STATISTICAL REPORTS

Ecology, 86(7), 2005, pp. 1751–1756

A STATISTICAL TEST TO SHOW NEGLIGIBLE TREND

PHILIP M. DIXON[1,3] AND JOSEPH H. K. PECHMANN[2]

[1]*Department of Statistics, Iowa State University, 120 Snedecor Hall, Ames, Iowa 50011-1210 USA*
[2]*Department of Biological Sciences, University of New Orleans, New Orleans, Louisiana 70148 USA*

Abstract. The usual statistical tests of trend are inappropriate for demonstrating the absence of trend. This is because failure to reject the null hypothesis of no trend does not prove that null hypothesis. The appropriate statistical method is based on an equivalence test. The null hypothesis is that the trend is not zero, i.e., outside an a priori specified equivalence region defining trends that are considered to be negligible. This null hypothesis can be tested with two one-sided tests. A proposed equivalence region for trends in population size is a log-linear regression slope of (−0.0346, 0.0346). This corresponds to a half-life or doubling time of 20 years for population size. A less conservative region is (−0.0693, 0.0693), which corresponds to a halving or doubling time of 10 years. The approach is illustrated with data on four amphibian populations; one provides significant evidence of no trend.

Key words: Ambystoma; *amphibian decline;* Desmognathus; *equivalence tests; population trends; statistical power; testing for no effect.*

INTRODUCTION

Many discussions of ecological and environmental issues involve evaluating the evidence for or against a temporal trend. For example, is the abundance of a particular population increasing, remaining approximately constant, or declining over time? The data to answer this question are often a sequence of annual counts of individuals (e.g., Houlahan et al. 2000). Do the observed counts represent random fluctuations around no trend or do they provide evidence of some trend? If it is reasonable to assume a linear trend, the usual statistical analysis is to fit a linear regression and test the null hypothesis that the slope is zero. This analysis is appropriate to identify a non-zero trend because a statistically significant result provides good evidence that the trend is not zero. It is not appropriate for identifying the absence of an important trend. Failure to reject the null hypothesis of no trend does not imply that the null hypothesis is true (Anderson and Hauck 1983, Millard 1987, Dixon 1998, Johnson 1999, Parkhurst 2001, Cole and McBride 2004). A nonsignificant result may be due to a small sample size, large random fluctuations in abundance, a poor choice of test, a trend that is close to zero in a practical sense, or the true absence of trend.

Manuscript received 30 August 2004; accepted 27 October 2004; final version received 3 February 2005. Corresponding Editor: K. L. Cottingham.

[3] E-mail: pdixon@iastate.edu

Previous approaches to the interpretation of nonsignificant results have focused on statistical power (Thomas 1997), the probability that a statistical-hypothesis test will reject the null hypothesis of no trend when that hypothesis is false (i.e., the true trend is not zero). A typical use of power calculations is to find a sample size (e.g., number of survey years), given a specified trend and random variation, for which a trend test is likely (e.g., power > 0.8) to give a statistically significant result. While power calculations are invaluable for designing a study (Cohen 1988), they are less useful for interpreting nonsignificant results once obtained (Mead 1988, Gerard et al. 1998, Hoenig and Heisey 2001). One problem is that power calculations should be based on a priori specification of the trend and error variance, derived from the literature, preliminary data, or biological principles (Thomas 1997). If power is calculated after data are collected and the observed estimates of trend and variance used in the power calculation (post hoc power), the estimated power is simply a function of the P value (Mead 1988). For example, if the estimated trend and its variance are such that the P value is exactly 5%, the post hoc power of an $\alpha = 5\%$ test is approximately 50% (Mead 1988). If the P value is smaller ($P < 5\%$), the post hoc power is larger; if the P value is larger than 5%, the post hoc power is less than 50%. Such post hoc power calculations provide no additional insight into the nature of nonsignificant results (Thomas 1997, Hoenig and Heisey 2001).

The usual test of no trend can also be too powerful, although this rarely happens with ecological data. If the sample size is large or the residual variation small, then a biologically insignificant trend (e.g., numerically close to zero) can be statistically significant. A statistical test of no trend is not a test of whether the trend is biologically important.

