Frontiers in Ecology and the Environment

Statistical inference, Type II error, and decision making under the US Endangered Species Act

Berry J Brosi and Eric G Biber

Front Ecol Environ 2009; 7, doi:10.1890/080003

Statistical inference, Type II error, and decision making under the US Endangered Species Act

Berry J Brosi[1] and Eric G Biber[2]

Critical conservation decisions have been made based on the spurious belief that "no statistically significant difference between two groups means the groups are the same". We demonstrate this using the case of the Preble's meadow jumping mouse (*Zapus hudsonious preblei*), an endangered species in the US. Such faulty statistical logic has been recognized before, but ecologists have typically recommended assessing post hoc statistical power as a remedy. Statisticians, however, have shown that observed power will *necessarily* be low when no differences are found between two populations. Alternatives to assessments of statistical power include equivalence testing (a method rarely used by ecologists) and Bayesian or likelihood methods. Although scientists play a central role in ameliorating this problem, the courts could also assist by requiring litigated federal agency decisions to consider the risks of both Type I and Type II errors.

Front Ecol Environ 2009; 7, doi:10.1890/080003

Conservation decision making can involve irreversible risks (such as species extinction) and is often carried out based on scarce data. Conservation scientists use statistical tests to decide if two or more groups, such as two different fire management regimes in a protected area, are significantly different from one another (see Panel 1 for a primer on classical hypothesis-testing frameworks). In this paper, we focus on when such tests fail to find a statistically significant difference. Scientists routinely draw inappropriate conclusions from this particular statistical outcome, with potentially serious outcomes for management, policy, and the environment. Here, we outline an illustrative example and offer suggestions for scientists, managers, and the judiciary regarding how this problem could be more effectively addressed.

We focus here on the Endangered Species Act (ESA), the primary law protecting biodiversity in the US and, in particular, on statistical distinctions between what should or should not constitute a protected taxon. The ESA's unit of protection is a "species", including "subspecies" and "distinct population segments" of vertebrate species (16 USC §§ 1532[16] and 1533[a]). Listing a "species" as protected under the ESA carries high stakes, because it involves an investment of limited funds that could protect other taxa, and because it triggers legal protection of habitat that may limit economic activity.

The agencies in charge of implementing the ESA have relied increasingly on genetics in defining "species" (eg Goldstein *et al.* 2000; Fallon 2007), and have issued policy statements where the designation of a "subspecies" or "distinct population segment" depends largely on genetic differences (US FWS and NOAA 1996). Accordingly, scientists' conclusions about whether populations are genetically distinct have become extremely important for the regulatory process, and ultimately affect both species survival and economic activity.

Time and funding for species conservation are always limited. Protection of a spurious subspecies (ie a population that is not truly biologically distinct from its abundant and widespread conspecifics) takes away resources from other species, subspecies, or populations that need protection. On the other hand, the ESA was drafted with the precautionary principle in mind (Ruhl 2004), so management decisions under the ESA should err on the side of caution. This is especially true, given the asymmetric risk in conservation, since extinction is irreversible. Accordingly, the decision to list or delist a taxon is of critical importance.

In a nutshell:

- Ecologists often assume populations are identical ("null hypothesis"), and then statistically test to identify differences; if a significant difference is *not* found, ecologists usually accept the null hypothesis as true, when in fact it *cannot be rejected*
- The statistical tools commonly recommended for use in such situations are inadequate; a better option in some contexts is *equivalence testing*, which assumes that populations are different, and attempts to prove they are the same
- Scientists, land managers, and courts must work together to avoid these pitfalls in conservation management

[1]*Department of Biological Sciences, Stanford University, Stanford, CA* [*]*(bbrosi@stanford.edu)*; [2]*Boalt Hall School of Law, University of California, Berkeley, CA*

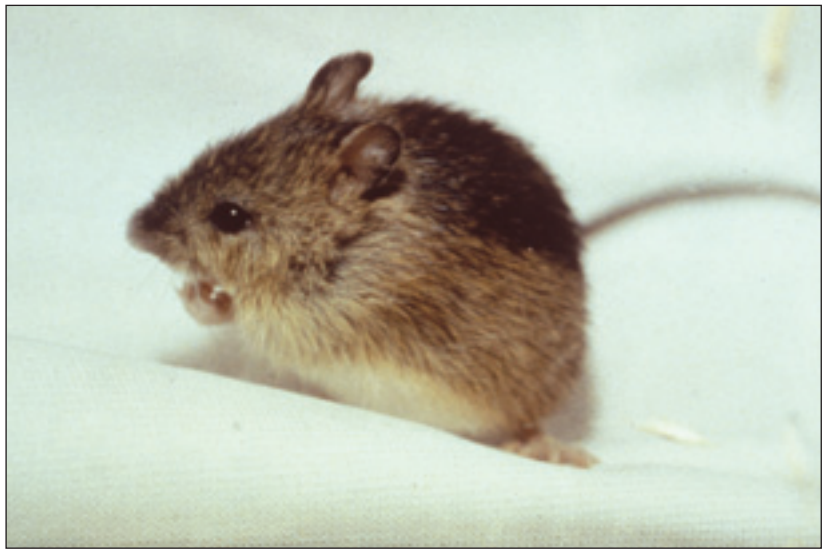

Figure 1. *Preble's meadow jumping mouse*, Zapus hudsonious preblei.

