Ecological Applications, 21(1), 2011, pp. 281–289
© 2011 by the Ecological Society of America

Estimating equivalence with quantile regression

Brian S. Cade[1]

Fort Collins Science Center, U.S. Geological Survey, 2150 Centre Avenue, Building C, Fort Collins, Colorado 80526 USA

Abstract. Equivalence testing and corresponding confidence interval estimates are used to provide more enlightened statistical statements about parameter estimates by relating them to intervals of effect sizes deemed to be of scientific or practical importance rather than just to an effect size of zero. Equivalence tests and confidence interval estimates are based on a null hypothesis that a parameter estimate is either outside (inequivalence hypothesis) or inside (equivalence hypothesis) an equivalence region, depending on the question of interest and assignment of risk. The former approach, often referred to as bioequivalence testing, is often used in regulatory settings because it reverses the burden of proof compared to a standard test of significance, following a precautionary principle for environmental protection. Unfortunately, many applications of equivalence testing focus on establishing average equivalence by estimating differences in means of distributions that do not have homogeneous variances. I discuss how to compare equivalence across quantiles of distributions using confidence intervals on quantile regression estimates that detect differences in heterogeneous distributions missed by focusing on means. I used one-tailed confidence intervals based on inequivalence hypotheses in a two-group treatment-control design for estimating bioequivalence of arsenic concentrations in soils at an old ammunition testing site and bioequivalence of vegetation biomass at a reclaimed mining site. Two-tailed confidence intervals based both on inequivalence and equivalence hypotheses were used to examine quantile equivalence for negligible trends over time for a continuous exponential model of amphibian abundance.

Key words: amphibian decline; arsenic remediation; bioequivalence tests; confidence intervals; equivalence tests; mine land reclamation; negligible trends; quantile regression; quantile treatment effects; rank score tests.

Introduction

The concept of equivalence testing has been introduced to ecologists and biologists as a method for using common statistical procedures to provide more enlightened statements about parameter estimates in relation to effect sizes deemed to be of scientific importance rather than just to zero effect (McBride et al. 1993, McBride 1999, Parkhurst 2001, Brosi and Biber 2009). An equivalence region that is an interval of nonzero effects sizes that are considered indistinguishable for the process being investigated is specified a priori. Equivalence tests and associated confidence interval estimates can be based on a null hypothesis that a parameter estimate is either outside (inequivalence hypothesis) or inside (equivalence hypothesis) an equivalence region, depending on the question of interest and desired assignment of risk (McBride 1999, McBride and Ellis 2001, Camp et al. 2008). The equivalence testing framework that uses a null hypothesis of inequivalence, often referred to as bioequivalence testing, reverses the burden of proof from conventional hypothesis tests when sample sizes are small such that small probability given the null hypothesis ($P \mid H_0$) implies accepting the alternative hypothesis of equivalence (Erickson and McDonald 1995, Manly 2004). This approach minimizes risk (type I error under H_0) in claiming equivalence when it is not true (often termed consumer's risk) and often is used when the equivalence region is a more desirable environmental state because it invokes a precautionary principle (McBride 1999, McBride and Ellis 2001, Parkhurst 2001, Manly 2004). An alternative approach that is less commonly employed posits a null hypothesis of equivalence such that small probability given the null hypothesis ($P \mid H_0$) implies accepting the alternative hypothesis of inequivalence (McBride 1999, McBride and Ellis 2001, Camp et al. 2008). This approach minimizes risk in claiming inequivalence when it is not true (often termed producer's risk), invoking a benefit of doubt principle (McBride and Ellis 2001).

The concept of equivalence testing has been applied to practical problems of regulatory evaluation of environmental damage and remediation in a bioequivalence formulation (Erickson and McDonald 1995, Manly 2004), for evaluating negligible trend in animal population growth (Dixon and Pechmann 2005, Camp et al. 2008), evaluating assumptions of equal detectability in animal sampling (MacKenzie and Kendall

Manuscript received 4 November 2009; revised 18 March 2010; accepted 19 March 2010. Corresponding Editor: J. Van Buskirk.

[1] E-mail: cadeb@usgs.gov

2002), and for model validation (Robinson and Froese 2004). All these examples were based on the concept of detecting average equivalence, where tests and estimates were functions of means. A common problem with testing average equivalence as with other statistical comparisons is that data distributions often violate assumptions of Normality (a weak assumption) and homogeneous variances (a stronger assumption) associated with tests of differences in means (Manly 2004, Gruman et al. 2007). Manly (2004) discussed some alternatives for one-sided tests of equivalence for nonnormal distributions with heterogeneous variances. He also developed an approach for comparing means of a selected subset of the data (e.g., those observations ≥median in the treatment group) because comparisons of means for the entire sample were not completely informative for establishing equivalence when heterogeneous variances were identified. Here, I discuss a more general approach to solving the equivalence testing issue identified by Manly (2004) based on estimating quantile treatment effects (Doksum 1974, Lehmann 1975) with quantile regression (Koenker 2005). Furthermore, quantile regression easily extends the concept of quantile treatment effects for two categorical groups to more complicated designs including those with continuous predictors.

Anytime there is heterogeneity in variances across any predictor in a linear model, then all quantiles of the response distribution cannot change at the same rate and summarizing them by the mean rate can miss important information (Cade and Noon 2003, Koenker 2005). This is true whether the predictors are indicator variables for categorical groupings, as in the treatment-control design used by Manly (2004), or continuous measures as used by Dixon and Pechmann (2005) for trends in animal abundance over time. Quantile regression provides a comprehensive approach for estimating and comparing quantiles of data distributions in a linear model without making unnecessary parametric distributional assumptions and has a well developed arsenal of procedures for hypothesis testing and constructing confidence intervals (Koenker 2005). The quantile regression approach to evaluating equivalence is related to ideas embodied in testing population equivalence (Food and Drug Administration 2001, Chow et al. 2003) rather than average equivalence. Unlike other approaches to testing population equivalence, the quantile regression comparisons are based on all or a selected subset of quantiles from conditional cumulative distributions rather than based on simultaneous differences in the mean and variance of a normal distribution.