A better approach for testing the absence of trend is motivated by the idea that the true trend is unlikely to be exactly zero. The important question is whether the true trend is negligible. This requires defining an equivalence region (b_l, b_u) that includes all values of the trend parameter that are considered negligible. The lower bound of the equivalence region, b_l, separates larger declines (i.e., more-negative trends) that are biologically important from smaller declines that are considered negligible. The upper bound, b_u, separates larger biologically important increases from smaller positive trends. An equivalence test assumes that the trend is large, i.e., outside the equivalence region, unless the data suggest otherwise. If β is the true, but unknown, trend, the null hypothesis of non-equivalence is that

$$H_0: \beta \leq b_l \text{ or } \beta \geq b_u. \quad (1)$$

The alternative hypothesis is that the true trend is within the equivalence region: H_a: $b_l < \beta < b_u$. The usual null and alternative hypotheses are reversed, so that a trend is considered negligible only if there is sufficient evidence that it is close to zero.

Testing the Null Hypothesis of Non-equivalence

Most statistical research on equivalence testing has focused on equivalence tests for two means. One primary motivation was to compare properties of generic and name-brand drugs (Wellek 2003:6). Equivalence tests are relatively unknown in ecological and environmental applications, although they have been applied to assess remediation success (McDonald and Erickson 1994), the assumption of equal detectability (MacKenzie and Kendall 2002), and the lack of environmental impact (Erickson and McDonald 1995, Cole and McBride 2004).

Many different equivalence tests have been suggested (e.g., Westlake 1979, Anderson and Hauck 1983, Schuirmann 1987, Dannenberg et al. 1994, Hsu et al. 1994). There is no optimal test. Instead there is a trade-off between three characteristics of the equivalence test: the Type I error rate, the power, and the shape of the rejection region (Chow and Liu 1992, Berger and Hsu 1996, Perlman and Wu 1999). The rejection region of a statistical test is the set of sample statistics that lead to rejecting the null hypothesis. For t tests of trend, the relevant sample statistics are the estimated slope and standard error of the slope. All equivalence tests reject the non-equivalence hypothesis when the observed trend is close to zero and precisely known (small SE). Some equivalence-test procedures also conclude that the trend is negligible when the slope is poorly known (large SE), although this is counterintuitive. The two one-sided tests method (Schuirmann 1987) is widely used because it has a bounded type I error rate, good power, and a well-behaved rejection region (Hsu et al. 1994).

The two one-sided tests method separately tests each part of the non-equivalence hypothesis given by Eq. 1 (Schuirmann 1987, Parkhurst 2001). Two one-sided null hypotheses are tested: H_{0a}: $\beta \leq b_l$ and H_{0b}: $\beta \geq b_u$. Non-equivalence (Eq. 1) is rejected only if both subhypotheses, H_{0a} and H_{0b}, are rejected. The details of each one-sided test depend on the properties of the data. This flexibility permits generalization to many approaches, including tests for data with unequal variances (Dannenberg et al. 1994), nonparametric tests (Hauschke et al. 1990), and complex experimental designs (Chow and Liu 1992). A single one-sided test can be applied when the original hypothesis is one sided, i.e., only positive or negative trend is important (Parkhurst 2001). Here we extend equivalence testing to the question of whether a linear trend is close to zero.

Using an equivalence test requires an a priori specification of b_l and b_u, the bounds of the equivalence region. These values should represent biological knowledge and informed judgment about trends that are considered small for a specific population over a specific time frame. One approach is based on the doubling time for the population. Here we operationally define a trend as small if the associated population doubling time is longer than 20 years, i.e., the log-linear slope is smaller than 0.0346. The comparable criterion for declining populations is a half-life longer than 20 years, i.e., a log-linear slope larger than -0.0346. A related approach is to consider the time to reach 1% of the starting size (pseudo-extinction). A consistent annual decline of -0.0346 translates into a pseudo-extinction time of 133 years. The bounds of the equivalence region may vary with species characteristics, e.g., life history and current population size. Looser bounds on the equivalence region, e.g., $(-0.0693, 0.0693)$ that correspond to a doubling or halving time of 10 years, might be appropriate for populations of shorter lived species with larger annual fluctuations in abundance.

