Opinion is intended to facilitate communication between reader and author and reader and reader. Comments, viewpoints or suggestions arising from published papers are welcome. Discussion and debate about important issues in ecology, e.g. theory or terminology, may also be included. Contributions should be as precise as possible and references should be kept to a minimum. A summary is not required.

Escaping the Bonferroni iron claw in ecological studies

Luis V. García, Depto de Geoecología. IRNAS-CSIC. P.O. Box 1052, ES-41080 Sevilla, Spain (ventura@cica.es).

I analyze some criticisms made about the application of alpha-inflation correction procedures to repeated-test tables in ecological studies. Common pitfalls during application, the statistical properties of many ecological datasets, and the strong control of the tablewise error rate made by the widely used sequential Bonferroni procedures, seem to be responsible for some 'illogical' results when such corrections are applied. Sharpened Bonferroni-type procedures may alleviate the decrease in power associated to standard methods as the number of tests increases.

More powerful methods, based on controlling the false discovery rate (FDR), deserve a more frequent use in ecological studies, especially in those involving large repeated-test tables in which several or many individual null hypotheses have been rejected, and the most significant p-value is relatively large.

I conclude that some reasonable control of alpha inflation is required of authors as a safeguard against striking, but spurious findings, which may strongly affect the credibility of ecological research.

Moran (2003) recently suggested rejecting the application of the sequential Bonferroni rule in ecological studies. He based his proposal on certain mathematical, logical, and practical objections which led him to conclude that it would be better for ecological research to abandon the awkward constraints derived from the sequential Bonferroni rule, allowing the researcher to interpret more freely the multiple test outcomes without testing for alpha inflation. Thereby, detailed ecological research would be stimulated, while avoiding the loss of potentially relevant results, which are at risk of remaining unknown when authors are required to adhere strictly to the sequential Bonferroni rule. The likely increase in the frequency of 'false positives' in the ecological literature would be of minor importance, since these spurious results will not be confirmed by subsequent experiments.

In other contexts, even stronger claims against alpha corrections have recently been the subject of controversy (Perneger 1998, 1999, Feise 2002). Surprisingly few people have questioned the same corrections which are implicit in the standard *post hoc* methods routinely applied to perform multiple comparisons between treatments for a single dependent variable. Accepting Moran's arguments, it could be argued that relevant research results are perhaps not being published because people use these alpha-corrected methods instead of looking directly at the individual pairwise-test p-values.

There is an apparent inconsistency between the unquestioned acceptance of the "alpha inflation under repeated test" principle in the univariate case, and the controversy about the convenience or not of applying the same statistical principle in the multivariate case.

Arbitrary rejection of the application of a well-founded statistical principle does not seem an acceptable scientific solution for a problem. If the way in which alpha-inflation corrections are routinely applied in multivariate ecological studies does not work, it seems more reasonable to analyze and improve the procedures rather than simply 'kill the principle'.

The frustrating repeated $p < 0.045$

The above-cited paper of Moran specifically addresses the certainly frustrating situation of having several 'relatively large' p-values in a table, all of which have a clear logical link (as in Moran's Table 1, the "grazed forbs and grasses" example), but – after applying the sequential Bonferroni rule — you get nothing. In such cases, the proposal was made to 'follow your own logic and abandon the illogical sequential'.

Caution is needed when those 'relatively large' p-values are found. Time and routine have apparently diluted the origin of the 'universal' $p < 0.05$ criterion, proposed by R.A. Fisher more than 60 years ago, as the limit around which it *is difficult to conclude something against or in favour of the null hypotheses*. For some of these 'relatively large' p-values, the correct conclusion would be to repeat the experiment, *whether you are performing repeated tests or only one* (Sterne and Smith 2001). According to Fisher's criterion, a $p < 1/50$ would

be required to conclude consistent evidence against the null hypotheses, and, for publication, even a $p < 1/100$ was recommended (Ibbetson 2001, Sterne and Smith 2001).

