ATTACHMENT 16

Ecology, 67(4), 1986, pp. 929–940
© 1986 by the Ecological Society of America

ENVIRONMENTAL IMPACT ASSESSMENT: "PSEUDOREPLICATION" IN TIME?[1]

ALLAN STEWART-OATEN AND WILLIAM W. MURDOCH
Department of Biological Sciences, University of California, Santa Barbara, California 93106 USA

AND

KEITH R. PARKER
Marine Review Committee, 531 Encinitas Boulevard, Encinitas, California 92024 USA

Abstract. A recent monograph by Hurlbert raised several problems concerning the appropriate design of sampling programs to assess the impact upon the abundance of biological populations of, for example, the discharge of effluents into an aquatic ecosystem at a single point. Key to the resolution of these issues is the correct identification of the statistical parameter of interest, which is the mean of the underlying probabilistic "process" that produces the abundance, rather than the actual abundance itself. We describe an appropriate sampling scheme designed to detect the effect of the discharge upon this underlying mean. Although not guaranteed to be universally applicable, the design should meet Hurlbert's objections in many cases. Detection of the effect of the discharge is achieved by testing whether the *difference* between abundances at a control site and an impact site changes once the discharge begins. This requires taking samples, replicated in time, Before the discharge begins and After it has begun, at both the Control and Impact sites (hence this is called a BACI design). Care needs to be taken in choosing a control site so that it is sufficiently far from the discharge to be largely beyond its influence, yet close enough that it is influenced by the same range of natural phenomena (e.g., weather) that result in long-term changes in the biological populations. The design is not appropriate where local events cause populations at Control and Impact sites to have different long-term trends in abundance; however, these situations can be detected statistically. We discuss the assumptions of BACI, particularly additivity (and transformations to achieve it) and independence.

Key words: environmental monitoring; impact assessment; independence; pollutants; power plants; replication; serial correlation; statistical transformations; statistics.

INTRODUCTION

Assessing the environmental effects of a single source of pollution within the constraints of a feasible sampling regime can pose difficult statistical problems. Some of these are raised by Hurlbert (1984) in his discussion of "pseudoreplication" in ecology, under the headings of "optimal impact design" and "temporal pseudoreplication" (Hurlbert 1984:203–205). Analogous problems can also occur in the design of ecological experiments unrelated to pollution. Here we use Hurlbert's paper as a starting point to present an impact assessment procedure that can solve the statistical problems in many situations. Since Hurlbert contends that the problems are insurmountable, we begin by disputing several of his claims.

Hurlbert discusses Green's (1979:24) example in which a pollutant is discharged into a river and the biological species of interest is sampled upstream and downstream of the discharge point, both before and after the discharge flow begins. Hurlbert makes the following assertions concerning a "situation where only a single control area and a single impact area are available."

1) ANOVA (and by implication other procedures of inferential statistics) cannot be validly used to test whether the discharge has affected the downstream abundance. (a) It "can only demonstrate significant differences between locations, not significant effects of the discharge." (b) "Since the treatments cannot be . . . assigned randomly to experimental plots . . . the experiment is not controlled except in a subjective and approximate way." (c) A significant "'areas-by-times interaction' can be interpreted as an impact effect *only* if we assume that the differences between upstream and downstream locations will remain constant over time" if the discharge has no effect. "This is unreasonable. The magnitude of the true difference . . . changes constantly over time." In any case "we would have to make arbitrary decisions about how to measure difference." E.g., if X_I and X_C are the densities at Impact and Control areas, should the "difference" be $X_I - X_C$ or X_I/X_C or something else?

Hurlbert appears here to be dismissing the design described as "optimal" by Green (1979:161 and columns 3–5 of Fig. 3.2). In this design, observations are made at one or more sites in the Impact and Control areas on a *single* occasion Before the disturbance and on another single occasion After. A significant areas-by-times interaction is taken to imply a discharge effect. Such an interaction implies that the *difference* between the Impact and Control (or downstream and upstream) abundances changed after the discharge began.

[1] Manuscript received 21 June 1984; revised 15 October 1985; accepted 18 October 1985.

2) Hurlbert claims that the "optimal" design described by Green cannot be salvaged by taking repeated observations over time, treating the times as replicates, and testing to see if the difference between the Impact and Control means (over time) undergoes a change at the time of the introduction of the discharge. "[S]uccessive samples . . . are so obviously going to be correlated with each other" that analyzing the successive dates "as if they were independent replicates of a treatment . . . is invalid." (Hurlbert's Fig. 5c legend adds that this analysis assumes, implicitly, that data from the same site at different times have come from independent replicates.)

Although Hurlbert does not say so, objections (a)–(c) in (1) seem intended to apply also when there is repeated sampling over time. (The "areas-by-times" interaction would now be an "areas-by-periods" interaction, referring to the periods Before and After discharge.

3) For the design described by Green, Hurlbert says "the best one can do . . . is to develop graphs and tables that clearly show both the approximate mean values *and* the variability of the data on which they are based."

Hurlbert's claims under point (1) relate to Green's "optimal" design, and we first stress that he is correct in rejecting that design, which does not solve the assessment problem.

However, we believe that there is a design that will permit a valid impact assessment in many cases. This design involves using the sampling times as replicates, as in (2) above, but with the important qualification that Impact and Control sites are sampled simultaneously and each sampling time is represented in the analysis by only one number, the difference between the Impact and Control samples for that time. One of our main aims is to explore the extent to which this design overcomes Hurlbert's objections. To do this, we need first to clarify some aspects of claims (a), (b), and (c) in point (1).

Claim (a) is wrong if it means the test can demonstrate only that one site is different from the other. It has force if it means that it is hard to distinguish a discharge effect from some other change occurring at the same time. This force is much weakened by the replication-in-time design, as we discuss in A Proposed Solution (Power Plant Example): Interpreting the Test Results.

Both (b) and (c) concern the statistical "population" and parameters to which our inferences are to refer. We believe (b) is misleading in implying that impact assessment is like analyzing a badly designed experiment in which treatment allocation has not been randomized: assessments and experiments ask different questions, as we discuss in The "Statistical" Population and Parameters in Question. Claim (c) identifies the crucial error in Green's design, but describes it less precisely than we need here. The error arises because Green misidentifies the population and parameters of concern. These need to be carefully described before we can judge whether (c) applies to the replication-in-time design. The "Statistical" Population and Parameters in Question is devoted to this description.

With this clarification, we then argue that the assessment problem can indeed be solved by taking replicates over time. We illustrate this in A Proposed Solution by describing an analysis of the impact of a power plant on the open ocean. We use this example because we are familiar with it and with the biology involved. This is essential, since an important point we want to make is that *all* statistical procedures require assumptions, and these assumptions must be justified by reference both to the data (by plots and formal tests) and to a priori knowledge of the physical and biological system generating the observations.

