Judicious Use of Multiple Hypothesis Tests

PAUL J. ROBACK* AND ROBERT A. ASKINS+

* Department of Mathematics, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057, U. S. A.

+ Department of Biology, Connecticut College, New London, CT 06320, U.S.A.

Running Head: Multiple Testing Methods

Key words: multiple hypothesis testing, Bonferroni Method, False Discovery Rate, Holm's method, hypothesis generation, q-value

Word count: 4,438

Send correspondence and proofs to Robert Askins, Department of Biology, Connecticut College, 270 Mohegan Avenue, New London, CT 06320-4196;

Phone: 860-439-2149; e-mail raask@conncoll.edu

Abstract

When analyzing a table of statistical results, one must first decide whether adjustment of significance levels is appropriate. If the main goal is hypothesis generation or initial screening for potential conservation problems, then it may be appropriate to use the standard comparisonwise significance level to avoid Type 2 errors (not detecting real differences or trends). If, however, the main goal is rigorous testing of a hypothesis, then an adjustment for multiple tests is needed. To control the familywise Type 1 error rate (the probability of rejecting at least one true null hypothesis), sequential modifications of the standard Bonferroni Method, such as Holm's method, will provide more statistical power than the standard Bonferroni method. Additional power may be achieved by using procedures that control the False Discovery Rate (the expected proportion of false positives among tests found to be significant). When the Holm's method and two different false discovery rate procedures (FDR and pFDR) were applied to the results of multiple regression analyses of the relationship between habitat variables and abundance for 25 species of forest birds in Japan, the pFDR procedures provided the greatest statistical power.

Introduction

After the publication of Rice's (1989) paper on analyzing tables of statistical tests, reviewers and editors for field biology journals became more concerned about the problem of significant results that occur by chance when a large number of statistical tests are completed. If the significance level for each test (α) is set at

0.05, then one in every 20 tests in which there is actually no difference or effect will be significant by chance. Hence, if scores of tests are completed, a large number of significant results may be spurious. The recommended solution is applying the Bonferroni Method or related procedures that set the "familywise" or "experimentwise" significance level at α rather than using the standard "comparisonwise" significance level appropriate for a single, isolated statistical test.

As an example, Kurosawa and Askins (2003) recently investigated the effect of forest fragmentation on bird populations in southern Hokkaido, Japan. In one analysis, the authors assessed the effects of forest area and isolation on the abundance of 25 common species using multiple regression methods to control for habitat variables such as canopy height, herb cover, shrub cover, and conifer cover. When testing model significance for each species, the authors used Holm's Sequential Bonferroni Method to adjust their levels of significance. As a result, they found only 2 species in which the multiple regression model significantly explained variability in species abundance, despite having 9 p-values below 0.05, 6 of which were below 0.01.

Conventional wisdom demands the sort of multiple testing adjustments performed by Kurosawa and Askins; otherwise, spurious conclusions (declaring results significant when they really are not) become all too common. The standard Bonferroni method controls the familywise error rate by simply dividing

the level of significance α by the number of tests *n*. Modifications such as Holm's Sequential Bonferroni Method can increase the statistical power of multiple tests (Holm 1979; Rice 1989; Wright 1992; Shaffer 1995). Nonetheless, researchers often suspect that these procedures are too conservative, making it difficult to detect valid differences (Saville 1990; Benjamini et al. 2001). More recently developed alternatives that control the False Discovery Rate (Benjamini & Hochberg 1995; Storey 2002) can provide marked increases in power over sequential Bonferroni methods.

A more fundamental consideration is whether adjustments in the significance level are appropriate in all cases of multiple testing (Saville 1990; Crabbe et al. 1999). As Tukey (1991) wrote: "We do not dare work at very high error rates. We should not try to work at very low ones. We need to work in the range where error rates make an appreciable contribution to the "fuzz" that is always involved in our knowledge or belief". This "fuzz" is tolerable because, as Tukey emphasized, "truly solid knowledge" comes from repeated confirmation from numerous studies. Decisions about whether adjusted significance levels in multiple tests are appropriate and, if so, how testing power can be maximized, require careful consideration. In particular, attention must be paid to the nature of the multiple tests (hypothesis generating or hypothesis confirming), the responses of interest (specific items or general patterns), and the error rate the researchers desire to control (comparisonwise, familywise, or false discoveries).

