A Stepwise Resampling Method of Multiple Hypothesis Testing

James F. TROENDLE*

This article introduces a method of multiple hypothesis testing that combines the idea of sequential multiple testing procedures with the structure of resampling methods. The method can be seen as an alternative to the analytic method of Dunnett and Tamhane, which requires a specific distributional form. Resampling incorporates the covariance structure of the data without the need for distributional assumptions. Recent work by Westfall and Young has shown that a step-down resampling method is asymptotically consistent when adjusted p values can be obtained exactly for continuous data. This article shows that in the case of a comparison of two groups on multiple outcomes, those results are generalizable to discrete data where exact adjusted p values are not available. It is shown that the method asymptotically attains the desired level for controlling the experimentwise probability of a type I error.

KEY WORDS: Experimentwise; P value; Test statistic; Type I error.

1. INTRODUCTION

When an experiment records several different variables for each subject in two (or more) groups, it is often desired that a statistical test identify any variable for which the group means differ significantly. If the researchers record enough different variables, it is apparent that they will likely find one significant comparison by chance alone. Many different methods of treating such simultaneous testing situations have been proposed. One such method is concerned with controlling the probability of at least one type I error at or below a specified level α. This is called the experimentwise error rate control approach (Dunnett and Tamhane 1992). In some cases (e.g., a particular comparison has been isolated prior to the experiment) it may be unnecessary to make any adjustments to the individual tests. But if the goal of the experiment is to select a few significant variables from a large pool, then I recommend the adoption of the experimentwise approach. In any case, in this article I shall assume that this approach has been taken.

The simplest and most general multiple hypothesis testing methods are the Bonferroni and its improvements (Hochberg 1988; Holm 1979; Simes 1986). The primary advantage of methods based on the Bonferroni inequality is that they are applicable to any multiple hypothesis testing situation; they require no assumption about the data or dependence between comparisons. Another advantage of such methods is their simplicity in use. They require very little computation once the individual comparisons have been completed to yield p values. Unfortunately, they are ultraconservative in that no attempt is made to incorporate the dependence between tests. A second group of testing procedures is centered around the assumption that the data are multivariate normal (Dunn 1959; Dunn and Massey 1965; Dunnett and Tamhane 1992; Sidak 1967, 1971). Dunnett and Tamhane (1992) analytically approximated the joint distribution of k student t variables used to compare normally distributed estimates. The method requires assumptions about the structure of the data that may be quite strong. Another approach to the multiple testing problem is to use resampling techniques to incorporate the dependence structure of the tests. Westfall and Young (1989) showed how resampling can be useful in a single-step manner to adjust p values. Analogous to the improvements of the Bonferroni procedure, stepwise resampling is presented as an improvement to their single-step resampling method of p-value adjustment. Stepwise resampling can be thought of as a resampling approximation to the joint distribution analytically approximated by Dunnett and Tamhane (1992), which does not require the structural assumptions yet retains a large portion of the increased power. In this article I introduce a method of sequential testing that relies on sample reuse to estimate probabilities needed to control the experimentwise error rate. The method is illustrated in the setting of comparing k treatment and control means. The method is shown quite generally to be asymptotically conservative; that is, the probability that any type I error is committed is asymptotically bounded above by α.

2. BONFERRONI PROCEDURES

Before introducing the stepwise resampling method, we shall review the Bonferroni procedure and its improvements. Consider the problem of simultaneously testing k univariate null hypotheses $H_1, H_2, \ldots, H_k$ based on the observed values $t_1, t_2, \ldots, t_k$ of some test statistics $T_1, T_2, \ldots, T_k$. If p_i is the p value computed from the observed value t_i for $i = 1, \ldots, k$, then the Bonferroni procedure rejects any H_i with $p_i \leq \alpha/k$. The factor $1/k$ accounts for the fact that there are k possible true null hypotheses, and the rejection of any one of these could cause a type I error. The Bonferroni procedure can be improved upon by realizing that once one has rejected one null hypothesis (assuming it was false), there are only $k-1$ possible true null hypotheses to guard against rejecting, and so one can reduce the factor $1/k$ to $1/(k-1)$ in the Bonferroni procedure. This gives the step-down procedure of Holm (1979):

HM Algorithm

1. Order the p values and hypotheses

$$P_{(1)} \geq \cdots \geq P_{(k)}$$

corresponding to $H_{(1)}, \ldots, H_{(k)}$.

* James F. Troendle is Staff Fellow, National Institute of Child Health and Human Development, Bethesda, MD 20892. The author thanks James Mills for the use of the data presented here and Young Jack Lee for his helpful suggestions.

2. Let $i = 1$.
3. If $P_{(k-i+1)} > \alpha/(k-i+1)$, then accept all the remaining hypotheses $H_{(k-i+1)}, \ldots, H_{(1)}$ and STOP.
4. If $P_{(k-i+1)} \leq \alpha/(k-i+1)$, then reject $H_{(k-i+1)}$, increment i, and RETURN to Step 3.