Finally, in Assumptions, Graphs, and Tables, we criticize Hurlbert's claim (3) on the grounds that graphs and tables that are used to justify conclusions (rather than to suggest patterns or appraise assumptions) require at least as many assumptions as inferential statistical procedures do. Since their assumptions are frequently implicit, these graphs and tables usually provide a less reliable basis for conclusions, rather than a more reliable one. In particular, graphs and tables are more likely than inferential procedures to lead to false conclusions from a design like that described by Green, because they make misidentification of the appropriate parameters even easier than analytical methods do.

We focus our discussion of environmental impact assessment on one question: "Has the impact altered the local abundance of species *X*?" We believe this question can be satisfactorily answered in some cases, and A Proposed Solution outlines how. However our procedure answers the question only if there exists a suitable control area, as defined in that section.

The "Statistical" Population and Parameters in Question

First, contrary to the implication of (b) in what we have labeled Hurlbert's point (1), our concern in most environmental impact problems is with a *particular* impact in a *particular* place resulting from a *particular* facility. It is not the general problem of determining the effect of impacts of this kind in places of this kind. This is an important difference between impact assessment, where the effect of an intervention in a particular instance is at issue, and most basic scientific studies, where we are interested in the average or "usual" effect of an intervention over a large population of possible instances. The general question would require the selection of a set of sites representative of the kind of places we want to study, and random choices to decide which of these sites will be subjected to the power plant or discharge and which will be controls. But the particular impact question does not require such randomized choices.

The next point is as important, but less obvious. To

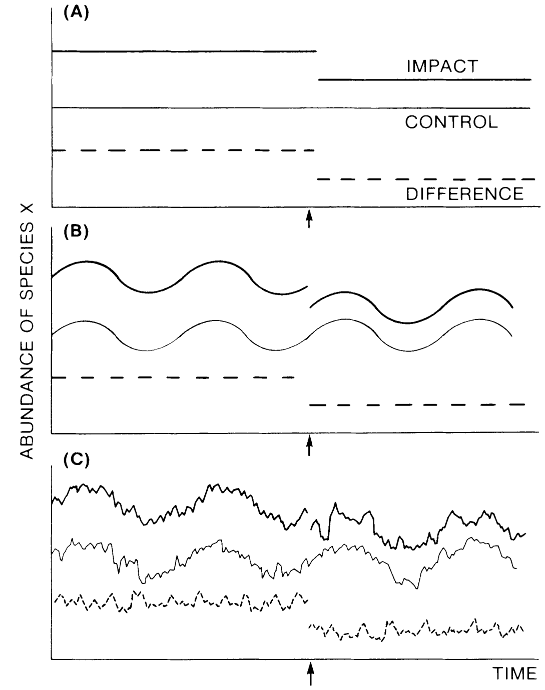

FIG. 1. The abundances of "Species *X*" at the Impact and Control stations, and the difference of the abundances, as functions of time, in three versions of impact assessment. (A) In the most naive view, each station's abundance is constant except for a drop in the Impact station's abundance when the power plant starts up. (B) In a more plausible but still naive view, the abundances fluctuate (e.g., seasonally), but the difference still remains constant except at start-up of the power plant. (C) In a more realistic view, the abundances fluctuate partly in synchrony and partly separately; the former fluctuations disappear in the differences but the latter remain, and the power plant effect must be distinguished from them.

illustrate it, we imagine we are observing the abundances in two locations (Impact and Control), and that these abundances are observed over time without any sampling error. A naive view of impact detection is shown in Fig. 1A, where the abundance is constant at each location (though possibly different for different locations) except for a drop in the Impact abundance at the time of "start-up" (the beginning of operation of the power plant or the pollutant discharge). This naive view is made slightly more realistic in Fig. 1B, where seasonal or other regular movements are introduced, but it remains naive, as the difference in abundance (Impact minus Control) remains absolutely constant except for a drop at start-up. (The change in the Impact-minus-Control value between Before and After start-up reflects the "areas-by-periods" interaction.)

The view in Fig. 1A and B is naive because, in nature, abundances and their differences do not remain constant but vary in response to what can usefully be thought of as random influences: random births and deaths, encounters between individuals, behavioral choices (e.g., local movement), weather changes (storms), water changes (eddies, upwellings), and so on. Some of these influences (e.g., large storms) might have essentially the same effect at both locations, thus not affecting the difference, but others (e.g., individual movements) affect the two locations differently, and thus affect the difference. As a result, both the abundances and (perhaps to a lesser extent) the difference vary, as in Fig. 1C.

It is presumably such varying values that Hurlbert has in mind when he says "The magnitude of the true difference . . . changes constantly over time," thereby making it impossible to evaluate any times-by-location interaction. The important point, however, is that this "true difference" is *not* the difference about which our inferences are to be made. The difference of relevance in impact studies is that between the *mean* heights of the dashed curves in Fig. 1C in the Before and After periods. The fluctuations in the dashed lines within each period do indeed show that the difference between locations is constantly changing, but one can still distinguish an additional difference between locations, not present in the Before period, that is present in the After period. The statistical problem is to characterize the differences between locations within both periods, so that the added difference due to impact can be distinguished.

To explore the characterization of these differences between locations, we present a more detailed picture in Fig. 2. Here, the smooth curve represents the true mean abundance of the population. The aim of assessment is to determine if this mean abundance at the Impact location has been altered by the power plant or discharge. The jagged line represents the path traced out by the actual population. This actual population will differ from the mean as a result of a variety of chance factors, as previously discussed. Finally, the dots represent estimates, based on replicate samples, made of the actual population by the investigator.

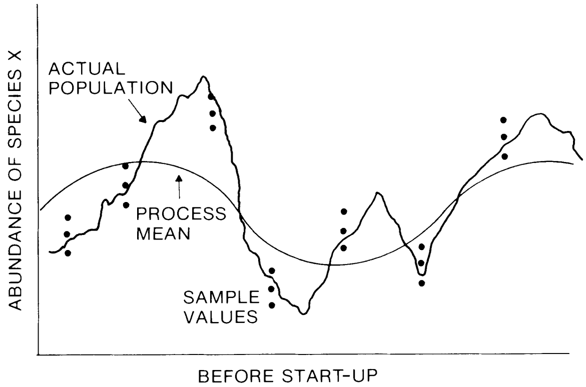

FIG. 2. The three functions to be distinguished: the mean of the population-producing process (considered as a set of random and deterministic elements as inputs, with the population abundance as output); the actual population, which constitutes a single realization of this process; and the sample estimates ●, which (if unbiased) are centered on the actual population.