While much of the literature discusses establishing equivalence in a hypothesis testing framework, the use of appropriate one- or two-tailed confidence intervals greatly facilitates implementing, understanding, and interpreting the evidence for establishing equivalence (McBride 1999, Dixon and Pechmann 2005, Limentani et al. 2005, Brosi and Biber 2009). Most equivalence testing is presented in the context of establishing average equivalence where tests or confidence intervals are based on differences in means. A common framework for establishing average equivalence based on a null hypothesis of inequivalence uses the two one-sided test (TOST) approach on differences in means (Schuirmann 1987). The TOST approach requires $P \leq \alpha$ for both tests; e.g., H_{0l}: $\mu_T - \mu_C \geq \Delta_U$ and H_{0l}: $\mu_T - \mu_C \leq \Delta_L$, where μ_T and μ_C are means of treated and control sites, to establish that the estimated differences are within two-sided equivalence regions defined by $[\Delta_L, \Delta_U]$, where Δ_L and Δ_U are the lower and upper bounds, respectively. Equivalently, a $100\% \times (1 - 2\alpha)$ two-tailed confidence interval on the parameter that is completely contained within the two-sided equivalence region also establishes equivalence. A two-tailed confidence interval that is outside the two-sided equivalence region at all (this is the precautionary principle) results in a determination of no equivalence. An alternative framework for establishing average equivalence based on a null hypothesis of equivalence uses two one-sided tests on differences in means (McBride 1999), requiring $P \leq \alpha$ for either test, e.g., H_{0E}: $\mu_T - \mu_C < \Delta_U$ and H_{0E}: $\mu_T - \mu_C > \Delta_L$, to establish that estimates are outside the equivalence region. A $100\% \times (1 - 2\alpha)$ two-tailed confidence interval on the parameter that is completely outside the two-sided equivalence region results in a determination of no equivalence, identical to McBride's (1999) interpretation based on lower or upper $100\% \times (1 - \alpha)$ one-tailed confidence intervals. If the confidence interval overlaps the equivalence region at all then a determination of equivalence is made. Similar relationships apply to one-sided equivalence regions (e.g., $[-\infty, \Delta_U]$, where Δ_U is the upper endpoint of the equivalence region and $-\infty$ implies no lower bound), $100\% \times (1 - \alpha)$ one-tailed confidence intervals, and one-sided null hypotheses of inequivalence or equivalence (McBride and Ellis 2001).

Here, I present the use of one-tailed confidence intervals of quantile regression estimates for a two-group design for establishing equivalence for remediation of soils contaminated with arsenic and reclamation of mine land vegetation from Manly (2004). The two-group examples also serve as a simple heuristic device to demonstrate the relationship between differences in cumulative distributions and their inverse the quantiles and estimating those differences with categorical predictors in a linear model, expanding on the concepts presented for continuous predictors in Cade and Noon (2003). I examine negligible trend for an exponential regression model of temporal changes in amphibian populations from Dixon and Pechmann (2005) using two-tailed confidence intervals of quantile estimates for a continuous predictor. I also explore alternative one-sided equivalence regions for species vulnerability listing criteria and the use of different one-tailed confidence intervals that are associated with minimizing different risks for the amphibian trend example.

Bioequivalence in Treatment–Control Examples

Arsenic contamination in soils

The one-sided equivalence examples in a two-group, treatment-control design used by Manly (2004) included an investigation of remediation of soils that were contaminated with arsenic at a former ammunition testing site. It was assumed that expensive and potentially environmentally destructive remediation wouldn't be required if the arsenic contaminated site (treated) had concentrations <125% of an uncontaminated reference site (control). The nature of the multiplicative comparison required using natural logarithms (ln) of arsenic concentrations (mg/kg) in a linear model with treated and control groups. Quantile plots of the cumulative distributions of ln arsenic concentrations and summary statistics indicated that the majority of the distribution of arsenic concentrations from the ammunition testing site was less than in the reference site (Fig. 1A). However, a small proportion of the sample (four of $n = 32$) from the ammunition testing site had arsenic concentrations exceeding any levels measured in the reference site ($n = 6$), leading to lower mean arsenic concentration but higher variances in the treated compared to control groups.

The simple test for average equivalence for the arsenic contamination used by Manly (2004) posits a one-sided null hypothesis of inequivalence, H_{0I}: $\mu_T \geq 1.25 \times \mu_C$, where μ_T and μ_C are the means of the treated and control sites, respectively. The one-sided alternative hypothesis of equivalence is H_{1I}: $\mu_T < 1.25 \times \mu_C$. Equivalently in the linear form with logarithmically transformed concentrations, the null and alternative hypotheses are H_{0I}: $\ln(\mu_T) \geq \ln(1.25) + \ln(\mu_C)$ and H_{1I}: $\ln(\mu_T) < \ln(1.25) + \ln(\mu_C)$. One of the tests suitable for this comparison of means with heterogeneous variances is Welch's t test on the logarithmically transformed concentrations with ln(1.25) added to each of the control observations (Manly 2004). This test yielded a one-sided $P < 0.001$ and an upper one-tailed 95% confidence interval $[-\infty, -0.202]$ around the estimated difference in means of -0.426. Exponentiating this confidence interval to return it to the multiplicative scale provides an interval of [0.000, 0.817], establishing average equivalence since the one-tailed interval is contained within the equivalence region defined by 0.00 and $\Delta_U = 1.25$. But how do we reconcile this test with those few observations of arsenic concentrations greatly exceeding those in the control site?

