Controlling the false discovery rate and increasing statistical power in ecological studies

Author(s): Thomas A. Waite and Lesley G. Campbell
Source: Ecoscience, 13(4):439-442.
Published By: Centre d'études nordiques, Université Laval
https://doi.org/10.2980/1195-6860(2006)13[439:CTFDRA]2.0.CO;2
URL: http://www.bioone.org/doi/full/10.2980/1195-6860%282006%2913%5B439%3ACTFDRA%5D2.0.CO%3B2

ECOSCIENCE 13 (4): 439-442 (2006)

Controlling the false discovery rate and increasing statistical power in ecological studies[1]

Thomas A. WAITE[2] & Lesley G. CAMPBELL, Department of Evolution, Ecology, and Organismal Biology,
Ohio State University, 318 W. 12th Avenue, Columbus, Ohio 43210, USA, e-mail: waite.1@osu.edu

Abstract: Ecologists routinely use Bonferroni-based methods to control the alpha inflation associated with multiple hypothesis testing, despite the aggravating loss of power incurred. Some critics call for abandonment of this approach of controlling the familywise error rate (FWER), contending that too many unwary researchers have adopted it in the name of scientific rigour even though it often does more harm than good. We do not recommend rejecting multiplicity correction altogether. Instead, we recommend using an alternative approach. In particular, we advocate the Benjamini–Hochberg and related methods for controlling the false discovery rate (FDR). Unlike the FWER approach, which safeguards against falsely rejecting even a single null hypothesis, the FDR approach controls the rate at which null hypotheses are falsely rejected (*i.e.*, false discoveries are made). The FDR approach represents a compromise between outright refusal to control for multiplicity, which maximizes alpha inflation, and strict adherence to FWER control, which minimizes power. We review the multiplicity problem, illustrate the advantage of the FDR approach, and promote this approach for widespread adoption in ecology.
Keywords: Benjamini–Hochberg method, false discovery rate, familywise error rate, multiple comparisons, sequential Bonferroni.

Résumé : Les écologistes emploient régulièrement les méthodes de Bonferroni pour contrôler l'inflation du risque alpha associé aux tests d'hypothèses multiples, et ce malgré la perte de puissance qui s'en suit. Certains critiques suggèrent d'abandonner complètement cette approche de contrôle du taux d'erreur global (FWER : Family Wise Error Rate), soutenant que trop de chercheurs imprudents l'ont adoptée au nom de la rigueur scientifique, bien qu'elle fasse souvent plus de mal que de bien. Nous ne recommandons pas d'abandonner entièrement la correction de multiplicité. Au lieu de cela, nous recommandons d'utiliser une approche alternative. En particulier, nous plaidons pour l'approche de Benjamini-Hochberg et les méthodes apparentées pour contrôler le taux de fausses découvertes (FDR : False Discovery Rate). Contrairement au contrôle du FWER, qui protège contre le rejet à tort de même une seule hypothèse nulle, l'utilisation du FDR contrôle le taux auquel des hypothèses nulles sont faussement rejetées (c.-à-d., le taux de fausses découvertes). Le contrôle du FDR représente un compromis entre le refus absolu de contrôler la multiplicité, ce qui maximise l'inflation du risque alpha, et l'adhésion stricte au contrôle du FWER, ce qui minimise la puissance. Nous réexaminons le problème des tests multiples, illustrons les avantages du contrôle du FDR, et encourageons une vaste adoption de cette approche en écologie.
Mots-clés : Comparaisons multiples, correction séquentielle de Bonferroni, méthode de Benjamini-Hochberg, taux de fausses découvertes, taux d'erreur global.

Introduction

Most ecologists are all too familiar with the so-called multiplicity problem. Having performed multiple statistical tests, we routinely use a criterion for significance more stringent than the conventional critical α (*i.e.*, 0.05). If we don't, we run the risk of being cajoled by reviewers and editors to do so. Routinely, this multiplicity correction uses the sequential Bonferroni procedure (Holm, 1979), popularized among ecologists by Rice (1989). This procedure simply involves ordering observed *P*-values from smallest to largest and then evaluating the potential significance of each test. If the smallest observed *P*-value is less than $0.05/m$, where m is the number of tests performed, this test is considered to be significant. If the next smallest *P*-value is less than $0.05/(m\text{-}1)$, this test is also considered to be significant, and so on. The advantage of this procedure is that it controls the familywise error rate (FWER) at 0.05, permitting the practitioner to conclude with 95% certainty that none of the rejected null hypotheses has been wrongly rejected.