The concept of a true mean population density perhaps needs clarification. We provide this before discussing the implication of Figs. 1 and 2 for Green's proposed optimal design.

The sequence of abundances in Fig. 1C is generated by the interaction of a large set of factors, some random (like those mentioned) and some systematic (e.g., seasons, local topography and current patterns, and the presence or absence of the power plant). The interaction of these factors constitutes an abundance-producing process, and the sequence of abundances that arises is part of a single outcome (or "realization") of this process. It is one such outcome that we actually sample over any period of time at a given place. In principle, if we could go back in time, we could fix the systematic factors at the same values and re-run the process: because the random factors would take different values, we would expect to get a different sequence of abundances, a different realization. Thus the sequence of abundances that actually occurred was not the only one possible: there is a vast collection of sequences that could have occurred. Given the systematic factors, including presence or absence of the power plant, it is the random factors that determine which of these sequences will be the one that actually arises. Thus the random factors impose a probability distribution on the collection of possible sequences. The parameters of this probability distribution are determined by the systematic factors.

Our task is to use the observed sequence to estimate aspects of the probability distribution; in particular, we aim to infer the effect of the power plant on the mean.

This mean, or mean function, represented by the smooth line in Fig. 2, can be thought of as the average of the jagged lines obtained by a large number of reruns of the process.

A change in the Impact area's smooth line between Before and After might indicate a power plant effect. But the line may be changing anyway, e.g., with seasons and perhaps also with regional long-term trends. For this reason (and others related to independence and discussed later) we search instead for a change in the *difference* between the Impact area's smooth line and that of a nearby Control. This is the situation shown in Fig. 1B, but now with the lines interpreted as the two mean functions, not as the paths of the actual populations. The important assumption that the difference between the smooth lines is constant within each period (Before and After) is discussed in the next section (A Proposed Solution).

These concepts are central to the error in Green's (1979) proposed design. To make inferences about the smooth lines we must estimate them and estimate the variance of our estimates. A natural estimate of the smooth line value at the time of sampling is the average of the sample values (the dots in Fig. 2). This is reasonable because the dots approximate the jagged line, which in turn approximates the smooth line. Green's crucial error is to use the variance among the dots to estimate the variance of this average. But the variance of the dots estimates variance about the jagged line; it gives no information on the variance of the jagged line about the smooth line. Consequently, observations taken at only one Before and one After time are useless for assessing impact, unless supported by other information: neither Green's inferential methods nor Hurlbert's recommended graphs and tables take account of the full variability of the data.

Only by sampling at many different times, both Before and After start-up, can we hope to estimate the variability due to all sources, both sampling error and random population fluctuations. However, problems remain in this case, too. First, even knowledge of the entire jagged line is insufficient for inference about the smooth line unless some assumptions are made: in essence, we need the long-run averages of the deviations of the jagged line from the smooth line, and of the squares of these deviations, to be close to zero and to some constant (σ^2), respectively. The usual assumption is some form of ergodic stationarity in which the process "forgets" its remote past—i.e., the correlation between deviations that are far apart in time is close to zero (see, e.g., Parzen 1962:69 and following; Breiman 1968:chapter 6; Priestley 1981:chapter 5). Second, if our conclusions are to apply to times outside the study period, the systematic factors contributing to the difference (Impact − Control) between the mean abundances (average physical conditions, etc.) must not change much over time.

Both of these considerations, especially the first, play a role in the next section. The main conclusion of this section is that the correct assessment viewpoint is that of Fig. 1B, with the important qualification that the functions of interest are the means and their difference, not the actual populations and their difference (Fig. 1C). The latter, which must themselves be estimated from samples, are only a guide to the former.

A Proposed Solution (Power Plant Example)

In this example, an Impact and a Control area are determined, and samples are taken simultaneously at both places at times t_{11}, t_{12}, . . . Before the start-up of a power plant, and at times t_{21}, t_{22}, . . . After start-up. We thus call this a "BACI" design, though this acronym omits the important fact that the two sites are sampled simultaneously. The object is to see whether the difference between Impact and Control abundances has changed as a result of the start-up. Since "replicates" taken at the same time do not help us estimate the variance that matters, we simply average them. We therefore assume only one observation, say X_{ijk}, for each time, t_{ij}, in period i (Before or After), at place k (Impact or Control). The plan is to compare the Before and After periods by a t test or a U test for a difference between the mean of the Before differences (estimated by $X_{1.1} - X_{1.2}$, which is the average, over j, of $X_{1j1} - X_{1j2}$) and the mean of the After differences.

These tests assume that, without the power plant, the difference between the means at the Impact and Control stations—the "smooth lines" (see Figs. 1B and 2)—would be constant. In other words, without the power plant, the effects of time and location on the mean local abundance of species X would be "additive." The tests also assume that the observed differences, calculated at different times, are independent. Finally, this impact detection procedure requires us to interpret a statistically significant result of the test as being due either to chance (whose probability is given by the significance level) or to the power plant. We discuss these issues in turn.

Constancy of differences: additivity of time and location effects

In Fig. 3A and B are illustrated ways in which the difference between the means (the smooth lines in Figs. 1B and 2) could be nonconstant.

In Fig. 3A, the problem is one of scale. Various physical and other factors (e.g., seasons) affect the two areas in the "same" way, but the effect is multiplicative, not additive. Consequently the arithmetic difference in abundance tends to be greater at times when both areas have abundant populations than when both have sparse ones.

This failure of the additivity assumption could have either of two consequences. If the fluctuations are periodic, as shown, and if the Before and After times are matched reasonably well (as *b*1, *b*2, . . . with *a*1, *a*2, . . .) the effect will be to produce a weaker, more conservative test: the observed differences will fluctuate more than they would have from random factors alone. If the sampling times are poorly matched (e.g., *b*1, *b*3, . . . to *a*2, *a*4, . . .), as could happen if the periodicity is not known or if its phase changes near the start-up time, then times with large smooth-line differences may be unequally represented in the two samples (Before and After), leading to a possible false finding.

This problem can be eliminated, or at least much reduced, if our measurements are made in the "right" scale: in Fig. 3A, the log of abundance rather than abundance. The right scale is unknown, as Hurlbert (quoting Eberhardt 1976) says in his point 1C. There are two related ways to approximate it. One is to derive it from a model of the actual process. This could be a detailed model, or it could be something as simple as "most environmental changes have approximately multiplicative effects, so let's try the log transformation." The success of such a transformation should then be tested, e.g., by the Tukey (1949) "one degree of freedom" test for nonadditivity. The other is to use the data in a formal procedure to estimate the right transformation. Box and Cox (1964) consider transformations of the form $Y = (X + c)^{\lambda}$, assume that Y is normally distributed, homoscedastic, and additive for the right c and λ, and propose estimating c and λ by maximum likelihood. However, the transformation that makes Y additive (our main goal) may not make it normal; the Box-Cox procedure may not work well in this case, so we suggest the procedure of Andrews (1971), in which λ and c are chosen to minimize the significance of a test for additivity of the Y's. (There is a vast literature on transformations. Some of it is reviewed by Hinkley and Runger 1984. A more informal discussion appears in Mosteller and Tukey 1977:chapter 5.)

