dramatic population cycles that may result from strong age class interactions (e.g., Aass 1972, Hamrin and Persson 1986, Townsend 1989), and the timing of these cycles may well vary from lake to lake. If variation in fish density cascades to lower trophic levels (Carpenter et al. 1987), then this could introduce local year (or even longer period) effects in a number of biological variables measured in a BACIP study.

There is no guaranteed resolution of these uncertainties. Whatever testing procedure is used should be derived from a model that is plausible and survives diagnostic checking against the data, both formal tests and informal inspection, especially plots. The plausibility is important. For example, a single year of data would be insufficient for a test of serial correlation in the examples just given, since the main source of variation, between years, is never observed. An implausible model might survive diagnostic checking in these cases, and could then be used to indicate a "perturbation effect" that was really natural year-to-year variation. In most cases, we would expect several years of Before and After data to be needed, with serial correlation of the Before differences checked by the Durbin-Watson (Durbin and Watson 1971) and Ljung-Box (Ljung and Box 1978) tests, and one-way ANOVA, using years as "treatments."

If serial correlation appears significant, either *a priori*

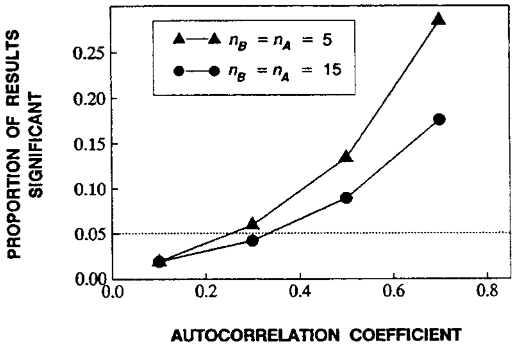

FIG. 2. An example of the effect of serial correlation on the randomization t test. The proportion of results that were significant (when H_0 is true and testing is done at the .01 level) is plotted against the value of the autocorrelation coefficient. Results are shown for sample sizes of 5 and 15 in each period. In all cases results are based on 5000 simulated trials, and randomization tests were based on the random selection of 5000 permutations. Data were generated from the same Gaussian autoregressive model of order one separately for each period.

or as a result of tests, the test for a change needs to be based on a model that includes plausible representations of the non-ignorable types of correlation (e.g., Box 1954, Box and Tiao 1965, 1975, Tiao et al. 1975, Jones 1980, 1981, McDowall et al. 1980), and is itself subjected to diagnostic checks (Box 1980).

Carpenter et al. (1989) recognize that serial correlation can inflate Type I error rates in randomization tests, but suggest that the rule "reject if the nominal P value is $< .01$" gives a conservative .05-level test. This rule lacks generality and seems to us undesirable. First, if the correlation is weak, this test is too conservative and is inefficient. Second, if the correlation is strong enough, the test is invalid. Fig. 2 shows that, for samples of 15 from a first-order autoregressive model with equal variances, the test is invalid if $r > 0.3$.

ADDITIVITY

Suppose the two populations vary but tend to track one another so that the density in the Impact area is typically 50% of that in the Control area. Then the difference between the raw Impact and Control densities will also vary. The effects of location and time on the means at the Impact and Control sites are not additive: the time effect does not cancel when we take the differences. Such non-additivity has three consequences.

Two arise when there is systematic (e.g., seasonal) variation in the overall density. The mean Impact–Control difference (the mean of the stochastic process mentioned in the previous section) then varies over time. The correct model for the data will not be the one the test is based on, i.e., $D_{Bi} = \mu_B + \epsilon_{Bi}$ and $D_{Ai} = \mu_A + \epsilon_{Ai}$, where the errors, ϵ_i, have mean zero, but $D_{Bi} = \mu_B + N_{Bi} + \epsilon_{Bi}$ and $D_{Ai} = \mu_A + N_{Ai} + \epsilon_{Ai}$ where the N_i's are non-random.

One consequence is that the test for an effect is not comparing μ_B with μ_A but comparing $\mu_B + N_{B\cdot}$. with $\mu_A + N_{A\cdot}$. These could differ solely because of the choice of sampling times, e.g., if the fraction of summer samples is higher in the Before period than in the After. Of course, we can balance the samples with respect to seasons, but there may be other cycles, perhaps unknown, that are not balanced.