Manly's (2004) approach was to use only a subset of the observations in the treated group that exceeded the median and then perform the test for average equivalence again. A simpler approach that has a well developed statistical theory recognizes that interest in subsets of the data distribution differences is tantamount to asking about quantile treatment effects (Koenker 2005). This approach avoids the arbitrariness of picking a single quantile as a dividing point for the data and consequent reduction in sample size associated with Manly's (2004) approach. The quantile treatment effect between a treated and control group is based on the horizontal distance $\Delta(y)$ between their cumulative distributions (Doksum 1974, Lehmann 1975, Koenker 2005). Thus, if $F_T(y)$ and $F_C(y)$ are the cumulative distribution functions for the treated and control groups, respectively, then $F_C(y) = F_T(y + \Delta(y))$, $\Delta(y) =$

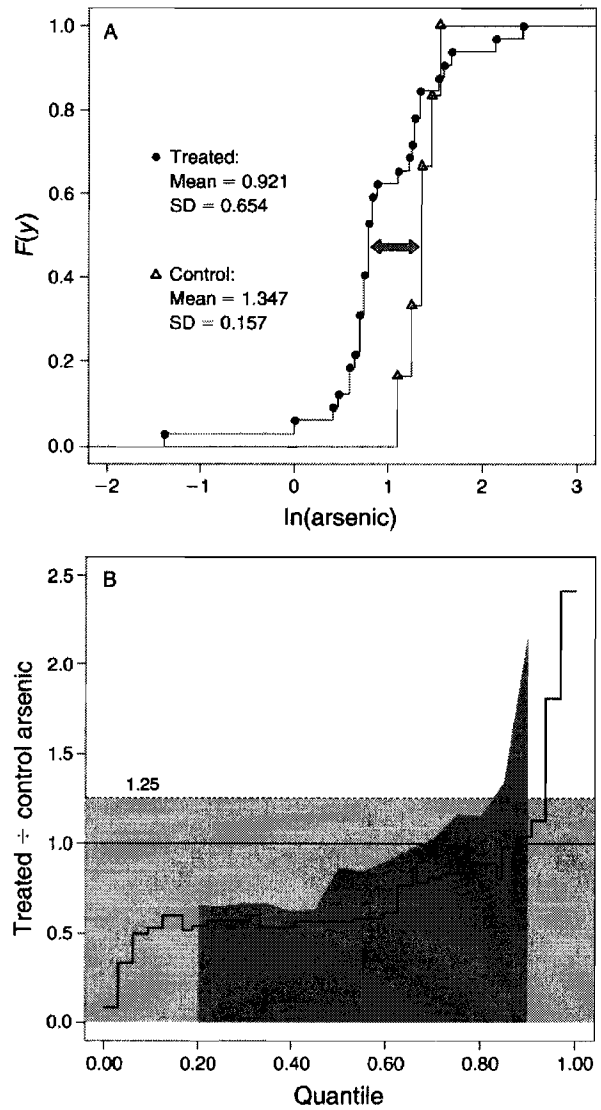

FIG. 1. (A) Cumulative distributions of ln(arsenic) (measured as mg/kg) in $n = 32$ locations in an ammunition testing site (treated) and $n = 6$ locations in a reference site (control); data are from Manly (2004). Quantile treatment effects are denoted by horizontal distances represented by gray arrow. (B) Estimated multiplicative quantile treatment effects (solid step function) made by exponentiating additive estimates in Fig. 2B, upper one-tailed 95% confidence interval (dark gray area), and one-sided equivalence region defined by 0.0 and upper bound $\Delta_U = 1.25$ (light gray area). The confidence intervals for 0.20 to 0.80 quantiles are within the equivalence region, but the upper 15th percentiles are outside the equivalence region and exceed the arsenic concentration standard.

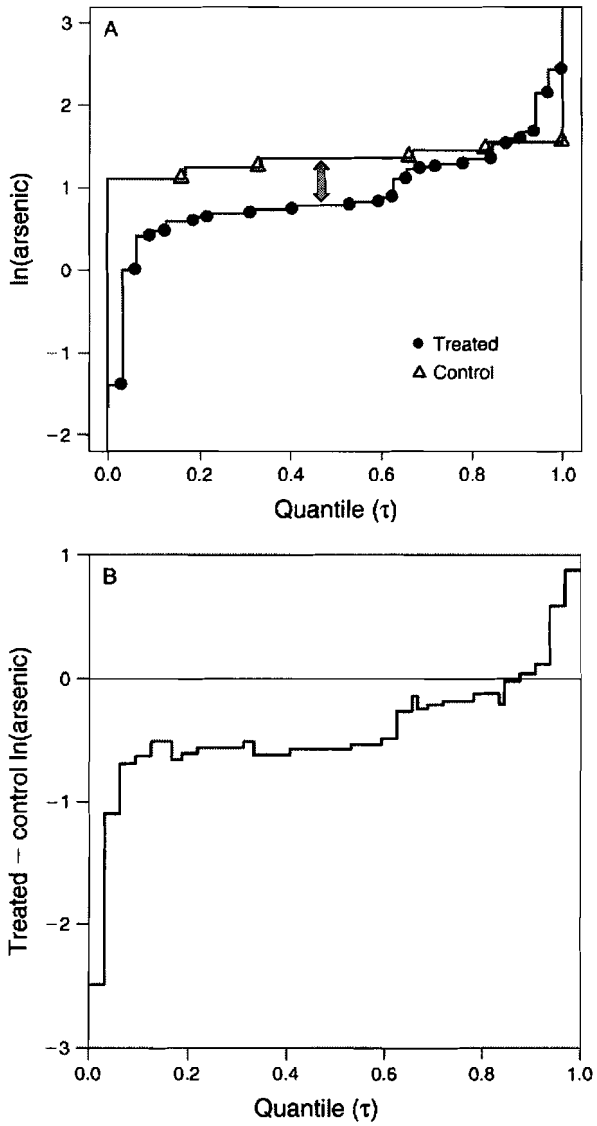

FIG. 2. (A) Additive quantile treatment effects from transposing horizontal distances (gray arrow) between treated and control cumulative distribution functions in Fig. 1B, and (B) estimated additive quantile treatment effect from quantile regression. Arsenic data are from Manly (2004).