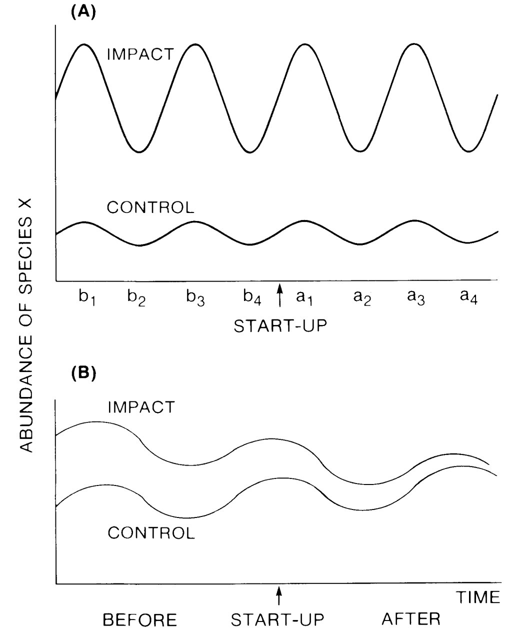

FIG. 3. How natural nonconstancy of the difference of the process means influences the test for a power plant effect. (A) Nonadditive (e.g., multiplicative) effects cause the difference to be greater during periods of high abundance. If Before and After samples are reasonably matched for position on cycles (as *b*1, *b*2, . . . are matched with *a*1, *a*2, . . .) the effect is an artificially inflated estimate of variance and a weak test. If samples are poorly matched (as *b*1, *b*3, . . . are with *a*2, *a*4, . . .) the effect is to bias the test. (B) Impact and Control sites change relative to each other in the "Before" period, without the power plant. If this trend continues, After differences should be smaller than Before ones, even if the plant has no effect.

We now present an example using real data: 9 yr of observations on the arthropod *Acuminodeutopus heteruropus* near the San Onofre Nuclear Generating Station, in California, Before start-up of the station (Fig. 4). The Tukey test for additivity (Fig. 4C), which in this case amounts to a test for zero slope in the regression of the differences against the averages, reveals a significant ($P < .0001$) departure from additivity; it remains significant ($P \simeq .0003$) when the extreme point on the upper right in Fig. 4C is removed. Rather than use formal procedures to estimate λ and c, we have chosen the commonly used transformation $\log(X + 1)$, the natural transformation for multiplicative effects but

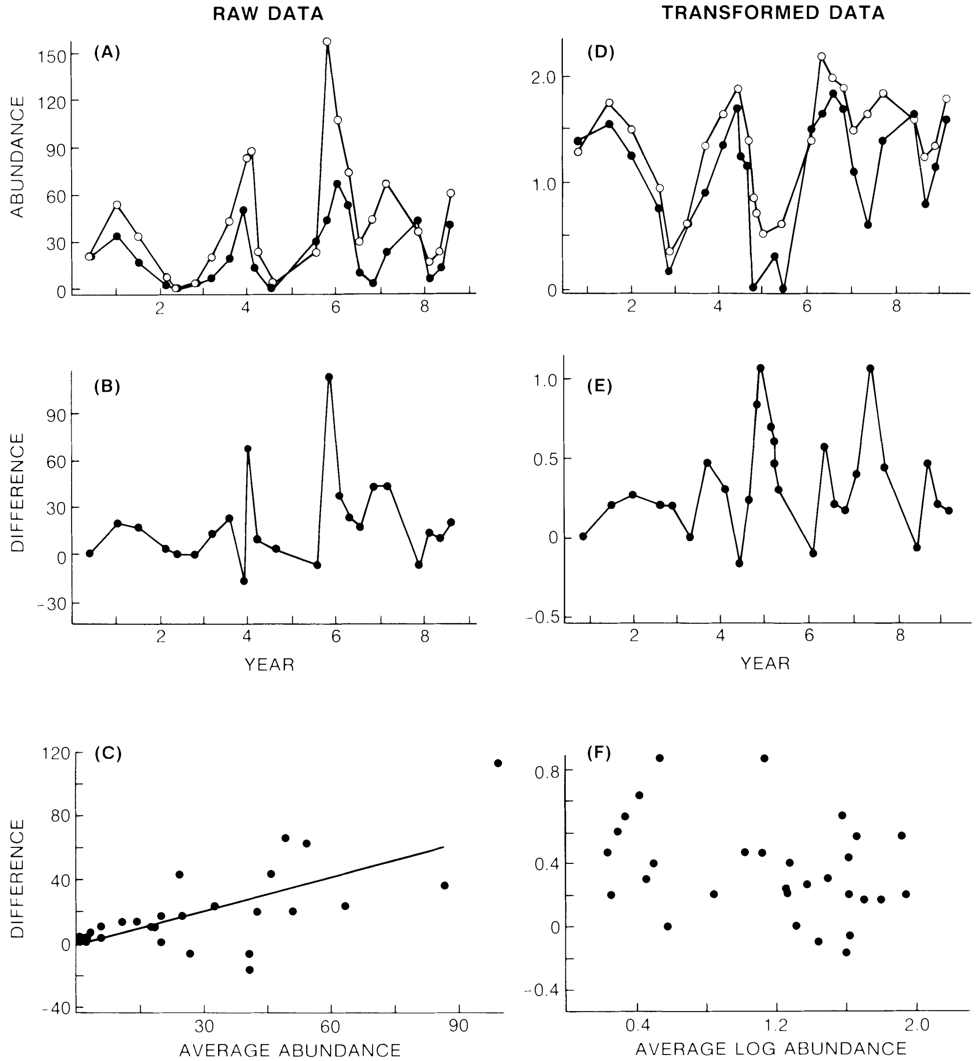

FIG. 4. Example of real nonadditive data made additive by a transformation. (A) Raw abundances at Control (○) and Impact (●). (B) Difference in raw abundances between Control and Impact. (C) Tukey test for additivity, which is equivalent, in this case, to a test for zero slope in the regression of differences against averages. Test is significant ($P < .0001$). (D) Log (abundance + 1) plotted against time. (E) Difference in log(abundance + 1) between Control and Impact. (F) Tukey test for nonadditivity on log(abundance + 1) is nonsignificant ($P \simeq .16$). The data are total counts from samples of the arthropod *Acuminodeutopus heteruropus* over a 9-yr period near the site of the San Onofre Nuclear Generating Station, California. (For the sake of clarity, a dense cluster of points that should have appeared near zero at about year 5 of the time plots has been reduced to a single point. All points appear in the plots of difference vs. average.)

perturbed to avoid problems with zeros. With the transformed data (Fig. 4D and E), the departure from additivity is not significant ($P \simeq .16$, Tukey test; Fig. 4F).

