Invasive Plant Science and Management 2010 3:106–112

Review

Invasive Plant Researchers Should Calculate Effect Sizes, Not P-Values

Matthew J. Rinella and Jeremy J. James*

Null hypothesis significance testing (NHST) forms the backbone of statistical inference in invasive plant science. Over 95% of research articles in *Invasive Plant Science and Management* report NHST results such as P-values or statistics closely related to P-values such as least significant differences. Unfortunately, NHST results are less informative than their ubiquity implies. P-values are hard to interpret and are regularly misinterpreted. Also, P-values do not provide estimates of the magnitudes and uncertainties of studied effects, and these effect size estimates are what invasive plant scientists care about most. In this paper, we reanalyze four datasets (two of our own and two of our colleagues; studies put forth as examples in this paper are used with permission of their authors) to illustrate limitations of NHST. The re-analyses are used to build a case for confidence intervals as preferable alternatives to P-values. Confidence intervals indicate effect sizes, and compared to P-values, confidence intervals provide more complete, intuitively appealing information on what data do/do not indicate.

Key words: Estimation, statistics, confidence interval, P-values, Null hypothesis significance testing.

P-values …, after being widely seeded by statisticians, are now well established in the wild, and often appear as pernicious weeds rather than a useful crop (comment by I. M. Wilson in Nelder 1999).

Contrary to common dogma, tests of statistical null hypotheses have relatively little utility in science and are not a fundamental aspect of the scientific method (Anderson et al. 2000).

… p values are neither objective nor credible measures of evidence in statistical significance testing. Moreover, the authenticity of many published studies with $p < .05$ findings must be called into question. Rather than the preoccupation with p values … the goal … should be the estimation of sample statistics, effect sizes and the confidence intervals (CIs) surrounding them (Hubbard and Lindsay 2008).

Null hypothesis significance testing (NHST) is very widely used in invasive plant science. Of the papers in *Invasive Plant Science and Management* that present data amenable to NHST, over 95% report NHST results such as P-values or statistics closely related to P-values such as least significant differences. Despite its widespread use, NHST has serious practical limitations. For one thing, P-values are hard to interpret and often are misinterpreted. Also, P-values do not answer the two questions invasive plant scientists care about most: (1) What is the magnitude of the treatment effect? and (2) With what level of precision did the study estimate the treatment effect? Over the past half century, these and other limitations have prompted harsh critique of NHST from scientists in virtually every field that analyzes data (see 402 citations at http://welcome.warnercnr.colostate.edu/~anderson/thompson1.html), with critics in ecology and allied disciplines growing increasingly adamant over the past decade (Anderson et al. 2000; Fidler et al. 2006; Martinez-Abrain 2007, and many others).

This paper advocates for confidence intervals as alternatives to NHST. We begin with a brief review of NHST and confidence intervals. Then, we address the following four points in the course of illustrating advantages of confidence intervals over NHST: (1) Nonsignificant statistical tests are not evidence for null hypotheses; (2) Effect sizes are uncertain, even when statistical tests are significant; (3)

DOI: 10.1614/IPSM-09-038.1

* First author: Rangeland Ecologist, United States Department of Agriculture, Agricultural Research Service, 243 Fort Keogh Road, Miles City, MT 59301; second author: Plant Physiologist, United States Department of Agriculture, Agricultural Research Service, Eastern Oregon Agricultural Research Center, 67826-A Hwy 205, Burns, OR 97720. Corresponding author's E-mail: matt.rinella@ars.usda.gov

Significance thresholds are arbitrary; and (4) One minus the P-value is not the probability the alternative hypothesis is true. To illustrate these points, we reanalyzed four published datasets from the plant science literature (two of our own and two of our colleagues'). We fit the same or very similar models that the authors originally fit, but instead of P-values, we present confidence intervals on effect sizes. For example, in some cases we present confidence intervals on effect size = treatment − control.