Preble's meadow jumping mouse: a case study of a common inferential error

A high-profile example of the challenges in species listing decisions is the Preble's meadow jumping mouse (*Zapus hudsonious preblei*), a federally endangered subspecies in Colorado and Wyoming (Figure 1). Attention has focused on the listing decision because protection of this animal has impacted development in a rapidly growing region (Holthouse 2005). In 2004, a scientific group released a study that compared genetic sequences and morphometric data between the Preble's jumping mouse and other, more abundant and widespread subspecies of *Z hudsonious*, especially *Z hudsonious campestris* (Ramey *et al.* 2004). The study found no statistically significant genetic differences between the two groups of mice. Based on this finding, they reported, "Our analysis of mtDNA [mitochondrial DNA] sequence data refutes… that *Z h preblei* is a unique taxon" (Ramey *et al.* 2004).

When a statistical hypothesis test such as this one (see Panel 1) finds no significant difference between two populations (Figure 2, bottom row shaded in yellow), the test outcome could be correct (Figure 2, bottom left cell) or, alternatively, could represent a Type II error (Figure 2, bottom right cell; explained in Panel 1). The key problem with Ramey *et al.*'s conclusion is that it fails to distinguish between these two possibilities. A lack of statistical significance does support the statement that the researchers "cannot reject the null hypothesis". But that statement is subtly, yet crucially, different from the statement that the researchers should "accept the null hypothesis as true" (eg Taylor and Dizon 1996), which is the statement that Ramey *et al.* in essence made in their paper. As Wellek (2002) put it, "A *non-significant difference must not be confused with significant homogeneity*" between two populations (emphasis in original). Indeed, this problem is highlighted by a later study that examined more specimens and a larger genetic sequence, and concluded that Preble's jumping mouse was indeed a separate subspecies (King *et al.* 2006). While policy makers could consider this outcome typical of the inconsistencies between scientific studies, a more important message is the importance of sample size, which we address later.

The fundamental cause of this confusion is that standard statistical tests are set up in a way that gives researchers relative certainty about the result only when the test shows a significant difference between two groups. By definition, you can be at least 95% sure that you are correct when a hypothesis test finds a significant difference (assuming the standard $\alpha = 0.05$; Figure 2, top row). But when the test outcome is not significant (Figure 2, bottom row), there is no way to reliably estimate how likely you are to be wrong (from Type II error) if you conclude that the populations in question are homogeneous (eg Hoenig and Heisey 2001). Thus, if a significant difference is not found in a statistical test, the only appropriate conclusion is that the *null hypothesis cannot be rejected*.

Concern about non-significant test results is not new, and has largely been framed in the context of Type II

Panel 1. A primer on classical hypothesis-testing statistical frameworks

Classical hypothesis-testing frameworks are typically structured with two hypotheses: a "null hypothesis" (symbolized as H_0; left column of Figure 2) that is assumed as the background state of reality and which usually states that there is no difference between two groups being analyzed, and an "alternative hypothesis" (H_I; right column of Figure 2) that generally states that there is a detectable difference between the two groups. In the context of conservation genetics, H_0 might be that a putative subspecies is taxonomically the same as a more common subspecies, with H_I stating that the putative subspecies is distinct from other subspecies. If evidence for H_I is not sufficiently strong, then the researcher concludes that H_0 cannot be rejected.

The four possible outcomes of a hypothesis test are presented in Figure 2. Type I error (a "false positive", with probability α) is the chance of accepting H_I when the groups are not different (Figure 2, top left cell). Reporting of Type I errors is essentially mandatory in scientific venues where statistics are used, presented as the ubiquitous *P* value. Scientists generally now accept a threshold of $\alpha < 0.05$, or a less than 1-in-20 chance, as representing a statistically significant difference. Type II error (a "false negative", with probability β) is the chance of not rejecting H_0, or finding no difference between two groups, when a difference does, in fact, exist (Figure 2, bottom right cell). Probabilities of Type II error are much less commonly reported, and are usually presented as its inverse, *statistical power* ($1-\beta$), which is the probability of detecting a significant difference if one were truly there (Figure 2, bottom left cell).

error and statistical power, the two cells comprising the bottom row of Figure 2 (eg Toft and Shea 1983; McGarvey 2007). The recognition of this problem, however, has not lessened the incidence of faulty inference from non-significant statistical outcomes. Indeed, a review of studies published in the journals *Conservation Biology* and *Biological Conservation* in 2003 found that nearly two-thirds of manuscripts with non-significant results inappropriately interpreted these as evidence for significant homogeneity (Fidler *et al.* 2006). In response, there have been regular calls for scientists and managers to rethink non-significant statistical results in decision-making (eg Taylor and Dizon 1996; Palsbøll *et al.* 2006) and to consider alternatives to hypothesis testing (eg Fidler *et al.* 2006).