$F_T^{-1}(F_C(y)) - y$, and the quantile treatment effect $\delta(\tau) = \Delta(F_C^{-1}(\tau)) = F_T^{-1}(\tau) - F_C^{-1}(\tau)$, for τ in the interval [0, 1] and $\tau = F_C(y)$ (Fig. 1A). If the cumulative distribution functions and quantile treatment effect plots are transposed so that $\tau = F_C(y)$ is on the horizontal axis as is conventionally done for quantile regression estimates (Fig. 2A), it is easy to visualize that the two-group quantile treatment effects (Fig. 2B) are just the estimates for the parameters $\beta_1(\tau)$ in the linear quantile regression model $Q_Y(\tau \mid X) = \beta_0(\tau) + \beta_1(\tau)X$, where X is an indicator variable taking the values 0 for control and 1 for treated groups (Koenker 2005). The estimates for the parameters $\beta_0(\tau)$ are the quantiles for the control group. If the treated and control groups differed only in their means, then the plot of the estimated quantile treatment effects would vary randomly around a constant (flat horizontal line) with respect to the quantiles τ, something that clearly is not true for the log transformed arsenic concentrations (Fig. 2B).

To perform an equivalence test for the quantile treatment effects similar to that done for average equivalence, I specified the null hypothesis of inequivalence for the quantile treatment effect as H_{01}: $Q_Y(\tau \mid X = 1) \geq 1.25 \times Q_Y(\tau \mid X = 0)$ and the alternative hypothesis of equivalence as H_{11}: $Q_Y(\tau \mid X = 1) < 1.25 \times Q_Y(\tau \mid X = 0)$, where $X = 1$ for treated and $X = 0$ for control groups. The quantile treatment effect model was estimated in its linear form by taking logarithms of arsenic concentrations. Quantile estimates (and confidence intervals) were readily exponentiated back to the multiplicative scale (Fig. 2B to 1B) because quantiles are equivariant to nonlinear, monotonic transformations such as the logarithmic (Koenker 2005). I used a weighted quantile rank score test inversion method for estimating two-tailed 90% confidence intervals for the quantile regression estimates (Koenker 2005, Cade et al. 2006). Weights used were the inverse of the group sample standard deviations, although the default quantile bandwidth weights provided in the "quantreg" package for R (Koenker 2005) yielded very similar results. By using only the upper tail of the two-tailed 90% confidence interval, I obtained an upper one-tailed 95% confidence interval. Confidence intervals were estimated by increments of 0.05 from 0.20 to 0.90 quantiles (Fig. 1B). Code to estimate the quantile regression models, confidence intervals, and plot results using the "quantreg" package in R are provided for this (Supplement 1) and the following examples (Supplements 2 and 3).

The upper one-tailed confidence intervals for the quantile treatment effect indicated equivalence for the lower 80th percentiles because the intervals were contained within the equivalence region (Fig. 1B). However, the intervals for the 85th and 90th percentiles were outside the equivalence region, indicating that the upper 15th percentiles were outside of the equivalence region. The quantile treatment effects provided a stronger statement of inequivalence for more extreme arsenic concentrations than made by Manly (2004) and a more focused interpretation of where differences occurred in the distributions. The average equivalence test for the treated arsenic values exceeding the median used by Manly (2004) had $P = 0.066$ for the null hypothesis of inequivalence for this subset of data. But this analysis provided no information on where the differences occurred in the distribution of arsenic values > median.

Because of the relatively small sample sizes in this two-group comparison (treated $n = 32$, control $n = 6$), it was not possible to get reliable confidence intervals for $\tau < 0.20$ or $\tau > 0.90$. Although $n = 6$ for the control group is small, the differences in cumulative distributions

modeled by $\beta_1(\tau)$ in the quantile regression model actually results in 25 differences (the flat horizontal increments in Fig. 1B), making the problem less discrete than it might first appear. Excessive discreteness due to many tied values associated with smaller samples leads to unreliable inferences for quantile regression estimates. Some guidance on sample size requirements relative to quantiles and number of model parameters for the quantile rank score tests are given in Cade et al. (2006). Increased sample sizes would reduce the length of the confidence intervals for the quantile effects (Fig. 1B), suggesting that with larger samples sizes we might find <10% of the site exceeds the remediation goals.

Mine land reclamation

A second example provided by Manly (2004) considered whether vegetation production (g/m^2) on a mine reclamation site was sufficient to be considered restored for regulatory purposes. Here vegetation production on the reclamation site was considered restored if it was greater than 90% of an adjacent reference (control) site (Fig. 3). So the one-sided null hypothesis of average inequivalence was H_{0I}: $\mu_T \leq 0.90 \times \mu_C$, where μ_T and μ_C are the means of the treated and control sites, respectively. The one-sided alternative hypothesis of equivalence is H_{1I}: $\mu_T > 0.90 \times \mu_C$. My corresponding one-sided hypotheses for the quantile treatment effects were H_{0I}: $Q_Y(\tau \mid X = 1) \leq 0.90 \times Q_Y(\tau \mid X = 0)$ and H_{1I}: $Q_Y(\tau \mid X = 1) > 0.90 \times Q_Y(\tau \mid X = 0)$, again where $X = 1$ or 0 for treated and control groups, respectively. Again, estimates and confidence intervals were made in the linear scale by using logarithms of vegetation biomass and then exponentiated back to the multiplicative scale. Establishing equivalence in this example requires lower one-tailed 95% confidence intervals, either obtained by Welch's t-test on the means or the weighted quantile rank-score test for quantile treatment effects. Confidence intervals for the quantiles were obtained by increments of 0.05 from 0.10 to 0.90, based on estimates weighted by the inverse of the group standard deviations and the rank score test inversion.

The lower one-tailed 95% confidence interval on the difference in means of −0.034 was [−0.278, +∞] in the additive scale on the log transformed data, which corresponds to an interval of [0.757, +∞] in the multiplicative scale. This confidence interval for average equivalence extends outside the equivalence region defined by $\Delta_L = 0.90$ and +∞, indicating no evidence of average equivalence, consistent with the one-sided Welch's t-test with $P = 0.315$ for the null hypothesis of inequivalence (Manly 2004). The lower one-tailed 95% confidence intervals for the quantile treatment effects all extended outside the equivalence region except at the 90th percentile (Fig. 3B). The trend in the quantile treatment effects and associated confidence intervals indicated a lack of equivalence for the majority of the biomass distribution but approaching or slightly exceeding equivalence in the upper 20th percentiles. Even if the length of the confidence intervals were reduced by having larger sample sizes, the estimated quantile effects (Fig. 3B) suggest that a large proportion (lower 40th percentiles) of the site has vegetation production that is unlikely to meet the restoration standard.