Equivalence Tests for Trend

Many different models could be used to estimate trends. We will use a log-linear model in which the slope, β, describes the linear trend in the log-transformed abundance, N_t:

$$\ln(N_t + 1) = \alpha + \beta t + \varepsilon_t. \quad (2)$$

We chose to log transform abundances to linearize an exponential growth model and to stabilize the error variances. A constant of 1 is added to all values of N_t to avoid ln(0). When N_t is large, Eq. 2 describes a

population with exponential growth or exponential decline at a rate given by β. When N_t is small, the growth or decline is approximately linear, because of the added constant.

The choice of method to estimate the trend, $\hat{\beta}$, and its standard error, $s_{\hat{\beta}}$, depends on the characteristics of the errors, i.e., the deviations from the specified model. If the errors are additive, independent, normally distributed, with equal variances, then least-squares regression (Draper and Smith 1981) is appropriate and inference about the slope can be based on a Student's t distribution. If errors are correlated, either because of autocorrelation between observations in consecutive years or because of subsampling (e.g., more than one count in the same year), the annual trend and its SE can be estimated using a linear mixed model (Schabenberger and Pierce 2002: chapter 7). Inference about the slope is based on an approximate t distribution with estimated degrees of freedom (Kenward and Roger 1997).

In either case, the subhypothesis H_{0a}: $\beta \leq b_l$ is rejected if the t statistic $T_l = (\hat{\beta} - b_l)/s_{\hat{\beta}}$ is larger than the one-sided critical value for a t distribution with the appropriate degrees of freedom. The second sub-hypothesis H_{0b}: $\beta \geq b_u$ is rejected if the t statistic $T_u = (b_u - \hat{\beta})/s_{\hat{\beta}}$ is larger than the same t critical value. If the P values for both sub-hypotheses are less than α (e.g., 5%), then the data provide evidence that the trend is negligible. Although this decision requires two hypothesis tests, a multiple testing adjustment is not necessary because rejecting the non-equivalence hypothesis requires that both tests are significant.

Equivalence can also be based on a confidence interval. The hypothesis of non-equivalence (Eq. 1) is rejected at $\alpha = 5\%$ if and only if a 90% confidence interval for the trend lies entirely within the equivalence region (Schuirmann 1987). If the usual least-squares assumptions are appropriate, a 90% confidence interval for the trend is $\hat{\beta} \pm ts_{\hat{\beta}}$, where t is the 0.95 quantile of a t distribution with the appropriate degrees of freedom. The size of the confidence interval is $100\% - 2\alpha$ not the usual $100\% - \alpha$ because each tail of the confidence interval is based on a one-sided α-level test.

Amphibian Examples

Equivalence tests for trend will be illustrated with four long-term data sets on amphibian (salamander) population sizes. Complete counts of all breeding females of two *Ambystoma* species, *A. talpoideum* and *A. tigrinum*, have been made at Rainbow Bay, South Carolina, USA since 1979 (Semlitsch et al. 1996). Estimates of abundance of *Desmognathus monticola* and *D. ochrophaeus* at Coweeta Hydrological Laboratory North Carolina, USA have been made by constant-effort searches since 1976 (Hairston 1996). The data used here include the population counts until 2002. The number of searches for *Desmognathus* varied between one and three per year; for this paper, we consider the average count for each year. Two populations (*Ambystoma* spp.) have large annual variation; two (*Desmognathus* spp.) have small annual variation (Fig. 1). The four were selected from the larger number of amphibian species monitored in these community surveys.

AIC statistics were used to choose an appropriate model for the variability of observations around the log-linear regression line (Verbeke 1997:113–115). For all four species a first-order autoregressive error model was more appropriate than the independence model. For the two *Desmognathus* species, an equal-variance model was more appropriate than a weighted model that assumed the variance was a function of the number of counts made each year. Diagnostic plots indicate little to no evidence of unequal variances or non-normality in the residuals from the log-linear model. The degrees of freedom were estimated using the Kenward-Roger's (1997) approximation. The degrees of freedom differ between species, partly because of the larger sample size for *Desmognathus* and partly because of different autocorrelation coefficients. SAS version 8.2 (SAS Institute 1999) was used for all computations. The *Ambystoma* data and the code used to estimate the slopes and their standard errors and then calculate P values for equivalence tests is given in the Supplement.