Therefore, I conclude that Moran's grass-forb example is not appropriate to express the well-founded concern arising when you get *several clearly significant* individual results (say $p = < 0.02$, Table 3) in a battery of many (say ten) simultaneous tests (having a first sequential Bonferroni threshold of 0.005).

Some remarks on alpha-inflation control in multivariate ecological studies

In my opinion, to achieve a reasonable solution for the multiplicity problem in multivariate ecological studies – which, in Moran's words, 'will affect ecological research indefinitely' – several aspects warrant comment:

(1) Omnibus tests

Unlike the univariate multiple testing situations (such as those in which ANOVA is usually applied), omnibus multivariate tests are seldom applied in the multivariate case. Although these omnibus multivariate tests are available to most ecologists, few people use them (Espinar et al. 2002 for an exception). A multivariate ANOVA table may be first analyzed using an omnibus MANOVA, or by alternative non-parametric permutation tests (Anderson 2001). A correlation matrix may be first tested against the identity matrix by the omnibus Bartlet sphericity test (Bartlet 1954), and so on.

In general, omnibus multivariate tests allow 1) a control of the overall, experiment-wide error rate, and 2) the finding of differences between treatments based on combinations of variables (which may remain undetected when dependent variables are examined separately), making them suitable for testing whether any significant effect should be expected in the data, provided that the sample size is greater than the number of dependent variables being analyzed (Huberty and Morris 1989, Hair et al. 1995).

If, after performing an omnibus test on meaningful variables, you get a significant result, the problem newly arises of how to reach a satisfactory conclusion about the individual univariate tests (the unsolvable sequential Bonferroni 'dilemma'). However, after a significant result has been obtained in the omnibus test, the question is not whether some significant difference exists in the table, but whether you have enough power and ability to select the appropriate statistical procedure to specify where the differences are.

(2) Power of individual tests

According to Moran (2003), ecological studies often have a 'small number of replicates, high variability, and (subsequently) low statistical power'. In such circumstance it is difficult to get any relatively small individual p-value. In fact, this problem is different from applying (or not) some alpha-inflation correction, since the fact here may be that power is insufficient even for consistently rejecting the hypotheses in *any* of the individual tests performed, according to the above-cited criterion of Fisher. In such cases, either a more powerful individual test and/or a higher sample size should be used. As Sterne and Smith (2001) have recently pointed out, the maximum increase in size required for a move from $p < 0.05$ to the more conclusive $p < 0.01$ is by a factor of only 1.75.

(3) Sharpened Bonferroni methods

The standard Bonferroni-type procedures (described in points 1 and 2 of Table 1) are designed to control the type I error by assuming that all the tested null hypotheses are true, which in many cases may be quite unrealistic and make them too conservative. A strategy to increase power – in both the single-step and the sequential Bonferroni methods – is to estimate the number of true null hypotheses (n_0) and use it to sharpen the standard methods. Table 1 (point 3) summarizes the P-plot estimation method for n_0, and includes two main references for applying these corrections. Figure 1 shows an example of the application of the first step of this procedure to the p-vectors corresponding to the hypothetical experiments in Table 3. Results of applying the sharpened Bonferroni-type corrections are shown in Table 3.

(4) Dependence

A prime argument against applying alpha–inflation procedures in multivariate ecological studies is that several relatively high p-values (of say 0.02) is stronger evidence against the null hypothesis than one moderately low value, since the probability of finding several simultaneous significant tests only by chance is very low. A problem with this reasoning is that it implicitly assumes that *variables being tested are independent*, which is probably not true in many multivariate ecological studies.

When repeated tests are performed on redundant variables, you are, to some extent, *repeating the same test several* (or many) *times*. One reason why you may have several 'logically linked' significant p-values, but not a significant result after correcting for alpha inflation, is that you are performing the test on highly

Table 1. Different procedures for controlling (at the α level) the familywise error rate (FWER: 1 to 3), or the false discovery rate (FDR: 4), when n null hypotheses ($H_1 \ldots H_n$) are simultaneously tested. All of them use only the p-values from the individual tests ($p_1 \ldots p_n$) to perform the corrections. For other resampling-based approaches, which require the raw data, see text.