By starting with the largest p value and sequentially accepting hypotheses, one gets the procedure of Hochberg (1988):

HM Algorithm

1. Order the p values and hypotheses

$$P_{(1)} \geq \cdots \geq P_{(k)}$$

corresponding to $H_{(1)}, \ldots, H_{(k)}$.

2. Let $i = 1$.
3. If $P_{(i)} \leq \alpha/i$, then reject all the remaining hypotheses $H_{(i)}, \ldots, H_{(k)}$ and STOP.
4. If $P_{(i)} > \alpha/i$, then accept $H_{(i)}$, increment i, and RETURN to Step 3.

Hochberg's procedure is called a step-up algorithm. All three procedures control the probability of a type I error at level α. But it can be easily seen that Hochberg's procedure is uniformly more powerful than Holm's, which is uniformly more powerful than Bonferroni's.

The sequential multiple hypothesis testing procedures just described, though not dependent on any particular distribution form, are limited by their reduction of the data to p values. Information in the data is lost when only the p values are used. Sequential procedures can be improved by methods that incorporate the covariance structure of the data. Dunnett and Tamhane (1992) used an analytic approach to approximate the multivariate distribution of the test statistics when the data have a special form. This approach requires multivariate normal estimates with equal marginal variances and a known common correlation coefficient. Resampling is a computational approach geared toward making use of the dependence structure of the data without distribution restrictions. Westfall and Young (1989) used the resampling approach to compute adjusted p values for multivariate binomial data, but considered only a single-step adjustment.

3. STEPWISE RESAMPLING METHODS

Suppose in the comparison of a treatment group and a control group, k outcome variables are observed. Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be independent k-dimensional random variables, representing the control group, with common joint distribution

$$\mathbf{X}_i \sim F(\boldsymbol{\mu}_1; \mathbf{x}), \qquad i = 1, \ldots, N,$$

where $\boldsymbol{\mu}_1$, the mean of the distribution, consists of the component means

$$\boldsymbol{\mu}_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1k} \end{pmatrix}.$$

For the treatment group, let $\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{2N}$ be independent k-dimensional random variables, with

$$\mathbf{X}_i \sim F(\boldsymbol{\mu}_2; \mathbf{y}), \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2k} \end{pmatrix}, \qquad i = N+1, \ldots, 2N.$$

The hypothesis tests of interest may be either one-sided or two-sided comparisons of the component means

$$H_i : \mu_{1i} = \mu_{2i} \quad \text{versus} \quad K_i : \mu_{1i} < \mu_{2i}$$

or

$$H_i : \mu_{1i} = \mu_{2i} \quad \text{versus} \quad K_i : \mu_{1i} \neq \mu_{2i}$$

for $i = 1, \ldots, k$, and it is desired to have no more than an α chance of committing a type I error. Typically, test statistics $T_1, \ldots, T_k$ are available for testing each individual hypothesis separately. For example, if the data are multivariate normal with a common variance, then each T_j would be a t statistic with $2N - 2$ degrees of freedom. Let the observed values of $T_1, \ldots, T_k$ be $t_1, \ldots, t_k$. These values are ordered along with the hypotheses

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(k)}$$

corresponding to

$$H_{(1)}, H_{(2)}, \ldots, H_{(k)}.$$

A conservative approach would be to reject each $H_{(i)}$ as long as

$$P_{H_{(k)}}\{\max_{1 \leq j \leq k} T_j \geq t_{(i)}\} \leq \alpha, \tag{1}$$

where $P_{H_{(k)}}$ is the probability under $H_{(1)}, \ldots, H_{(k)}$. In this case we may estimate the probability given in (1) by resampling, because under $H_{(1)}, \ldots, H_{(k)}$ the control and treatment groups are identically distributed. Let '$*$' denote a random variable whose value is obtained by random sampling from the entire original data. Then a resample

$$\mathbf{X}_1^*, \ldots, \mathbf{X}_{2N}^*$$

is a sample from

$$\mathbf{x}_1, \ldots, \mathbf{x}_{2N}.$$

Sampling could be done either with or without replacement. Also, define T_i^*, $i = 1, \ldots, k$, to be the test statistics corresponding to the resampled data. For each resample, we calculate whether

$$\max_{1 \leq j \leq k} T_j^* \geq t_{(i)}, \qquad i = 1, \ldots, k \tag{2}$$

and keep track of the proportion of resamples for which (2) holds. If after many resamples this proportion is less than or equal to α, then we reject $H_{(i)}$. In fact, the proportion of times that (2) holds gives us an adjusted p value in the sense of Westfall and Young (1989). This method of testing is easily seen to be overly conservative in the case of more than one false null hypothesis. To illustrate this point, consider the following example.

