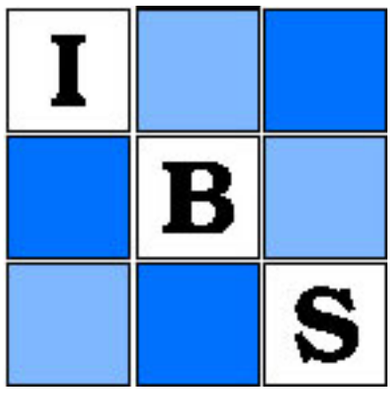

BIOMETRICS 49, 941–945
September 1993

READER REACTION

On Adjusting *P*-Values for Multiplicity

P. H. Westfall and S. S. Young
Texas Tech University, College of Business Administration, Box 42101,
Lubbock, Texas 79409-2101, U.S.A.

1. The Development of PROC MULTTEST

While controversial, the use of multiplicity adjustments has gained acceptance recently in varied fields of scientific endeavor and their associated publications. Multiplicity adjustments have been considered important in pharmaceutical safety determinations involving multiple endpoints, such as in adverse events analysis of clinical trials, and in animal carcinogenicity studies. In situations where the compound tested is completely safe, it is likely to observe false positive indications of one or more untoward outcomes when unadjusted testing methods are used. Multiplicity adjustment is also important in epidemiology and other complicated areas of data analysis, as it offers protection against conclusions that are driven by excessive data mining.

Multiplicity concerns in toxicology and clinical trials prompted the development of the SAS® procedure PROC MULTTEST which calculates adjusted *P*-values for a user-supplied family of tests in a wide variety of applications. *Biometrics* readers will be interested to know that this software is readily available to calculate many (but not all) of the multiplicity-adjusted *P*-values described by Wright (1992). In addition, the software incorporates many improvements and enhancements, such as the ability to incorporate correlations and nonnormal distributions.

The development of PROC MULTTEST started in May 1987. One of us (Westfall) presented a resampling approach to calculating adjusted *P*-values for multiple tests in multivariate binomial models, with special application to the animal carcinogenicity problem. The talk was entitled "Multivariate Binomial Testing," and was given as the keynote address for the Midwestern Biopharmaceutical Statistics Workshop (MBSW) held in Muncie, Indiana. Similar approaches to the same problem were developed concurrently and independently, all of which have been published (Farrar and Crump, 1988; Heyse and Rom, 1988; Westfall, and Young, 1989).

[An interesting *Biometrics* connection should be mentioned here. Dr Young thought that Westfall's (1985) publication concerning simultaneous inference with multivariate binary data might be applied to the multiplicity problem in animal carcinogenicity studies. Dr Young therefore invited Dr Westfall to speak at the MBSW conference on this application.]

The initial precursor to PROC MULTTEST, called PROC MBIN, was developed in 1988. We wrote the specifications for the software, and the coding was performed by Youling Lin at Texas Tech University. Under the auspices of the Pharmaceutical Manufacturing Association, a consortium of drug companies offered financial and intellectual support for the project; specific individuals and companies who contributed are listed below in the Acknowledgements section. The development proceeded through a series of meetings at national and regional statistics conferences; attendees included representatives from the funding organizations and from the United States Food and Drug Administration. From the outset, it was decided that the primary form of output from the software would be the adjusted *P*-value, for the reasons Wright describes in his opening paragraph.

The procedure PROC MBIN was described by Westfall, Lin, and Young (1989) and was donated to SAS Institute Inc. The software performed multiplicity adjustments for multiple tests (*z*-score or exact permutation tests) in multivariate (possibly stratified), multiple group binary outcome situations. The main feature of the output of this software was the use of adjusted *P*-values to summarize many statistical tests. Single-step resampling (bootstrap or permutation), Bonferroni, and Šidák methods were used to calculate the adjustments. The tabular form of the output was much like Tables 2 and 3 of Wright, showing unadjusted *P*-values (which we called "raw" *P*-values) side-by-side with various types of adjusted *P*-values.

Key words: Resampling; Statistical software.