1. One-step Bonferroni.
   While $p_i \leq \alpha/n$ reject H_i, otherwise accept H_i.
2. Stepwise Bonferroni.
   p-values are ranked in ascending order, j being the resulting rank.
   A. Step-down sequential (Holm 1979).
      Testing is conducted in decreasing order of significance of the ordered hypotheses (i.e. proceed from $j = 1$ to $j = n$). While $p_j \leq \alpha/(n - j + 1)$ reject H_j, otherwise accept H_j and all the remaining null hypotheses. It is uniformly more powerful than the one-step test.
   B. Step-up sequential (Hochberg 1988).
      Testing is conducted in increasing order of significance of the ordered hypotheses (i.e. proceed from $j = n$ to $j = 1$). While $p_j > \alpha/(n - j + 1)$ accept H_j, otherwise reject H_j and all the remaining null hypotheses. It is uniformly more powerful than the Holm test.
3. Increasing power of the Bonferroni-type procedures.
   A. P-plots and sharpened Bonferroni methods
      The number of 'true' null hypotheses (n_0) is estimated by plotting the $1 - p_i$ values, sorted in ascending order, versus their rank (Fig. 1). The points corresponding to true null hypothesis (large p-values) tend to fall along a straight line passing through the origin, whose estimated slope (b_1*) gives an estimate of n_0 (n_0*), calculated as $n_0^* = (1/b_1^*) - 1$.
      The use of n_0* values in conjunction with Bonferroni-type methods allows for increased power, especially when $n_0 \ll n$. See Schweder and Spjotvøll (1982) and Hochberg and Benjamini (1990) for a detailed description of these sharpened procedures.
   B. Empirical corrections for dependence.
      A new adjusted critical α-level (α_i) is calculated for the i-th hypotheses after correcting for correlation (r) or by shared variance (R^2) between variables. Sankoh et al. (1997) for a description and evaluation of several different algorithms.
4. Step-up FDR (Benjamini and Hochberg 1995).
   p-values are ranked in ascending order, j being the resulting rank.
   Proceed from $j = n$ to $j = 1$, until finding a first p-value, ranked k, satisfying $p_k \leq k \times \alpha/n$. Then reject H_j for $j \leq k$ and accept all the remaining null hypotheses.

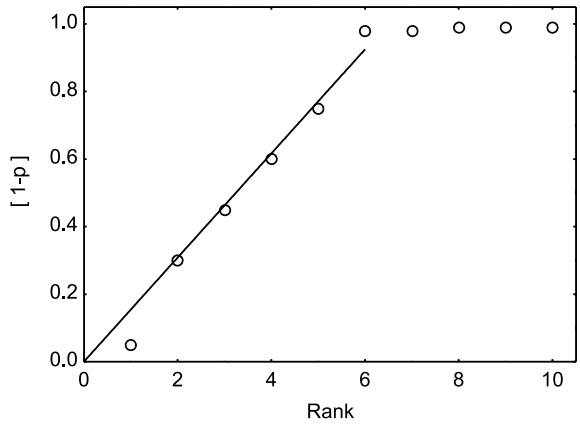

Fig. 1. P-plot of the p-values corresponding to the hypothetical experiment Exp. 1 in Table 3. 1-p values were rank ordered and plotted versus their ranks. The estimated slope of the line fitted to the linear portion of the curve was used to estimate the number of "true" null hypotheses (n_0), as the inverse of the slope minus 1. For Exp. 1 in Table 3, $n_0 = 5$, while for Exp. 2 (not shown), $n_0 = 1$. The estimated value of n_0 was then used for sharpening the Bonferroni-type corrections (point 3A in Table 1).

correlated variables, which are expressing a very similar response to some underlying factor, as could be the case of forbs in Moran's hypothetical example. For instance, in a recent study involving many repeated univariate tests on environmental variables (García et al. 2002), we realized that directly correcting alpha inflation for all the measured variables led to a dramatic loss of power. Therefore, we analyzed the dependence structure in the data and excluded from the repeated univariate ANOVA table some variables which were statistically redundant, and conceptually related, with other variables which remained included in the analysis (this was the case for the soil electrical conductivity and for the concentration of some dominant ions, such as Cl and Na).