Example. Suppose that one is testing one-sided hypotheses on bivariate normal data with $\mu_1 = \binom{0}{0}$, $\mu_2 = \binom{50}{5}$, and identity covariance matrix. Suppose that $N = 5$ and the data turn out to be

$$\text{control group} \quad \binom{0}{1}, \binom{1}{0}, \binom{-1}{0}, \binom{-1}{-1}, \binom{1}{0}$$

$$\text{treatment group} \quad \binom{50}{4}, \binom{49}{5}, \binom{52}{5}, \binom{48}{6}, \binom{51}{5}.$$

In this case,

$$t_2 = 5 - 0 = 5,$$

and if resampling is done without replacement, then the event

$$t_1^* \geq 29 - 21 = 8$$

will occur with probability .5. Therefore, the probability of rejecting H_2 by the foregoing resampling method with $\alpha < .5$ will converge to zero as the number of resamples goes to infinity. Note that if the resampling is with replacement, then the same conclusion holds for $\alpha < .25$.

Consider the point at which we have decided to reject $H_{(k)}$, the hypothesis corresponding to the largest test statistic. If $H_{(k)}$ is actually true, then we have already committed a type I error. Therefore, from the standpoint of controlling the probability of committing any type I error, we can assume that $H_{(k)}$ is false. Thus in deciding whether to reject $H_{(k-1)}$, we should not let the component corresponding to $t_{(k)}$ influence our decision. Consequently, I propose resampling the data with the component corresponding to $t_{(k)}$ deleted. Then in this partial resample, we calculate whether

$$\max T_j^* \geq t_{(k-1)},$$

where the maximum extends over all $1 \leq j \leq k$ except for the index of the component corresponding to $t_{(k)}$. By continuing in this manner, we arrive at a stepwise resampling (SR) method:

SR Algorithm 1

1. Order the observed test statistics and hypotheses

$$t_{(1)} \leq \cdots \leq t_{(k)}$$

$$\text{corresponding to} \quad H_{(1)}, \ldots, H_{(k)}.$$

2. Let $i = 1$.
3. By repeatedly resampling the data, estimate

$$\alpha_i = P_{H_{(k-i+1)}}\{T_{(k-i+1)}^{(k-i+1)} \geq t_{(k-i+1)}\}, \tag{3}$$

where $T_{(k-i+1)}^{(k-i+1)}$ is the largest of the $(k - i + 1)$ test statistics corresponding to $t_{(1)}, \ldots, t_{(k-i+1)}$.
4. If $\alpha_i \geq \alpha$, then accept the remaining hypotheses $H_{(1)}, \ldots, H_{(k-i+1)}$ and STOP.
5. If $\alpha_i < \alpha$, then reject $H_{(k-i+1)}$, increment i, and RETURN to Step 3.

When we estimate α_i in (3), we are essentially using the estimated distribution of $T_{(k-i+1)}^{(k-i+1)}$ from resampling. This gives rise to a second version of SR:

SR Algorithm 2

1. Order the observed test statistics and hypotheses

$$t_{(1)} \leq \cdots \leq t_{(k)}$$

$$\text{corresponding to} \quad H_{(1)}, \ldots, H_{(k)}.$$

2. Let $i = 1$.
3. By repeatedly resampling the data, estimate $\varphi_\alpha^{(k-i+1)}$ such that

$$\alpha = P_{H_{(k-i+1)}}\{T_{(k-i+1)}^{(k-i+1)} \geq \varphi_\alpha^{(k-i+1)}\}, \tag{4}$$

where $T_{(k-i+1)}^{(k-i+1)}$ is the largest of the $(k - i + 1)$ test statistics corresponding to $t_{(1)}, \ldots, t_{(k-i+1)}$.
4. If $t_{(k-i+1)} \leq \varphi_\alpha^{(k-i+1)}$, then accept the remaining hypotheses $H_{(1)}, \ldots, H_{(k-i+1)}$ and STOP.
5. If $t_{(k-i+1)} > \varphi_\alpha^{(k-i+1)}$, then reject $H_{(k-i+1)}$, increment i, and RETURN to Step 3.

The two SR algorithms are equivalent in the sense that they are identical in the idealized case where the sample is so large and the number of resamples so large that $\varphi_\alpha^{(k-i+1)}$ is determined exactly.

In general, we must consider the error of estimating α_i by an estimate based on a finite number of resamples of the data. Suppose now that M resamples are used, each of size $2N_0$, to estimate α_i given by (3). Let T_{ij}^* be the value of T_i^* in the jth resample. We propose using the estimate

$$\alpha_w^* = \frac{1}{M}\sum_{j=1}^{M} I[\max_i T_{ij}^* \geq t_{(k-w+1)}], \tag{5}$$

where the maximum extends over all i corresponding to $t_{(1)}, \ldots, t_{(k-w+1)}$, and $I[A]$ is the indicator function of the event A. We now can state the practical SR algorithm.