Null Hypothesis Significance Testing and Confidence Intervals

Consider a hypothetical experiment designed to evaluate the effect of a weed control treatment on desired species productivity. The hypothetical treated and untreated plots were arranged in a completely randomized design. An expression for the data is:

$$y_i = b_0 + b_1 x_i + \varepsilon_i \quad [1]$$

where y_i is desired plant biomass (g m^{-2}) in plot i, x_i equals 0 if plot i was not treated and 1 if plot i was treated, and ε_i is normally distributed random error. Equation 1 can be thought of as a simple ANOVA model. With this scenario for the x_i values, b_0 represents the mean of untreated plots and b_1 describes the effect of treatment. As is typical in invasive plant science, the hypothetical authors conduct a point null hypothesis test of $b_1 = 0$; i.e., weed control had no effect. The authors use the F-test to calculate a P-value, which has the following interpretation: Pr (observed data or data more extreme| $b_1 = 0$). In other words, a P-value is the probability of witnessing the observed data, or data more extreme, given that $b_1 = 0$. If this probability is lower than some threshold statistical significance level (e.g., 0.05), the authors will conclude that the null hypothesis is unlikely in light of the data; i.e., they will reject the null hypothesis that $b_1 = 0$. It is important to comprehend what the P-value is not. It is not the probability that $b_1 = 0$.

An increasing number of authors avoid P-values and instead use confidence intervals to describe the range of likely values of parameters such as b_1 (e.g., Nickerson 2000; Stephens et al. 2007). Strictly defined, a 95% confidence interval is one realization of a procedure (collect data, calculate interval) that has a 95% success rate at bracketing the true parameter value (Berry and Lindgren 1996). (The interpretations for other confidence intervals such as 50%, 68%, etc. are analogous.) The reason for the complicated interpretation of confidence intervals is that parameters are not viewed as random variables in classical statistics; a confidence interval either brackets a parameter or it does not. Many researchers and statisticians employ an alternative interpretation of confidence intervals; they interpret 95% confidence intervals simply as having a 0.95 probability of bracketing the true parameter value. Although not technically correct, this interpretation is sometimes acceptable because it is the correct interpretation for Bayesian confidence intervals, and Bayesian and classical confidence intervals are identical under particular sets of assumptions/conditions. Specifically, Bayesian and classical confidence intervals for normally distributed data are identical when a particular noninformative Bayesian prior distribution is used (p. 355, Gelman et al. 2004). We present several 50 and 95% confidence intervals in this paper, and given the noninformative priors we used, it is acceptable to view these intervals as having a 0.50 and 0.95 probability of bracketing the parameter.

Nonsignificant Statistical Tests Are Not Evidence for Null Hypotheses

Failing to reject a null hypothesis (large P-value) does not suggest the null hypotheses is true or even approximately true (Fisher 1929). We reanalyzed one of our published datasets to illustrate this point.

To evaluate the ability of plant groups to repel weed invasions, James et al. (2008) sowed medusahead (*Taeniatherum caput-medusae* (L.) Nevski) seeds in plots after removing either (1) nothing, (2) annual forbs, (3) perennial forbs, or (4) bunchgrasses. The authors tested null hypotheses that weed densities did not differ between treatments, and instituted Bonferonni corrections to maintain an experiment-wise error rate of 5%. The null hypothesis was rejected only for bunchgrasses and it was concluded that ''Bunchgrasses were the only functional group that inhibited *T. caput-medusae* establishment.''