That is, when analyzing highly redundant multivariate data, the approach used should account for variable dependence while correcting for alpha inflation. Uitenbroek (1997) has implemented an empirical approach in the web-based statistical program SISA, which estimates the value of the critical p-level for a set of repeated univariate tests, considering not only the fixed overall tablewise error rate and the number of tests being performed, but also the average correlation between the analyzed variables. When a strong average correlation is found in the data, the significance threshold may be considerably increased (Table 2). Comparison of the critical values in Table 2 with the p-values included in Moran's Table 1 leads to the conclusion that several of the studied forb responses to grazing could have been declared significant if the average correlation among variables had been about 0.6, while maintaining overall experiment error rate at the 0.05 level. Even at a very high value of average correlation (0.9), all forb responses could have been declared significant.

Another possible approach to cope with variable redundancy is estimating the number of significant eigenvalues which may be extracted from the dependent variable set. This number is a reasonable approximation to the effective number of independent tests being performed (Chevereud 2001). Thus, the critical p-value for the univariate tests may be calculated using this number, instead of the overall number of tests.

A more exact way of adjusting p-value for experimentalwise error rate, regardless of the dependence structure in the data, is provided by resampling techniques (Westfall and Young 1993, Bender and Lange 2001), now available in some widely used statistical software (Westfall et al. 1999). For normally distributed

Table 2. Change in critical p-values for the individual tests, taking into account the average correlation between variables. Tablewise type I error rate is fixed at the 0.05 level. N is the number of repeated tests. Calculations were performed using the program SISA (Uitenbroek 1997).

	Average correlation				
N	0	0.3	0.5	0.7	0.9
5	0.010	0.016	0.023	0.031	0.043
10	0.005	0.010	0.016	0.025	0.040
25	0.002	0.005	0.010	0.019	0.036
100	5×10^{-4}	0.002	0.005	0.013	0.032

data with known covariance, Dunnet and Tamhane (1995) have proposed another approach which takes into account the dependence structure among variables.

An alternative formulation of the multiple testing problem: the false discovery rate (FDR)

Traditionally, procedures for controlling alpha inflation have been focused on controlling the familywise error rate (FWER), that is, the probability of wrongly rejecting one or more null hypotheses. This is in fact suitable when the prime interest of the researcher is to avoid *any* wrong rejections. However, in many cases the researcher is more interested in controlling the *fraction of wrong rejections* among the rejected hypotheses, rather than the occurrence of one or more of them. This led Benjamini and Hochberg (1995) to define the false discovery rate (FDR) as the expected proportion of true null hypotheses that are erroneously rejected, out of the total number of hypotheses rejected (i.e. the proportion of false positives among all significant hypotheses). In fact, the FDR idea seems to be an elaboration of an older proposal of Eklund (1961–1963) which was first published by Seeger (1968).

The FDR criterion has the advantage of being less restrictive (and more powerful) than the FWER one, and is of particular interest when analyzing large multiple-test tables in which several or many null hypotheses have been rejected. In fact, it seems to be the most satisfactory approach for coping with multiplicity when hundreds or thousands of simultaneous tests are performed, as occurs in genetic microarray experiments (Reiner et al. 2003), in neuroimage analysis (Genovese et al. 2002), in astrophysics (Miller et al. 2001) or in the so-called 'exploratory' or 'data mining' procedures.

Many observational or experimental field ecological studies in which a large number of dependent variables are measured have some degree of 'screening', since it may be virtually impossible 'a priori' to predict exactly the final number of variables that will have to be analyzed. Here, the goal is primarily detecting which, among all 'positive findings', may be considered 'false positives', rather than control of any wrong rejection (García 2003).