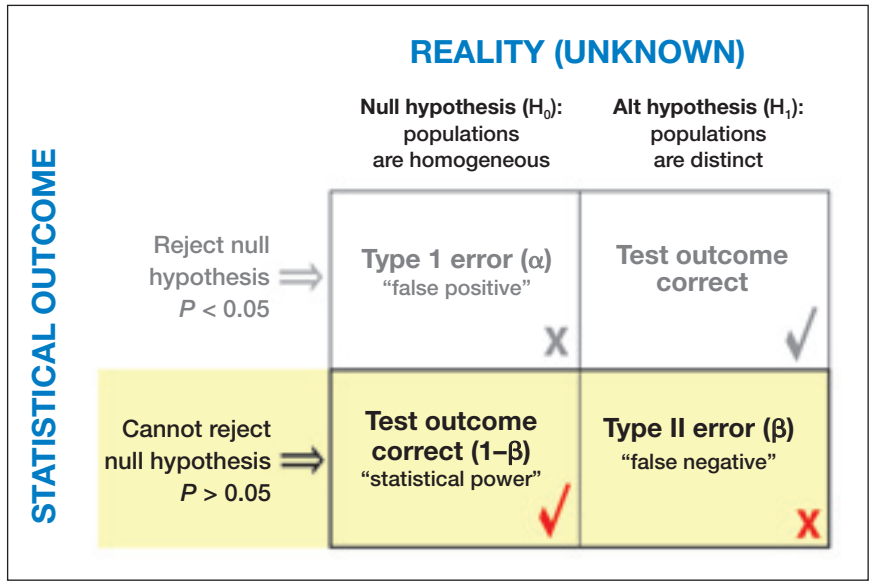

Figure 2. *Outcomes of statistical tests. In this 2 x 2 table, the null hypothesis (no difference between groups) is true of cells in the left column, and the alternative hypothesis (groups are different) is true for cells in the right column. The top row represents cases where the statistical test found a significant difference between groups. The bottom row (shaded in yellow) is for cases when the null hypothesis cannot be rejected. Cells whose outcome matches the underlying reality are marked with a "✓", and those with errors are marked with an "x".*

Emblematic of both the subtle nature and the widespread occurrence of the error of inferring significant homogeneity from a non-significant test result is the fact that almost no one in the Preble's debate, on *any* side, recognized it as a problem, despite the intense scrutiny of the report. This includes the authors of the two studies, relevant officials at the US Fish and Wildlife Service (FWS), any of the interest groups concerned about the conservation status of this organism, and all but two of the 15 scientists who peer-reviewed the studies. We acknowledge that there were other issues in the debate over Ramey *et al.*'s study, including potential genetic contamination (Colorado Division of Wildlife 2004; King *et al.* 2006), and it is also true that Ramey *et al.* (2005) qualified their statements to some degree in a later paper, though the inferential error remained. Moreover, in focusing on this particular example, we emphasize that we do not endorse any side in the controversy; given the widespread nature of this inferential error in conservation biology, our selection of the Preble's jumping mouse case as an example should not be taken as a critique of the overall quality of the work of Ramey and his co-authors. Nonetheless, we think this example is enlightening because so many parties made the inferential error, despite the high profile of the case.

The larger problem here is not limited to the Preble's case, or even conservation genetics. For example, when the FWS analyzed whether low water flows were correlated with endangered fish mortality in the Klamath River of Oregon and California, the agency required a statistically significant connection between the two before it would commit to action to increase water flows, even though (as discussed above) the lack of a statistically significant correlation would provide little or no information about whether such a correlation existed (McGarvey 2007). Likewise, in the Grand Canyon, FWS planned on reducing ESA protection for the humpback chub (*Gila cypha*), absent a statistically significant decline in population levels over time (US FWS 2002), again even though a hypothesis-testing result of no statistical significance would provide little or no information about whether the population was, in fact, declining. In contexts such as these, the classical hypothesis-testing framework that scientists turn to by default may be inappropriate for addressing conservation issues, precisely because the inability to determine the risk of Type II errors will leave decision makers with little additional information if no statistically significant results are identified. The end result may be that major policy decisions will be made, with potentially irreversible risks, based on fundamentally incorrect conclusions.

What to do with a non-significant test result

There is clearly room for improvement in dealing with non-significant test results in conservation, particularly when this affects the legal designation of protected species. Although ecologists have recognized the flaws of interpreting a non-significant test finding as proving the null hypothesis for the past two decades, the typical recommendation has been to use power analyses to assess Type II errors (eg Toft and Shea 1983; Taylor and Dizon 1996). However, most ecologists are unaware of the serious problems with such post hoc (also called "observed" or "retrospective") power calculations.