Some critics call for the abandonment of this approach (Perneger, 1998; 1999; Moran, 2003; see also Cabin & Mitchell, 2000; Feise, 2002). At the top of their list of grievances is the following claim. Whether a truly significant phenomenon is interpreted as such depends on the number of tests conducted, m. Any given biologically significant outcome is increasingly likely to be dismissed as m increases. Stated another way, multiplicity correction inflates the probability of type II error, and potentially leads to flawed statistical inference. These critics call for an outright rejection of multiplicity correction, suggesting that simply reporting which tests were done and allowing the reader to make judgments is a preferable alternative.

We recommend a less drastic measure. Rather than supporting this call to abandon multiplicity correction, we join others in promoting a more powerful alternative to Bonferroni-based correction (Benjamini *et al.*, 2001; García, 2003; 2004; Reiner, Yekutieli & Benjamini, 2003; Storey & Tibshirani, 2003; van den Oord & Sullivan, 2003; Verhoeven, Simonsen & McIntyre, 2005a,b). Specifically, we promote a procedure (Benjamini & Hochberg, 1995) that controls the false discovery rate (FDR) rather than FWER. This

[1]Rec. 2005-07-01; acc. 2006-02-21.
Associate Editor: Marco A. Rodríguez.
[2]Author for correspondence.

Benjamini–Hochberg (BH) procedure controls for multiplicity while maintaining greater power than Bonferroni-based procedures. In the next section, we will illustrate the power advantage and discuss some limitations of the BH procedure.

Methods for coping with multiplicity

Bonferroni-based methods for controlling FWER ensure with, say, 95% certainty that all rejected null hypotheses are correctly rejected. That is, these methods allow the practitioner to say, with 95% certainty, that there are no false discoveries whatsoever. By safeguarding against wrongly rejecting even a single null hypothesis, these methods suffer from reduced power. The probability that any given test will be deemed significant is compromised. By contrast, the Benjamini–Hochberg (BH) method for controlling FDR ensures that only, say, 5% of rejected null hypotheses are false discoveries (for proof see Benjamini & Hochberg, 1995). This procedure represents a compromise between the need to correct for multiplicity and the need to conserve power.

Trading power against false discoveries

The increased power of the BH method comes at a price. This method is not designed to ensure with some high degree of certainty that zero false discoveries are made. Instead, the number of false discoveries will tend to rise with m. To clarify this point, recall that the sequential Bonferroni method ensures with, say, 95% certainty that no false discoveries are made, regardless of the number of tests conducted. So, the probability of making even a single false discovery is held at 5%, regardless of the number of tests conducted. By contrast, the BH method holds the percentage of false discoveries at some preset level, say, 5% and thereby permits the number of false discoveries to increase with the number of tests conducted. More precisely, the probability of making at least one false discovery increases with the number of nominally significant P-values. (According to the cumulative binomial distribution, this probability reaches 50% if there are 14 such P-values, 90% if there are 45 such values, and so on.) Naturally, the probabilities of making larger numbers of false discoveries likewise increase predictably with the number of nominally significant Ps. The BH method should be used judiciously, with this caution in mind.

Recipe for computation and a simple hypothetical example

The basic Benjamini–Hochberg algorithm for calculating sequential thresholds of significance is simply iq/m, where i is the i^{th} observed P-value (ordered from smallest to largest, 1, 2,..., m), q is the assigned false discovery rate (*e.g.*, 0.05), and m is the number of tests conducted (see comments below regarding how to define m). We provide an example below.

Consider the following set of observed P-values: 0.002, 0.012, 0.020, 0.030, 0.040, and 0.060. The thresholds for significance are identical for the smallest observed P-value, P_1. That is, the Bonferroni threshold is $0.05/m = 0.05/6$ and the BH threshold is $iq/m = (1)(0.05)/6 = 0.05/6$. The thresholds are also identical for the largest observed P-value, P_m: $0.05/(m - 5) = 0.05/(6 - 5) = 0.05$ and $(6)(0.05)/6 = 0.05$. However, for intermediate P-values, the thresholds are more stringent for the (sequential) Bonferroni than BH procedure. For example, for P_2 the Bonferroni threshold is $0.05/(m - 1) = 0.05/5 = 0.01$, whereas the BH threshold is $(2)(0.05)/6 = 0.0167$. The hypothetical observed P-value (0.012) falls between these thresholds. Likewise, each of the next three P-values falls between the thresholds. Thus, five of six tests are significant when FDR is controlled at 0.05 and just one of six tests is significant when FWER is controlled at 0.05.

How to determine m?