PROC MTEST subsumed PROC MBIN in 1989. The new procedure allowed multiple tests with continuous as well as binary endpoints, and also computed step-down adjustments with or without resampling. One of the step-down options was the adjusted *P*-value based on Holm's (1979) method. These adjustments appear in Wright (1992), Table 3, in the column headed "p_{Holm}." The software also calculated adjusted *P*-values using the step-down Šidák method, reported in Wright's Table 3 under the column headed "$r_{i(\text{Šidák})}$." Wright did not impose monotonicity enforcement on the Šidák adjustments of Table 3, leaving the adjusted *P*-value for test 5 smaller than that of test 4. PROC MTEST performed such adjustments automatically; for example, the value $r_{5(\text{Šidák})} = .25873$ in Wright's Table 3 would have been reported by PROC MTEST as $r_{(\text{Šidák})} = .36179$.

We presented the adjusted *P*-value methodology used in PROC MTEST in December 1989, at the Conference on Applied Statistics and Quality Control, Atlantic City, New Jersey; we also distributed software to perform these analyses to the conference attendees. Later, the methodology was presented at the SAS® User's Group International Conference, Nashville, Tennessee, April 1990, with the research published in the proceedings (Westfall, Lin, and Young, 1990). PROC MTEST was donated to SAS Institute Inc. in 1990.

PROC MULTTEST is currently available as a SAS/STAT® procedure, and it has been available as of the Version 6.07 release. It is essentially identical to PROC MTEST, but contains syntactical modifications. Further details concerning the use and capabilities of the software may be found in SAS Institute Inc. (1992), and Westfall and Young (1993).

2. Resampling-Based Adjusted *P*-Values

The resampling paradigm provides a convenient framework for calculating, interpreting, and explaining adjusted *P*-values. Typical model-based parametric adjustments such as Tukey's method are interpretable in the resampling framework—one simply resamples *parametrically* (e.g., using normal random numbers), rather than *nonparametrically* (e.g., using bootstrap or randomized data). To make this correspondence exact, one must resample an infinite amount of data; in practice, one resamples a large but finite number of data sets. Extremely large simulations are feasible using modern computing machinery, making simulation error negligible.

A natural definition of a resampling-based adjusted *P*-value is simply

$$\tilde{p}_i = \Pr\left(\min_{1 \leq j \leq k} P_j \leq p_i \mid H_0^C \right), \tag{1}$$

where $\tilde{p}_i$ denotes the adjusted *P*-value for test i, p_i is the observed *P*-value for test i, H_0^C symbolizes the complete null hypothesis (all null hypotheses are true), and P_j is the *random* *P*-value for test j, considered under the complete null hypothesis. Westfall et al. (1990) note that this method leads to the adjusted *P*-values denoted p_{Tukey} in Wright's Table 2, provided resampling is done parametrically in the balanced homoscedastic normal model.

Use of definition (1), with simple modifications and probability inequalities, leads naturally to Holm's (1979), Shaffer's (1986), and Holland and Copenhaver's (1987) methods as special cases. In addition, its close connection with simultaneous confidence intervals allows directional decisions, all discussed by Westfall and Young (1993).

Definition (1) also leads to easy calculation using resampling—one simply simulates the distribution of the vector $\{P_1, \ldots, P_k\}$ under the complete null hypothesis. The application of (1) is situation-dependent; for example, in MANOVA data one may estimate the adjustments by resampling the pooled and centered data vectors (this is the bootstrap method). Parametric analyses are accommodated by resampling from an appropriate parametric (usually normal) distribution rather than by resampling the observed data vectors.

The operational interpretation of (1) is simple: the adjusted *P*-value measures how extreme a given *P*-value is, relative to the distribution of the most extreme *P*-value. When faced with a long list of *P*-values, an analyst's natural inclination is to search the list for the smallest one. The resampling framework allows raw *P*-values to be judged against the result of such scanning and selection, when repeated samples are drawn from the same distribution. Thus, the adjusted *P*-value captures the degree of "surprise" that one should feel after having selected one *P*-value from a long list of *P*-values.

