HOGENBOOM. 1986. Pollen selection in breeding glasshouse tomatoes for low energy conditions, pp. 125–130. *In* D. Mulcahy, G. Bergamini Mulcahy, and E. Ottaviano (eds.), Biotechnology and Ecology of Pollen. Springer-Verlag, N.Y.

SARI-GORLA, M. C. FROVA, AND R. REDAELLI. 1986. Extent of gene expression at the gametophytic phase in maize, pp. 27–32. *In* D. Mulcahy, G. Bergamini Mulcahy, and E. Ottaviano (eds.), Biotechnology and Ecology of Pollen. Springer-Verlag, N.Y.

SEARCY, K., AND D. MULCAHY. 1986. Gametophytic expression of heavy metal tolerance, pp. 159–164. *In* D. Mulcahy, G. Bergamini Mulcahy, and E. Ottaviano (eds.), Biotechnology and Ecology of Pollen. Springer-Verlag, N.Y.

SHIVANNA, K. R., AND J. HESLOP-HARRISON. 1981. Membrane state and pollen viability. J. Ann. Bot. 47:759–766.

SIMON, J., AND J. C. SANFORD. 1986. Induction of gametic selection *in situ* by stylar application of selective agents, pp. 107–112. *In* D. Mulcahy, G. Bergamini Mulcahy, and E. Ottaviano (eds.), Biotechnology and Ecology of Pollen. Springer-Verlag, N.Y.

SNOW, A. A. 1986. Pollination dynamics in *Epilobium canum* (Onagraceae): Consequences for gametophytic selection. Amer. J. Bot. 73:139–151.

SOKAL, R. R., AND F. J. ROHLF. 1981. Biometry, 2nd Ed. Freeman, San Francisco, CA.

STEAD, A. D., I. N. ROBERTS, AND H. G. DICKINSON. 1979. Pollen-pistil interactions in *Brassica oleracea*: Events prior to pollen germination. Planta 146: 211–216.

TANKSLEY, S., D. ZAMIR, AND C. M. RICK. 1981. Evidence for extensive overlap of sporophytic and gametophytic gene expression in *Lycopersicon esculentum*. Science 213:453–455.

THOMSON, J. D. 1986. Pollen transport and deposition by bumble bees in *Erythronium*: Influences of floral nectar and bee grooming. J. Ecol. 74:329–341.

WEEDEN, N. F. 1986. Identification of duplicate loci and evidence for post-meiotic gene expression in pollen, pp. 9–14. *In* D. Mulcahy, G. Bergamini Mulcahy, and E. Ottaviano (eds.), Biotechnology and Ecology of Pollen. Springer-Verlag, N.Y.

WILLING, R. P., AND J. P. MASCARENHAS. 1984. Analysis of the complexity and diversity of mRNAs from pollen and shoots of *Tradescantia palludosa*. Plant Physiol. 75:865–868.

Corresponding Editor: A. G. Stephenson

Evolution, 43(1), 1989, 223–225

ANALYZING TABLES OF STATISTICAL TESTS

WILLIAM R. RICE
Department of Biology, University of New Mexico, Albuquerque, NM 87131

Received March 18, 1988. Accepted August 10, 1988

Tables of statistical tests are commonly analyzed in evolutionary studies. These include analysis-of-variance and regression tables as well as tables of correlation coefficients, chi-square values, G values, Student's t values, etc. To see the prevalence of such tables, one need only refer to a recent issue of *Evolution* (e.g., Evolution 41(6), November 1987, where such tables appeared in 14 of 22 empirical articles). Here, I point out that testing for the statistical significance of component tests is routinely carried out in a biased fashion that liberally judges far too many tests to be significant. I then describe a nonparametric technique, originally proposed by Holm (1979), to eliminate this bias.

So as not to single out any one person unfairly and use his published results as a straw man, consider a hypothetical correlation table examining five variables. The procedure standardly used to evaluate such a table is to carry out an individual significance test on each of the ten correlation coefficients and then denote those significant at the 5% level with an asterisk, those significant at the 1% level with two asterisks, etc. Suppose that two of the ten correlation coefficients were found to be individually significant ($P < 0.05$). Using the "individual significance method," a researcher might spend several journal pages explaining the evolutionary ramifications of the two individually significant correlations observed in the table. Yet there may be insufficient evidence to be 95% confident that there are any nonzero correlations. Appropriate probability values must adjust for the number of simultaneous tests.

One can solve for the probability of observing at least one individually significant correlation (P value less than 0.05) in the above, hypothetical correlation table on the composite null hypothesis ($H_{0,c}$) that all the component correlations are zero. In computer simulations (Appendix), this probability is approximately 40%. Moreover, the probability of observing two or more individual P values less than or equal to 0.05 is about 7%. If a dozen variables were correlated, we would be more than 95% certain, on $H_{0,c}$, that at least one correlation would be judged individually significant by chance alone. Even very small P values are expected in moderately large correlation tables. With a dozen variables, chance alone would produce a P value less than or equal to 0.001 about 7% of the time. The marking of component tests as statistically signif-

icant based on their single-test significance values is therefore clearly inappropriate, yielding far too many significant results. Yet this is standard procedure in articles published in *Evolution* and related journals.