In these cases, the FDR approach has several advantages over the classical FWER approaches: (1) it enables controlling the proportion of false positives among the rejected null hypotheses; (2) it avoids performing individual tests at very low p-levels in large problems; (3) it is more powerful than the sequential Bonferroni procedures (such as those proposed by Holm 1979, Hochberg 1988 and Hommel 1988) used up to now for a so-called 'strong control' of familywise error rates; (4) when no actual true positive findings exist (i.e. all the null hypotheses are true), the FDR method has the same control as the previous methods; that is, FDR methods have a so-called 'weak control' of familywise type I error; (5) the FDR threshold may be determined from the observed p-value distribution, and hence is adaptive to the 'amount of signal' in the data (Genovese et al. 2002); and (6) FDR is familywise robust (i.e. tends to be far more consistent than procedures controlling FWE in terms of whether a particular hypothesis is rejected, as the family in which this hypothesis is located changes in size, Holland and Cheung 2002). Additionally, FDR may account for the exact dependence structure of the data, via resampling-based procedures (Yekutieli and Benjamini 1999) or under the normal assumption (Troendle 2000).

Table 1 shows several widely used algorithms for controlling FWER and the one proposed by Benjamini and Hochberg (1995) for controlling FDR. In Table 3, the outcomes of all these procedures are compared, using two p-vectors resulting from two hypothetical ecological experiments. The first has five clearly significant and five clearly non-significant univariate p-values, on an individual basis. Unlike the one-step Bonferroni and the sequential Bonferroni methods of strong FWER control, the FDR-calculated minimum p-value for rejection (0.02) enables declaring as significant the five rejected null hypotheses, while controlling both the proportion of false positives at the 0.05 level and the overall probability of confusing a random finding with a meaningful one when the null hypotheses are all true. The second p-vector (Exp. 2) illustrates well the differences between the three FWER-controlling procedures (that of Hochberg is the most powerful), and the fact that the increased power under the FDR approach is more apparent when the number of tests is large.

Several improvements of the original FDR methodology of Benjamini and Hochberg (1995), which have extended its applicability to dependency situations (Benjamini and Yekutieli 2001), together with the implementation by the authors of several stepwise algorithms for controlling FDR in an easy-to-use standalone program (available free at http://www.math.

Table 3. Result of applying different procedures for controlling alpha inflation described in Table 1 to outcomes of two hypothetical ecological experiment involving 10 (Exp. 1) and four (Exp. 2) repeated tests. Methods 1, 2A, and 2B, which strongly control the FWER (at the 0.05 level), lead to an overall acceptance of all null hypotheses in Exp. 1, but differences exist in Exp. 2. Sharpened one-step (3A1) and sequential (3A2: Holm's, 3A3: Hochberg's) Bonferroni methods appreciably increase the power with respect to the corresponding standard methods, detecting three significant effects in Exp. 1. Method 4, which controls the FDR rate at the 0.05 level, while maintaining a weak control (when all null hypotheses are true) on the overall FWER error rate at the same level, leads to a rejection of the first five null hypotheses in Exp. 1. (ns = non-significant result, s = significant result).

	FWER						FDR
Exp. 1	(1)	(2A)	(2B)	(3A1)	(3A2)	(3A3)	(4)
0.01	ns	ns	ns	s	s	s	s
0.01	ns	ns	ns	s	s	s	s
0.01	ns	ns	ns	s	s	s	s
0.02	ns	ns	ns	ns	ns	ns	s
0.02	ns	ns	ns	ns	ns	ns	s
0.25	ns	ns	ns	ns	ns	ns	ns
0.40	ns	ns	ns	ns	ns	ns	ns
0.55	ns	ns	ns	ns	ns	ns	ns
0.70	ns	ns	ns	ns	ns	ns	ns
0.95	ns	ns	ns	ns	ns	ns	ns
Exp. 2							
0.009	s	s	s	s	s	s	s
0.020	ns	ns	s	s	s	s	s
0.024	ns	ns	s	s	s	s	s
0.650	ns	ns	ns	ns	ns	ns	ns

tau.ac.il/ ~ roee/ FDR_Downloads2.htm), allow wide application of these procedures in ecological studies.