On one hand, James et al. (2008) are to be congratulated for reporting dispersion statistics (i.e., standard errors), a practice advocated by many (e.g., Anderson et al. 2001; Nagele 2001) (Figure 1). On the other hand, the large standard errors were clearly ignored in blithely concluding forb removals had no effect. Instead, it is far more likely that every removal treatment had some effect, and that measurement error and small sample size prevented the authors from detecting these effects. Many authors believe it is practically impossible for treatments to have no effect to an infinite number of decimal points (i.e., a true point null hypothesis) (e.g., Cohen 1994; Kirk 1996; Tukey 1991; but see Guthery et al. 2001). In a reanalysis, we calculated confidence intervals for the treatment effects (Figure 1). For Figure 1 to support the no effect of forb removal conclusion there would need to be extremely narrow confidence intervals centered on the 0 line in place of the wide confidence intervals for forbs. Like the significance tests from the original analysis, the forb confidence intervals do not rule out a lack of effect because they overlap 0. But in addition to 0, the confidence intervals encompass a wide range of values, indicating possibly large effects.

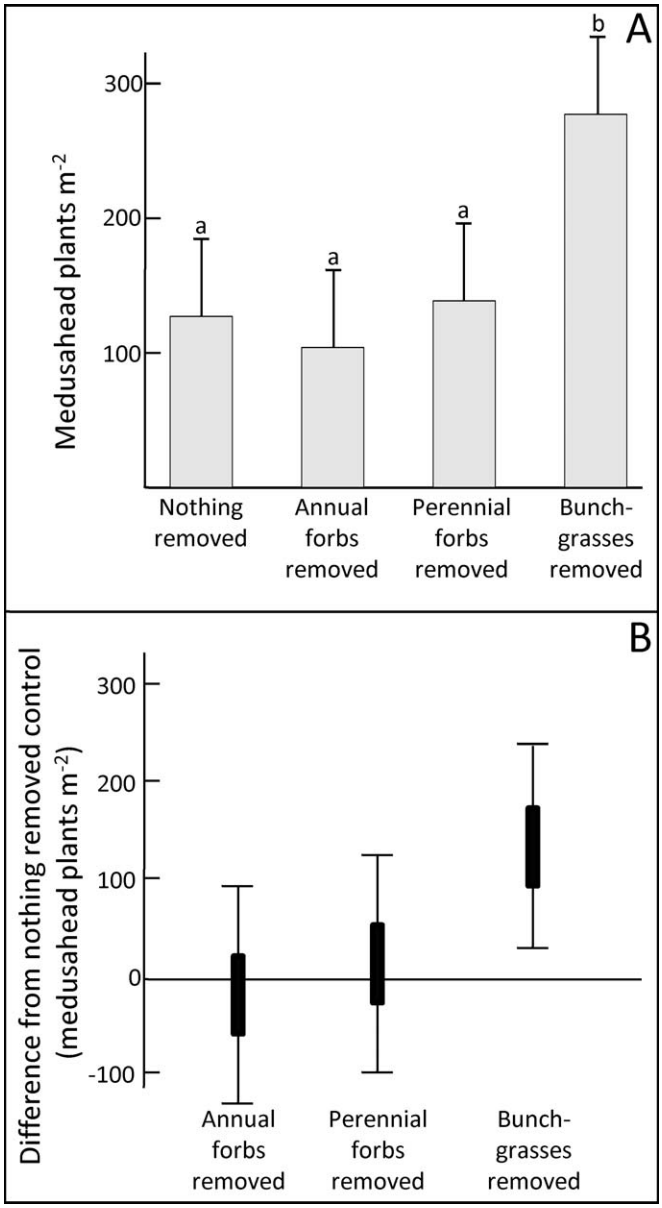

Figure 1. (A) Data and results of analysis from James et al. (2008). Medusahead densities (mean + SE, $n = 8$) resulting from treatments that sowed the weed after imposing plant group removal treatments. The authors tested all of the means against each other, and letters indicate significant differences ($P < 0.05$). (B) Confidence intervals (50% [bars] and 95% [lines]) resulting from a reanalysis of the James et al. (2008) data. Confidence intervals estimate effect size of treatment that removed plant groups (i.e., effect size = plant group removed − nothing removed).