Estimated trends are superimposed on the data in Fig. 1. The t tests of the null hypothesis that the trend equals 0 indicate strong evidence of a decline in *A. tigrinum* ($\hat{\beta} = -0.16$, $P = 0.0044$), weak evidence of a decline in *A. talpoideum* ($\hat{\beta} = -0.076$, $P = 0.051$), and no evidence of a trend in the two *Desmognathus* species (Table 1). The nonsignificant (defined as P value > 0.05) results for *A. talpoideum*, *D. ochrophaeus*, and *D. monticola* are not convincing evidence of no trend. An equivalence test is needed to support the claim of no trend.

We report the equivalence test for *D. monticola* in detail. The estimated slope is -0.0074, with a standard error of 0.0096. There are 27 years of data but a relatively large lag-1 autocorrelation ($\hat{\rho}_1 = 0.43$). The approximate t distribution has 3.34 degrees of freedom. The t values for each subhypothesis are: $T_l = (-0.0074 - -0.0346)/0.0096 = 2.83$ and $T_u = (0.0346 - -0.0074)/0.0096 = 4.37$. Both subhypotheses are rejected with $P < 0.05$, and we reject the null hypothesis of "non-equivalence." The P value for the overall equivalence test is the larger of P values for T_l and T_u, i.e., 0.029 (Table 1). There is evidence that the trend in *D. monticola* is negligible, according to our choice of equivalence region. For each of the other three species, at least one of the two subhypothesis is not rejected, so one cannot conclude that the trend is within the equivalence region (Table 1).

Examining the 90% confidence intervals for the trends provides exactly the same conclusions. The 90% CI for *D. monticola* is contained in the equivalence region of $(-0.0346, 0.0346)$, so the trend for that species is negligible (Table 1). The 90% confidence in-

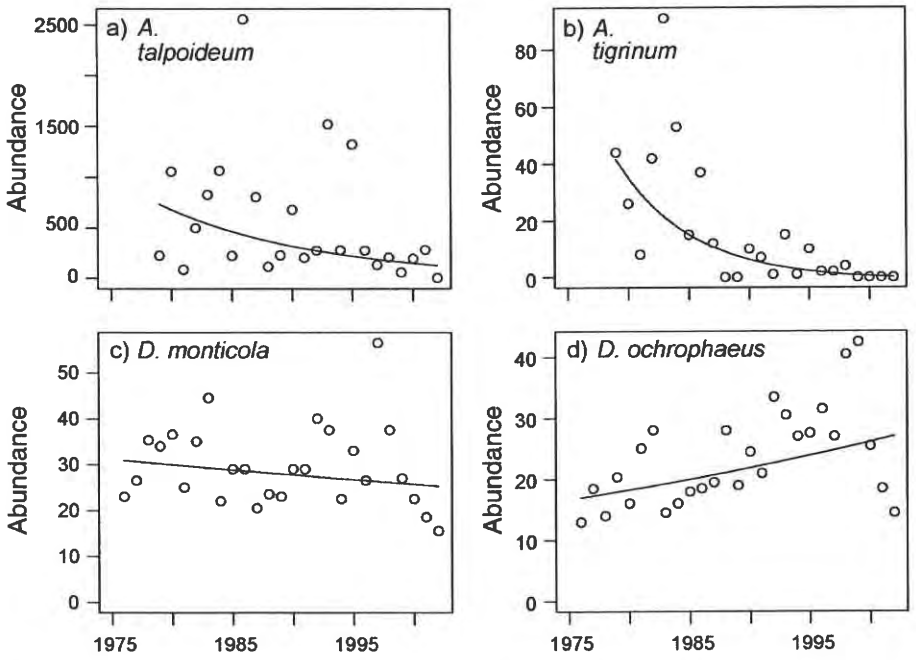

FIG. 1. Trends in number of breeding females for (a) *Ambystoma talpoideum* and (b) *Ambystoma tigrinum* over 24 years at Rainbow Bay, Aiken County, South Carolina USA, and in number of amphibians seen in constant-effort searches for (c) *Desmognathus monticola* and (d) *Desmognathus ochropheaus* over 27 years at Coweeta Hydrological Laboratory, Macon County, North Carolina, USA. Lines indicate predicted values from log-linear regressions.

tervals for the other three species fall at least partly outside the equivalence region (Table 1).