Research on the "FDR family" is currently very active, and new concepts, which apparently may improve the original idea, are currently emerging (Efron and Tibshirani 2002, Genovese and Wasserman 2002, Sarkar 2002, Fernando et al. 2004). For example, Storey (2002) recommended the use of the so-called positive false discovery rate (pFDR), an estimate of the rate of discoveries that are false (in contrast to FDR, which is the rate of false discoveries that occur) and of the related q-value (which is the minimum pFDR above which that statistic can be rejected). Here the rejection region is fixed and the error rate estimated, instead of setting the error rate and estimating the rejection region. This apparently more powerful approach has also been implemented and documented in the Q-value program (available free, as an R-package function, at http://faculty.washington.edu/ ~ jstorey/qvalue).

More recently, Bickel (2003) proposed the decisive false discovery rate (dFDR) – the ratio of the expected number of false discoveries to the expected total number of discoveries – as being advantageous over both the "traditional" FDR and the newer pFDR. The practical advantages of the various (frequentist, Bayesian and decision-theory-based) working approaches will probably be clarified in the near future.

Is your whole scientific career a single experiment?

One of the challenges for the multiple-testing alpha-inflation procedures is to define what is a "family" of tests, or what are the exact limits of an "experiment". Certainly, there is no statistical theory giving a definitive answer for this question, and, in fact, "increases in the scale of Bonferroni corrections can quickly degenerate into the absurd" (Cabin and Mitchell 2000).

Significantly, advocates of exclusively applying "common sense" when interpreting multiple-test outcomes (on an individual p basis) do not seem to follow the same rule in determining when it may be "meaningful to take into account some combined measure of errors" (Hochberg and Tamhane 1987).

For testing the relationship between some (more or less) controlled factors and some measured response variables using samples extracted from a population, most authors would agree that some kind of correction for alpha inflation should be done if (1) you continue investigating the relationships between other factors and response variables in the same samples, or (2) you repeat the same tests in different subsamples, or (3) you make sequential tests on the same set of subjects. However, if you plan a new study and gather new datasets on different subjects for data analysis, most people would think that you are performing different "experiments", and, consequently, independent corrections should be performed on each to prevent alpha inflation.

As Proscham et al. (2000) have recently suggested, a consensus could be reached to correct for multiplicity when the repeated tests are performed on the same units of analysis, considering sequential (monitoring) and subsample tests as a part of the same overall experiment.

It is possible that somebody has spent their whole scientific career continuously testing the same individuals or plots. In such case, an overall "careerwise" correction for alpha inflation should probably be performed.

However, many researchers might be interested in estimating how many false discoveries (above the 5% level reference) they may have made during their whole scientific career. It will depend on the number, dependence structure, and power of the individual tests made, which in turn are related with the sample sizes and number of variables tested in the various experiments performed, and with the frequency of "fishing expeditions". Undoubtedly, the more hypotheses you have tested in your career, the more 'false positives' you may have obtained, but also the more true interesting findings you may have reported. The key point is whether you have or do not have an unacceptable proportion of false discoveries among your reported findings, which largely depends on the quality and rigour, rather than on the amount or detail, of your research.

The benefits of preserving rigour in ecological research

Lottery tickets should not be free

In such purely random and independent events as the lottery, the probability of having a winning number depends directly on the number of tickets you have purchased. When one evaluates the outcome of a scientific work, attention must be given not only to the potential interest of the 'significant' outcomes but also to the number of 'lottery tickets' the authors have 'bought'. Those having many have a much higher chance of 'winning a lottery prize' than of getting a meaningful scientific result. It would be unfair not to distinguish between significant results of well-planned, powerful, sharply focused studies, and those from 'fishing expeditions' (Cormier and Pagano 1999), with a much higher probability of catching an old truck tyre than of a really big fish.

Are 'false positives' equivalent to 'false negatives'?