The mistake of construing fail-to-reject decisions as evidence for null hypotheses is pervasive in invasive plant science. Consider these examples from the literature: (1) ''yellow starthistle plant density did not differ from the untreated control in plots treated with imazapic;'' (2) ''Mowing had no effect on either *B. ischaemum* or other dominant species at either site…;'' (3) ''Grazing did not increase or decrease density of mature leafy spurge stems…;'' (4) ''Burning at tiller emergence did not affect smooth brome…'' These examples could be easily developed into management recommendations. Misinterpreting significance tests not only misdirects science, it misdirects management. De-emphasizing P-values and instead focusing on confidence intervals would limit the potential for misinterpretation.

Effect Sizes Are Uncertain, Even When Statistical Tests Are Significant

When authors detect treatment effects (e.g., $P \leq 0.05$), they often erroneously assert that the effect size precisely equals the difference between means (e.g., treatment − control) or the ratio of means (e.g., treatment / control). Here are examples from the literature: (1) ''chaining + fire treatments reduced juniper cover from 32% to <6%;'' (2) ''grazing decreased vegetative stem density from 104 … to 20 stems m^{-2};'' (3) ''CO_2 enrichment increased mass of cotyledons, leaves, and total aboveground plant material by 33% to 35%;'' and (4) ''defoliation reduced … densities … 55% (below) nontreated controls.'' There can be large discrepancies between estimated and actual effect sizes, and we reanalyzed data from Rinella et al. (2001) to illustrate this point.

In Rinella et al. (2001), we measured spotted knapweed (*Centaurea stoebe* L., synonym *C. maculosa* auct. non Lam.) densities after applying mowing treatments. After rejecting the null hypothesis of no effect of mowing ($P \leq 0.05$), we concluded (based on sample means) that: ''Fall mowing decreased adult (knapweed) density 85 and 83% below that of the control at Sites 1 and 2, respectively.'' In retrospect, the recklessness of this statement is striking. For one thing, given the high spatial variability of the study areas, the probability is essentially 0 that mean knapweed densities in mowed and nonmowed plots were equivalent prior to treatment. So for a given site, the probability is roughly 0.5 that nonmowed plots contained more knapweed plants than fall-mowed plots merely due to sampling variability. Furthermore, in addition to sampling variability, there was a high potential for measurement error; imagine multiple observers counting tightly clumped plants.

In a reanalysis, we calculated confidence intervals for the fall mowing effects, and unlike the original analysis we analyzed the data on the percent scale (i.e., effect size = $-100 \times [1 -$ mowed/ nonmowed]) (Figure 2). These intervals are asymmetric because they involve a ratio of normally distributed treatment means. Had the intervals of Figure 2 been reported in Rinella et al. (2001), we would not have reported the 85% point estimate as if it were precisely correct. Instead, we would have reached the more logical conclusion that fall mowing likely reduced weed

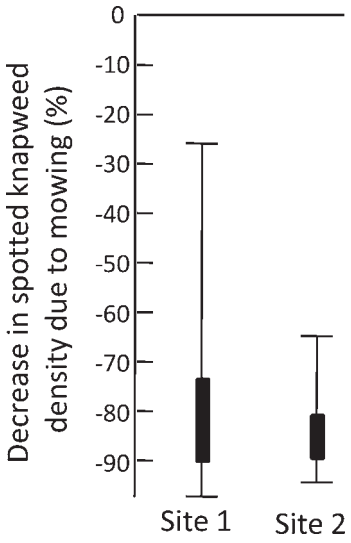

Figure 2. Reanalysis of data from Rinella et al. (2001). Confidence intervals (50% [bars] and 95% [lines]) estimate effect size based on ratio of treatment means (i.e., effect size = $-100 \times [1 -$ mowed / nonmowed]).

density somewhere between 22 and 96%. Unfortunately, the 85% estimate was published in a widely distributed Extension publication. These kinds of mistakes could be avoided if invasive plant scientists used confidence intervals instead of P-values.