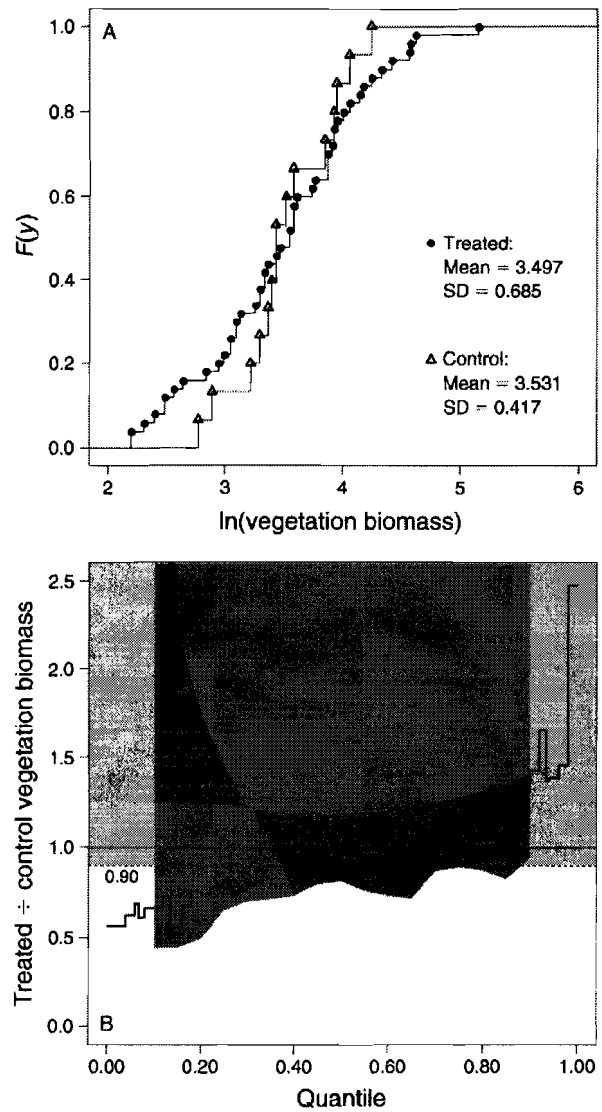

FIG. 3. (A) Cumulative distributions of ln(vegetation production) (measured as g/m^2) in $n = 50$ locations in a formerly mined site (treated) and $n = 15$ locations in a reference site (control); data are from Manly (2004). (B) Estimated multiplicative quantile treatment effects (solid step function), lower one-tailed 95% confidence interval (dark gray band), and one-sided equivalence region defined by lower bound $\Delta_L = 0.90$ and ∞ (light gray area). Confidence intervals for 0.10 to 0.85 quantiles are outside the equivalence region, but the upper 10th percentiles are within the equivalence region and have met the revegetation standard for biomass.

TRENDS IN AMPHIBIAN ABUNDANCE

Two-sided equivalence regions and intervals

Exponential trends over time in four amphibian species were used as examples by Dixon and

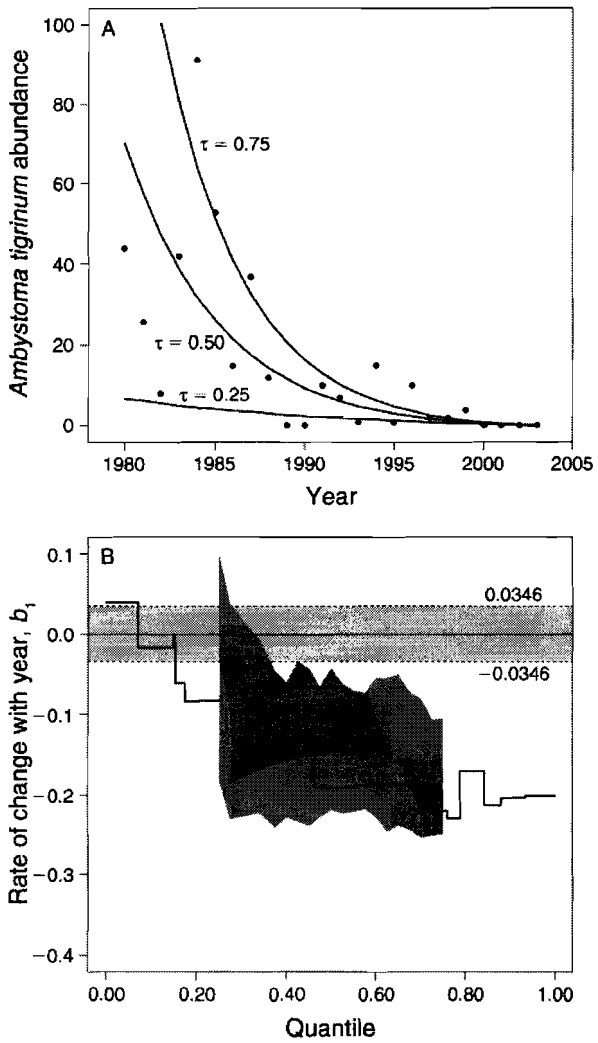

FIG. 4. (A) Trends in number of breeding female *Ambystoma tigrinum* over 24 years (data are from Dixon and Pechmann [2005]) and estimated 0.75, 0.50, and 0.25 quantile regressions for a log-linear model with a first-order autoregressive term set at an average prior year abundance of 16.5. (B) Estimated quantiles of temporal trend parameter β_1 (solid step function), two-tailed 90% confidence intervals (dark gray band), and two-sided equivalence region defined by $\Delta_L = -0.0346$ and $\Delta_U = 0.0346$ (light gray area). Intervals for 0.35 to 0.75 quantiles are outside the equivalence region indicating that upper 65th percentiles of trend were not negligible.