The failure of additivity in Fig. 3B is of a different kind. Here the Impact and Control areas are changing relative to each other, even without the power plant. Our test for a power plant effect assumes there is no such change. More elaborate tests can be devised to allow for the change, but these tests must make assumptions about how the change would have proceeded if there had been no power plant. In most cases, it seems safer to decide that, with this design, it will not be possible to detect a power plant effect on the species in question. Thus no attempt should be made to choose a transformation to correct this change. Rather, such correction should be avoided, the trend should be tested for, and the species not used in this analysis if the trend is present. In real-time plots of the differences of raw abundances and of log-transformed abundances for the polychaete *Aricidea wassi* during 30 mo before start-up near San Onofre, both sets of differences appear to be decreasing over time (Fig. 5A and B).

A third type of violation of the additivity assumption, not described in Fig. 3, can be dealt with. This is the problem of covariates. In the analysis we are describing, the Control area is not expected to be identical with the Impact area, but is intended to satisfy two much weaker requirements. One is that there exists a transformation such that the corresponding smooth lines at Impact and Control differ by a constant amount (not necessarily zero) in the Before period. The other is that not only regular effects (e.g., the seasonal fluctuations in the smooth lines) but also long-term random effects should be approximately equal at the two areas. By "long-term random effects" we mean effects that perturb the jagged line in Fig. 2 away from the smooth line by large amounts for long periods. Major storms might do this. Unequal long-term effects will lead to violations of independence assumptions, which we discuss in Independence of Temporal "Replicates." However, we note here that some violations of these assumptions can be dealt with by including the perturbing influences as covariates. For example, if the Control area is nearer the mouth of a creek it may be more affected by runoff—potentially a long-term, random influence whose effect could be felt through several sampling times. In this case, one might include the volume of flow in the creek (possibly lagged) as a covariate, and use analysis of covariance to test for the power plant effect.

We have assumed here that any transformation should be based on the Before data, since these are known to be unaffected by the power plant. A power plant effect could cause the mean difference in the After period to change over time. This would not invalidate the test, since the null hypothesis is "no effect," though the power of the test will be less than it would have been had the After mean difference remained constant at its average value over the study period.

Independence of temporal "replicates"

Our observation or sample average at time t_{ij} in place k is equal to the corresponding smooth line value plus the deviation from it:

$$X_{ijk} = \mu_{ijk} + E_{ijk}. \tag{1}$$

The aim of the transformation procedure was to make the μ's additive, at least approximately, so that

$$\mu_{ij1} - \mu_{ij2} = \mu_i, \tag{2}$$

where μ_i is constant within period i (Before or After). Thus the transformed differences D satisfy

$$D_{ij} = X_{ij1} - X_{ij2} = \mu_i + E_{ij1} - E_{ij2}, \tag{3}$$

where the E's are errors, having means of zero (since the μ's are defined to be the means of the X's). To carry out a t test, U test, or any other standard two-sample test comparing the D_{1j}'s with the D_{2j}'s, we require that the errors

$$\delta_{ij} = E_{ij1} - E_{ij2} \tag{4}$$

be independent.

Note that it is the *differences* that are to be uncorrelated, not the successive samples themselves, as the legend to Hurlbert's (1984) Fig. 5c indicates. If the samples are uncorrelated (i.e., E's for different times are uncorrelated) then the differences will also be uncorrelated, but the latter condition can also hold without the former.

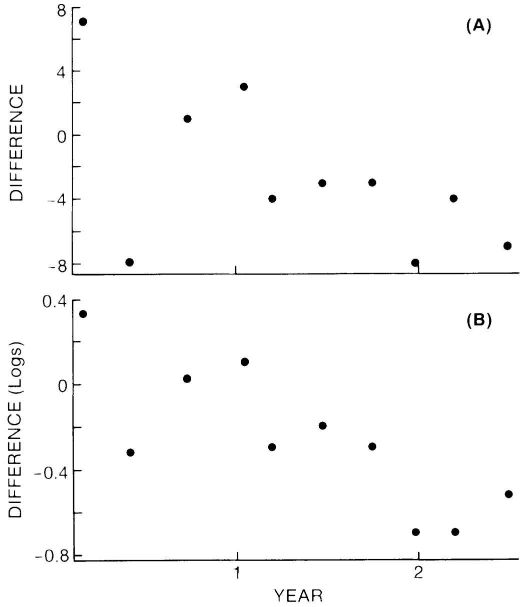

FIG. 5. Example of real data showing a trend over time in the differences of both (A) raw and (B) log-transformed abundances. The species is the polychaete *Aricidea wassi* over 2½ yr, near San Onofre, California.

For example, each error E_{ijk} consists of two parts: sampling error (the difference between the average of the dots and the jagged line in Fig. 2), and what we might call "process error" (the difference between the jagged line and the smooth line). Thus

$$E_{ijk} = u_{ijk} + v_{ijk}, \tag{5}$$

where u and v are sampling and process errors, respectively. The independence of sampling errors is a standard assumption in virtually every study, randomized or not. (This does not mean the assumption is always true, merely that it needs no special attention here.) If it should happen that the process errors are not independent over time, but are identical at the two locations, then the E's will not be independent but the process errors will vanish in the differences; i.e., if $v_{ij1} = v_{ij2}$ then $\delta_{ij} = u_{ij1} - u_{ij2}$, so the differences will be independent. Note also that interdependence of the sampling errors among locations does not affect the independence of the differences over time.

It is not realistic to assume identical process errors, but considerably weaker assumptions that achieve essentially the same result are plausible. We have suggested earlier some of the random events that might cause the (jagged line) actual population to deviate from the (smooth line) mean. Such events may have effects that are small or large, local or area-wide (including both the Impact and the Control locations), long-lasting or short-term. Random births and deaths, or encounters between individuals, are likely to have small, local, short-term effects. Movement by schools of fish, or the activities of a single fishing boat, might have large local effects, but these would often be short-term as immigrants arrive from surrounding neighborhoods. Large storms may have large area-wide effects that may be long-lasting if no sources of immigrants are nearby and population processes are slow to respond to the lower density levels. Broadscale weather changes, like El Nino events on the West Coast, are likely to have large, area-wide, long-lasting effects.