What is m? Is it the number of tests (P-values) in a single statistical table? Is it the total number of tests in a single paper? Is it the cumulative number of tests conducted over the course of one's career? These questions have been discussed extensively elsewhere (Perneger, 1998). We do not intend to resolve the debate over m's definition here. Instead, we simply recommend that however one might ordinarily determine m for a Bonferroni-based correction for multiplicity, the same m could be used for a BH correction.

Why have ecologists failed to reject Bonferroni?

If the FDR approach is really a preferable alternative, why does Bonferroni continue to dominate in ecological journals? Why has the BH procedure caught on in some fields but, until very recently, not in ecology? Here, we entertain three explanations, beyond the obvious possibility that the vast majority of ecologists remain unaware of BH. First, some ecologists may argue that we routinely do so few multiple tests (*i.e.*, m is so small) that the BH alternative would have little to offer. However, a casual inspection of recent papers in ecology journals suggests otherwise, with m routinely exceeding 10 and often 100. Thus, it appears that ecologists could substantially improve their statistical power by controlling FDR rather than FWER.

Second, some ecologists may argue that the FDR approach is philosophically inappropriate for our field. To counter we simply ask whether we routinely wish to safeguard against even a single false positive, thereby minimizing our ability to identify true positives? We concede that there will be some scenarios in which controlling FWER is more compelling. However, we think that the less conservative approach of controlling the rate of false positives, and thereby improving our ability to detect true positives, will generally provide more meaningful inferences.

Finally, we contend that the Bonferroni method has been a successful meme in ecology, so successful that the field has been resistant to invasion by the Benjamini–Hochberg mutant. Adoption of BH has probably been met with less resistance in various fields (journals) lacking a strong history of multiplicity correction or where the FDR approach is so obviously appropriate (*e.g.*, papers on DNA microarray analysis [Reiner, Yekutieli & Benjamini, 2003] in Bioinformatics or functional neuroimaging in Neuroimaging, Neuron, and others). Even so, we think this inertia can be easily overcome and that we can induce a successful invasion by the FDR approach.

Potential for widespread application

What can Benjamini–Hochberg do for ecology? Many empirical studies published in ecology journals use the

sequential Bonferroni procedure to correct for multiplicity. Consider, for example, Molecular Ecology. This journal has cited Rice (1989), the citation classic promoting this procedure, more than any other ecology journal. Many of the papers published in this journal include multiple testing aimed at determining on a locus-by-locus basis whether Hardy–Weinberg equilibrium is satisfied. Many other papers include multiple testing aimed at detecting pairwise genetic differentiation between populations. These are general situations in which the BH method could be applied. Other potential applications in molecular ecology include studies aimed at (1) detecting specific quantitative trait loci (QTL) that are under selection by using coalescent-based simulation and applying an outlier test for each of many markers (Storz & Nachman, 2003), (2) determining degree of kinship for many pairs of individuals within a population (Gautschi *et al.*, 2003), and (3) identifying episodes of poaching by using assignment tests for many individuals (Paetkau *et al.*, 2004). We think these applications are compelling because they typically involve large numbers of tests and because they may be exploratory in nature. However, we emphasize that any study involving multiple testing, in any ecological subdiscipline, could benefit from controlling FDR rather than FWER.

Indeed, since we produced the first draft of this paper, the BH method has made its debut in a few published ecological studies, spanning a diverse array of topics. Examples include studies on spatial distributions of bird species (Roback & Askins, 2005), food preferences of insects (Bluthgen, Metzner & Ruf, 2006), ecological effects of toxins on multiple trophic levels (Raimondo & McKenney, 2006), fluctuating asymmetry in multiple traits in isopods (Vilisics, Solymos & Hornung, 2005), taxon-specific responses by moths to a successional gradient (Hilt & Fiedler, 2006), geographic variation in interactions among parasitoid species (Batchelor, Hardy & Barrera, 2006), and paradoxical decision-making in birds (Waite & Passino, 2006). Thus, it appears that the BH approach is on the verge of making its mark across the full spectrum of ecological subdisciplines.

When controlling FDR may be unnecessary or inappropriate

In promoting the BH method as an alternative to Bonferroni-based methods, we do not mean to imply that correcting for multiplicity is universally appropriate. In some cases, it may be compelling to control neither FWER nor FDR (Roback & Askins, 2005). Instead, it may be appropriate to simply interpret each nominal *P*-value without regard to the number of tests conducted. (It may even be appropriate to set the critical α-level higher than the conventional 0.05.) Imagine, for instance, an exploratory analysis aimed at identifying correlates of extinction risk. The very nature of such an analysis, where the aim might be to screen some large number of potential correlates, suggests that it could be philosophically inappropriate and even unethical to correct for multiplicity. Doing so would minimize false positives, but could obscure potentially causal relationships and indirectly lead to misguided conservation action. However, we do recommend BH as an alternative to Bonferroni-based methods in studies involving rigorous hypothesis testing rather than exploratory analysis.