Pechmann (2005) to demonstrate the use of equivalence testing for establishing whether or not temporal trends in abundance were negligible. Here, I used data for one species, *Ambystoma tigrinum*, to demonstrate how to obtain comparable information from quantile regression using two-tailed confidence intervals (Fig. 4A). I initially used the two-sided equivalence region defined by Dixon and Pechmann (2005) as $\Delta_L = -0.0346$ and $\Delta_U = 0.0346$, which was considered a negligible trend where population size would not halve or double in <20 years. The exponential model of change in abundance (N) over time (t) for the quantiles of abundance was estimated in its linearized form, $Q_{\ln(N_t+1)}(\tau \mid t, N_{t-1}) = \beta_0(\tau) + \beta_1(\tau)t + \beta_2(\tau)N_{t-1}$, by adding 1 to abundances before taking logarithms and including abundances in the prior year to account for the first-order autocorrelation of the serial measures of abundance. Dixon and Pechmann (2005) used a similar model for mean abundances, except that they used a mixed model approach to incorporate the first-order autoregressive process in the error term. The inclusion of the first-order autoregressive term [AR(1)] as a parameter in the quantile regression model was done for comparable reasons as the mixed model approach with the AR(1) term for the mean regression model: it increased the sampling variation of the estimated temporal trend $\beta_1(\tau)$ by accounting for the serial dependency of abundances. I estimated $100\% \times (1 - 2\alpha) = 90\%$ two-tailed confidence intervals for 0.25 to 0.75 quantiles by increments of 0.025 using the quantile rank-score test inversion approach with the default quantile bandwidth weighting to account for heterogeneity in slopes. Lower and higher quantiles had unreliable 90% confidence intervals in this model with three estimated parameters and $n = 24$.

My model of abundance over time for *Ambystoma tigrinum* had modest heterogeneity in the temporal trend parameter, varying around −0.10 for lower 45th percentiles and varying around −0.20 for upper 55th percentiles (Fig. 4B). The two-tailed confidence intervals overlapped the equivalence region for quantiles ≤0.35, providing inconclusive evidence about whether trends in lowest quantiles were negligible or not which corresponds to inconsistent conclusions from null hypotheses of equivalence and inequivalence (McBride 1999, Camp et al. 2008). For quantiles >0.35, the intervals for the temporal trend were all outside the equivalence zone indicating they were not negligible. The 90% confidence interval for the mean regression model used by Dixon and Pechmann (2005) was [−0.23, −0.09], clearly outside the two-sided equivalence region. The 90% confidence interval estimated from a mean regression model corresponding to my quantile regression model with the first-order autoregressive term was [−0.20, −0.06], also clearly outside the equivalence region. The mean regression estimates can be regarded as an average across all the quantile estimates. Decomposing the regression model into quantiles demonstrated that heterogeneity in temporal trends led to higher quantiles of abundance decreasing more rapidly than the lower quantiles. Negligible trends in the lower quantiles of abundance are to be expected when many of the annual abundances approach or are zero as in the *Ambystoma tigrinum* example.

One-sided equivalence regions and intervals

In their discussion, Dixon and Pechmann (2005) considered other alternatives to setting equivalence regions by referencing IUCN, The World Conservation Union (2001) categories of threatened or endangered species. A decline in abundance >50% in

10 years defined an "endangered" species ($\Delta_L = -0.0693$), and a decline in abundance of 30% in 10 years defined a "vulnerable" species ($\Delta_L = -0.0346$). Unlike Dixon and Pechmann (2005) though, I believe that an analysis with respect to equivalence regions for defining "not vulnerable" or "not endangered" categories requires one-sided equivalence regions and lower one-tailed $100\% \times (1 - \alpha) = 95\%$ confidence intervals. An equivalence region corresponding to a statement of "not vulnerable" ("not endangered") would include all temporal trends exceeding −0.0346 (−0.0693), not just those in the interval [−0.0346, 0.0346] ([−0.0693, 0.0693]). Because the lower one-tailed 95% confidence intervals for all the 0.25 to 0.75 quantiles were outside one-sided equivalence regions for either "not vulnerable" ($\Delta_L = -0.0346$, $\Delta_U = +\infty$) or "not endangered" ($\Delta_L = -0.0693$, $\Delta_U = +\infty$) designations, it is reasonable to conclude that *Ambystoma tigrinum* was "endangered" (Fig. 5A) as there was no evidence in support of "not endangered."

The lower one-tailed confidence intervals for the previous approach (Fig. 5A) are associated with a one-sided null hypothesis test of inequivalence, exemplifying a precautionary approach to species protection (McBride and Ellis 2001) because it minimizes the risk of declaring the species "not vulnerable" ("not endangered") when that is not true. An alternative approach would be to use upper one-tailed confidence intervals associate with a one-sided null hypothesis of equivalence (Fig. 5B), exemplifying a benefit of doubt approach (McBride and Ellis 2001) because it minimizes the risk of declaring the species "vulnerable" ("endangered") when that is not true. The combined results of these two approaches were inconsistent only for the quantiles $\leq$ 0.35, consistent with the previous interpretation of the two-tailed confidence intervals.