Screening and Confirmatory Testing

The appropriateness of adjustments for multiple tests for clinical trials has been the subject of ongoing debate and study by pharmaceutical researchers. Following ICH (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use) guidelines, pharmaceutical researchers make a distinction between screening analyses for safety and confirmatory analyses for efficacy, and have reached a consensus that the Bonferroni method and other multiple-testing adjustments generally should apply only to the latter (Food and Drug Administration 1998:33-37). Screening tests for safety provide information about whether a compound produces deleterious side effects. Typically, a large number of potential side effects are recorded and analyzed, resulting in a large table of related test results. This is the type of situation in which the Bonferroni correction is often applied, but in this case it would result in few side effects showing significant results at the familywise level. Detecting false positives is a less serious issue than disregarding potential side effects, so application of the Bonferroni method is considered inappropriate. Researchers are aware that some of the apparently significant results will be due to chance, but all side effects that show a significantly higher frequency in the test group than in the control group will become the subject of more focused research, resulting in a higher level of drug safety.

In confirmatory analyses for efficacy, where the effectiveness of a compound is evaluated, the philosophy is different. In such analyses it is important to show rigorously that the compound has a substantial positive effect on health. These analyses may draw on data from the same trials as screening tests for safety, but typically they are more focused statistical analyses with fewer variables or comparisons, and a Bonferroni or similar multiple-comparison adjustment in significance levels is applied (Food and Drug Administration 1998:33-37). In this case, false positives are a serious problem because they could result in the production of a useless pharmaceutical.

Ecological research involves analogues to safety (screening) and efficacy (confirmatory) testing. For example, researchers in the Hokkaido study compared the distributions of a large number of species in habitat patches of different sizes to determine if there is a set of species that are potentially affected by habitat fragmentation. After this broad-scale screening process, the next step would be more intensive and focused studies of the life history, demography, and distribution of these species to more rigorously test whether they are affected by habitat fragmentation and, if so, to determine the causes. While the Bonferroni correction would be appropriate in these focused studies, it is unnecessarily conservative for the initial screening. Insisting on Bonferroni-type corrections for all tables of statistical test results will prevent the type of large-scale, exploratory studies that help identify important questions for more intensive studies. The distinction between "screening" and "confirmatory" studies is not normally made

by ecologists, who tend to either ignore multiple-testing adjustments altogether or apply them to all tables of test results.

The distinction between "screening" and "confirmatory" studies applies particularly well to research that has immediate relevance to conservation. "Screening" might involve the detection of potential population declines among a large group of species or of potential negative impacts on ecosystem functions following an environmental change. Using confidence intervals or comparisonwise significance levels for each test will reveal potential problems, each of which can be studied more intensively. In this situation it is important to reduce the frequency of Type 2 errors (false negatives) which might result in not detecting a serious population decline or ecological problem. The inevitable "false positives" (Type 1 errors) can be screened out later with more intensive studies.

Avoiding Type 2 errors appears to be less critical in basic ecological research than in safety testing of drugs or general surveys geared to early detection of environmental problems, but there may be cases in which the screening approach without multiple hypothesis testing would be appropriate in basic research. For example, in the early stages of a new research program (Saville 1990; Cobb 1998:453-455), a researcher might legitimately engage in hypothesis generation (screening) rather than hypothesis testing (confirmation). Numerous regression analyses on different species might be used to determine whether

there are some key habitat variables that determine the species composition of a community. If this approach is used without multiple hypothesis testing, however, the researcher should explicitly explain that the intent is generation of hypotheses, not hypothesis testing. In many studies these two processes are confounded (Saville 1990); hypotheses generated by screening with p-values are implicitly assumed to have been adequately confirmed without further testing, leading to unreliable conclusions. To be confirmed, the hypotheses generated from large sets of statistical tests must later be tested in more focused studies with a small number of tests and with the application of multiple testing adjustment procedures. Often both hypothesis generation and confirmation can be accomplished in the same study and explained in the same paper, but this is not always practical. When the goal of a study is hypothesis generation, this must be emphasized clearly in the abstract, introduction, and discussion section of any paper describing the results, and significant relationships should be labeled as "potential" or "tentative" until confirmed by further testing to reduce the chance that the results will be misapplied.

