RESAMPLING TESTS FOR META-ANALYSIS OF ECOLOGICAL DATA

Dean C. Adams, Jessica Gurevitch, and Michael S. Rosenberg

*Department of Ecology and Evolution, State University of New York at Stony Brook,
Stony Brook, New York 11794-5245 USA*

Abstract. Meta-analysis is a statistical technique that allows one to combine the results from multiple studies to glean inferences on the overall importance of various phenomena. This method can prove to be more informative than common ''vote counting,'' in which the number of significant results is compared to the number with nonsignificant results to determine whether the phenomenon of interest is globally important. While the use of meta-analysis is widespread in medicine and the social sciences, only recently has it been applied to ecological questions. We compared the results of parametric confidence limits and homogeneity statistics commonly obtained through meta-analysis to those obtained from resampling methods to ascertain the robustness of standard meta-analytic techniques. We found that confidence limits based on bootstrapping methods were wider than standard confidence limits, implying that resampling estimates are more conservative. In addition, we found that significance tests based on homogeneity statistics differed occasionally from results of randomization tests, implying that inferences based solely on chi-square significance tests may lead to erroneous conclusions. We conclude that resampling methods should be incorporated in meta-analysis studies, to ensure proper evaluation of main effects in ecological studies.

Key words: bootstrapping; meta-analysis; randomization tests; resampling statistics vs. standard methods; statistical techniques.

Introduction

There is a compelling need for new methods for combining ecological data from different experimental studies in order to reach general conclusions. While conventional reviews and syntheses of ecological data have relied on subjective, narrative methods, or ''vote-counting'' approaches, ecologists have recently begun to explore the use of meta-analysis to address this need. Meta-analysis is a set of statistical methods that provides a rigorous framework for the quantitative synthesis of the results of independent studies. The use of these techniques has become widespread and even routine in medicine and in the social sciences, but their potential for integrating ecological data is just beginning to be realized (Arnqvist and Wooster 1995). Recent applications to ecological problems include syntheses of the experimental evidence for competition (Gurevitch et al. 1992), the responses of woody plant species to elevated CO_2 (Curtis 1996), and the effectiveness of crop diversification in deterring herbivorous insects (Tonhasca and Byrne 1994).

Most researchers use parametric fixed-effects mod-

els, although random-effects models exist (Raudenbush 1994), and a mixed model has been proposed (Gurevitch and Hedges 1993). These procedures are summarized elsewhere (Hedges and Olkin 1985, Cooper and Hedges 1994). The parametric model used to derive the effect size d used in meta-analysis relies on the assumption that the observations in the experimental and control groups are normally distributed for each study. The large-sample distribution of d tends to normality, and the large-sample approximation to the distribution of the effect size estimator is fairly reliable for sample sizes that exceed 10 in each of the control and experimental groups. The large-sample approximation is likely to be less accurate when sample sizes are small, when there are large differences in sample size between the experimental and control groups, and with very large effect sizes (Hedges and Olkin 1985). Unfortunately, it is not uncommon for ecological data to violate each of these conditions (e.g., Gurevitch et al. 1992). In addition, the test statistic used to assess the homogeneity of the effect sizes among studies, Q, is approximately chi-square distributed when the above assumptions of normality are met. If they are violated, the conventional tests of homogeneity may be flawed (Hedges and Olkin 1985).

One alternative to traditional parametric and non-

TABLE 1. Equations used in the calculation of mean effect sizes and homogeneity components in meta-analysis. Symbols in equations follow Gurevitch and Hedges (1993) and are as follows: $\overline{X}_{ij}{}^C$ = mean of control group, $\overline{X}_{ij}{}^E$ = mean of the experimental group, s_{ij} = pooled standard deviation of the control and experimental groups, J = correction term for bias because of small sample size, w = weighting for each study [see *Methods: Calculations* for parametric and nonparametric weighting schemes].