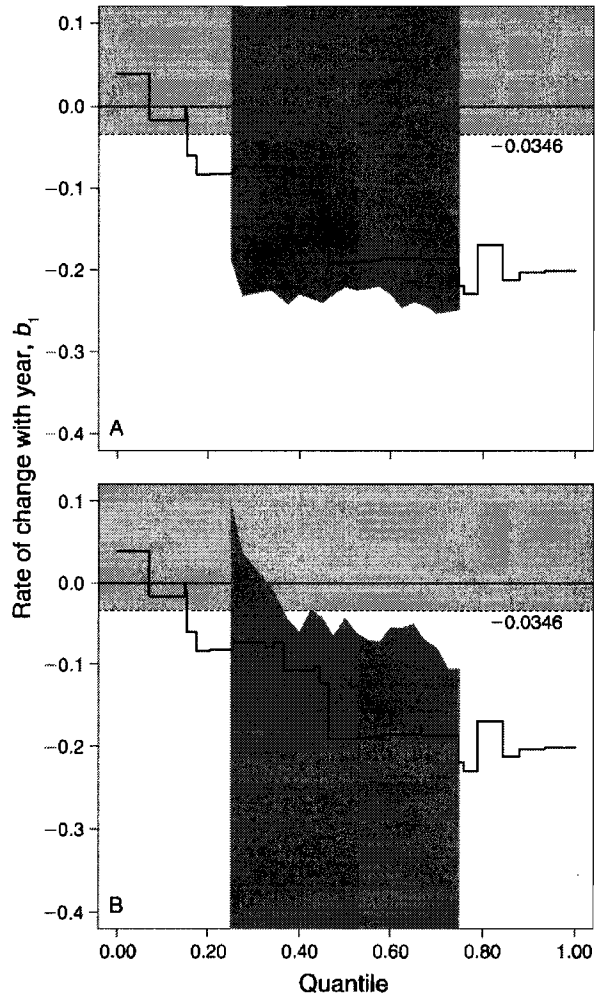

FIG. 5. (A) Estimated quantiles of temporal trend parameter β_1 (solid step function) for a log-linear model of breeding female *Ambystoma tigrinum* abundance over 24 years as in Fig. 4, but with lower one-tailed 95% confidence intervals (dark gray band), and one-sided equivalence region defined by $\Delta_L = -0.0346$ and ∞ (light gray area). Interval endpoints for 0.25 to 0.75 percentiles are outside the one-sided equivalence region of negligible decreasing to increasing trends indicating the estimated trends support "vulnerable" status under a precautionary principle. (B) As above but now with upper one-tailed 95% confidence intervals (dark gray band), indicating only trends for quantiles >0.35 support "vulnerable" status under a benefit of doubt principle.

DISCUSSION

It is most convenient that equivalence can be evaluated by simply estimating one- or two-tailed confidence intervals as appropriate for the problem, and then referencing them with respect to the corresponding one- or two-sided equivalence regions. This is as true for conventional models of means in the linear model as it is for quantiles. The quantile estimates and confidence intervals simply provide a more comprehensive view of where equivalence is or is not occurring for distributional changes that are not homogeneous. Although my preference is to perform these evaluations of quantile equivalence with confidence intervals, it is certainly possible to obtain *P*-values from the rank score tests of the null hypotheses of inequivalence or equivalence by quantile (Koenker 2005, Cade et al. 2006). The use of confidence intervals, however, provides direct graphical interpretation and helps identify regions where the null hypotheses of inequivalence and equivalence provide inconsistent, inconclusive results. A two-tailed confidence interval completely contained within or completely outside the equivalence region provides unambiguous determination of either equivalence or inequivalence, respectively. Therefore, a reasonable interpretation of a two-tailed confidence interval that overlaps the equivalence and inequivalence regions is that a determination is inconclusive (Limentani et al. 2005, Brosi and Biber 2009). An inconclusive interpretation based on confidence intervals is consistent with results where testing both the inequivalence and equivalence null hypotheses yields conflicting determinations (McBride 1999, Camp et al. 2008). An inconclusive finding suggests the need for

larger sample sizes, improved models, and so on, to improve the precision of the estimates to allow an unambiguous finding. However, an inconclusive finding may be difficult to eliminate for estimates that approach one of the boundaries of the equivalence region (McBride 1999).

The only other example I'm aware of for establishing quantile equivalence for differences between groups relied on assuming a Normal distribution of the responses (Pei and Hughes 2008), an assumption that will usually be violated in practice and is not required by quantile regression (Cade and Noon 2003, Koenker 2005). Although the examples presented here involved fairly simple designs, quantile regression can be used with more complicated multi-factorial designs or with multiple continuous covariates similar to any other linear model. For those preferring a Bayesian approach to estimating quantile equivalence, it also is possible to estimate quantile regression and associated credible intervals in a Bayesian context (Yu and Moyeed 2001). McBride and Ellis (2001) describe the use of one-tailed confidence intervals based on a binomial distribution that can be used for establishing quantile equivalence in one-sample problems either in a frequentist or Bayesian approach.

The choice of one- or two-tailed confidence intervals, one- or two-sided equivalence regions, and whether risk should be minimized in detecting equivalence or inequivalence is not always obvious. The use of two-tailed vs. one-tailed confidence intervals and equivalence regions has more flexibility in basic scientific investigations depending on the intentions of the investigators, whereas their use in environmental regulatory settings will often be dictated by current accepted standards. In the environmental regulatory examples I used from Manly (2004), the one-sided equivalence regions were defined based on regulatory policies defining acceptable versus unacceptable environmental conditions and there would seem to be little argument for anything but one-tailed intervals minimizing the risk of declaring compliance (equivalence) when that was not true (Lombardi and Hurlbert 2009). However, the amphibian abundance example from Dixon and Pechmann (2005) is a case where arguments could be made for either or both one- or two-tailed intervals, one- or two-sided equivalence regions, and minimizing risk in either a precautionary or benefit of doubt principle depending on your objectives.

If your objective is to establish that the estimated temporal trends are within or outside an equivalence region that defines negligible changes in abundance over time, as originally formulated by Dixon and Pechmann (2005), then two-tailed confidence intervals and equivalence regions are appropriate. These can be interpreted both with respect to null hypotheses of inequivalence (minimizing risk in declaring trends negligible) and equivalence (minimizing risk in declaring trends not negligible). However, if your objective is to determine whether the trend is less than some negligible negative trend over time, as implied by the discussion in Dixon and Pechmann (2005) relative to IUCN "vulnerable" and "endangered" categories, and you want to minimize the risk of declaring positive trends when that is not true, then it would be more appropriate to use lower one-tailed confidence intervals associated with an inequivalence null hypothesis and a one-sided equivalence region encompassing trends greater than a negligible decrease. For a listing decision of "vulnerable" or "endangered" this precautionary approach avoids wasting small sample sizes on bounding sampling variation in the direction of positive trends that can only be consistent with temporal trends indicating "not endangered" or "not vulnerable." Thus, we focus on the sampling variation in the direction of negative temporal trends, e.g., the lower one-tailed confidence intervals of the temporal trends of 25th to 75th percentiles of *Ambystoma tigrinum* abundance that were less than −0.0693. In the *Ambystoma tigrinum* example the lower one-tailed intervals provided enough precision to be confident that trends could be less than the "endangered" category. However, the two-tailed confidence intervals indicated a lack of precision for determining precisely how negative the trends were in the lower quantiles. Inconsistencies of this nature have caused some frustration with the use of one-tailed tests and confidence intervals (Lombardi and Hurlbert 2009).