Statisticians argue that observed statistical power

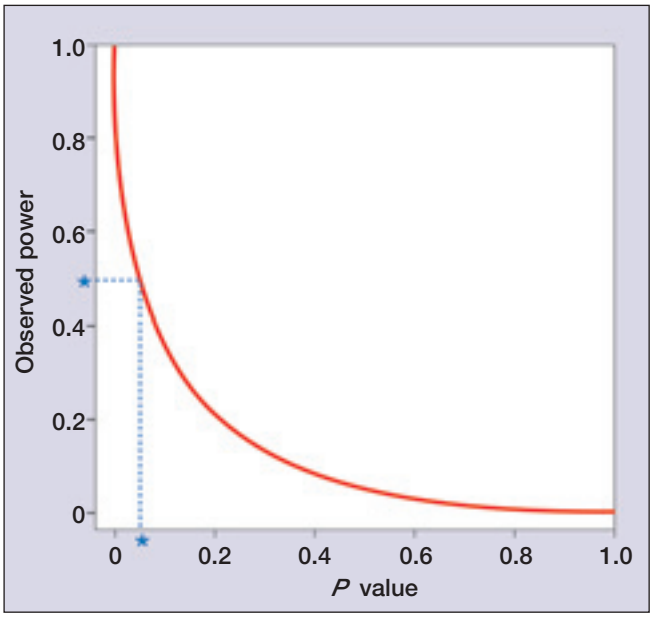

Figure 3. *Observed statistical power and P value, redrawn from Hoenig and Heisey (2001). The curve is based on a one-tailed z test where α = 0.05; the blue dashed lines indicate that where P= 0.05, observed power = 0.5. Thus, when P > 0.05, the test is not significant and observed power is lower than 50%.*

should never be used in data analysis (eg Goodman and Berlin 1994; Hoenig and Heisey 2001), for the fundamental reason that power is directly related to the *P* value of the test (Figure 3). When a *P* value is not significant, power will necessarily be low. In nearly all cases, when *P*>0.05, power will be 0.5 or lower. As such, observed power gives essentially no additional information: if an experimenter finds a significant difference, the power of the test to resolve differences was high, and if no significant difference was found, then the power was low. We note, however, that statistical power can be very useful for planning *future* experiments (eg Hoenig and Heisey 2001).

Since using post hoc statistical power in data analysis is inappropriate, what are the alternatives? One is to reverse the conventional null and alternative hypotheses through *equivalence testing* – that is, setting a null hypothesis that two groups are different and then testing whether or not they are the same (Hoenig and Heisey 2001; Wellek 2002; McGarvey 2007; see Panel 2 for a basic explanation of a simple equivalence *t* test). One challenge of this technique is that the investigator must set a minimum difference (△) between samples that is assumed as the null hypothesis, analogous to *effect size* in assessing the power of a standard hypothesis test. While this requirement could be thought to introduce an element of subjectivity into the analysis, it also forces researchers to make their underlying assumptions more explicit. For example, what *is* the minimum genetic difference that a researcher is using to define taxa as distinct enough to qualify for legal protection? The result of equivalence testing, properly used, is that the risk of a Type II error (as defined in a traditional hypothesis test) is reduced to 5% or less.

Scientists could also consider adopting a Bayesian or likelihood-based statistical framework; these come with their own limitations, but do reduce the problems with frequentist statistics (the hypothesis-testing paradigm, based on the probabilities of events occurring over many trials) discussed here, by avoiding the hypothesis-testing paradigm. See Wade (2000) for guidelines on using Bayesian statistics in conservation biology. For Bayesian readers, we suggest consideration of Bayesian equivalence testing (eg Wellek 2002); likelihood ratio tests can also be reversed for likelihood-based equivalence testing. Given the reluctance of many scientists and federal agency officials to utilize Bayesian and likelihood methods, however, for the foreseeable future conservation biologists will probably have to rely in part on frequentist methods.

Finally, one major factor contributing to the problems of interpreting non-significant test results is sample size (Figure 4). On the one hand, with small sample sizes it is very difficult to find a statistically significant difference between two populations (low power). On the other hand, with a very large sample size one can often establish a sig-