Invasive plant-infested lands are highly variable, both spatially and temporally, and invasive plant studies often suffer perilously small sample sizes because of constraints on land, labor, and other resources. Therefore, point estimates of effect sizes are of questionable value in invasive plant science, even when effects are deemed statistically significant. Our statistical reporting should characterize uncertainty about parameter values.

Significance Thresholds Are Arbitrary

Significance thresholds (e.g., $P \leq 0.05$) are arbitrary in the sense that small changes to thresholds and to measured variables can cause dramatic changes to conclusions. We reanalyze two published datasets to illustrate this point.

Heitschmidt and Vermeire (2006) applied drought and irrigation treatments and measured biomass production of perennial cool-season grasses in invasive annual grass-infested rangeland. After failing to reject the null hypothesis of no effect at the $P \leq 0.05$ threshold, the authors concluded that production was "similar among treatments." Figure 3 shows 50 and 95% confidence intervals for the difference between the two most extreme treatments (i.e., effect size = moderate drought with irrigation − severe drought without irrigation). The 95% confidence interval barely overlaps 0, which indicates the

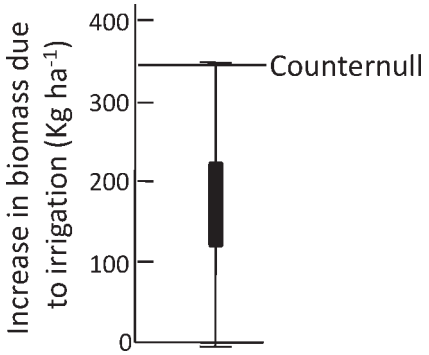

Figure 3. Reanalysis of data from Heitschmidt and Vermeire (2006). Confidence intervals (50% [bars] and 95% [lines]) and counternull value describing effect size of irrigation treatment on cool-season perennial grass production (i.e., effect size = irrigated − nonirrigated). There is equal evidence for the counternull value and the null hypothesis of 0 effect.

treatment difference is not significant at the $P \leq 0.05$ level (Figure 3). However, adding just 1.0 kg ha^{-1} to just one data point renders the difference significant. Given the size of the author's sampling frames (0.025 m^2), 1.0 kg ha^{-1} corresponds to a few blades of grass. So had a few additional grass blades been collected, the authors would have concluded (implicitly or explicitly) that irrigation increased cool-season perennial grass by 285% (97 vs. 273 kg ha^{-1}) rather than concluding production was "similar among treatments." We can be certain this would have been the author's conclusion because they concluded the following regarding another treatment that was significant: "irrigation increased warm-season perennial grass production by 251%." It is curious that just a few blades of grass would arbitrate between such sharply contrasting conclusions (i.e., 0 vs. 285%). Had a confidence interval been used, Heitschmidt and Vermeire (2006) would have been compelled to settle on a range of plausible conclusions/hypotheses.

To further explore results from Heitschmidt and Vermeire (2006), we calculated the counternull value (Rosenthal and Rubin 1994). The counternull value corresponds to the effect size that is equally as likely as the null hypothesis value. The counternull serves as a reminder that alternative hypotheses can be more likely than the null hypothesis, even when there is a failure to reject the null. For the hypothesis of no effect of Heitschmidt and Vermeire (2006) discussed above, the counternull value is approximately 350 kg ha^{-1} (Figure 3). Therefore, the author's conclusion of no effect of irrigation, and the conclusion that irrigation increased cool-season grasses 350 kg ha^{-1} are equally plausible. In northern mixed-grass prairie where the study was conducted, 350 kg ha^{-1} is a considerable quantity of biomass. Because of the shape of the normal distribution, values between 0