The overlap of the equivalence region by the upper tails of the $100\% \times (1 - 2\alpha)$ two-tailed intervals (Fig. 4B) correspond to the same overlap indicated by upper $100\% \times (1 - \alpha)$ one-tailed intervals associated with a one-sided null hypothesis of equivalence (Fig. 5B). We might prefer this latter one-tailed confidence interval for trends in abundance if we wanted to minimize risk in declaring them declining when that was not true. This interpretation and risk minimization is the antithesis of a precautionary principle for listing a species as "vulnerable" or "endangered." But as a benefit of doubt principle it might be appropriate if deciding whether a listed species exhibited a sufficiently precise decline such that we might reasonably anticipate detecting a positive response to environmental policies and management alternatives designed to increase its abundance. We also might prefer to minimize risk in this latter fashion if evaluating a negative trend in a "noxious" species that was being controlled to reduce its abundance.

An issue that is evident in my evaluation of quantile equivalence is defining the equivalence regions. I have used rectangular regions with constant boundaries (horizontal lines) across quantiles because they correspond to regions consistent with the comparisons for average equivalence that assumed homogeneity of effects. However, quantile regression provides a method for quantifying heterogeneous effects so it would be reasonable to consider equivalence regions that don't have constant boundary values across quantiles. For example, in Fig. 1 you might envision a nonlinear equivalence region boundary where Δ_U decreased from

1.25 at lower quantiles to 1.05 at higher quantiles. This would allow higher concentrations of arsenic in the treated site to differ less from highest concentrations in the reference control site compared to allowable differences in lower concentrations. As with any definition of equivalence regions, defining appropriate boundary values (constant or varying across quantiles) will require substantial deliberation by subject matter specialists.

ACKNOWLEDGMENTS

The manuscript benefited from reviews by P. M. Dixon, B. F. J. Manly, and L. L. McDonald, and two anonymous referees. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

LITERATURE CITED

Brosi, B. J., and E. G. Biber. 2009. Statistical inference, Type II error, and decision making under the US Endangered Species Act. Frontiers in Ecology and the Environment 7:487–494.

Cade, B. S., and B. R. Noon. 2003. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment 1:412–420.

Cade, B. S., J. D. Richards, and P. W. Mielke, Jr. 2006. Rank score and permutation testing alternatives for regression quantile estimates. Journal of Statistical Computation and Simulation 76:331–355.

Camp, R. J., N. E. Seavy, P. M. Gorresen, and M. H. Reynolds. 2008. A statistical test to show negligible trend: comment. Ecology 89:1469–1472.

Chow, S.-C., J. Shao, and H. Wang. 2003. Statistical tests for population bioequivalence. Statistica Sinica 13:539–554.

Dixon, P. M., and J. H. K. Pechmann. 2005. A statistical test to show negligible trend. Ecology 86:1751–1756.

Doksum, K. 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. Annals of Statistics 2:267–277.

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.

Food and Drug Administration. 2001. Guidance for industry: statistical approaches to establishing bioequivalence. Center for Drug Evaluation and Research, FDA, Rockville, Maryland, USA.

Gruman, J. A., R. A. Cribbie, and C. A. Arpin-Cribbie. 2007. The effects of heteroscedasticity on tests of equivalence. Journal of Modern Applied Statistical Methods 6:133–140.

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

Koenker, R. 2005. Quantile regression. Econometric Society Monographs. Cambridge University Press, New York, New York, USA.

Lehmann, E. 1975. Nonparametrics: statistical methods based on ranks. Holden-Day, Oakland, California, USA.

Limentani, G. B., M. C. Ringo, F. Ye, M. L. Bergquist, and E. O. McSorley. 2005. Beyond the *t*-test: statistical equivalence testing. Analytical Chemistry, June 1 2005:221–226.

Lombardi, C. M., and S. H. Hurlbert. 2009. Misprescription and misuse of one-tailed tests. Austral Ecology 34:447–468.

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

Manly, B. F. J. 2004. One-sided tests of bioequivalence with nonnormal distributions and unequal variance. Journal of Agricultural, Biological, and Environmental Statistics 9:270–283.

McBride, G. B. 1999. Equivalence tests can enhance environmental science and management. Australian and New Zealand Journal of Statistics 41:19–29.

McBride, G. B., and J. C. Ellis. 2001. Confidence of compliance: a Bayesian approach for percentile standards. Water Research 35:1117–1124.

McBride, G. B., J. C. Loftis, and N. C. Adkins. 1993. What do significance tests really tell us about the environment? Environmental Management 17:423–432.

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

Pei, L., and M. D. Hughes. 2008. A statistical framework for quantile equivalence clinical trials with application o pharmacokinetic studies that bridge from HIV-infected adults to children. Biometrics 64:1117–1125.

Robinson, A. P., and R. E. Froese. 2004. Model validation using equivalence tests. Ecological Modelling 176:349–358.

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

Yu, K., and R. A. Moyeed. 2001. Bayesian quantile regression. Statistics and Probability Letters 54:437–447.

SUPPLEMENT 1

Scripts and data for estimating quantile equivalence of the arsenic contamination data (Manly 2004) with the quantreg package in R (*Ecological Archives* A021-013-S1).

SUPPLEMENT 2

Scripts and data for estimating quantile equivalence of the vegetation reclamation data (Manly 2004) with the quantreg package in R (*Ecological Archives* A021-013-S2).

SUPPLEMENT 3

Scripts and data for estimating quantile equivalence for *Ambystoma tigrinum* trends (Dixon and Pechmann 2005) with the quantreg package in R (*Ecological Archives* A021-013-S3).