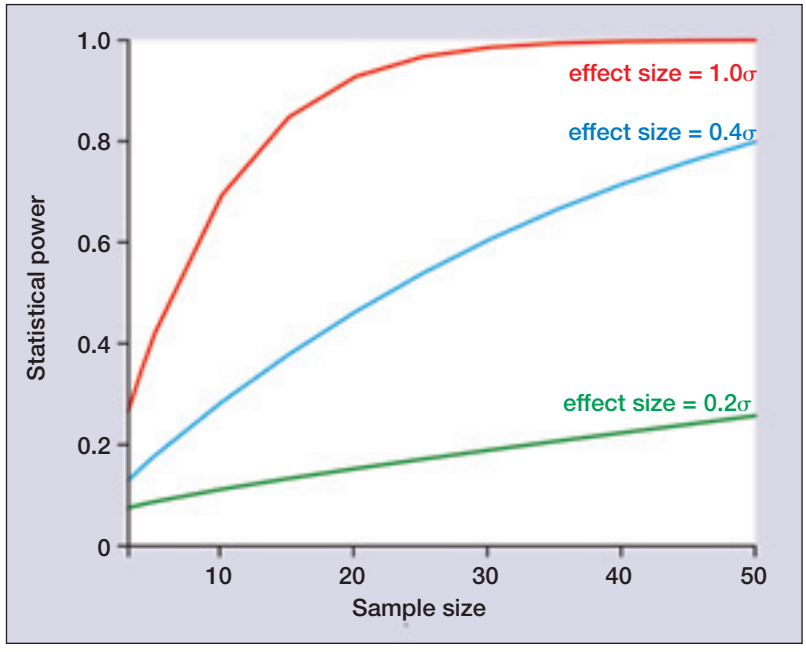

Figure 4. *Sample size, effect size, and statistical power. Here, we show statistical power, or the probability that a statistical comparison will find a significant difference when one is present, as a function of sample size (on the x-axis) and effect size (the different colored lines, representing the magnitude of difference between the two groups in a comparison, shown in standard deviation equivalents). Power was calculated for two-sample, two-tailed* t *tests, using the "PWR" package for the R statistical language. With small effect sizes, a very large sample size is needed to reliably find a significant difference between groups.*

Panel 2. Equivalence testing

Here, we present a practical method for calculating a two-sample *t* test for equivalence in means using confidence intervals, following Jones *et al.* (1996) and assuming equal variance in the two samples. Readers who use the R statistical programming language can easily conduct the same equivalence test using the "tost.data" function in the "equivalence" package (Robinson 2008). Readers wishing to conduct calculations by hand, or wanting details on equivalence testing in other contexts, should refer to Wellek (2002).

In equivalence testing, the experimenter must define an a priori minimum difference Δ between studies that is assumed as the null hypothesis. This interval must be considered carefully as an expected effect size, minimum functional difference, or minimum detectable biological difference. Wellek (2002) sets out general guidelines for defining Δ with "strict" and "liberal" criteria; these guidelines are quite generic, however, and researchers should work to identify the most relevant biological difference between the two study groups in setting equivalence intervals.

Whatever the context, *t* tests are parametric, and the researcher must assess whether or not the data meet the necessary assumptions of sample independence, normal data, and so forth. Equivalence tests have been developed for non-parametric tests (eg the Mann-Whitney test; Wellek 2002) if data do not meet these assumptions.

Confidence intervals have the advantage of allowing the experimenter to *simultaneously* conduct an equivalence test and a traditional hypothesis test (Jones *et al.* 1996). To show significant equivalence for a two-tailed test, the confidence interval of the difference in means between the two groups must fall completely within the equivalence interval $-\Delta$ to $+\Delta$. Confidence intervals are familiar to most scientists (the calculations are the same as in a traditional test), and are good statistical practice, because visualizing the variation in the data generally translates to a more holistic view of the problem than will be provided by a binary test result (eg Fidler *et al.* 2006).

To conduct a *t* test for equivalence using confidence intervals:

(1) Check that the data meet assumptions for *t* tests.

(2) Set the equivalence interval $-\Delta$ to $+\Delta$ based on biological understanding of the system at hand (for advanced applications, an asymmetric equivalence interval can be defined; see Wellek 2002).

(3) Calculate the confidence interval:

confidence limits = *estimate* ± *critical value* x *standard error*

(a) The estimate is the difference between the means of the two sample distributions, which can be set in units of the data (eg meters) or in standard deviation units.

(b) The *critical value* is the value that defines the limits of a standard normal distribution containing $(1-\alpha)$% (typically 95%) of the curve's area. For a two-tailed test with $\alpha = 0.05$, this critical value = 1.96. See Sokal and Rohlf (1995) for calculations in other contexts.

(4) Plot the confidence and equivalence intervals and assess significance following Figure 5 (Jones *et al.* 1996).

The interpretation of confidence intervals in the context of equivalence testing is usually straightforward, but can, in some cases, lead to conflicting test results. For example, in Figure 5, the confidence interval for example D does not cross zero, indicating a significant difference between the two groups. However, the confidence interval is also completely contained within the equivalence interval, indicating significant homogeneity between the two groups.

How is this result – that two groups are both significantly different and significantly homogeneous – possible? Such a result can occur with a very small but significant difference between two groups, when the difference is small enough to be deemed not important by the researcher. It can, in part, be explained by the concept that, with enough samples from two groups, one is likely to find a significant difference between them, however small its magnitude. For example, imagine that scientists or other US FWS personnel have deemed a difference of ≥ 1% of genomic variation to constitute a taxonomic difference worthy of protected legal status (this is simplistic, but illustrates the point). Thus, two populations of the same species, if sampled enough, might be significantly different from one another using a traditional hypothesis test, while still having less than 1% genomic difference (and thus also being significantly homogeneous using an equivalence test). In such a case, we would generally suggest using the result from the equivalence procedure, since it relates back to the difference that scientists or policy makers care about; it can also help to prevent problems from significant results in traditional tests that come about only because of extremely large sample sizes. Note that such an outcome (when a result is both significantly different and significantly homogeneous) is unlikely in most ecological contexts.