Alternatives to conventional statistical inference

In recent years, ecologists have begun to embrace statistical paradigms celebrated for their putative superiority to traditional inferential statistics (*e.g.*, Bayesian [Anderson, 1998] and likelihood methods [Royall, 1997]). In doing so, we have adopted, intentionally or otherwise, alternative approaches to coping with the multiplicity problem. For example, consider the rapidly growing popularity of the information-theoretic approach to model selection (Burnham & Anderson, 2002). By using the Akaike Information Criterion (AIC) to identify the best combination(s) of predictor variables, we not only avoid conventional *P*-values altogether, but we also invoke the principle of parsimony. By penalizing each additional predictor variable, this approach effectively minimizes the number of falsely positive predictors included in models.

However, these alternatives to conventional inferential statistics are no panacea to the multiplicity problem. Imagine, for instance, an AIC-based study aimed at evaluating whether climate change has contributed to the demographic decline of some species. Even though models with multiple predictor variables would be penalized, the chance of finding spurious correlates would increase with the number of climatic variables considered (Burnham & Anderson, 2002). Likewise, such fishing expeditions can compromise Bayesian inferences (Klugkist, Laudy & Hoijtink, 2005). By exploring too many possibilities, the likelihood of finding something interesting but spurious increases. Thus, concerns over false discoveries are not limited to conventional statistical analysis.

Final remarks and recommendations

Beyond recommending the Benjamini–Hochberg procedure for controlling false discoveries, we alert the reader to the recent development of new theory and variants (Black, 2004; Genovese & Wasserman, 2004; Storey, Taylor & Siegmund, 2004), including those designed to cope with multiple dependent tests (Benjamini & Yekutieli, 2001; Fernando *et al.*, 2004). Regardless of the specific algorithm used, the FDR approach provides greater power than the conventional Bonferroni method and even its refined variants (see Table I in García, 2004 for a list of algorithms; see also Nichols & Hayasaka, 2003).

What course of action would help accelerate the adoption of FDR-controlling methods in ecology? First, it would be helpful if authority figures such as journal editors and graduate advisors were to promote the FDR approach. The information-theoretic approach to model selection (Burnham & Anderson, 2002) has successfully invaded ecological analysis, and we anticipate similar success for FDR methods. Second, although free software for FDR methods is already available, it would be helpful if software for analyzing ecological data were to incorporate FDR options. At a minimum, it would be useful if various programs were modified to provide exact *P*-values. With these in hand, it would be straightforward to apply FDR correction. We

foresee these developments leading to improved inference in ecology. By overcoming our fear of false discoveries, we stand to achieve greater power without damaging the credibility of our work.

Acknowledgements

We thank L. Gibbs for encouraging this contribution, M. Bélisle and P. Peres-Neto for helpful comments, and Stauf's, Seven Main, and the University of Michigan Biological Station for providing pleasant workspace.

Literature cited

Anderson, J. L., 1998. Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. Conservation Ecology, 2(1): 2. [Online] URL: http://www.consecol.org/vol2/iss1/art2/

Batchelor, T. P., I. C. W. Hardy & J. F. Barrera, 2006. Interactions among bethylid parasitoid species attacking the coffee berry borer, *Hypothenemus hampei* (Coleoptera: Scolytidae). Biological Control, 36: 106–118.

Benjamini, Y. & Y. Hochberg, 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B, 57: 289–300.

Benjamini, Y. & D. Yekutieli, 2001. The control of the false discovery rate under dependency. Annals of Statistics, 29: 1165–1188.

Benjamini, Y., D. Drai, G. Elmer, N. Kafkafi & I. Golani, 2001. Controlling the false discovery rate in behavior genetics research. Behavioural Brain Research, 125: 279–284.

Black, M. A., 2004. A note on the adaptive control of false discovery rates. Journal of the Royal Statistical Society B, 66: 297–304.

Bluthgen, N., A. Metzner & D. Ruf, 2006. Food plant selection by stick insects (Phasmida) in a Bornean rain forest. Journal of Tropical Ecology, 22: 35–40.

Burnham, K. P. & D. R. Anderson, 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd Edition. Springer-Verlag, New York, New York.