DISCUSSION

For *Ambystoma tigrinum* and *Desmognathus monticola*, the conclusions from the equivalence test agree with those from the *t* test of slope equal to 0. The trend in *A. tigrinum* is not 0 using the *t* test; the null hypothesis of non-negligible trend is accepted using the equivalence test (Table 1). The trend in *D. monticola* is not significantly different from 0; the equivalence test indicates a negligible trend. The two tests provide complementary rather than redundant insights, however, because they address different questions. This is illustrated by *D. ochrophaeus*. The trend is not significantly different from 0, but the equivalence test fails to support the opposite conclusion that the trend is near 0. Together, the two tests suggest that there is insufficient evidence to decide whether the *D. ochrophaeus* population is increasing slowly or remaining the same. The evidence is also inconclusive for *A. talpoideum* although there is borderline support for a decline.

The two tests do not always agree because the rejection regions for the two tests are quite different. The rejection region for a test is the set of observed summary statistics for which that test rejects the null hypothesis at a specified α level. For tests of trend, the two summary statistics are the estimated trend, $\hat{\beta}$, and the standard error of that estimate. The orientations of the boundaries of the rejection region depend on the *t* quantile, i.e., they are related to the error degrees of freedom. The rejection region for the usual test of no

TABLE 1. Log-linear trends, $\hat{\beta}$, for each of the four amphibian data sets.

Species	No. years	β	s_β†	df	*P* value‡			90% confidence limit
					H_0: $\beta = 0$	H_0: $\beta < b_l$	H_0: $\beta > b_u$	
Ambystoma talpoideum	24	−0.076	0.031	5.78	0.051	0.88	0.0062	(−0.14, −0.016)
A. tigrinum	24	−0.162	0.034	5.43	0.0044	0.99	0.00089	(−0.23, −0.09)
Desmognathus monticola	27	−0.0074	0.0096	3.34	0.49	0.029	0.0089	(−0.029, 0.014)
D. ocrophaeus	27	0.017	0.010	1.98	0.23	0.018	0.11	(−0.012, 0.046)

† Standard error of the trend, β.

‡ *P* values are given for the usual test of no trend (H_0: $\beta = 0$) and the two parts of the equivalence test. The equivalence region is (lower bound $b_l = -0.0346$, upper bound $b_u = 0.0346$), which corresponds to a doubling or halving time of 20 years.

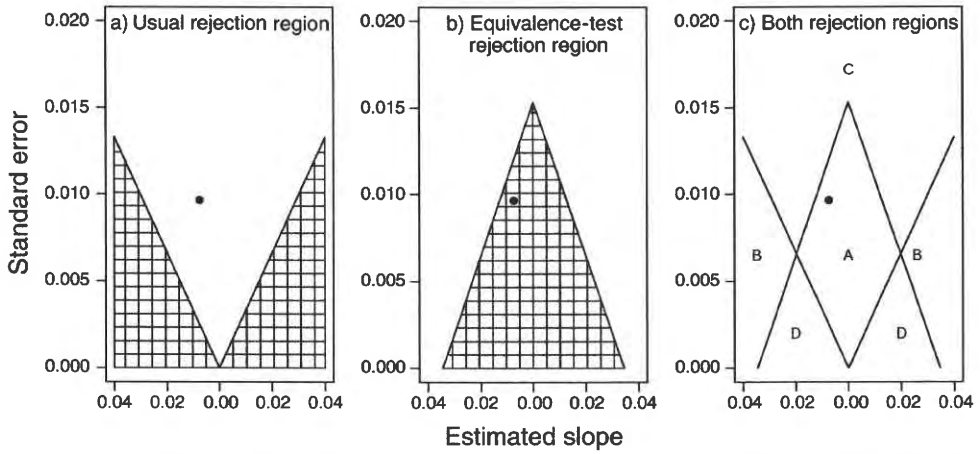

FIG. 2. Rejection regions for the usual test of no trend and the equivalence test. The rejection region is the set of estimated slopes and their standard errors for which the null hypothesis is rejected. These are shown for error df = 3.34, the estimated error df for *Desmognathus monticola*. (a) The cross-hatched area is the rejection region for the usual test of H_0: $\beta = 0$. (b) The cross-hatched area is the rejection region for the equivalence test, using an equivalence region with lower bound $b_l = -0.0346$ and upper bound $b_u = 0.0346$. (c) Overlay of the two rejection regions. The areas labeled A, B, C, and D are described in the *Discussion*. In all panels, the dot indicates the estimated slope and standard error for *D. monticola*.

difference (H_0: $\beta = 0$) for *D. monticola* is the cross-hatched area in Fig. 2a. The rejection region for the equivalence test for this species is the region inside the crosshatched triangle in Fig. 2b.