Of these kinds of effects, only those that are large, local, and long-lasting will lead to serious violations of the independence assumption. Small effects will be overwhelmed by the noise from other errors, including sampling error. Short-term effects, by definition, will not affect samples spaced far enough apart. Area-wide effects may not lead to identical errors but will lead to highly correlated errors, whose difference is small with high probability.

A simple model may clarify these assertions. A common time series model is the Markovian one in which the deviation at time t_{ij} is affected by previous deviations only through the deviation at time $t_{i,j-1}$. Such a model is

$$v_{ijk} = e^{-\alpha T} v_{i,j-I,\mathrm{k}} + A(1 - e^{-2\alpha T})^{1/2} w_{ijk}, \qquad (6)$$

where the w_{ijk}'s have mean zero and variance 1 and are independent over time (i.e., w's for different t_{ij}'s are independent), A and α are constant, and $T = t_{ij} - t_{i,j-1}$, the time between samples. Thus the deviation at time $t_{i,j-1}$ affects that at time t_{ij}, but as the time gap increases its effect decreases, while the effect of events occurring between $t_{i,j-1}$ and t_{ij} increases. If the sampling errors are independent of the process errors and their difference, $u_{ij1} - u_{ij2}$, has variance σ^2, then Eqs. 4, 5, and 6 can be used to show that the differences, δ_{ij}, have mean zero, variance $\sigma^2 + 2A^2(1 - r)$, and covariances given by $\mathrm{cov}(\delta_{ij}, \delta_{i,j+n}) = 2A^2e^{-\alpha T}(1 - r)$, where r is the correlation between w_{ij1} and w_{ij2}, the error at the two locations for the same time, and T is now $t_{i,j+n} - t_{ij}$. The correlation (=covariance/variance) between differences at different times will be small if A is small (small perturbations), αT is large (short-term effects; i.e., rapid decay, with sufficient spacing between samples), or r is large (high correlation between locations; i.e., area-wide effects).

Of course, this model and argument do not justify a claim of strict independence between observations, only one of low correlation. The effect of positively correlated observations is to inflate the value of the t statistic because the denominator of t (e.g. $[s^2/n]^{1/2}$) underestimates the variance of the numerator (e.g., $\bar{X}$). In a Markovian model like the one above, if the spacings are equal (i.e., $t_{ij} - t_{i,j-1} = T$ is the same for all i and j), then the t statistic for the two-sample test on the differences is inflated by a factor close to the square root of $(1 + \rho)/(1 - \rho)$, where ρ is the correlation between consecutive observations, $e^{-\alpha T}(1 - r)(1 - r + \sigma^2/2A^2)^{-1}$. The effect on more robust tests, like the U test, is similar but seems to be somewhat smaller (Serfling 1968, Gastwirth and Rubin 1971, 1975). Thus, decisions based on decisive t values (greater than 3, say) will not be affected by even moderate ρ values (less than 0.2, say). Borderline t values will be suspect, but these will need scrutiny anyway because of possible nonnormality, influential outliers, etc. Low correlation is an additional nuisance but not a qualitatively different one except that its effect does not decline as sample size increases.

The independence (or low correlation) assumption is plausible if large, local, long-lasting random effects are unlikely. Whether or not they are depends on the situation being studied. We suspect the assumption would hold for many species in the pollutant discharge problem, but do not feel we know this well enough to be confident. Even for the open ocean power plant, one can think of difficulties. We have mentioned runoff as one. Another is the possibility that the long-term abundances of some sedentary species may depend heavily on events occurring in short time periods, e.g., recruitment or winter storms. The local effects of these random events may last for a long time. Nevertheless, these are rather special cases, and they are not hard to identify. For most species, the sources of local variation will be events at the individual level, very small-scale weather or water disturbances, or one-time shocks, all of which are likely to be short-term just because the ocean is big and contains powerful diffusing mechanisms.

But plausibility is not enough. The independence assumption should also be checked against the data. There are, in fact, powerful tests for doing so. Many of these are reviewed by King (1986). The most popular (and in many ways the best) is the Durbin-Watson test. This is a generalization, to a regression situation, of the von Neumann ratio test, which is suitable for our case. To carry out the test on the differences, $D_{ij} = X_{ij1} - X_{ij2}$, we compute $Q = \Sigma(D_{i,j+1} - D_{ij})^2/\Sigma(D_{ij} - D_{i\cdot})^2$. Small values of Q indicate serial correlations. Significance levels for $nQ/(n - 1)$ are given by Hart (1942). The differences of the transformed abundances in Fig. 4E may be suspected to have serial correlation, although it should be noted that some observations that are close in time are quite different in value. In fact, for the data of Fig. 4E, Q is significant ($P \simeq .03$).

If additivity and independence are plausible assumptions, and also satisfy the tests, then we suggest using the t, U, and perhaps other standard two-sample tests (normal scores, trimmed t) to decide whether the power plant (for example) has an effect. Contradictory results and borderline values should lead to further scrutiny of the data. Possible serial correlation could then be dealt with more directly by dropping some observations, averaging some neighboring values, transforming to approximate independence via $Z_{ij} = D_{ij} - (1 - Q/2)D_{i,j-1}$ (since $1 - Q/2$ estimates ρ if the sampling times are evenly spaced), or using a more elaborate "intervention analysis" (Box and Tiao 1975).

Interpreting the test results

If our transformations have rendered the difference between the smooth lines constant in the Before period, and if our observations are independent, a statistically significant test result is due either to chance (with the stated probability) or to a change (Before vs. After) in the mean difference between locations.

It is the second possibility, a change in the difference, that we want to attribute to the power plant. What we have labeled Hurlbert's assertion (1a) seems to be that this cannot be done with certainty: the After differences might misbehave for some other reason.

This possibility is an important reason in basic studies for preferring experiments, involving randomized allocation of treatments, to natural observations. But Impact assessment does not fit the experiments/observations dichotomy. The main objection to nonrandomized experiments and to observational studies is that the units receiving the treatment are not a representative sample of the "unit" population. Bias may enter either through the investigator's treatment allocation or through spurious effects due to causes other than the treatment (e.g., stress may cause both smoking and heart attacks). This objection does not apply with full force to impact studies. The Impact location abundances, Before and After the actual impact, are the *only* populations of interest: neither the time nor the place of impact can be unrepresentative in this sense, even if the power company has cleverly chosen the only site in the world that is invulnerable to power plant effects. The Control location also does not represent a population of locations, though it does play the usual role of helping distinguish between treatment effects and time effects. Lastly, the times at which samples are taken could give us an unrepresentative view of the continuous sequence of abundances: some randomization of sampling times may thus be advisable. However, we note again that the parameter of inteiest is the mean of the hypothetical population of all possible re-runs of the abundance process, not the mean, over the time of the study, of the particular run that happened to occur, so random sampling times do not give us randomly chosen population members. More often, sampling times will be chosen to minimize serial correlation (e.g., by equal spacing) and to maximize useful information (e.g., sampling in seasons when the target species should be reasonably abundant).