A repeatedly used argument in favour of abandoning the alpha-inflation corrections when multiple testing is performed is more or less the following: (1) let authors freely interpret their results on an individual test basis, thereby giving the chance of possibly relevant achievements – derived from complex and detailed multivariate investigations – that will become known to the scientific community; (2) if these 'significant results' were, in fact, spurious, there is no reason to worry: somebody will carry out some similar experiments elsewhere, and will demonstrate the original author's error.

The trade-offs are, in essence, (1) to use some of the available alpha-inflation correction procedures, thus protecting the whole scientific community against an excessive proportion of 'false positives', while penalizing some possibly interesting results; or (2) to stimulate author creativity – but also the most-profitable 'fishing expeditions' – at the price of increasing the false discovery rate.

In my opinion, in accord with Arndt and Bartko (2003), it is the responsibility of the researcher to provide an estimate of the likelihood that results are chance findings. Since an underestimation of type I error rates can lead to false impression, this issue is of serious concern. While it may be comforting to speculate that follow-up studies will fail to replicate the spurious finding – and hence eventually set the record straight – this attitude is becoming an increasingly shallow reassurance. In fact, once some relevant results have been published, the public's knowledge about them is seldom corrected, since follow-up negative studies are not deemed newsworthy. This is not only misleading, it also unfavourably affects scientific credibility.

Acknowledgements – I thank Alma Alada-Blanca, Fernando Ojeda and Bob O'Hara for their comments and suggestions while preparing and correcting the manuscript.

References

Anderson, M. J. 2001. A new method for non-parametric multivariate analysis of variance. – Aust. Ecol. 26: 32–46.
Arndt, S. and Bartko, J. J. 2003. Why you need to correct for multiple tests. Part 2. The solutions. Statistics for readers and writers, 9. – Working paper (http://sarndt.psychiatry.uiowa.edu/Webpage/methresources/statarticles/Part9-web.htmHYPERLINK, accessed 12–22–03).
Bartlet, M. S. 1954. A note on multiplying factors for various chi-squared approximations. – J. R. Stat. Soc. B 16: 296–298.
Bender, R. and Lange, S. 2001. Adjusting for multiple testing: when and how? – J. Clinical Epidemiol. 54: 343–349.
Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. – J. R. Stat. Soc. B 57: 289–300.
Benjamini, Y. and Yekutieli, D. 2001. The control of the false discovery rate under dependency. – Ann. Stat. 29: 1165–1188.
Bickel, D. 2003. Error-rate and decision-theoretic methods of multiple testing. – Working paper. Medical College of Georgia. Available at http://arxiv.org/ftp/math/papers/0212/0212028.pdf (accessed 12–22–03).
Cabin, R. J. and Mitchell, R. J. 2000. To Bonferroni or not to Bonferroni: when and how are the questions. – ESA Bull. 81: 246–248.
Chevereud, J. M. 2001. A simple correction for multiple comparisons in interval mapping genome scans. – Heredity 87: 52–58.
Cormier, K. D. and Pagano, M. 1999. Multiple comparisons: a cautionary tale about dangers of fishing expeditions. – Nutrition 15: 332–333.
Dunnet, C. W. and Tamhane, A. C. 1995. Step-up multiple testing of parameters with unequally correlated estimates. – Biometrics 51: 217–227.
Efron, B. and Tibshirani, R. 2002. Empirical Bayes methods and false discovery rates for microarrays. – Genet. Epidemiol. 21: 70–86.
Espinar, J. L., García, L. V., García-Murillo, P. et al. 2002. Submerged macrophyte zonation in a Mediterranean salt marsh: a facilitation effect from established helophytes? – J. Veg. Sci. 13: 831–840.
Feise, R. J. 2002. Do multiple outcome measures require p-value adjustment? – Br. Med. C. Med. Res. Meth. 2: 8.
Fernando, R. L., D. Nettleton, B. R. Southey, J. C. M. et al. 2004. Controlling the proportion of false positives (PFP) in multiple dependent tests. – Genetics 166, in press.
García, L. V. 2003. Controlling the false discovery rate in ecological research. – Trends Ecol. Evol. 18: 553–554.
García, L. V., Marañón, T., Ojeda, F. et al. 2002. Seagull influence on soil properties, chenopod shrub distribution, and leaf nutrient status in semi-arid Mediterranean islands. – Oikos 98: 75–86.
Genovese, C. R. and Wasserman, L. 2002. Operating characteristics and extensions of the false discovery rate procedure. – J. R. Stat. Soc. B 64: 419–437.
Genovese, C. R., Lazar, N. A. and Nichols, T. 2002. Thresholding of statistical maps in functional neuroimage analysis using the false discovery rate. – Neuroimage 15: 870–878.
Hair, J. F., Anderson, R. E., Tatham, R. L. et al. 1995. Multivariate data analysis, 5th edn. – Prentice-Hall.
Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. – Biometrika 75: 800–803.
Hochberg, Y. and Tamhane, A. C. 1987. Multiple comparison procedures. – John Wiley.