If results of the two tests are considered together, there are four possible outcomes (Fig. 2c). If the trend is significantly different from zero and not significantly inside the equivalence region, both tests provide evidence of an ecologically significant trend (areas labeled B, Fig. 2c). The trend in *A. tigrinum* illustrates this case. The other consistent pair of results is when the trend is not significantly different from 0 and significantly inside the equivalence region (Fig. 2c: area A). This provides strong evidence of no ecologically significant trend. The trend in *D. monticola* illustrates this case. A third case occurs when the trend is not significantly different from zero and also not significantly inside the equivalence region (Fig. 2c: area C; e.g., *D. ocrophaeus*). This indicates that the trend is not estimated well enough to make strong conclusions. The sample size is insufficient relative to the residual variation (and perhaps also autocorrelation). A fourth case, trend both significantly different from 0 and significantly negligible, is possible (Fig. 2c: areas labeled D). This case is most likely when the standard error of the trend is small. One interpretation of this fourth case is that the trend is not 0, but is so small that it is biologically unimportant. None of the species considered here illustrate this case.

An alternative to the three hypothesis tests is to calculate confidence intervals around estimated trends. Two intervals must be calculated. A $1 - \alpha$ confidence interval is appropriate to evaluate whether the trend differs from 0. A $1 - 2\alpha$ confidence interval is appropriate to evaluate whether the trend is negligible. Both the hypothesis test and confidence-interval methods of evaluating equivalence require the definition of an equivalence region.

The proposed equivalence regions can be related to IUCN—The World Conservation Union categories of threatened species (IUCN 2001). Simplifying the definitions slightly, a decline in numbers of >50% in 10 years defines an "endangered" species. So, the equivalence region of (−0.0693, 0.0693) corresponds to "not endangered." A decline of 30% in 10 years defines a "vulnerable" species, so the equivalence region of (−0.0346, 0.0346) corresponds to "not vulnerable." Results from equivalence tests depend critically on the choice of equivalence region. The 90% CI of the trend in *D. ochrophaeus* (−0.012, 0.046) falls entirely within the larger equivalence region of (−0.0693, 0.0693), indicating that we have sufficient evidence to conclude that the species is "not endangered," even though there was insufficient evidence to conclude that it is "not vulnerable," i.e., that the trend lies within (−0.0346, 0.0346).

Equivalence methods provide a way to evaluate the absence of trends after data are collected. They complement power analyses, which are most useful for designing a study. As always, summarizing a trend and understanding its cause(s) are separate issues.

ACKNOWLEDGMENTS

Research was conducted with support of the Environmental Remediation Sciences Division of the Office of Biological and Environmental Research, U.S. Department of Energy, through the Financial Assistance Award number DE-FC09-96SR18546 to the University of Georgia Research Foundation, and by Iowa State University and the University of New Orleans. We thank Kathy Cottingham, David Parkhurst, and an anonymous reviewer for their insightful and helpful comments. The *Desmognathus* data were kindly provided by N.

G. Hairston, Sr., and R. H. Wiley. We thank D. E. Scott, J. W. Gibbons, J. L. Greene, and B. S. Metts for access to recent *Ambystoma* data and the many other colleagues who have helped with the Rainbow Bay project.

Literature Cited

Anderson, S., and W. W. Hauck. 1983. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Communications in Statistics, A: Theory and Methods **12**:2663–2692.

Berger, R. L., and J. C. Hsu. 1996. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science **11**:283–319 (with discussion).

Chow, S-C., and J. P. Liu. 1992. Design and analysis of bioavailability and bioequivalence studies. Marcel Dekker, New York, New York, USA.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

Cole, R., and G. McBride. 2004. Assessing impacts of dredge spoil disposal using equivalence tests: implications of a precautionary (proof of safety) approach. Marine Ecology Progress Series **279**:63–72.

Dannenberg, O., H. Dette, and A. Munk. 1994. An extension of Welch's approximate *t*-solution to comparative bioequivalence trials. Biometrika **81**:91–101.