While the adjusted *P*-value definition (1) is useful for computation and interpretation, its greatest benefit is that it incorporates distributional characteristics, thereby improving power and/or robustness characteristics of the multiple testing procedure. Typically, the *P*-values $\{P_1, \ldots, P_k\}$ are correlated, and definition (1) naturally accounts for such correlations. In parametric analysis, analytic evaluation of (1) is often difficult (if not impossible) because of such correlation structures, as for example, when performing pairwise comparisons of adjusted means in an ANCOVA model. In such cases, use of

parametric resampling for the calculation of adjusted *P*-values is quite appealing. This approach was described in unpublished work by Westfall and Rom (1990), who used Shaffer's (1986) restricted step-down method as a basis. The research was presented in May 1989, to the Cincinnati Chapter of the American Statistical Association, and in August 1989, at the Joint Statistical Meetings in Washington, D.C.

While it is important to account for correlations among variables and tests, it is often *far more important* to account for distributional characteristics. Wright's comment on our (Westfall and Young, 1989) analysis of Brown and Fears' (1981) data misses this point: he implies that the drastic improvement of our adjustments over the Bonferroni adjustments is mainly due to correlations. This is not true. While incorporating correlations does have an impact on the adjustments, the main contribution is from nonnormality of the data. As we discussed in our 1989 article, binary variables with small total occurrences have little or no opportunity to contribute to multiplicity adjustment, so the effective number of tests is greatly reduced. This point is also developed by Haseman, Winbush, and O'Donnell (1986), Heyse and Rom (1988), Soper and Westfall (1990), and Tarone (1990).

Interested readers may consult Westfall and Young (1993) for a more complete development of these and other issues. Topics developed therein include the following: use of bootstrap, permutation, and parametric resampling; adjusted *P*-values when there are logical restrictions; control of familywise error rates under complete *and partial* null hypotheses; control of *directional* error rates; asymptotic consistency; the general multivariate regression model (includes ordinary regression, ANOVA, ANCOVA, MANOVA, and MANCOVA all as special cases); application to multivariate binary and categorical data; use of historical data; weighted adjustments; comparison of variances and correlations; infinitely many tests; power; and much more. Dozens of worked examples are provided using real data sets.

Acknowledgements

The following individuals (in alphabetical order) provided intellectual support in various stages of the development of the procedures PROC MBIN and PROC MTEST, precursors to PROC MULTTEST: R. M. Bittman, W. R. Fairweather, J. F. Heyse, R. Langston, W. C. Louv, M. Murphy, D. Rom, H. J. Rostami, A. J. Roth, S. J. Ruberg, L. Sanathanan, K. A. Soper, R. N. Tamura, and R. Wolfinger.

Development of the software would not have been possible without the financial support of Eli Lilly and Co., Glaxo Inc., Merck Sharp & Dohme Co., Merrell Dow Co. (now Marion Merrell Dow Co.), G. D. Searle and Co., Smith Kline & French Co. (now Smith Kline & Beacham), and the Upjohn Corporation.

References

Brown, C. C. and Fears, T. R. (1981). Exact significance levels for multiple binomial testing with application to carcinogenicity screens. *Biometrics* **37,** 763–774.

Farrar, D. B. and Crump, K. S. (1988). Exact tests for any carcinogenic effect in animal bioassays. *Fundamental and Applied Toxicology* **11,** 652–663.

Haseman, J. K., Winbush, J. S., and O'Donnell, M. W. (1986). Use of dual control groups to estimate false positive rates in laboratory animal carcinogenicity studies. *Fundamental and Applied Toxicology* **7,** 573–584.

Heyse, J. F. and Rom, D. (1988). Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biometrical Journal* **30,** 883–896.

Holland, B. S. and Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics* **43,** 417–424.

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

SAS Institute Inc. (1992). *SAS/STAT® Software: Changes and Enhancements, Release 6.07.* SAS® Technical Report P-229. Cary, North Carolina: SAS Institute Inc.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association* **81,** 826–831.

Soper, K. A. and Westfall, P. H. (1990). Monte Carlo estimation of significance levels for carcinogenicity tests using univariate and multivariate models. *Journal of Statistical Computation and Simulation* **37,** 189–209.

Tarone, R. E. (1990). A modified Bonferroni method for discrete data. *Biometrics* **46,** 515–522.

Westfall, P. H. (1985). Simultaneous small-sample multivariate Bernoulli confidence intervals. *Biometrics* **41,** 1001–1013.

Westfall, P. H., Lin, Y., and Young, S. S. (1989). A procedure for the analysis of multivariate binomial data with adjustments for multiplicity. *SAS Users Group International Conference Proceedings* **14,** 1385–1392.