Improving the Statistical Power of Multiple Hypothesis Tests

If a researcher determines that an adjustment for multiple hypothesis tests is appropriate, then he or she should strive to use a procedure that produces maximum power while all its assumptions are satisfied. Most adjustments for multiple testing attempt to control the familywise error rate—the probability of making at least one error by rejecting a true null hypothesis. The standard

Bonferroni method controls the familywise error rate by simply dividing the level of significance α by the number of tests *n*. Many researchers find this adjustment to be highly conservative, but fortunately there are modifications of the basic Bonferroni method that increase power while maintaining familywise significance.

One such modification is Holm's Sequential Bonferroni Method (Holm 1979; Rice 1989). First, p-values corresponding to the *n* tests are ordered from smallest to largest: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$. In stage one, if $p_{(1)} \; ? \; \alpha / n$, then the null hypothesis $H_{(1)}$ associated with $p_{(1)}$ is not rejected and all other null hypotheses are not rejected without further test; otherwise, $H_{(1)}$ is rejected and one moves to stage two. In stage two, if $p_{(2)} \; ? \; \alpha /(n-1)$, then the null hypothesis $H_{(2)}$ associated with $p_{(2)}$ is not rejected and null hypotheses $H_{(3)}, H_{(4)}, \ldots, H_{(n)}$ are not rejected without further test; otherwise, $H_{(2)}$ is rejected and one moves to stage three. Continuing in this manner, at any stage *j*, $H_{(j)}$ is rejected if and only if all $H_{(i)}, i < j$, have been rejected and $p_{(j)} \leq \alpha /(n-j+1)$. Later, Hochberg (1988) and Hommel (1988) both provided procedures which, based on a result of Simes (1986), modify Holm's method. Hommel's method is more powerful than Hochberg's, which is more powerful than Holm's (Shaffer 1995) for independent tests and most dependent test scenarios, although improvement typically is minor. In addition, simulations in Simes (1986) suggest that dependency among tests leads to a conservative multiple testing procedure.

Benjamini and Hochberg (1995) have developed an alternative approach to multiple hypothesis testing that controls the expected proportion of false positive findings among all rejected hypotheses (the False Discovery Rate). In many studies with multiple tests, the FDR is the more relevant error rate; out of all the rejected hypotheses, what proportion do we expect were incorrectly rejected? For example, if there are 100 *significant* tests and one is willing to incur a False Discovery Rate of 5%, about 5 of these significant results would be false positives. As the number of tests increases, control of the familywise error rate can become overly restrictive, and few significant tests are noted as a result. If we expect more than just a few null hypotheses to be truly false, then controlling the familywise error rate is impractical. In response, FDR-controlling methods have started to appear in studies covering such diverse topics as plant breeding (Basford & Tukey 1997), education (Williams et al. 1999), and genetic mapping (Weller et al. 1998).

Consider the following simulated example in which we compare controlling methods for the familywise error rate and the False Discovery Rate. This simulation was designed as an illustrative example, not as proof of general properties, although the results here generally agree with those of more complete simulation studies (see, for example, Benjamini & Hochberg 1995 or Storey 2002). P-values corresponding to 1000 independent tests of significance were generated; 500 were randomly sampled decimal values between 0 and 1, and

500 were randomly sampled values from an exponential distribution with mean 0.02. The first set of 500 random p-values reflect p-values under the null distribution (i.e., those corresponding to true null hypotheses) while the 500 p-values from the exponential distribution correspond to null hypotheses that are actually false. Under the familywise error controlling methods discussed above (Bonferroni, Holm, and Hommel), only 1 of the 1000 null hypotheses is rejected, despite the fact that 184 p-values are below 0.01 and 483 are below 0.05. Note, for instance, that the standard Bonferroni method would only reject p-values below $0.05/1000 = 0.00005$, while Holm's method requires the minimum p-value to fall below the same 0.00005 level and the next smallest p-value to fall below 2*0.00005 = 0.0001. In this simulation, the minimum p-value was 0.0000413 and the next smallest was 0.0001092. On the other hand, Benjamini and Hochberg's linear step-up procedure for controlling the FDR (1995; see below) rejects 114 of the 1000 null hypotheses, essentially setting a p-value cut-off of 0.00564. In this example, their procedure has succeeded in controlling the False Discovery Rate at 5%; of the 114 rejections, only 5 represent false discoveries (4.39%).