The opposite result – that the two groups are neither significantly different nor significantly homogeneous – is also possible (Figure 5, examples E and F), and indicates that more data are needed to reach a reliable conclusion. The possibility of such a result – not just in the context of equivalence testing – should be communicated to policy makers as a normal, even common outcome of scientific studies.

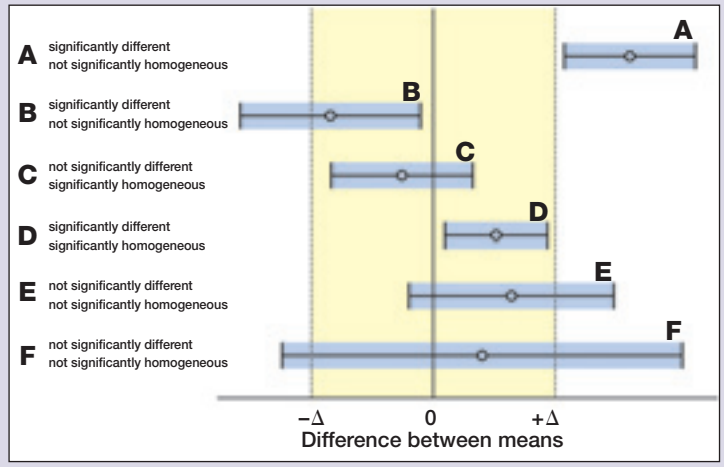

Figure 5. *Hypothetical confidence intervals for the difference between two means, following Jones* et al. *(1996). Each example (A–F) represents a confidence interval derived from the difference between two sample means. Confidence intervals contained completely within the equivalence region $-\Delta$ to $+\Delta$ are significantly homogeneous; confidence intervals not overlapping zero are significantly different. Examples A and B have confidence intervals that do not overlap zero and are thus significantly different; neither confidence interval is contained completely within the equivalence region, and therefore neither is significantly homogeneous. In example C, the confidence interval is contained within the equivalence interval and also overlaps zero, and thus is significantly homogeneous (and not significantly different). The confidence interval of D is contained within the equivalence region but does not cross zero, meaning that it is both significantly different and significantly homogeneous; see the text of Panel 2 for interpretation. In examples E and F, the confidence intervals overlap both zero and the edges of the equivalence interval; therefore, neither result is significantly different or significantly homogeneous.*

nificant difference between two essentially identical populations (see also Panel 2 on equivalence testing). Usually, this isn't an issue in ecological studies, because such large sample sizes are hard to come by in ecological contexts. As a matter of course, scientists should conduct a priori estimates of statistical power to assess the proper and practical level of sampling necessary to answer the question at hand; such methods are widely available for classical hypothesis tests and even for equivalence tests (eg Jones *et al.* 1996).

Incentives and inferential error

Given the lengthy training of scientists and managers and the quality control of peer-review systems, why do we still see inappropriate interpretations of non-significant statistical test results? Much of the answer lies in the institutional incentives for both research scientists and regulatory agencies.

In "basic" or "fundamental" studies – the historical model underlying most scientific endeavors – the social benefit arises from the advance of knowledge, and avoiding Type I errors is the primary concern. A Type II error might result from the improper rejection of a new theory when statistical differences are not found. While slowing down the march of scientific progress, the Type II error does no other direct harm, and generates no benefit to the scientist who committed the error. On the other hand, a Type I error could lead to the allocation of scientific resources down an unpromising pathway, and would simultaneously generate undeserved benefits of increased prestige and recognition for the scientist(s) responsible. Accordingly, the scientific establishment has placed an emphasis on reducing Type I error to reduce perverse incentives for scientists proposing theories that are supported only by spurious statistical findings (NRC 1995; Lemons *et al.* 1997; Doremus 2005; Doremus and Tarlock 2005).

In contrast, in applied scientific studies, a Type II error can have substantial repercussions, for example when a life-saving drug is not found to have a benefit in a small sample-size study or when a species is lost forever because it is not found to be genetically distinct. The consideration of Type II error has therefore been connected by many commentators to a more robust use of the "precautionary principle" in policymaking (eg Lemons *et al.* 1997). The medical community has developed ways of addressing Type II error, including, among others, defining "clinical" significance as opposed to "statistical" significance (eg Kendall and Grove 1988), but the conservation science community has yet to follow suit.

Regulatory agencies such as the FWS also face pressures that discourage the consideration of Type II errors. In situations such as the Preble's debate, the consideration of Type II errors might generate regulatory restrictions that carry considerable economic impacts. Such impacts are often concrete, substantial, and borne by small groups (eg businesses faced with increased production costs because of regulation), while benefits from regulatory restrictions are often abstract (eg public gain from preventing an extinction). Thus, stakeholders facing restrictions have a greater incentive to organize and lobby the agencies or the US Congress to influence its decisions and prevent regulatory restrictions (Biber 2007). Accordingly, regulatory agencies (such as FWS) will systematically face pressure not to regulate, and to avoid or discount the analysis of Type II statistical errors that would provide support for regulation.