The purpose of the proposed method of calculating table-wide significance levels is to 1) control the probability of incorrectly rejecting one or more true null hypotheses (component $H_{0,i}$), and 2) simultaneously maintain substantial power in detecting one or more component false $H_{0,i}$. The rationale for the method is that a researcher typically uses the minimum significance levels (P values) of component tests to resolve which among a group of $H_{0,i}$ are false. Such a procedure necessarily results in a posteriori significance testing. If no adjustment is made for the number of tests included in the group, then there is no control over the group-wide type-I error rate.

It would appear that most evolutionary biologists are aware, in principle, of the above problem, since simultaneous inference techniques (i.e., multiple comparison techniques such as the Scheffe, Tukey, and Student-Newman-Keuls methods) are routinely used when comparing groups of means. My conjecture is that simultaneous inference techniques are not used when analyzing tables of test statistics because most evolutionary biologists are unaware of a proper procedure. A nonparametric technique that can be used in virtually all applications is the sequential Bonferroni test, originally developed by Holm (1979). A general treatment of simultaneous statistical inference is reviewed in Miller (1981).

The standard Bonferroni technique is described in many general statistics texts. It can be readily shown (e.g., see Miller, 1981) that, if a collection of k tests is simultaneously carried out at the α/k significance level, the probability, on $H_{0,c}$, that at least one component $H_{0,i}$ will be erroneously rejected is less than or equal to α. This inequality does not require that component tests be independent. A major disadvantage of the standard Bonferroni method, however, occurs when more than one component $H_{0,i}$ is false. For example, suppose in the hypothetical correlation table described above that four of the ten correlations were actually different from zero. The standard Bonferroni test has substantially reduced power in detecting more than one false $H_{0,i}$ (see Holm, 1979).

To increase power in detecting more than one false $H_{0,i}$, Holm (1979) introduced the sequential Bonferroni technique. To begin the test, select a significance level (α). Next, replace each test statistic by its corresponding P value and rank the P values from smallest (P_1) to largest (P_k). First consider the smallest P value (P_1). If $P_1 \leq \alpha/k$, then judge that the corresponding test indicates significance at the "table-wide" α level; if the inequality is not met, declare that all tests indicate nonsignificance at the table-wide α level. If and only if $P_1 \leq \alpha/k$, proceed to the second smallest P value (P_2). If $P_2 \leq \alpha/(k-1)$, then judge this test also to indicate statistical significance at the α table-wide level of significance and proceed to the third smallest P value (P_3). If $P_2 > \alpha/(k-1)$, then declare the corresponding test and all other tests with larger P values to indicate nonsignificance at the table-wide α level. Continue in this fashion until the inequality $P_i \leq \alpha/(1+k-i)$ is not met.

The sequential Bonferroni test does not require that component tests be independent. A small gain in power can be achieved when the component tests can be assumed to be independent. In this case, the test criterion becomes, $P_i \leq (1-[1-\alpha]^{1/(1+k-i)})$.

One of the problems with reporting the results of significance tests is that nonsignificance is frequently reported by $P > 0.05$. This does not tell the reader how closely significance was approached. A more informative means of describing the results of a test is to report the minimum significance level at which the test would be judged significant i.e., the P value). A minimum table-wide significance value can be calculated for the sequential Bonferroni test by iteration, but the calculations can be tedious even for small tables. To eliminate this problem, an interactive computer program, written in standard Pascal, is available from the author upon request. The program calculates the minimum table-wide significance of component test statistics.

The advantage of the sequential Bonferroni test over the standard Bonferroni test is increased statistical power. To illustrate, suppose that five allozymes were measured and that each was tested for deviations from Hardy-Weinberg ratios, resulting in P values of 0.4, 0.02, 0.015, 0.012, and 0.01. In this case, only one component test would be judged to be significant at the 5% significance level with the standard Bonferroni test. With the sequential Bonferroni test, four of the tests would be found to be significant. The standard and sequential Bonferroni tests have identical power in detecting a single false $H_{0,i}$, but the sequential technique improves power in detecting any additional false $H_{0,i}$. The sequential Bonferroni test is more powerful than the standard Bonferroni test for the same reason that the Student-Newman-Keuls multiple-range test is more powerful than the Tukey test, i.e., the rejection criteria are less stringent for all tests other than the test with the smallest P value. The increased power of the sequential Bonferroni test is not due to this test being in any way liberal, however, since the probability of a type-I error for the entire test as well as each step of the test is less than or equal to α. Power is gained in the sequential Bonferroni test by eliminating much of the conservativeness found in the standard Bonferroni test when more than one null hypothesis is false. See Miller (1981) for a general discussion of the increased power of sequential tests when more than one $H_{0,i}$ is false.