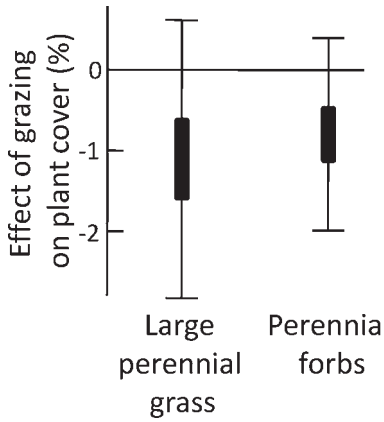

Figure 4. Reanalysis of data from Bates (2005). Confidence intervals (50% [bars] and 95% [lines]) describing effect size of grazing treatment on plant cover (i.e., effect size = grazed cover − not-grazed cover).

and 350 are more plausible than either 0 or 350 and the most likely value is in the center of the confidence interval (i.e., 175).

As another example, consider data from Bates (2005) evaluating effects of cattle grazing on several plant groups after cutting juniper (*Juniperus occidentalis* Hook.) plants. Based on nonsignificant tests at the 5% level, it was concluded that ''Canopy cover … did not differ between not-grazed cut and grazed cut treatments in any year of the study.''

We used data from the last year of the study to compute confidence intervals on the grazing effect for large perennial grasses and perennial forbs (Figure 4). P-values calculated from the confidence intervals were 0.06 for large perennial grasses and 0.09 for perennial forbs. These small P-values cast serious doubt on the conclusion that ''Grazing in the cut treatment did not limit herbaceous recovery.'' Furthermore, simple calculations illustrate the grazing effects might have been quite severe. The least squares parameter estimate for perennial grasses is −1%, and according to the mean of ungrazed plots, perennial grass cover was 3.6%. Therefore a −1% change would imply that grazing reduced grasses by $100 \times (1 - (3.6 - 1)/3.6) = 38\%$. The same calculation for perennial forbs yields a decrease of 57%. Significance thresholds often lull researchers into believing treatments have little or no effect. When correctly interpreted, confidence intervals do not suffer this limitation.

Invasive plant scientists should refrain from hinging important conclusions on P = 0.04 vs. P = 0.06, or similarly small differences in P. P-values promote dichotomous conclusions (e.g., grazing is/is not deleterious) based on arbitrary cutoffs, whereas confidence intervals compel analysts to consider the range of conclusions their data do/do not support (e.g., the grazing effect was somewhere between trivial and highly deleterious). And it is not only small changes in measurements that push P-values slightly above or below significance thresholds; modeling assumptions can do the same thing. Whether or not an effect is significant ($P \leq 0.05$) often comes down to modeling choices such as which covariates to include, whether or not a factor is considered fixed or random, or whether or not covariances among measurements are assumed to be 0. Compared to NHST, conclusions based on confidence intervals are more robust to changes in modeling assumptions. Often, modest changes to models induce only modest changes in confidence intervals; the basic message behind the data often remains unaltered. Conversely, when significance cutoffs are used, slightly altering the model can push the P-value above or below the cutoff, which dramatically changes the message behind the data.

One Minus the P-value is Not the Probability the Alternative Hypothesis is True

We showed previously that large P-values are routinely misconstrued as evidence for null hypotheses. It is even more common to construe small P-values as evidence for alternative hypotheses, and this too can be a mistake. A P-value is strictly the probability of observed or more extreme data, given a true null hypothesis: P(observed or more extreme data| null hypothesis). A small P-value indicates the data are improbable if the null hypothesis is true. But this is not how P-values are generally interpreted in invasive plant science. Instead, as statisticians Berger and Sellke (1987) note, ''most nonspecialists interpret (P-values) precisely as P(hypothesis| data).'' That is, in invasive plant science one minus the P-value is often erroneously interpreted as the probability the alternative hypothesis is true. The probability the hypothesis is true clearly is what scientists want, but in using P-values they settle for a tangentially related probability. It is worth noting that the probability that a hypothesis is true is only calculable via Bayes formula.