How might we overcome these institutional challenges? One option is through the judiciary, especially since most ESA decisions are litigated at some stage. As a general rule, courts could require agencies that have relied on a quantitative statistical analysis to explain the uncertainties in their analysis in both directions (ie the risk of both false positives and false negatives). For instance, where an agency has relied on the lack of statistical significance to make a management decision (a Type I error analysis), a court should require some consideration of the Type II error (eg with an equivalence test where appropriate). Agencies that failed to provide such an explanation in a clear manner, accessible to non-specialist judges, would be required to reconsider their decision.

There is precedent for such a course in the courts. In one recent decision, a court struck down the National Oceanic and Atmospheric Administration's (NOAA) decision to weaken the "dolphin-safe" tuna labeling standard, because the sample sizes in the agency's research were too small to provide adequate power (*Earth Island Institute v Hogarth*, 494 F 3d 757 [9th Circuit 2007]). In general, courts have the authority to strike down agency decisions that are irrational, fail to consider important factors, or are incoherent or inexplicable (eg *Motor Vehicle Manufacturers Association v State Farm Mutual Automobile Insurance Company*, 463 US 29 [1983]; *Federal Power Commission v Texaco*, 417 US 380 [1974]).

We emphasize that we are not calling for courts to require wildlife agencies to provide particular types of any statistical analysis in their decisions, but rather that agencies be required to explain the risk that their analysis failed to find differences that would have resulted in an alternative policy choice. In many cases, this could be done through Bayesian methods or equivalence tests. If special cases arise, where such relative uncertainties cannot be calculated quantitatively, agencies should give a clear and thorough qualitative assessment of the risks associated with both sides of their decision.

The benefits of a more active judicial role here are twofold. First, agencies would be provided with at least some incentive to be more attentive to the risk of Type II errors in their decision making. Second, it could help to encourage more thorough consideration of Type II errors by the scientific community.

Conclusion

Scientists play the most important role in reducing errors of interpretation from non-significant test results. People

conducting conservation studies or making management decisions should remember the importance of both Type I and II errors, the pitfalls of using post hoc power analysis to assess Type II errors, and the need to use alternatives such as equivalence testing. Editors and peer reviewers should carefully examine submissions for the inferential errors that we discuss here and, where appropriate, encourage the use of alternative statistical techniques. And educators should reinforce in their students an understanding of the limitations of statistical analyses in this area.

Ecologists can also play a role in helping educate policy makers, lawyers, and even the courts, about the statistical pitfalls discussed in this paper. For instance, scientists who are working on wildlife populations that are the subject of litigation can inform all parties about the risks of statistical inferential error through, for example, communications with the lawyers in the case, or even the submission of an amicus brief (a brief filed in court on behalf of outside groups who are not parties to the case, but seek to inform the court about important, relevant policy issues).

An increased focus on better statistical inference will have broad benefits by providing an impetus for a more intelligent and productive debate about our conservation priorities. In most cases, both power analysis and equivalence testing require a determination about what is the minimum level of change or difference ("effect size") that the analyst wishes to detect (Sokal and Rohlf 1995; Wellek 2002; Adelman 2004). In the context of decision making, a determination must be made of the *meaningful* effect size that could influence a management or policy change.

For instance, for agencies that rely on genetic differences to make decisions about protected taxa, equivalence testing *requires* setting a numeric threshold for the level of genetic differentiation that we are concerned about. Defining such a threshold is at some level subjective (eg Waples and Gaggiotti 2006); thus, methods such as equivalence testing force us to recognize that genetics alone cannot answer the question of "how much" difference is enough to warrant protection.

Accordingly, proper consideration of Type II errors opens the door to discussions about a range of other values besides genetic variation and evolutionary trajectory. On the one hand, selecting a very low threshold of difference between taxa could result in the identification of many more conservation units (taxa) for protection, with potentially important economic and political consequences. At the same time, reliably resolving small differences statistically involves large sample sizes. On the other hand, selection of a higher threshold to justify protection might result in the disappearance of not just valuable genetic diversity, but also ecosystems that might otherwise be protected through the conservation of the population in question, and the loss of important aesthetic and cultural values associated with certain plant and animal populations (eg, the persistence of bald eagles, grizzly bears, and gray wolves in the conterminous US). We are not arguing that genetic information is irrelevant, only that better statistical inference might bring more balance to a debate that, in the case of the Preble's meadow jumping mouse, was too focused on genetics and not on the other important values that necessarily inform our conservation efforts (Taylor and Dizon 1996; Doremus 2005; Doremus and Tarlock 2005).

To the extent that policy makers and scientists rely on the reporting of only Type I errors and levels of statistical significance to make these types of decisions, they allow themselves to avoid these hard discussions. While that may be the easy road in the short term, we believe that in the long run it ill serves efforts to conserve our valuable biodiversity resources.