Specifically, Benjamini and Hochberg's False Discovery Rate

is $E(\frac{V}{R} \mid R > 0) ?\Pr(R > 0)$, where R = number of rejected null hypotheses, and V = number of false positives among rejected null hypotheses. In their 1995 paper, Benjamini and Hochberg provided a linear step-up procedure for controlling the FDR at a given level α for independent test statistics. Simply let *k* be the largest *i* for which $p_{(i)} \le i\alpha / n$, then reject $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$. Later, Benjamini and Yekutieli

(2001) showed that this linear step-up procedure controls the FDR under positive regression dependency on a subset, which, in simplified terms, means that knowing that a certain p-value is small does not decrease the chances of any other p-value being small. Several alternatives to the linear step-up procedure have subsequently been developed (Benjamini & Liu 1999, Benjamini & Hochberg 2000), including a procedure that controls the FDR under general dependency (Benjamini & Yekutieli 2001), although this procedure tends to be highly conservative. Several FDR-controlling procedures are incorporated into a stand-alone Windows program for computing the FDR minimum p-value for rejection and an analogous S-Plus function, both available at http://www.math.tau.ac.il/%7Eroee/index.htm.

Storey (2002, 2003) introduced a related quantity of interest called the positive false discovery rate (pFDR) and a new approach for controlling the pFDR and the FDR. In most instances of multiple testing, where the probability of rejecting at least one hypothesis is near 1, the pFDR is essentially equivalent to the FDR. However, Storey's approach to controlling the FDR provides a potentially powerful alternative to Benjamini and Hochberg's approach.

One contribution of Storey's approach, in addition to increased power in many instances, is the definition of the q-value, an analogue of the p-value. As Storey and Tibshirani (2003) nicely summarize, "Given a rule for calling features significant, the false positive rate [the basis for traditional p-values] is the rate

that truly null features are called significant. The FDR is the rate that significant features are truly null." The q-value attempts to measure each feature's significance while taking into account that many features are being simultaneously tested, and a threshold placed on the q-value limits the proportion of significant features that turn out to be false leads. The q-value is precisely defined as the minimum FDR at which the test can be called significant; it can be thought of as the expected number of false positives among all tests with results as or more extreme than the observed one. A researcher interested in controlling the FDR at, say, 5%, can simply convert ordered p-values to q-values and reject hypotheses associated with q-values below 0.05.

Since it is not a sequential rejection method like others previously discussed, Storey's algorithm for estimating q-values cannot be adequately summarized here. However, a key step is to estimate the true proportion of all null hypotheses which are true rather than assuming it is 1 as in many FDR-controlling methods (Storey 2002). To illustrate Storey's q-value approach and compare it with other approaches, we applied it to our earlier example with 1000 simulated p-values representing 500 true null hypotheses and 500 false null hypotheses. For this example, Storey's q-value procedure rejected 456 hypotheses with an effective p-value cut-off of 0.0407 and false discovery rate of 21/456 = 4.61%. Thus, in this simulated example where 50% of the null hypotheses were truly false, Storey's q-value procedure displayed the most power, rejecting the most hypotheses while still achieving the desired control

over the FDR. In general, Storey's approach becomes more advantageous as the proportion of truly false null hypotheses grows.

Scientists working in large studies with many tests are therefore turning to the q-value as a more appropriate measure of statistical significance; one notable example is in genome studies (e.g., Storey & Tibshirani 2003). Although research into q-values and Storey's procedure for controlling the FDR and the pFDR is ongoing, it has been shown that q-value estimates are conservative under weak dependence, especially the type of "clumpy" (local) dependence typically found in genomewide studies (Storey & Tibshirani 2001; Storey 2003; Storey et al. 2004). The software QVALUE takes a list of p-values and calculates their estimated q-values, an estimate of the proportion of tests truly following the null hypothesis, and some useful diagnostic plots; it is available at http://faculty.washington.edu/~jstorey/qvalue/.

The Hokkaido example: adjustments for multiple tests

To provide an actual example of the different approaches to multiple hypothesis testing discussed here, Table 1 shows results from the Hokkaido study (Kurosawa & Askins 2003). For each of 25 common bird species, a multiple regression model was fit using abundance as a response variable, and forest area, forest isolation, and three principal components summarizing several vegetation variables as predictor variables. In this study, it is reasonable to assume that each species is a genetically independent population with a

separate evolutionary trajectory, so that each species responds to habitat and landscape variables independently. Because adaptations to habitats are flexible and rapid in birds, this assumption of independence is even reasonable for species with relatively recent common ancestors. The following significance levels are provided: comparisonwise (unadjusted) p-values, adjusted p-values based on Holm's (1979) Sequential Bonferroni Method, adjusted p-values based on Benjamini and Hochberg's (1995) controlling method for the FDR, and estimated q-values based on Storey's (2002) approach for controlling the pFDR.