Statistic	Symbol	Equation
Study effect size	d_{ij}	$\dfrac{\overline{X}_{ij}{}^E - \overline{X}_{ij}{}^C}{s_{ij}}J$
Class effect size	d_{i+}	$\dfrac{\sum\limits_{j=1}^{k_i} w_{ij}d_{ij}}{\sum\limits_{j=1}^{k_i} w_{ij}}$
Grand mean effect size	d_{++}	$\dfrac{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k_i} w_{ij}d_{ij}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k_i} w_{ij}}$
Homogeneity within classes	Q_W	$\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k_i} w_{ij}(d_{ij} - d_{i+})^2$
Homogeneity between classes	Q_B	$\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{k_i} w_{ij}(d_{i+} - d_{++})^2$

parametric statistical tests is the use of resampling methods. These computer-intensive techniques are now beginning to gain wider application in single-study analyses in ecology and evolution (Manly 1991, Crowley 1992). Such methods have not previously been applied to meta-analysis. Resampling methods test the significance of a statistic by generating a distribution of that statistic by permuting the data many times, each time recalculating the statistic. By comparing the original statistic to this generated distribution, a significance level can be determined (Kempthorne and Doerfler 1969, Manly 1991). Resampling methods such as the bootstrap can also be used to estimate confidence limits for statistics. Because they generate their own distributions, resampling methods are free from the distribution assumptions of parametric tests, and, in many cases, may be more powerful than conventional nonparametric ranking approaches (Manly 1991, Adams and Anthony 1996).

Despite their growing popularity in primary analyses, resampling and randomization techniques have not been used in meta-analysis in any field, and their application to this secondary level of analysis raises questions that have not been addressed before by statisticians. Because ecological data may violate some of the assumptions for common meta-analysis statistics, we propose an approach by which resampling methods can be applied to the statistical tests of significance and to the calculation of confidence limits in meta-analysis. We then compare the results of these analyses to the results from standard meta-analytic methods for three ecological data sets.

METHODS

Calculations

In conventional meta-analysis, effect sizes are calculated from the means, sample sizes, and standard deviations of the experimental and control groups in each study (Hedges and Olkin 1985), and are then combined to obtain an estimate of the mean effect size for each class, d_{i+}, as well as the grand mean effect size for all studies, d_{++}. It is often of interest to test whether classes of studies differ in their effect sizes. A homogeneity statistic, Q_B, can be used to assess whether the classes of studies differ significantly from one another, and the statistic Q_W can be used to test for within-class homogeneity (see Table 1 for formulas). This method of determining within- and between-class homogeneity is analogous to the partitioning of variance into within- and between-group components in an analysis of variance (see Gurevitch and Hedges [1993] for a more detailed explanation).

Studies are typically weighted by an estimate of the precision of the effect size, based on the reasonable assumption that more-precise studies (e.g., those with larger sample sizes) should be weighted more heavily than those that are less precise. The parametric weights usually used are inversely proportional to the estimated sampling variance, and are calculated as $w_{ij} = 1/v_{ij}$, where

$$v_{ij} = \frac{N_{ij}{}^{E} + N_{ij}{}^{C}}{N_{ij}{}^{E} N_{ij}{}^{C}} + \frac{d_{ij}{}^{2}}{2(N_{ij}{}^{E} + N_{ij}{}^{C})}$$

(Hedges and Olkin 1985). N^{E} and N^{C} are the experimental and control group sample sizes, and d_{ij} is the effect size for that study. This weighting minimizes the variance of d_{i+}, and is the most precise weighting estimate when the assumptions based on large-sample theory are satisfied. An alternative weighting that makes fewer assumptions, but still incorporates the desired property of counting larger studies more heavily than small ones is

$$w_{ij} = \frac{N_{ij}{}^{E} N_{ij}{}^{C}}{N_{ij}{}^{E} + N_{ij}{}^{C}}$$

(Hedges and Olkin 1985:110). We used this weighting for all resampling tests to reduce our reliance on parametric assumptions. This choice of weights will result in slightly less efficient estimates than the parametric weights when all of the assumptions for the parametric models are met, but the resampling analysis will be valid even when the parametric analysis is not.