Acknowledgements

We thank H Doremus, S Fallon, D Farber, C Lunch, M Pinsky, J Sax, and D Skelly for helpful discussion and comments on earlier drafts of the manuscript. G Daily and her lab at Stanford also provided stimulating ideas and discussion. BJB gratefully acknowledges support from the Moore Family Foundation.

References

Adelman DE. 2004. Scientific activism and restraint: the interplay of statistics, judgment, and procedure in enviromental law. *Notre Dame Law Rev* **79**: 497–583.

Biber E. The importance of resource allocation in administrative law: a case study of agency inaction under the Administrative Procedure Act. *Admin Law Rev* **60**: in press.

Colorado Division of Wildlife. 2004. Peer review of Ramey *et al.* 2004. www.fws.gov/mountain-prairie/endspp/peerreview/PeerReviewPrebles.htm. Viewed 16 Dec 2007.

Doremus H. 2005. Science plays defense: natural resource management in the Bush Administration. *Ecol Law Quart* **32**: 249–305.

Doremus H and Tarlock AD. 2005. Science, judgment, and controversy in natural resource regulation. *Public Land Resour Law Rev* **26**: 1–37.

Fallon SM. 2007. Genetic data and the listing of species under the US Endangered Species Act. *Conserv Biol* **21**: 1186–95.

Fidler F, Burgman MA, Cumming G, *et al.* 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv Biol* **20**: 1539–44.

Goldstein PZ, LaSalle R, Amato G, and Vogler AP. 2000. Conservation genetics at the species boundary. *Conserv Biol* **14**: 120–31.

Goodman SN and Berlin JA. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* **121**: 200–06.

Hoenig JM and Heisey DM. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat* **55**: 19–24.

Holthouse D. 2005. Building a better mousetrap. Denver Westword. Jan 20: www.westword.com/2005-01-20/news/building-a-better-mousetrap/. Viewed 18 Aug 2008.

Jones B, Jarvis P, Lewis JA, and Ebbutt AF. 1996. Trials to assess equivalence: the importance of rigorous methods. *Brit Med J* **313**: 36–45.

Kendall PC and Grove WM. 1988. Normative comparisons in therapy outcome. *Behav Assess* **10**: 147–58.

King TL, Switzer JF, Morrison CL, *et al.* 2006. Comprehensive genetic analyses reveal evolutionary distinction of a mouse

(*Zapus hudsonius preblei*) proposed for delisting from the US Endangered Species Act. *Mol Ecol* **15**: 4331–59.

Lemons J, Shrader-Frechette K, and Cranor C. 1997. The precautionary principle: scientific uncertainty and Type I and Type II errors. *Found Sci* **2**: 207–36.

McGarvey DJ. 2007. Merging precaution with sound science under the Endangered Species Act. *BioScience* **57**: 65–70.

NRC (National Research Council). 1995. Science and the Endangered Species Act. Washington, DC: National Academies Press.

Palsbøll PJ, Berube M, and Allendorf FW. 2006. Identification of management units using population genetic data. *Trends Ecol Evol* **22**: 11–16.

Ramey RR, Liu HP, Carpenter LM, and Epps CW. 2004. Testing the uniqueness of *Z h intermedius* relative to *Z h campestris*. Denver, CO: Denver Museum of Nature and Science. Technical report 2004–8.

Ramey RR, Liu HP, Epps CW, *et al.* 2005. Genetic relatedness of the Preble's meadow jumping mouse (*Zapus hudsonius preblei*) to nearby subspecies of *Z hudsonius* as inferred from variation in cranial morphology, mitochondrial DNA and microsatellite DNA: implications for taxonomy and conservation. *Anim Conserv* **8**: 329–46.

Robinson A. 2008. Equivalence: provides tests and graphics for assessing tests of equivalence. Package for the R Statistical Computing Language. http://cran.r-project.org/web/packages/equivalence/index.html. Viewed May 2008.

Ruhl JR. 2004. The battle over Endangered Species Act methodology. *Environ Law* **34**: 555–603.

Sokal RR and Rohlf FJ. 1995. Biometry: the principles and practice of statistics in biological research, 3rd edn. New York, NY: WH Freeman and Company.

Taylor BL and Dizon AE. 1996. The need to estimate power to link genetics and demography for conservation. *Conserv Biol* **10**: 661–64.

Toft CA and Shea PJ. 1983. Detecting community-wide patterns: estimating statistical power strengthens inference. *Am Nat* **122**: 618–25.

US FWS (US Fish and Wildlife Service) and NOAA (National Oceanic and Atmospheric Administration). 1996. Policy regarding the recognition of distinct vertebrate population segments under the Endangered Species Act. Washington, DC: 61 Fed Reg 4722.

US FWS (US Fish and Wildlife Service). 2002. Recovery goals: amendment and supplement to the humpback chub recovery plan. Phoenix, AZ: US FWS.

Wade PR. 2000. Bayesian methods in conservation biology. *Conserv Biol* **14**: 1308–16.

Waples RS and Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol Ecol* **15**: 1419–39.

Wellek S. 2002. Testing statistical hypotheses of equivalence. Boca Raton, FL: CRC Press.