Westfall, P. H., Lin, Y., and Young, S. S. (1990). Resampling-based multiple testing. *SAS Users Group International Conference Proceedings* **15,** 1359–1364.
Westfall, P. H. and Young, S. S. (1989). *P*-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association* **84,** 780–786.
Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment.* New York: Wiley-Interscience.
Wright, S. P. (1992). Adjusted *P*-values for simultaneous inference. *Biometrics* **48,** 1005–1013.

Received February 1993; *revised March* 1993; *accepted March* 1993.

RESPONSE

I am pleased to have this opportunity to respond to the Reader Reaction of Westfall and Young. First of all, I would like to thank them for pointing out my errors. In particular, I direct the reader's attention to the last column ($r_{i(\check{S}id\acute{a}k)}$) of Table 3 in the original article. As they mention, I failed to impose monotonicity on the "*P*-values" therein. Actually, the omission was intentional, making this column analogous to an earlier, Bonferroni-based r_i column in the same table which also lacks monotonicity constraints. Strictly speaking, these columns do not contain adjusted *P*-values but are the "raw materials" from which adjusted *P*-values are obtained by imposing monotonicity. This should have been made clearer in the original article, and I thank Westfall and Young for bringing it up.

Mostly what I want to do in this response is to acknowledge the importance of the work of Westfall and Young, which they have so ably outlined in their Reader Reaction. I must say, frankly, that if I were sitting down today to write my article on adjusted *P*-values—I probably wouldn't. There would be no need to. Westfall and Young have literally written the book on the subject. Obviously I didn't feel this way at the time I wrote. My perception then was that adjusted *P*-values, as a way to report results from simultaneous inference, had tremendous but largely unrecognized potential. I thought an article on the subject was worth doing to bring it to the attention of the statistical community. (I am glad to learn that, even at that time, one segment of the statistical community, the pharmaceutical industry, was already making use of adjusted *P*-values.) The work of Westfall and Young at that time was, at least in my view, relatively new and not widely known. And I must confess that during the revision of the article (by which time I was more familiar with their work and, frankly, excited by it), I limited myself to satisfying the reviewers who, it should be noted, made no reference to the work of Westfall and Young. I mention this only as a way of indicating that their work was then, and perhaps still is, less well known than it deserves to be in some parts of the statistical community. I would like to think that my article has contributed in some way to raising the level of awareness about the applicability of adjusted *P*-values. It may turn out that the most valuable result of my article is that it elicited a reaction from Westfall and Young, thereby bringing more attention to their work and to adjusted *P*-values in general.

As a service to readers, I would like to mention some additional articles that refer to adjusted *P*-values but using different terminology. Dunnett and Tamhane (1992) use the term "joint *P*-value" for what I have called an adjusted *P*-value, but their definition is the same (the smallest *overall* significance level at which a hypothesis can be rejected). In a different setting—that of sequential rather than simultaneous inference—Jennison and Turnbull (1989) have championed the use of repeated confidence intervals. But after all, a *P*-value is just the significance level (for which there is a corresponding confidence level) for which the null-hypothesized-value of a parameter falls at the limit of the confidence interval. Using this connection between *P*-values and confidence intervals, Jennison and Turnbull (1990) define a "repeated *P*-value," a sequential-testing version of an adjusted *P*-value. The point is that adjusted *P*-values, whatever they are called, are useful in a wide variety of settings.

It is probably fair to say that the new statistical procedures (such as adjusted *P*-values and related confidence intervals) that become widely used are not the ones about which journal articles are written but the ones that get incorporated into popular statistical packages. Westfall and Young realized this early on, and they are to be commended as much for their efforts in making software available as for their theoretical work. I hope to see more of the procedures they discuss in their book incorporated into software in the near future.

References

Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association* **87,** 162–170.

Jennison, C. and Turnbull, B. W. (1989). Interim analyses: The repeated confidence interval approach (with Discussion). *Journal of the Royal Statistical Society, Series B* **51,** 305–361.

Jennison, C. and Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science* **5,** 299–317.

S. Paul Wright
University of Tennessee
328 SMC, Statistics Department
Knoxville, Tennessee 37996-0532, U.S.A.