Second, if we have balanced cycles, the N_i's will not bias the estimates of the means, but they will add to the estimated variances: the test will be more conservative (and less efficient) than it should be.

The third consequence arises from random natural variation, such as major storms or long spells of unusual weather. This changes densities in both areas; without additivity, it also changes their difference. Thus region-wide, long-lasting random variation may not tend to be cancelled when we take differences. The assumption that the observed differences are independent is then less plausible.

For hypothesis testing, the obvious way to satisfy the additivity assumption is to transform the data. If the data are multiplicative, as in the example above, we would expect to transform to logs. In practice, the "right" transformation may not be known, and various methods have been suggested for choosing a transformation in this situation (Tukey 1949, Box and Cox 1964, Andrews 1971, Carroll and Ruppert 1981, 1984, Hinkley and Runger 1984).

It may be that there is no monotone transformation for which the data (or the underlying process that produced them) are additive. For example, it may be that Impact densities are higher than Control in winter, but are lower in summer. In such cases a different analysis may be better. We return to this below. The main message is that the problem of non-additivity cannot be ignored, regardless of whether the final test is a t test, a randomization test, or something else.

EFFICIENCY

Validity is not the only important consideration in the choice of tests. We also want a test that is efficient, i.e., which has good power.

All three of the tests discussed here can be inefficient, because they are based on the Before and After sample averages. The average is, in some non-Normal cases, an inefficient estimator: for a given sample size, there are other unbiased estimators with much smaller variances for non-Normal distributions and only slightly larger variances for Normal distributions (Andrews et al. 1972). These "efficiency robust" estimators maintain small variances against a range of distributions by reducing the influence of the extreme observations.

A major virtue of randomization tests is the possibility of greater efficiency, from the use of robust estimates whose distributions are hard to determine, e.g., the median. However, as we have seen, these tests are likely to be invalid when the null distributions are not identical, as in the assessment problem.

Fortunately, there are robust estimators whose variances can be estimated. These can be used as the basis for "t-like" tests (both standard and Welch). Examples include trimmed means (Yuen 1974), biweight estimators (Kafadar 1982), modified maximum likelihood estimators (Tiku and Singh 1982), and many others (Andrews et al. 1972). Perhaps the easiest to use are the trimmed means, although the biweight may be the most efficient overall (Gross 1976).

In many cases a reasonable approach is to use both a Welch t test and an efficient Welch t-like test. Only if they disagree is there a problem requiring a closer look at the data. Then the focus might well be on any skewness that might cause the tests to be testing different things: if so, the investigator needs to decide what kind of change is of concern.

Discussion

A main point of this paper is that there is no panacea for the difficult and messy technical problems in the analysis of data from assessments or unreplicated experiments using the BACIP design. Statistical analyses must be based on plausible models, themselves based on *a priori* empirical and theoretical arguments and checked by formal and informal methods.

In particular, randomization tests are likely to be invalid in assessment if sample sizes are unequal, because a crucial assumption, equal variances of the Before and After deviations, is likely to be violated. The Welch t test is more likely to be valid, because it does not require this assumption, and violation of its Normality assumption is not likely to be important to its validity. However, both tests also require the assumptions of independence and additivity.

We have concentrated on randomization t tests, but similar comments apply to virtually all "distribution free" and nonparametric tests. They require the additivity and independence assumptions and, contrary to frequent suggestions (e.g., Carpenter 1990, Jassby and Powell 1990), are less likely to be valid than are modifications of classical parametric tests, when distributions vary over time and sample sizes are unequal, as must be expected for assessment data.

For the remainder of this paper, we turn from the validity and efficiency of tests of "no effect" to the more important, if less technical, question of their proper role.

The "P value" is the probability that data indicating an effect as strongly as our data do, or more so, would arise by chance if in fact there was no effect. Reckhow (1990) asserts that it is often misinterpreted as the probability that there is no effect, and advocates direct calculation of this probability by Bayesian methods. We disagree.

First, the prudent solution to misinterpretation of classical P values is improved explication rather than dumping the methods.