Based on unadjusted p-values, there are several species for which the model seems to explain a significant portion of the variability in abundance; 9 of the 25 p-values are below 0.05, and 6 of these 9 are below 0.01. If Holm's Sequential Bonferroni Method is used to control the familywise error rate at 0.05, then the picture is much different: tests for only 2 of the 25 species (Oriental Cuckoo and Great Tit; see Table 1 for scientific names) are considered significant. If, however, the False Discovery Rate is controlled at the 0.05 level, then tests for 7 of the original 9 species (all except Coat Tit and Marsh Tit) can be considered significant while still maintaining the upper bound for the expected number of false positives among the significant findings at 5%. This is based on an FDR rejection value of 0.013 for Benjamini and Hochberg's linear step-up procedure; rejection values can depend on the FDR-controlling procedure used. Finally, if we require a q-value at or below 0.05, then tests for 10 of the 25 species can be considered significant (the original 9 species along with Oriental Greenfinch).

These q-values are based on an estimate (based on Storey 2002) of 86% for the proportion of false null hypotheses among all tests. Although this is merely one example, the gain in power from controlling the False Discovery Rate rather than the familywise error rate is clear. However, a researcher who decides that some adjustment for multiple testing is necessary must control the error rate that is most appropriate for the research question at hand.

Conclusions

When many hypothesis tests are performed, adjustments in the significance level may not always be warranted, especially when the purpose of the study involves screening for potential conservation problems or hypothesis generation. In cases where adjustment in the significance level is required, the False Discovery Rate may often be a more appropriate error rate to control than more traditional familywise and comparisonwise error rates. Controlling methods for the FDR offer powerful alternatives to controlling the familywise error rate with sequential Bonferroni methods, especially in cases where many independent tests are performed. Some work has been done on cases where dependencies exist among tests (Storey & Tibshirani 2001, Benjamini & Yekutieli 2001), so these FDR-controlling methods can be used under certain dependency structures (e.g., positive regression dependency or clumpy dependence). Dependency structures in specific problems must therefore be carefully considered. For example, independence might be a reasonable assumption for testing the effect of a habitat variable on different species, but not for assessing the effects of climate

change on different populations of the same species. At this point, if the False Discovery Rate is appropriate to control but the hypothesis tests have an uncommon dependence structure, one can use Benjamini and Yekutieli's (2001) conservative procedure for general dependency structures, or any procedure which controls the familywise error rate, since any procedure which controls the familywise error rate will also control the FDR (Benjamini & Hochberg 1995).

Regardless of the error rate controlled or the controlling procedure used, the application of multiple testing procedures to all tables of statistical tests involves reducing false positives (Type 1 errors) at the cost of not detecting real differences (Type 2 errors). This tradeoff becomes too costly when (a) adjustments for multiple testing are used when none are required, (b) an inappropriate error rate has been targeted for control, or (c) an underpowered adjustment procedure is applied. The consequent loss of statistical power may lead researchers to reduce the number of statistical tests and focus on only a small set of variables or populations, usually those that are already known to show important ecological patterns. The ultimate cost of using unnecessary or underpowered multiple testing adjustments is the reticence to explore new relationships or to screen for potential conservation problems, inhibiting a stage in scientific research that is critically important even though it is neither conclusive nor sufficient without further research.

Acknowledgments

We thank Naitee Ting (Pfizer Global Research and Development), Phillip Barnes (Connecticut College), and Julie Legler (St. Olaf College) for their comments on drafts of this paper. We thank Reiko Kurosawa for permitting us to use her data on Hokkaido birds. The paper was also greatly improved by recommendations from Philip Dixon and two anonymous reviewers.

Literature Cited

Basford, K. E. and J. W. Tukey. 1997. Graphical profiles as an aid to understanding plant breeding experiments. Journal of Statistical Planning and Inference **57**: 93-107.

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

Benjamini, Y. and 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. and Y. Hochberg. 2000. On adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics **25**: 60-83.

Benjamini, Y. and W. Liu. 1999. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. Journal of Statistical Planning and Inference **82**: 163-170.

Benjamini, Y. and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics **29**: 1165-1188.