Data

We compared the results of conventional meta-analysis to those obtained from resampling techniques for three ecological data sets. The first data set (Gurevitch and Hedges 1993) contained 43 experimental field studies of competition among plants, grouped into three classes according to system: terrestrial (19 studies), lentic (2 studies), and marine (22 studies). The responses to competitors were expressed as either an increase in the number of individuals or an increase in the size of individuals (i.e., growth). The second data set (Gurevitch et al. 1992) included 214 studies on the effect of competition on biomass, grouped into five classes: primary producers (74 studies), deposit feeders (3 studies), filter feeders (3 studies), herbivores (112 studies), and carnivores (22 studies). The third data set (Curtis 1996) examined the responses of plants from 69 studies to elevated atmospheric CO_2 levels, and included consideration of the effect of different stress (nutrient stress, 7 studies; general stress, 8 studies; unstressed, 54 studies), duration of exposure to elevated CO_2 (<50 d, 7 studies; 50–100 d, 19 studies; 101 d–1 yr, 21 studies; >1 yr, 6 studies), different experimental conditions (greenhouse, 12 studies; growth chamber, 12 studies; open-top chamber, 19 studies), and different pot sizes (small, 4 studies; large, 5 studies; in-ground, 10 studies). In all three data sets, several studies from each class had sample sizes <10, with the exception of deposit feeders (Gurevitch et al. 1992).

Statistics

All analyses were carried out using MetaWin (Rosenberg et al. 1997). For each data set, we calculated mean effect sizes and 95% confidence limits for each class using parametric, fixed-effects model and mixed-effects model meta-analytic techniques. We also calculated the between-class homogeneity (Q_B) for each data set and tested this against a chi-square distribution to determine if classes differed significantly from one another. We then calculated bootstrap confidence limits for the mean effect sizes for each class and for the grand mean effect sizes for comparison.

We first calculated bootstrap confidence limits for the mean class effect sizes using the conventional method, the percentile bootstrap (Efron 1979). For each class, we chose i studies with replacement and calculated a weighted mean effect size. We repeated this process 4999 times, ordered the output values sequentially, and chose the lowest and highest 2.5% values as our bootstrap confidence limits (Manly 1991, Dixon 1993). Classes whose cumulative effect size confidence limits contained zero were judged not to be significantly different from zero. One problem with percentile bootstrap confidence limits however, is that as sample sizes decrease, the lengths of percentile bootstrap confidence limits tend to become underestimated, because percentile bootstraps assume that the bootstrap distribution is centered around the observed value (Efron 1987, Dixon 1993). Bias-corrected percentile bootstrap confidence limits have been suggested to correct for distributions when >50% of the bootstrap replicates are larger or smaller than the observed value, which happens often with small samples (Efron 1987). As some of our classes contained a small numbers of studies, we also calculated the bias-corrected percentile bootstraps for comparison by calculating the normal cumulative distribution function for twice the probit transformation of the fraction of bootstrap replicates smaller than the observed value ± 1.96 (see Dixon 1993).

For all three data sets, we examined whether classes within each data set differed significantly from one another using both chi-square tests for between-class homogeneity (Q_B) and randomization tests for comparison. First we calculated the Q_B for each original data set using standard meta-analytic methods (see Table 1). We then conducted a randomization test by randomly re-assigning studies to classes, each time calculating a new Q_B. We repeated this procedure 4999 times to generate a distribution of Q_B values. The significance level of Q_B is the percentage of times that Q_B generated from the randomized replicates was more extreme than the Q_B calculated from the original data. One question of interest when using resampling techniques is how many iterations are necessary. To examine this we performed 30 meta-analyses on the Gurevitch and Hedges (1993) data set with 100, 250, 500, 1000, 2500, 5000, 7500, 10 000, 15 000, and 20 000

TABLE 2. Cumulative effect sizes and confidence limits based on parametric meta-analytic and resampling methods: d_{i+} = mean effect size for each class; d_{++} = grand mean effect size for all studies.