Second, Bayesian conclusions depend on subjective prior probabilities, which are likely to vary widely, especially in adversarial situations; there is a risk that debates about effects will focus less on the data and more on the credentials of the "experts" whose priors are invoked. For example, Reckhow (1990) claims that P values are misleading because Bayesian calculations of the probability of no effect by Berger and Sellke (1987) are usually much larger. But these calculations are based on a prior probability of ≈ 0.5 that there is indeed no effect. In most assessment problems we would regard this prior probability as quite unrealistic: there is almost certainly *some* effect, so the prior probability of no effect should be close to 0; the Bayesian posterior probability of no effect could then easily be smaller than the P value.

Third, and most important: neither a P value nor a Bayesian posterior probability, for a null hypothesis that is inherently implausible, is adequate for such purposes as making decisions about ending or mitigating the impact, resolving legal disputes, designing future power plants or sewage outfalls, managing ecosystems, or studying the biological mechanisms involved (e.g., National Research Council 1990:76). The important questions are how large the effects are, and whether they matter. The main statistical tasks are estimating effect sizes and estimating the precision of these estimates, not hypothesis testing.

For this, there is a standard classical format, confidence intervals. There are Bayesian alternatives, but the disagreement between the two is usually minor for large or moderate sample sizes (Pratt 1965), provided that the prior distribution does not have a sharp peak. In assessments, where there are usually many interacting species, environmental parameters and physiological processes, many of them poorly understood, we would expect honest prior distributions to be quite diffuse.

Any test can be used to form a confidence interval for the size of the effect: the confidence interval is the set of values, δ, for which the null hypothesis "the change in the difference of the means is δ" is accepted. For many parametric tests, this interval is as easily calculated as the test itself. Randomization tests are much harder to convert, although efficient algorithms exist for some special cases (Pagano and Tritchler 1983, Tritchler 1984).

But not all parametric tests will lead to useful estimates in the assessment problem. If a transformation is needed for additivity, the test will concern the mean difference of transformed data; the ecological significance of a change in this mean may be obscure. In some cases there may be no suitable transformation, e.g., if the "Control" population density is greater than the Impact density in winter but smaller in summer, no monotone transformation can achieve additivity.

Perhaps most important, real perturbation effects might not be constant, even if we have the correct transformation. They may vary seasonally or in response to other conditions. For example, the cooling water system of the San Onofre Nuclear Generating Station may reduce irradiance (and gametophyte survival) over the San Onofre Kelp bed when the current flows South, but increase it when the current flows North (Murdoch et al. 1989).

One way to deal with these problems is to think of the "Control" density and other variables (e.g., season, current direction, water temperature, etc.) as predictors. Using regression methods on the Before data, we could estimate the function that best predicts the Impact area density from these predictors. The perturbation effect could be estimated as the difference between this function and the corresponding function obtained from the After data. This approach allows for effects that vary with environmental conditions, includes quantitative estimates of uncertainty (via confidence bands), and is conducive to graphical presentation, which many audiences may find easier to understand. At least one successful example of this approach already exists (Mathur et al. 1980).

This would not usually be a "clean" approach. It would involve regressions based on guessed functional forms, which would be checked, in part, by formal statistical tests, themselves often approximate. Few, if any, of the confidence intervals could be regarded as exact. Increasing exactness, e.g., by incorporating the uncertainty over functional form into the confidence interval, would be a difficult task, requiring some arbitrary judgments, and probably of little use to readers.

But restricting consideration to questions that allow formally exact answers (or appear to), such as overall tests for an effect, risks losing the information of most value: "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise" (Tukey 1962).

Acknowledgments

We thank Bill Murdoch, Bill Lenarz, Alec MacCall, Russ Schmitt, and Keith Parker for helpful discussion and/or comments on earlier versions of this manuscript. This work was supported in part by NSF grant BSR8905867, the Minerals Management Service, United States Department of the Interior, under MMS Agreement Number 14-35-0001-30471 (The Southern California Educational Initiative); the University of California Coastal Toxicology Program; the Ecology Research Division Office of Health and Environmental Research, United States Department of Energy, grant DE-FG03-89-ER60885, and the Marine Review Committee. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either express or implied, of the United States Government or of the Marine Review Committee.

Literature Cited

Aass, P. 1972. Age determination and year-class fluctuations of cisco, *Coregonus albula* L., in the Mjosa hydroelectric reservoir, Norway. Report of the Institute of Freshwater Research Drottningholm **52**:5–22.