There still remains the possibility of a large, long-lasting but unpredictable effect occurring at about the same time as the start-up, and affecting one location much more than the other. It would be very difficult to detect such an effect from the data, because it would not have to arise simultaneously with start-up, merely near enough to affect most of the After sample and little of the Before. It would also be difficult to estimate the probability of such an effect on an objective basis. Thus, this possiblility must always be kept in mind, and efforts should be made to study the mechanisms both of a power plant effect and of any plausible alternative explanation for the results. However, randomized studies share these imperatives in many cases because their results may be due not to the treatment but to the way it is administered, e.g., placebo effects, cage effects in ecological studies, contaminants in drug studies. (For more subtle possibilities, see Bekan 1983 and Connell 1983, and the pressure vs. volume example in Pratt and Schlaifer 1984.) The difference seems to be of degree, not kind. In both cases, possible alternative mechanisms must be proposed (often by the study's critics) and their likelihoods assessed (often rather subjectively). Although such mechanisms might occur in our power plant example, we feel they are very unlikely to occur undetected. This is why we describe Hurlbert's point (1a) as having much reduced force in our setup.

Assumptions, Graphs, and Tables

Perhaps the most unfortunate of Hurlbert's assertions is that labeled (3) in our Introduction, advocating graphs and tables when there seems to be no analytic procedure whose assumptions can be guaranteed to hold.

The goals of statistical techniques can be classified into two fairly distinct groups. "Exploratory data analyses" (Tukey 1977) are attempts to reveal patterns and regularities in the data—to suggest hypotheses and potentially fruitful further studies. They require no formal attempt to distinguish chance patterns from systematic ones, or to quantify the confidence the conclusions warrant. A variety of imaginative, nonquantitative methods has developed (see Chernoff's 1973 proposal to cluster multivariate observations by converting them into computer-drawn faces). We have no objection to graphs and plots, unsupported by formal inference, in these cases where the "conclusions" are really suggestions.

But impact studies almost always belong to the second group, "confirmatory" rather than exploratory. There are clear (if not always clearly stated) hypotheses at issue—most frequently, perhaps, that "the intervention has reduced the abundance of species X in the Impact area." These hypotheses need to be resolved in as quantitative a way as possible: that is, the investigator is asked to give one or more numbers describing the extent to which the data support the hypothesis. There are plenty of arguments about what numbers should be given, but none about the need for a quantitative conclusion. Indeed, the final decision—e.g., to close down a power plant, alter its design, mitigate the effects, or do nothing—is inescapably quantitative in that it compares, implicitly or explicitly, the benefits (or costs) of the available options and chooses what is estimated to be the best.

In these cases, the unsupported use of graphs and tables is usually inappropriate. The investigator is required to use the data to decide whether the evidence for the hypothesis is strong enough for some action to be taken (or belief to be adopted). *Such a quantitative assessment necessarily requires assumptions about the relationship between the data and the hypothesis at issue.*

Not a single one of the objections Hurlbert raises against inferential statistics in this context is resolved by the presentation of graphs and tables. Indeed, virtually all appear with additional force. These objections all relate to the assumptions required for the inferential procedures to be valid. But the graphs and tables need assumptions too, if conclusions are to be drawn. Just because they are not stated does not mean that they are not there or are not needed. It is far more likely to mean that they have not been explicitly listed, examined for plausibility in the physical situation, or tested against the data.

In fact, it is likely to be generally the case that drawing conclusions from graphs and tables requires all the assumptions of inferential statistics (e.g., independence, additivity, no additional sources of error, etc.) together with some equally dubious new ones, e.g., that the viewer's interpretation of the graph is unaffected by the graph's color, scale, draftsmanship, position on the page, and other irrelevancies. See Wainer (1984) and the beautiful book by Tufte (1983).

Of course we have no objection to the use of graphs and tables in support, rather than instead, of inferential procedures, where they are essential for checking assumptions and for giving a clearer impression of the physical, rather than statistical, significance of our observations.

Conclusions

Restricting ourselves to only one section of Hurlbert's (1984) paper does not indicate agreement or disagreement with the rest of it. In fact, we agree with the main point, that statistical analyses in ecology too frequently underestimate error by ignoring some of its sources. In this section we offer some thoughts on what is to be done.

Part of the responsibility for the misuse of statistical procedures no doubt rests with statisticians, who sometimes do not appreciate the applied aspects of their work sufficiently to explain it clearly to users or to think about specific applications. We would especially like to see more attention paid to the meaning of the assumptions (or their violation) in practical situations.

Part of the responsibility may also rest with ecologists who do not take statistics seriously as a discipline: who either accept any and every analysis regardless of whether the assumptions make sense or fit the data; or reject virtually every analysis by insisting on strict compliance with all assumptions, without attempting to distinguish either essential assumptions (e.g., often, independence) from marginal ones (e.g., often, normality) or minor violations (a serial correlation coefficient of 0.05, say, when $t = 6$ or more) from major ones (a correlation coefficient of 0.7, say, when t is borderline).

We feel that Hurlbert's treatment of statistical issues is better than those of many who attempt to lay down the statistical law to biologists and other empirical scientists, because it goes deeper than a few quotes from standard textbooks. Nevertheless, a still more ecumenical spirit is called for.

Many of these issues have been discussed in detail by statisticians, economists, sociologists, philosophers, and others. The foundations and the meaning of data-based inference have been much discussed (e.g., Jeffreys 1961, Birnbaum 1962, Hacking 1965, Savage 1972, Good 1983; see also two collections: Kempthorne 1976, Neyman 1976, Pratt 1976, Roberts 1976; and Benenson 1977, Birnbaum 1977, Giere 1977, Kiefer 1977, Kyburg 1977, Le Cam 1977, Lindley 1977, Neyman 1977, Pratt 1977, Rosenkrantz 1977, Smith 1977, Spielman 1977; a stimulating introduction to some of the problems is in chapter 1 of Berger 1980). Another active topic is experimental design (e.g., Kempthorne 1952, Kiefer 1959, and the discussions of randomization in Rubin 1978, Basu 1980, Smith 1983, and Pratt and Schlaifer 1984). Since economists and sociologists often have at least as much trouble constructing "clean" experiments as ecologists do, some of their thoughts on these problems are helpful: Campbell and Stanley (1966), Riecken et al. (1974), and Cook and Campbell (1979) are examples.