Class	N	d_{i+} Fixed-effect model	d_{i+} Mixed-effect model	95% CL Fixed-effect model	95% CL Mixed-effect model
Gurevitch et al. 1992					
Primary producer	74	0.34	0.69	0.29–0.39	−0.05–0.63
Deposit feeder	3	0.06	0.12	−0.08–0.18	−0.71–0.96
Filter feeder	3	1.26	1.46	0.62–1.90	0.41–2.52
Herbivore	112	1.14	0.96	1.10–1.18	0.81–1.13
Carnivore	22	0.25	0.29	0.11–0.38	−0.05–0.21
Total d_{++}	214	0.80	0.78	0.77–0.83	0.72–0.84
Gurevitch and Hedges 1993					
Terrestrial organisms	19	1.141	1.09	0.92–1.37	0.75–1.43
Lentic organisms	2	4.107	4.12	2.37–5.84	2.26–5.98
Marine organisms	22	0.798	0.70	0.56–1.04	0.37–1.03
Total d_{++}	43	1.01	0.94	0.85–1.17	0.71–1.17
Curtis 1996					
Unstressed plant	54	1.08	1.47	0.94–1.25	1.16–1.79
Nutrient-stressed plant	7	0.68	0.76	0.29–1.09	−0.07–1.59
Generally stressed plant	8	1.66	1.80	1.15–2.17	0.96–2.65
Total d_{++}	69	1.09	1.42	0.95–1.23	1.12–1.56

replications, and calculated the variance at each of these levels to determine when it leveled off.

RESULTS

Confidence limits

For most classes, bootstrap confidence limits from the mixed-effects model were greater than those from the fixed-effects model, and bootstrap confidence limits tended to be slightly larger than the parametric confidence limits, suggesting that the resampling approach may be more conservative (Table 2). We also found that classes with a larger number of studies had more variation in their response, and thus tended to have wider percentile bootstrap confidence limits (Table 2). In general, confidence limits from the mixed-effects model were more similar to the bootstrap confidence limits than were the confidence limits from the fixed-effects model. For lentic organisms, with a sample size of two studies, we found narrower percentile bootstrap confidence limits than parametric confidence limits. In this case however, only three outcomes from resampling were possible: the smaller effect size being selected twice, each effect size being selected once, or the larger effect size being selected twice. These events occur in a 1:2:1 ratio, and the lower and upper percentile bootstrap confidence limits each represent one effect size being selected twice. Bias-corrected percentile bootstrap confidence limits were usually slightly wider than percentile bootstrap confidence limits (Table 2). In several cases we found slightly wider percentile bootstrap confidence limits. Three of these cases (deposit feeders, filter feeders, and lentic organ-

isms) had extremely small sample sizes (3, 3, and 2 respectively).

Mean effect sizes of all classes in the three studies were determined to be significantly different from zero using fixed-effects parametric confidence limits, with the exception of deposit feeders, where there was no difference in response to competition between experimental and control treatments. With the mixed-effects model, mean effect sizes for primary producers, carnivores, and nutrient-stressed organisms were not significantly different from zero. In one instance, primary producers, the effect size was considered to be different from zero using standard confidence limits, but not different from zero using bootstrap confidence limits, implying that, in some instances, confidence limits derived from standard methods may be less conservative in establishing the significance of small to moderate effects.

Homogeneity statistics

With the fixed-effects model, parametric tests of between-class homogeneity indicated highly significant differences between classes in all cases (Table 3). Results from the mixed-effects model were similar, except for stress and pot size, in which no significant differences among classes were found (Table 3). In general, results from randomization tests agree with those from the parametric tests. However, in two instances, stress level and pot size, chi-square tests from the fixed-effects model resulted in a significant difference between classes while randomization tests did not. In these cases, it is difficult to determine whether the underlying

Table 2. Continued.

Bootstrap CL		Bias-corrected CL	
Fixed-effect model	Mixed-effect model	Fixed-effect model	Mixed-effect model
−0.085–0.589	−0.082–0.659	−0.080–0.592	−0.529–0.901
−0.292–0.775	−0.292–0.775	−0.292–0.442	−0.292–0.775
0.158–3.426	0.158–3.426	0.158–2.805	0.158–2.805
0.769–1.528	0.769–1.209	0.774–1.530	0.774–1.530
0.276–0.722	−0.066–0.722	0.274–0.710	0.274–0.710
0.534–1.086	0.607–0.955	0.534–1.088	0.596–0.941
0.907–1.383	0.786–1.341	0.894–1.378	0.807–1.370
3.696–4.674	3.696–4.674	3.696–4.674	3.696–4.674
0.376–1.327	0.308–1.112	0.407–1.362	0.407–1.229
0.854–1.371	0.764–1.328	0.857–1.380	0.758–1.325
0.826–1.490	1.178–1.812	0.809–1.474	1.213–1.834
0.455–1.091	0.413–1.142	0.452–1.090	0.424–1.143
0.831–2.624	0.829–2.811	0.839–2.644	0.863–2.869
0.855–1.393	1.149–1.721	0.855–1.398	1.145–1.719

conditions of meta-analysis are being violated. However, this discrepancy suggests that conclusions based on resampling methods may be more conservative.