The sequential Bonferroni method can be used in a wide variety of applications, including all applications in which the standard Bonferroni method has traditionally been used (see Holm [1979] for a discussion). One of the most common situations where the need for simultaneous statistical inference is neglected, besides correlation tables and tables of independent tests, is the evaluation of regression, ANOVA, and ANCOVA coefficients. It is common for a researcher to analyze many Student's t tests from a single regression or ANOVA table. Just as one needs an a posteriori test when carrying out multiple comparisons of the component means in an ANOVA, one also needs an a posteriori test in evaluating the individual significance of variables in a regression, ANOVA, or ANCOVA table. The sequential Bonferroni test can be used for this purpose, although more complex parametric tests are available in some cases.

Many researchers choose not to use the standard Bonferroni test because it is considered to be "overly conservative." It appears to me that many people arrive at this view because they find, for example, that a correlation with a corresponding P value of 0.01 is not significant, on a table-wide basis, within a four-variable correlation table. This is not a fault of the Bonferroni method, however, since the exact probability, on $H_{0,c}$, of observing a P value of 0.01 can be shown to be about 6% (Appendix). Thus, it is not the conservativeness of the Bonferroni method but the number of correlation tests within the table that reduces power. It is true, however, that in some applications a parametric alternative that has higher statistical power can be found. Obviously, these should be used whenever appropriate. When no such alternative is available, the sequential Bonferroni is a useful choice, since much of the true conservativeness of the standard Bonferroni test is eliminated (Holm, 1979).

A question that frequently arises when contemplating the use of a simultaneous-inference test is: what constitutes a family of tests that needs to be analyzed collectively? Should, for example, all the tests within a manuscript be included, so that the reader can be 95% confident that not a single type-I error has been made in the entire manuscript? I think that most would agree that this is going too far. As pointed out by Miller (1981), there is no clear criterion for deciding when a simultaneous-inference significance test is required; it simply depends on how tightly one wants to control the group-wide type-I error rate. For example, suppose a researcher collects a sample of plants and tests for heritable variation for tolerance to three different pathogens under five different environmental conditions. Should the Bonferroni adjustment (k) be 5, for the five environmental conditions applied to each type of pathogen, or should it be 15 for all the tests combined? The answer depends on the probability statement desired. To control the type-I error rate for each individual pathogen, $k = 5$ is appropriate; to control it for all tests simultaneously, $k = 15$ is appropriate. The choice simply depends upon the group-wide type-I error rate desired. In this case it seems quite reasonable to make a separate probability statement for each pathogen.

I suggest that simultaneous inference be used whenever: 1) a group of two or more tests is scanned, and the P values of component tests are used to determine where significant differences occur (i.e., a posteriori testing); or 2) two or more tests (that cannot be pooled) address a common null hypothesis, and rejection of the null hypothesis is possible when only some of the tests are found to be individually significant. In these cases, if no adjustment is made for the number of tests performed, then the probability of a type-I error increases monotonically with the number of tests in the group. If we continue to use nonsimultaneous inference when analyzing data such as correlation tables, then we will spend many journal pages discussing spurious relationships that can be readily explained by chance alone. Clearly, we have more important things to discuss.

Acknowledgments

I thank S. Gaines, K. Ono, and an anonymous referee for helpful comments on the manuscript. This work was supported in part by grant BSR 8407440 from the National Science Foundation.

Literature Cited

Cooke, D., A. H. Craven, and G. M. Clarke. 1982. Basic Statistical Computing. Arnold, London, U.K.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6:65–70.

Miller, R. G., Jr. 1981. Simultaneous Statistical Inference. McGraw Hill, N.Y.

Corresponding Editor: M. K. Uyenoyama

Appendix

All of the probability estimates based on computer simulation were calculated as follows. The simulations were carried out using a microcomputer with a Turbo Pascal (version 3.0) compiler. A simulated $k \times k$-correlation table containing $k(k - 1)/2$ nonredundant correlation coefficients was generated by first using Turbo Pascal's random number generator to produce $N(k[k - 1])$ Uniform(0, 1) random variates, where k is the number of variables in the correlation matrix and N is the sample size for each of the variables. Next, the Uniform(0, 1) variates were used to generate a matrix of $N(k[k - 1]/2)$ standard normal variates using the Box-Muller technique (Cooke et al., 1982). These calculations generated k ordered samples of independent standard normal variates of size N each. Lastly, the product-moment correlation coefficients (r) were calculated between all $k(k - 1)/2$ pairwise combinations of the normal samples.

To calculate the probability that one or more r values exceeded a specified value, 20,000 tables were generated as described above, and the proportion of times that one or more of the $|r|$ values was greater than a specified value was recorded.