It is not reasonable to expect ecologists (or, perhaps, even statisticians) to become familiar with all the major ideas in these areas. The references cited are only a small sample, and few of them are easy reading. (The three sociology books, the discussions of randomization, and Berger 1980 are perhaps the most accessible of them.) Further, although the areas of agreement are larger than the noise would sometimes lead one to believe, none of these issues can be regarded as finally settled, or seems likely soon to be. Still, there is a real need both for some familiarity with these ideas (if only to prevent valuable but imperfect data being discarded as completely worthless) and for a greater respect for contributions from other fields (if only to avoid repeating the histories of their disputes).

Acknowledgments

Paul Switzer first suggested that the design discussed here would be suitable for a study of the effects of the cooling system of the San Onofre Nuclear Generating Station upon the marine environment. Detailed reviews, two anonymous and two by S. Hurlbert, greatly improved the focus, clarity, and accuracy of the paper. We thank Art Barnett and the Marine Review Committee for the data plotted in Figs. 4 and 5. Sue Bence provided technical assistance. The work was supported in part by NSF grant BSR 83-15235.

Literature Cited

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

Basu, D. 1980. Randomization analysis of experimental data. Journal of the American Statistical Association **75**:575–595.

Bekan, E. 1983. New ghosts of competition: reply to Connell. Oikos **41**:288–289.

Benenson, F. C. 1977. Randomness and the frequency definition of probability. Synthese **36**:207-233.

Berger, J. O. 1980. Statistical decision theory: foundations, concepts and methods. Springer-Verlag, New York, New York, USA.

Birnbaum, A. 1962. On the foundations of statistical inference. Journal of the American Statistical Association **57**: 269–306.

———. 1977. The Neyman-Pearson theory as decision theory and as inference theory; with a critique of the Lindley-Savage argument for Bayesian theory. Synthese **36**:19–49.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society (London), Series B **26**:211–252.

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.

Breiman, L. 1968. Probability. Addison-Wesley, Menlo Park, California, USA.

Campbell, D. T., and J. C. Stanley. 1966. Experimental and quasi-experimental designs for research. Rand McNally, Chicago, Illinois, USA.

Chernoff, H. 1973. The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association **68**:361–368.

Connell, J. H. 1983. Interpreting the results of field experiments: effects of indirect interactions. Oikos **41**:290–291.

Cook, T. D., and D. T. Campbell. 1979. Quasi-experimentation: design and analysis for field settings. Rand McNally, Chicago, Illinois, USA.

Eberhardt, L. L. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management **4**:27–70.

Gastwirth, J. L., and H. Rubin. 1971. Effect of dependence on the level of some one-sample tests. Journal of the American Statistical Association **66**:816–820.

Gastwirth, J. L., and H. Rubin. 1975. The behavior of robust estimators on dependent data. Annals of Statistics **3**:1070–1100.

Giere, R. N. 1977. Allan Birnbaum's conception of statistical evidence. Synthese **36**:1–13.

Good, I. J. 1983. Good thinking. The foundations of probability and its applications. University of Minnesota Press, Minneapolis, Minnesota, USA.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. Wiley, New York, New York, USA.

Hacking, I. 1965. Logic of statistical inference. Cambridge University Press, Cambridge, England.

Hart, B. I. 1942. Significance levels for the ratio of the mean square successive difference to the variance. Annals o Mathematical Statistics **13**:207–214.

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

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

Jeffreys, H. 1961. Theory of probability. Third edition. Oxford University Press, New York, New York, USA.

Kempthorne, O. 1952. The design and analysis of experiments. Wiley, New York, New York, USA.

———. 1976. Of what use are tests of significance and tests of hypothesis. Communications in Statistics—Theory and Methods **A5**(8):763–777.

Kiefer, J. 1959. Optimum experimental designs. Journal of the Royal Statistical Society (London), Series B **21**:272–319.

———. 1977. The foundations of statistics—are there any? Synthese **36**:132–176.

King, M. L. 1986. Testing for autocorrelation in linear regression models: a survey. *In* M. L. King and D. E. A. Giles, editors. Specification analysis in the linear model. Rutledge and Kegan Paul, London, England, *in press.*

Kyburg, H. E. 1977. Decisions, conclusions, and utilities. Synthese **36**:89–96.

Le Cam, L. 1977. A note on metastatistics or 'an essay toward stating a problem in the doctrine of chances.' Synthese **36**:133–160.

Lindley, D. V. 1977. The distinction between inference and decision. Synthese **36**:51–58.

Mosteller, F., and J. W. Tukey. 1954. Data analysis and regression. Addision-Wesley, Reading, Massachusetts, USA.

Neyman, J. 1976. Tests of statistical hypotheses and their use in studies of natural phenomena. Communications in Statistics—Theory and Methods **A5**(8):737–751.

———. 1977. Frequentist probability and frequentist statistics. Synthese **36**:97–131.

Parzen, E. 1962. Stochastic processes. Holden-Day, San Francisco, California, USA.

Pratt, J. W. 1976. A discussion of the question: for what use are tests of hypotheses and tests of significance. Communications in Statistics—Theory and Methods **A5**(8):779–787.

———. 1977. 'Decisions' as statistical evidence and Birnbaum's 'confidence concept.' Synthese **36**:59–69.

Pratt, J. W., and R. Schlaifer. 1984. On the nature and discovery of structure. Journal of the American Statistical Association **79**:9–33.

Priestley, M. B. 1981. Spectral analysis and time series. Academic Press, New York, New York, USA.

Riecken, H. W., R. F. Boruch, D. T. Campbell, N. Caplan, T. K. Glennan, J. W. Pratt, A. Rees, and W. Williams. 1974. Social experimentation: a method for planning and evaluating social intervention. Academic Press, New York, New York, USA.

Roberts, H. V. 1976. For what use are tests of hypotheses and tests of significance. Communications in Statistics—Theory and Methods **A5**(8):753–761.

Rosenkrantz, P. R. 1977. Support. Synthese **36**:181–193.

Rubin, D. B. 1978. Bayesian inference for causal effects: the role of randomization. Annals of Statistics **6**:34–58.

Savage, L. J. 1972. The foundations of statistics. Dover, New York, New York, USA.

Serfling, R. J. 1968. The Wilcoxon two-sample statistic on strongly mixed processes. Annals of Mathematical Statistics **39**:1202–1209.

Smith, C. A. B. 1977. The analogy between decision and inference. Synthese **36**:71–85.

Smith, T. M. F. 1983. On the validity of inferences from non-random samples. Journal of the Royal Statistical Society (London), Series A **146**:394–403.

Spielman, S. 1977. Physical probability and Bayesian statistics. Synthese **36**:235–269.

Tufte, E. R. 1983. The visual display of quantitative information. Graphics Press, Cheshire, Connecticut, USA.

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

———. 1977. Exploratory data analysis. Addison-Wesley, Reading, Massachusetts, USA.

Wainer, H. 1984. How to display data badly. American Statistics **38**:137–147.