Randomization tests at different levels of replication showed that the variance around the significance level of Q_B dropped substantially by 2500 and 5000 replications and approached zero as the number of replications increased, implying that Q_B is a consistent estimator of the true difference between classes. Because of this asymptotic decrease in Q_B past 2500 replications, we recommend that, for meta-analysis, at least 2500 replications be used in such randomization tests.

DISCUSSION

The purpose of meta-analysis is to provide researchers with a statistical tool to summarize, synthesize, and evaluate independent research studies in order to reach general conclusions. This approach is more powerful, more informative, more accurate, and less likely to be biased than simple narrative reviews and ''vote counting'' (e.g., see Hedges and Olkin 1985, Gurevitch et al. 1992, Arnqvist and Wooster 1995). Attempts to employ meta-analysis in ecological research bring into sharper focus a number of as-yet-unresolved issues and problems in synthesizing and summarizing the current state of knowledge in an area. These issues include publication bias (the tendency for journals to favor publication of papers with statistically significant results), research bias (the tendency to choose subjects for research where significant results are anticipated; J. Gurevitch, *unpublished manuscript*), non-independence and autocorrelation among studies, identifying the best metrics for meta-analysis, how to deal with studies of

varying quality, and other issues (Cooper and Hedges 1994).

One possible limitation of conventional parametric meta-analysis that has generally been overlooked is that if the distributional assumptions are not met, estimates of mean effect sizes and their confidence limits may not be accurate, and tests for the differences among classes may be unreliable. It is not known how robust the statistics of meta-analysis are to violations of these assumptions.

The general agreement between our results and those obtained using parametric calculations is reassuring in that it suggests either that ecological data commonly meet the assumptions of meta-analytic techniques, or that the statistics of parametric meta-analysis may be fairly robust to violations of these assumptions. Since the results from parametric meta-analysis and our resampling statistics were similar, one might ask, ''is this really necessary?'' In several instances the results of our resampling methods differed from those found with parametric techniques, resulting in different conclusions. Therefore, it may be reasonable to use more conservative approaches that do not rely on meeting distributional assumptions. This can be accomplished by incorporating bootstrap confidence limits in meta-analysis to help determine the magnitude of mean effect sizes, as well as using randomization tests to assess the between-class homogeneity.

Philosophically, we believe that it is important to be conscious of the underlying assumptions of any statistical technique, and that data subjected to meta-analysis may sometimes not meet the assumptions of common tests. Researchers performing meta-analysis are syn-

TABLE 3.   Homogeneity assessments based on chi-square and randomization tests. Q_B = between-class homogeneity statistic; NS = nonsignficant.

Study	Model	Group	Q_B	Chi-square P	Random-ization P
Gurevitch et al. 1992	Fixed-effect	All systems	793.430	0.0001	0.0358
	Mixed-effect	All systems	17.510	0.0015	0.0766[NS]
Gurevitch and Hedges 1993	Fixed-effect	All classes	16.497	0.0001	0.0006
	Mixed-effect	All classes	13.980	0.0001	0.0028
Curtis 1996	Fixed-effect	Stress	8.612	0.0135	0.3494[NS]
	Mixed-effect	Stress	3.290	0.1926[NS]	0.2180[NS]
Curtis 1996	Fixed-effect	Duration	47.949	0.0001	0.0034
	Mixed-effect	Duration	30.939	0.0001	0.0008
Curtis 1996	Fixed-effect	Experiment type	24.413	0.0001	0.0242
	Mixed-effect	Experiment type	19.641	0.0001	0.0004
Curtis 1996	Fixed-effect	Pot size	9.353	0.0093	0.6994[NS]
	Mixed-effect	Pot size	1.333	0.5130[NS]	0.7818[NS]

thesizing previously published studies, which may not provide all the necessary statistical information to assess whether assumptions are met. In many cases it is difficult to obtain even the basic statistics from published studies (Gurevitch et al. 1992). It is not unimaginable that other ecological data may violate these assumptions far more severely than did the data we analyzed, resulting in greater discrepancies between parametric and resampling results. Thus, distribution-free resampling methods may often be preferable for ecological meta-analysis.