Cabin, R. J. & R. J. Mitchell, 2000. To Bonferroni or not to Bonferroni: When and how are the questions. ESA Bulletin, 81: 246–248.

Feise, R. J., 2002. Do multiple outcome measures require *P*-value adjustment? BMC Medical Research Methodology, 2: 8. [Online] URL: http://www.biomedcentral.com/1471-2288/2/8

Fernando, R. L., D. Nettleton, B. R. Southey, J. C. M. Dekkers, M. F. Rothschild & M. Soller, 2004. Controlling the proportion of false positives (PFP) in multiple dependent tests. Genetics, 166: 611–619.

García, L. V., 2003. Controlling the false discovery rate in ecological research. TRENDS in Ecology and Evolution, 18: 553–554.

García, L. V., 2004. Escaping the Bonferroni iron claw in ecological studies. Oikos, 105: 657–663.

Gautschi, B., G. Jacob, J. J. Negro, J. A. Godoy, J. P. Muller & B. Schmid, 2003. Analysis of relatedness and determination of the source of founders in the captive bearded vulture, *Gypaetus barbatus*, population. Conservation Genetics, 4: 479–490.

Genovese, C. & L. Wasserman, 2004. A stochastic process approach to false discovery control. Annals of Statistics, 32: 1035–1061.

Hilt, N. & K. Fiedler, 2006. Arctiid moth ensembles along a successional gradient in the Ecuadorian montane rain forest zone: How different are subfamilies and tribes? Journal of Biogeography, 33: 108–120.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6: 65–70.

Klugkist, I., O. Laudy & H. Hoijtink, 2005. Bayesian eggs and Bayesian omelettes: Reply to Stern (2005). Psychological Methods, 10: 500–503.

Moran, M. D., 2003. Arguments for rejecting the sequential Bonferroni in ecological studies. Oikos, 100: 403–405.

Nichols, T. & S. Hayasaka, 2003. Controlling the familywise error rate in functional neuroimaging: A comparative review. Statistical Methods in Medical Research, 12: 419–446.

Paetkau, D., R. Slade, M. Burden & A. Estoup, 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: A simulation-based exploration of accuracy and power. Molecular Ecology, 13: 55-65.

Perneger, T. V., 1998. What's wrong with Bonferroni adjustments. British Medical Journal, 316: 1236–1238.

Perneger, T. V., 1999. Adjusting for multiple testing in studies is less important than other concerns. British Medical Journal, 318: 1288.

Raimondo, S. & C. L. McKenney, 2006. From organisms to populations: Modeling aquatic toxicity data across two levels of biological organization. Environmental Toxicology and Chemistry, 25: 589–596.

Reiner, A., D. Yekutieli & Y. Benjamini, 2003. Identifying differential genes using false discovery rate controlling procedures. Bioinformatics, 19: 368–375.

Rice, W. R., 1989. Analyzing tables of statistical tests. Evolution, 43: 223–225.

Roback, P. J. & R. A. Askins, 2005. Judicious use of multiple hypothesis tests. Conservation Biology, 19: 261–267.

Royall, R., 1997. Statistical evidence: A likelihood paradigm. Chapman & Hall/CRC, New York, New York.

Storey, J. D. & R. Tibshirani, 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Science, 100: 9440–9445.

Storey, J. D., J. E. Taylor & D. Siegmund, 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. Journal of the Royal Statistical Society B, 66: 187–205.

Storz, J. F. & M. W. Nachman, 2003. Natural selection on protein polymorphism in the rodent genus *Peromyscus*: Evidence from interlocus contrasts. Evolution, 57: 2628–2635.

Van den Oord, E. J. C. G. & P. F. Sullivan, 2003. False discoveries and models for gene discovery. TRENDS in Genetics, 19: 537–542.

Verhoeven, K. J. F., K. L. Simonsen & L. M. McIntyre, 2005a. Implementing false discovery rate control: Increasing your power. Oikos, 108: 643–647.

Verhoeven, K. J. F., K. L. Simonsen & L. M. McIntyre, 2005b. Implementing false discovery rate control: Increasing your power (vol 108, pg 643, 2005). Oikos, 109: 208.

Vilisics, F., P. Solymos & E. Hornung, 2005. Measuring fluctuating asymmetry of the terrestrial isopod *Trachelipus rathkii* (Crustacea: Isopoda, Oniscidea). European Journal of Soil Biology, 41: 85–90.

Waite, T. A. & K. M. Passino, 2006. Paradoxical preferences when options are identical. Behavioral Ecology and Sociobiology, 59: 777-785.