Hochberg, Y. and Benjamini, Y. 1990. More powerful procedures for multiple significance testing. – Stat. Med. 9: 811–818.
Holland, B. and Cheung, S. H. 2002. Familywise robustness criteria for multiple comparisons procedures. – J. R. Stat. Soc. B 94: 63–77.
Holm, S. 1979. A simple sequential rejective multiple test procedure. – Scand. J. Stat. 6: 65–70.
Hommel, G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. – Biometrika 75: 383–386.
Huberty, C. J. and Morris, J. D. 1989. Multivariate analysis versus multiple univariate analysis. – Psychol. Bull. 105: 302–308.
Ibbetson, D. 2001. What Fisher said. – Br. Med. J. rapid response to Steme & Smith paper (available at http://bmi.bmijournals.com/cgi/eletters/322/7280/226#12283).
Miller, C., Genovese, C. R., Nichol, R. C. et al. 2001. Controlling the false-discovery rate in astrophysical data analysis. – Astron. J. 122: 3492–3505.
Moran, M. D. 2003. Arguments for rejecting the sequential Bonferroni in ecological studies. – Oikos 100: 403–405.
Perneger, T. V. 1998. What's wrong with Bonferroni adjustments. – Br. Med. J. 316: 1236–1238.
Perneger, T. V. 1999. Multiple testing. – Br. Med. J. 322: 226–231.
Proscham, M. A. and Waclawiw, M. A. 2000. Practical guidelines for multiplicity adjustment in clinical trials. – Controlled Clinical Trials 21: 527–529.
Reiner, A., Yekutieli, D. and Benjamini, Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. – Bioinformatics 19: 368–375.
Sankoh, A. J., Huque, M. F. and Dubey, S. D. 1997. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. – Stat. Med. 16: 2529–2542.
Sarkar, S. K. 2002. Some results on false discovery rate in stepwise multiple testing procedures. – Ann. Stat. 30: 239–257.
Schweder, T. and Spjotvøll, E. 1982. Plots of p-values to evaluate many tests simultaneously. – Biometrika 69: 493–502.
Seeger, P. 1968. A note on a method for the analysis of significances en masse. – Technometrics 10: 586–593.
Sterne, J. A. C. and Smith, G. D. 2001. Sifting the evidence – what's wrong with significance tests? – Br. Med. J. 322: 226–231.
Storey, J. D. 2002. A direct approach to false discovery rates. – J. R. Stat. Soc. B 64: 479–498.
Troendle, J. F. 2000. Stepwise normal theory multiple test procedures controlling the false discovery rate. – J. Statist. Plann. Infer. 84: 139–158.
Uitenbroek, D. G. 1997. SISA Binomial. Southampton. (available at http://home.clara.net/sisa/bonfer.htm, accessed: 12–22–2003)
Westfall, P. H. and and. Young, S. S. 1993. Resampling-based multiple testing. – John Wiley & Sons.
Westfall, P. H., Tobias, R. B., Rom, D. et al. 1999. Multiple comparisons and multiple tests using the SAS system. – SAS Institute Inc., Cary, NC
Yekutieli, D. and Benjamini, Y. 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. – J. Statist. Plann. Infer. 82: 171–196.