An additional advantage of resampling methods in meta-analysis is that, unlike parametric methods, they do not require computation of the standard errors, used to calculate weights, of the individual effect size estimates. The randomization strategy advanced here uses weights that are similar but not identical to the parametric weights. When published research results fail to include the basic statistical information (in particular, the standard deviations of the means of the experimental and control group responses) needed to compute the parametric standard error of the effect size, they would ordinarily have to be omitted from any meta-analysis of the literature. The use of the approach suggested here would therefore potentially permit the inclusion of a larger number of studies in a meta-analysis than will the use of standard parametric methods. This is particularly true where, rather than *d,* an effect size is used that does not require an estimate of the standard deviation, such as the response ratio (the ratio of the means of the experimental and control groups; see Rosenberg et al. 1997; L. V. Hedges, P. Curtis and J. Gurevitch, *unpublished manuscript*). In ecology, where reporting of statistics is uneven, this may be a major advantage.

As more and more experimental studies are performed in ecology, the need for synthesizing information becomes increasingly important. We encourage researchers to explore the use of distribution-free re- sampling methods in conducting meta-analyses for both confidence-interval estimation and hypothesis testing, particularly where there is reason to believe that their data may not meet the assumptions of parametric methods.

LITERATURE CITED

Adams, D. C., and C. D. Anthony. 1996. Using randomization techniques to analyze behavioural data. Animal Behaviour **51**:733–738.

Arnqvist, G., and D. C. Wooster. 1995. Meta-analysis: synthesizing research findings in ecology and evolution. Trends in Ecology and Evolution **10**:236–240.

Cooper, H., and L. V. Hedges. 1994. The handbook of research synthesis. Russel Sage Foundation, New York, New York, USA.

Crowley, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. Annual Review of Ecology and Systematics **23**:405–447.

Curtis, P. 1996. A meta-analysis of leaf gas exchange and nitrogen in trees grown under elevated carbon dioxide. Plant, Cell and Environment **19**:127–137.

Dixon, P. M. 1993. The bootstrap and jackknife: describing the precision of ecological indices. Pages 290–318 in S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman & Hall, New York, New York, USA.

Efron, B. 1979. Bootstrap methods: another look at the jacknife. Annual Statistician **7**:1–26.

———. 1987. Better bootstrap confidence intervals (with discussion). Journal of the American Statistical Association **82**:171–200.

Gurevitch, J., and L. V. Hedges. 1993. Meta-analysis: combining the results of independent experiments. Pages 378–398 in S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman & Hall, New York, New York, USA.

Gurevitch, J., L. L. Morrow, A. Wallace, and J. S. Walsh.

1992. A meta-analysis of competition in field experiments. American Naturalist **140**:539–572.

Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, San Diego, California, USA.

Kempthorne, O., and T. E. Doerfler. 1969. The behaviour of some significance tests under experimental randomization. Biometrika **56**:231–248.

Manly, B. F. J. 1991. Randomization and Monte Carlo methods in biology. Chapman & Hall, New York, New York, USA.

Raudenbush, S. W. 1994. Random effects models. Pages 301–321 *in* H. Cooper and L. V. Hedges, editors. The handbook of research synthesis. Russel Sage Foundation, New York, New York, USA.

Rosenberg, M. S., D. C. Adams, and J. Gurevitch. 1997. MetaWin. Statistical software for meta-analysis with resampling tests. Version 1.0. Sinauer Associates, Sunderland, Massachusetts, USA, *in press.*

Tonhasca, A., and D. N. Byrne. 1994. The effect of crop diversification on herbivorous insects: a meta-analysis approach. Ecological Entomology **19**:239–244.