Andrews, D. F. 1971. A note on the selection of data transformations. Biometrika **58**:249–254.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. 1972. Robust estimates of location. Princeton University Press, Princeton, New Jersey, USA.

Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis: the irreconcilability of P-values and evidence. Journal of the American Statistical Association **82**:112–122.

Box, G. E. P. 1954. Some theorems on quadratic forms applied to the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics **25**:484–498.

———. 1980. Sampling and Bayes inference in scientific modelling and robustness (with discussion). Journal of the Royal Statistical Society **A143**:383–430.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society Series **B26**:211–252.

Box, G. E. P., and G. C. Tiao. 1965. A change in level of a non-stationary series. Biometrika **52**:1509–1526.

Box, G. E. P., and G. C. Tiao. 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association **70**: 70–79.

Carpenter, S. R. 1989. Replication and treatment strength in whole-lake experiments. Ecology **70**:453–463.

———. 1990. Large-scale perturbations: opportunities for innovation. Ecology **71**:2038–2043.

Carpenter, S. R., T. M. Frost, D. Heisey, and T. K. Kratz. 1989. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. Ecology **70**: 1142–1152.

Carpenter, S. R., J. F. Kitchell, and J. R. Hodgson. 1987. Cascading trophic interactions and lake productivity. BioScience **35**:634–639.

Carroll, R. J., and D. Ruppert. 1981. On prediction and the power transformation family. Biometrika **68**:609–616.

Carroll, R. J., and D. Ruppert. 1984. Comment on Hinkley and Runger 1984. Journal of the American Statistical Association **79**:312–313.

Cressie, N. A. C. 1982. Playing safe with misweighted means. Journal of the American Statistical Association **77**:754–759.

Cressie, N. A. C., and H. J. Whitford. 1986. How to use the two sample *t* test. Biometrical Journal **28**:131–148.

Durbin, J., and G. S. Watson. 1971. Testing for serial correlation in least squares regression. III. Biometrika **58**:1–19.

Efron, B. 1969. Student's *t* test under symmetry conditions. Journal of the American Statistical Association **64**:1278–1302.

Feller, W. 1966. An introduction to probability theory and its applications. Volume II. Wiley, New York, New York, USA.

Gans, D. J. 1981. Use of a preliminary test in comparing two sample means. Communications in Statistics, Simulation and Computation **B10**:163–174.

Glass, G. V., P. D. Peckham, and J. R. Saunders. 1972. Consequences of failure to meet the assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research **42**:237–298.

Gross, A. M. 1976. Confidence interval robustness with long-tailed symmetric distributions. Journal of the American Statistical Association **71**:409–416.

Hamrin, S. F., and L. Persson. 1986. Asymmetrical competition between age classes as a factor causing population oscillations in an obligate planktivorous fish species. Oikos **47**:223–232.

Hinkley, D. V., and G. Runger. 1984. The analysis of transformed data (with discussion). Journal of the American Statistical Association **79**:302–320.

Hoeffding, W. 1952. The large sample power of tests based on permutations of observations. Annals of Mathematical Statistics **23**:169–192.

Hurlbert, S. J. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**: 187–211.

Jassby, A. D., and T. M. Powell. 1990. Detecting change in ecological time series. Ecology **71**:2044–2052.

Jones, R. H. 1980. Maximum likelihood fitting of ARMA models to time series with missing observations. Technometrics **22**:389–395.

———. 1981. Fitting a continuous time autoregression to discrete data: applied time series analysis II. Pages 651–682 *in* D. F. Finley, editor. Academic Press, New York, New York, USA.

Kafadar, K. 1982. Using biweight m-estimates in the two sample problem. Part 1: symmetric populations. Communication in Statistics, Theoretical Methods **11**:1883–1901.

Ljung, G. M., and G. E. P. Box. 1978. On a measure of lack of fit in time series models. Biometrika **65**:297–304.

Mathur, D., T. W. Robbins, and E. J. Purdy. 1980. Assessment of thermal discharges on zooplankton in Conowingo Pond, Pennsylvania. Canadian Journal of Fisheries and Aquatic Science **37**:937–944.