Cobb, G. 1998. Introduction to the design and analysis of experiments. Key College Publishing, Emeryville, California.

Crabbe, J. C., D. Wahlsten, and B. C. Dudek. 1999. Genetics of mouse behavior: interactions with laboratory environment. Science **284**: 1670-1672.

Food and Drug Administration. 1998. Guidance for industry. E9 Statistical Principles for clinical trials. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Rockville, MD.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika **75**: 800-802.

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

Hommel, G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika **75**: 383-386.

Kurosawa, R. and R. A. Askins. 2003. Effects of habitat fragmentation on birds in deciduous forests in Japan. Conservation Biology **17**: 695-707.

Ornithological Society of Japan. 2000. Check-list of Japanese birds. 6th edition. Obihiro, Hokkaido.

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

Saville, D. J. 1990. Multiple comparison procedures: the practical solution. American Statistician **44**: 174-180.

Shaffer, J. P. 1995. Multiple hypothesis testing. Annual Review of Psychology **46**: 561-584.

Simes, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika **73**: 751-754.

Storey, J. D. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B **64**: 479-498.

Storey, J. D. 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Annals of Statistics **31**: 2013-2035.

Storey, J. D., J. E. Taylor, and 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, Series B **66**: 187-205.

Storey, J. D. and R. Tibshirani. 2001. Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical report 2001-28, Department of Statistics, Stanford University.

Storey, J. D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences **100**: 9440-9445.

Tukey, J. W. 1991. The philosophy of multiple comparisons. Statistical Science **6**: 100-116.

Weller, J. I., J. Z. Song, D. W. Heyen, H. A. Lewin and M. Ron. 1998. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics **150**: 1699-1706.

Williams, V. S. L., L. V. Jones, and J. W. Tukey. 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. Journal of Educational and Behavioral Statistics **24**: 42-69.

Wright, S. P. 1992. Adjusted *P*-values for simultaneous inference. Biometrics **48**: 1005-1013.

Table 1. Comparison of four approaches for multiple hypothesis testing applied to an analysis of data on the d

Habitat group Species [2]	*Model F*[3]	*Una p-*
Forest-interior species		
Oriental Cuckoo (*Cuculus saturatus*)	4.6	0
Grey Thrush (*Turdus cardis*)	0.7	0
Eastern Crowned Leaf Warbler (*Phylloscopus coronatus*)	4.0	0
Coat Tit (*Parus ater*)	2.4	0
Varied Tit (*Parus varius*)[4]	4.1	0
Japanese White-eye (*Zosterops japonicus*)[4]	0.8	0
Forest-generalist species		
Oriental Turtle Dove (*Streptopelia orientalis*)	0.3	0
Japanese Pygmy Woodpecker (*Dendrocopos kizuki*)	1.2	0
Brown-eared Bulbul (*Hypsipetes amaurotis*)	1.3	0
Siberian Blue Robin (*Luscinia cyane*)	1.6	0
Brown Thrush (*Turdus chrysolaus*)	0.6	0
Short-tailed Bush Warbler (*Urosphena squameiceps*)	1.6	0
Bush Warbler (*Cettia diphone*)	0.7	0
Arctic Warbler (*Phylloscopus borealis*)	1.3	0
Eastern Pale-legged Leaf Warbler (*Phylloscopus borealoides*)[4]	0.5	0
Narcissus Flycatcher (*Ficedula narcissina*)	4.1	0
Blue-and-white Flycatcher (*Cyanoptila cyanomelana)*	1.2	0
Brown Flycatcher (*Muscicapa dauurica*)	0.6	0
Long-tailed Tit (*Aegithalos caudatus*)[4]	1.7	0
Marsh Tit (*Parus palustris*)	2.7	0
Great Tit (*Parus major*)	5.1	0
Black-faced Bunting (*Emberiza spodocephala*)	1.2	0
Masked Grosbeak (*Eophona personata*)	3.2	0
Forest-edge species		

Oriental Greenfinch (*Carduelis sinica*)	2.2	0
Jungle Crow (*Corvus macrorhynchos*)	4.1	0

[1] See Kurosawa and Askins (2003) for a description of this study.

[2] Common and scientific names follow the Check-list of Japanese Birds (Ornithological Society of Japan 2000)
[3] The model F-test was based on results of multiple regression analyses with abundance of common species o
as

independent variables. Unless otherwise indicated, abundance was measured by the average of standardize

[4] Based on 1996 data.