P-value ≠ P(hypothesis| data) is not merely an esoteric inequality between two conditional probabilities. When the left hand probability is small, researchers routinely infer that the right hand probability, the probability of interest, must also be small; this is a mistake Falk and Greenbaum (1995) labeled ''the illusion of attaining improbability.'' This mistake would not be very damaging if the two probabilities were guaranteed to be somewhat similar, but they are not. See Cohen (1994) for a simple example of a large discrepancy between a P-value and the probability of a null hypothesis, and see Berger and Sellke (1987) for a dramatic example involving normal means where a P-value of 0.05 corresponds to P(null hypothesis| data) = 0.52. Similar examples of P-values distorting evidence against null hypotheses are not rare, and they derive from both a Bayesian (e.g., Diamond and Forrester 1983; Robert 2001)

and frequentist, or Neyman-Pearson, perspective (Sellke et al. 2001). Discrepancies between P-values and probabilities of hypotheses led Berger and Berry (1988) to question the validity of conclusions drawn from moderately small P-values. Moreover, these discrepancies prompted Hubbard and Lindsay (2008) to conclude that "*P* values are an inadequate measure of evidence."

This is not to imply that P-values always give different answers than P(hypothesis| data). For example, the 95% confidence intervals of this paper do not overlap the null hypothesis values when the P-values are less than 0.05 (also see Casella and Berger 1987). However, it is disconcerting to find that P-values are not a reliable source of evidence for all common testing problems (Hubbard and Lindsay 2008).

Concluding Remarks

The advantages of confidence intervals over NHST extend beyond interpreting results from individual studies. Confidence intervals promote what is often termed meta-analytic thinking; i.e., accumulation of evidence over multiple studies (Cumming and Finch 2001). Consider a collection of similar studies on a treatment with $P < 0.05$ reported for roughly half and $P > 0.05$ for the remainder. Based on P-values alone, it might be concluded that the studies are utterly inconclusive. But if confidence intervals were reported and each interval bracketed a similar range of values, although roughly half the intervals overlapped 0 in the lower tail (i.e., half were statistically nonsignificant), this would provide compelling evidence that the treatment had a consistent effect. Many authors have argued this and similar points (e.g., Fidler et al. 2006; Nakagawa and Cuthill 2007). Given the low power and imprecision (i.e., large error variances and small sample sizes) of many invasive plant studies, it seems we should base our understanding on data from multiple studies. This is what synthesis papers seek to do (e.g., D'Antonio et al. 1999), and confidence intervals are more appropriate for this cause than P-values.

In many cases, confidence intervals are easy to calculate. For example, when an effect is regulated by one regression coefficient, the confidence interval on the coefficient is all that is needed, and regression packages provide this interval automatically. Things become a little trickier when the confidence interval of interest involves multiple, correlated regression coefficients, but most statistics packages handle this situation as well via the Working-Hotelling procedure, delta method, or other procedures (Neter et al. 1996; SAS 1999). Many manuscripts provide help in computing and interpreting confidence intervals (e.g., Cumming and Finch 2001; Nakagawa and Cuthill 2007).

Invasive plant scientists work hard to design unbiased studies and gather reliable data. But in the end, too many hard-won datasets get distilled down to a dichotomous conclusion (i.e., effect/no effect). These binary conclusions promote confusion and information loss. Invasive plant scientists should use interval estimates to illustrate evidence for the range of conclusions their data support.

Literature Cited

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence and an alternative. J. Wildl. Manag. 64:912–923.

Anderson, D. R., W. A. Link, D. H. Johnson, and K. P. Burnham. 2001. Suggestions for presenting the results of data analysis. J. Wildl. Manag. 65:373–378.

Bates, J. D. 2005. Herbaceous response to cattle grazing following juniper cutting in Oregon. Rangeland Ecol. Manag. 58: 225–233.

Berger, J. O. and D. A. Berry. 1988. Statistical analysis and the illusion of objectivity. Am. Sci. 76:159–165.