McDowall, S. P., R. McCleary, E. E. Meidinger, and R. A. Hay. 1980. Interrupted time series analysis. Sage Publications, Beverly Hills, California, USA.

Mittelbach, G. G. 1988. Competition among refuging sunfishes and effects of fish density on littoral zone invertebrates. Ecology **69**:614–623.

Mittelbach, G. G., C. W. Osenberg, and M. A. Leibold. 1988. Trophic relations and ontogenetic niche shifts in aquatic ecosystems. Pages 219–235 *in* B. Ebenman and L. Persson, editors. Size-structured populations. Springer-Verlag, Berlin, Germany.

Murdoch, W. W., B. Mechalas, and R. C. Fay. 1989. Final report of the Marine Review Committee to the California Coastal Commission on the effects of the San Onofre Nuclear Generating Station on the Marine Environment. Available from the California Coastal Commission, San Francisco, California, USA.

Murphy, B. P. 1976. Comparison of some two sample means tests by simulation. Communications in Statistics, Simulation and Computation **B5**:23–32.

National Research Council. 1990. Managing troubled waters. National Academy Press, Washington, D.C., USA.

Osenberg, C. W., and G. G. Mittelbach. 1989. Effects of body size on the predator–prey interaction between pumpkinseed sunfish and gastropods. Ecological Monographs **59**: 405–432.

Osenberg, C. W., E. E. Werner, G. G. Mittelbach, and D. J. Hall. 1988. Growth patterns in bluegill (*Lepomis macrochirus*) and pumpkinseed (*L. gibbosus*) sunfish: environmental variation and the importance of ontogenetic niche shifts. Canadian Journal of Fisheries and Aquatic Sciences **45**:17–26.

Pagano, M., and D. Tritchler. 1983. On obtaining permutation distributions in polynomial time. Journal of the American Statistical Association **78**:435–440.

Posten, H. 1978. The robustness of the two-sample t-test over the Pearson system. Journal of Statistical Computation and Simulation **6**:295–311.

———. 1979. The robustness of the one-sample t-test over the Pearson system. Journal of Statistical Computation and Simulation **9**:133–149.

Pratt, J. W. 1965. Bayesian interpretation of standard inference statements (with Discussion). Journal of the Royal Statistical Society **B27**:169–203.

Pratt, J. W., and J. D. Gibbons. 1981. Concepts of nonparametric theory. Springer-Verlag, New York.

Reckhow, K. H. 1990. Bayesian inference in non-replicated ecological studies. Ecology **71**:2053–2059.

Romano, J. P. 1990. On the behavior of randomization tests without a group invariance assumption. Journal of the American Statistical Association **85**:686–692.

Snedecor, G. W., and W. G. Cochran. 1980. Statistical methods. Seventh edition. Iowa State University Press, Ames, Iowa, USA.

Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology **67**:929–940.

Stigler, S. M. 1976. The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation. Journal of the American Statistical Association **71**:956–960.

Tan, W. Y. 1982. Sampling distributions and robustness of t, F and variance ratio in two samples and ANOVA models with respect to departure from Normality. Communications in Statistics **A11**:2485–2511.

Tiao, G. C., G. E. P. Box, and W. J. Hamming. 1975. Analysis of Los Angeles photochemical smog data: a statistical overview. Journal of the Air Pollution Control Association **25**:260–268.

Tiku, M. L. 1980. Robustness of MML estimators based on censored samples and robust test statistics. Journal of Statistical Planning and Inference **4**:123–143.

Tiku, M. L., and M. Singh. 1982. Robust tests for means when population variances are unequal. Communications in Statistics **A10**:2057–2071.

Townsend, C. R. 1989. Population cycles in freshwater fish. Journal of Fish Biology **35** (supplement A):125–131.

Tritchler, D. 1984. On inverting permutation tests. Journal of the American Statistical Association **79**:200–207.

Tukey, J. W. 1949. One degree of freedom for non-additivity. Biometrics **5**:232–242.

———. 1962. The future of data analysis. Annals of Mathematical Statistics **33**:1–67.

Yuen, K. K. 1974. The two-sample trimmed t for unequal population variances. Biometrika **61**:165–170.

Yuen, K. K., and W. J. Dixon. 1973. Approximate behavior and performance of the two sample trimmed t. Biometrika **60**:369–374.