Dixon, P. M. 1998. Assessing effect and no effect with equivalence tests. Pages 273–301 *in* M. C. Newman and C. L. Strojan, editors. Risk assessment: logic and measurement. Ann Arbor Press, Chelsea, Michigan, USA.

Draper, N. R., and H. Smith. 1981. Applied regression analysis. John Wiley and Sons, New York, New York, USA.

Erickson, W. P., and L. L. McDonald. 1995. Tests of bioequivalence of control media and test media in studies of toxicity. Environmental Toxicology and Chemistry **14**: 1247–1256.

Gerard, P. D., D. R. Smith, and G. Weerakkody. 1998. Limits of retrospective power analysis. Journal of Wildlife Management **62**:801–807.

Hairston, N. G., Sr. 1996. Predation and competition in salamander communities. Pages 161–189 *in* M. L. Cody and J. A. Smallwood, editors. Long-term studies of vertebrate communities. Academic Press, New York, New York, USA.

Hauschke, D., V. W. Steinijans, and E. Diletti. 1990. A distribution-free procedure for the statistical analysis of bioequivalence studies. International Journal of Clinical Pharmacology, Therapy, and Toxicology **28**:72–78.

Hoenig, J. M., and D. M. Heisey. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. American Statistician **55**:19–24.

Houlahan, J. E., C. S. Findlay, B. R. Schmidt, A. H. Meyer, and S. L. Kuzmin. 2000. Quantitative evidence for global amphibian population declines. Nature **404**:752–755.

Hsu, J. C., J. T. G. Hwang, H-K. Liu, and S. J. Ruberg. 1994. Confidence intervals associated with tests for bioequivalence. Biometrika **81**:103–114.

IUCN—The World Conservation Union. 2001. IUCN red list categories and criteria, Version 3.1. IUCN, Gland, Switzerland.

Johnson, D. H. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management **63**:763–772.

Kenward, M. G., and J. H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics **53**:983–997.

MacKenzie, D. I., and W. L. Kendall. 2002. How should detection probability be incorporated into estimates of relative abundance. Ecology **83**:2387–2393.

McDonald, L. L., and W. P. Erickson. 1994. Testing for bioequivalence in field studies: has a disturbed site been adequately reclaimed. Pages 183–197 *in* D. Fletcher and B. F. Manly, editors. Statistics in ecology and environmental monitoring. University of Otago Press, Dunedin, New Zealand.

Mead, R. 1988. The design of experiments. Cambridge University Press, Cambridge, UK.

Millard, S. P. 1987. Proof of safety vs. proof of hazard. Biometrics **43**:719–725.

Parkhurst, D. F. 2001. Statistical significance tests: equivalence and reverse tests should reduce misinterpretation. BioScience **51**:1051–1057.

Perlman, M., and L. Wu. 1999. The emperor's new tests. Statistical Science **14**:355–369.

SAS Institute. 1999. Online documentation for SAS/Base and SAS/Stat, version 8.2. SAS Institute, Cary, North Carolina, USA.

Schabenberger, O., and F. J. Pierce. 2002. Contemporary statistical models for the plant and soil sciences. CRC Press, Boca Raton Florida, USA.

Schuirmann, D. J. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmokinetics and Biopharmaceutics **15**:657–680.

Semlitsch, R. D., D. E. Scott, J. H. K. Pechmann, and J. W. Gibbons. 1996. Structure and dynamics of an amphibian community: evidence from a 16-year study of a natural pond. Pages 217–248 *in* M. L. Cody and J. A. Smallwood, editors. Long-term studies of vertebrate communities. Academic Press, New York, New York, USA.

Thomas, L. 1997. Retrospective power analysis. Conservation Biology **11**:276–280.

Verbeke, G. 1997. Linear mixed models for longitudinal data. Pages 63–153 *in* G. Verbeke and G. Molenberghs, editors. Linear mixed models in practice: a SAS-oriented approach. Springer-Verlag, New York, New York, USA.

Wellek, S. 2003. Testing statistical hypotheses of equivalence. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.

Westlake, W. J. 1979. Statistical aspects of comparative bioavailability trials. Biometrics **35**:273–280.

SUPPLEMENT

SAS program code to estimate regression slopes and then test equivalence is available (along with *Ambystoma* data) in ESA's Electronic Data Archive: *Ecological Archives* E086-094-S1.