Berger, J. O. and T. Sellke. 1987. Testing a point null hypothesis: the irreconcilability of P values and evidence. J. Am. Statistical Assoc. 82: 112–122.

Berry, D. A. and B. W. Lindgren. 1996. Statistics, Theory and Methods. Belmont, CA: Wadsworth. 702 p.

Casella, G. and R. L. Berger. 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with comments). J. Am. Statistical Assoc. 82:106–139.

Cohen, J. 1994. The earth is round ($p < .05$). Am. Psychologist 49: 997–1003.

Cumming, G. and S. Finch. 2001. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. Educ. Psychol. Meas. 61:532–574.

D'Antonio, C. M. and Levine, J. M. 1999. Elton revisited: a review of evidence linking diversity and invasibility. Oikos 87:15–26.

Diamond, G. A. and J. S. Forrester. 1983. Clinical trials and statistical verdicts: probable grounds for appeal. Ann. Internal Med. 98: 385–394.

Falk, R. and C. W. Greenbaum. 1995. Significance tests die hard. The amazing persistence of a probabilistic misconception. Theory Psychol. 5:75–98.

Fidler, F., M. A. Burgman, G. Cumming, R. Buttrose, and N. Thomason. 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. Conserv. Biol. 20:1539–1544.

Fisher, R. A. 1929. The statistical method in psychical research. Proc. Soc. Psychical Res. 39:189–192.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis. Boca Raton, FL.: Chapman & Hall/CRC. 668 p.

Guthery, F. S., J. J. Lusk, and M. J. Peterson. 2001. The fall of null hypothesis: liabilities and opportunities. J. Wildl. Manag. 65: 379–384.

Heitschmidt, R. K. and L. T. Vermeire. 2006. Can abundant summer precipitation counter losses in herbage production caused by spring drought. Rangeland Ecol. Manag. 59:392–399.

Hubbard, R. and R. M. Lindsay. 2008. Why P values are not a useful measure of evidence in statistical significance testing. Theory Psychol. 18:69–88.

James, J. J., K. W. Davies, R. L. Sheley, and Z. T. Aanderud. 2008. Linking nitrogen partitioning and species abundance to invasion resistance in the Great Basin. Oecologia 156:637–648.

Kirk, R. E. 1996. Practical significance: a concept whose time has come. Educ. Psych. Meas. 56:741–745.

Martinez-Abrain, A. 2007. Are there any differences? A non-sensical question in ecology. Acta Ecol. Int. J. Ecol. 32:203–206.
Nagele, P. 2001. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. Br. J. Anaesth. 90:514–516.
Nakagawa, S. and I. C. Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol. Rev. 82:591–605.
Nelder, J. A. 1999. From statistics to statistical science. The Statistician 48:257–269.
Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models. New York: Irwin. 1408 p.
Nickerson, R. S. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psych. Methods 5:241–301.
Rinella, M. J., J. S. Jacobs, R. L. Sheley, and J. J. Borkowski. 2001. Spotted knapweed response to season and frequency of mowing. J. Range Manag. 54:52–56.
Robert, C. P. 2001. The Bayesian Choice. Paris, France: Springer. 604 p.
Rosenthal, R. and D. B. Rubin. 1994. The counternull value of an effect size. Psychol. Sci. 5:329–334.
SAS. 1999. Statistical software. Version 8.0. Cary, NC: SAS Institute.
Sellke, T., M. J. Bayarri, and J. O. Berger. 2001. Calibration of p values for testing precise null hypotheses. Am. Statistician 55:62–71.
Stephens, P. A., S. W. Buskirk, and C. Martinez del Rio. 2007. Inferences in ecology and evolution. Trends Ecol. Evol. 22:192–197.
Tukey, J. W. 1991. The philosophy of multiple comparisons. Statistical Sci. 6:100–116.

Received May 27, 2009, and and approved January 28, 2010.