SR Algorithm 3

1. Order the observed test statistics and hypotheses

$$t_{(1)} \leq \cdots \leq t_{(k)}$$

$$\text{corresponding to} \quad H_{(1)}, \ldots, H_{(k)}.$$

2. Let $w = 1$.
3. Generate T_{lj}^*, the lth test statistic in the jth resample, for $l = 1, \ldots, k$ and $j = 1, \ldots, M$.
4. Define α_w^* by (5).
5. If $\alpha_w^* \geq \alpha$, then accept the remaining hypotheses $H_{(1)}, \ldots, H_{(k-w+1)}$ and STOP.
6. If $\alpha_w^* < \alpha$, then reject $H_{(k-w+1)}$, increment w, and RETURN to Step 4.

4. SR SIGNIFICANCE LEVEL

I first consider the experimentwise type I error rate when using SR in the idealized case where resampling can be considered as perfect (i.e., SR Algorithm 2 is used where $\varphi_\alpha^{(j)}$ are known).

Theorem 1. In the multiple testing framework of Section 3, if each of k hypotheses are tested by an application of SR Algorithm 2 with $\varphi_\alpha^{(j)}$ known, then the probability of a type I error being committed is less than or equal to α.

Table 1. Proportion of Times That a Type I Error is Committed

Number of resamples	HB	SR
500	.0489	.0505
1,000	.0480	.0501

NOTE: Results for 500 resamples are based on 500,000 replications of the test, and results for 1,000 resamples are based on 200,000 replications. The nominal level is .05.

A proof of Theorem 1 can be found in the Appendix.

I now consider the practical case of applying SR Algorithm 3 to the multiple testing framework of Section 3. The theorem that follows treats the maximum of the test statistics $T_1, \ldots, T_k$ as a function of the observations. Let l be a given integer with $1 \le l \le k$ and suppose that a sample of size $2N_0$ is used to obtain $T_1, \ldots, T_l$. The maximum of $T_1, \ldots, T_l$ is a function of the random variables $X_1, \ldots, X_{2N_0}$:

$$\max_{1 \le i \le l} T_i = G(\mathbf{X}_1, \ldots, \mathbf{X}_{2N_0}).$$

The only assumption made about G is that its expectation exists and is finite,

$$E[G(\mathbf{X}_1, \ldots, \mathbf{X}_{2N_0})] < \infty. \tag{6}$$

For example, if $\mathbf{X}_1, \ldots, \mathbf{X}_{2N_0}$ are multivariate normal with known equal marginal variance σ^2, then T_i is the z statistic

$$T_i = \frac{(\sum_{j=1}^{N_0} X_{ji} - \sum_{j=N_0+1}^{2N_0} X_{ji})}{\sqrt{2N_0}\,\sigma}$$

and

$$G(\mathbf{X}_1, \ldots, \mathbf{X}_{2N_0}) = \max_i \left(\sum_{j=1}^{N_0} X_{ji} - \sum_{j=N_0+1}^{2N_0} X_{ji} \right),$$

which clearly satisfies (6).

Theorem 2. In the multiple testing framework of Section 3, suppose that each of k hypotheses is tested by an application of SR Algorithm 3. Assume that (6) holds and that the test statistics $t_1, \ldots, t_k$ are based on an initial sample of size $2N_0$, whereas resampling is done from the increasing total sample of size $2N$. If each resample consists of sampling with replacement a total of $2N_0$ times, then

$$\overline{\lim_{N \to \infty}} \lim_{M \to \infty} P\{\text{type I error}\} \le \alpha. \tag{7}$$

A proof of Theorem 2 can be found in the Appendix.

5. SIMULATED POWER

The nominal level and power of the SR method was checked by three simulation experiments and compared to that of the method due to Hochberg (1988), hereinafter denoted by HB, and to that of the step-up method of Dunnett and Tamhane (1992), denoted by SU. The SR method was programmed in Fortran; and a copy of the subroutine that requires the test statistics for each comparison as input is available from the author on request.

Table 2. Proportion of Times That Each Hypothesis is Rejected

Hypothesis	Effect size	HB	SR
1	.1	.0658	.0680
2	.2	.2877	.2956
3	0	.0136	.0135
4	0	.0137	.0135
5	0	.0132	.0131

NOTE. Results are with 1,000 resamples and are based on 100,000 replications of the test. The nominal level is .05.

5.1 SR versus HB

The first simulation set was performed with multivariate normal data, with $k = 5$ and covariance matrix

$$\begin{pmatrix} 1.0 & .2 & .6 & -.6 & .2 \\ .2 & 1.0 & -.2 & .2 & .2 \\ .6 & -.2 & 1.0 & -.2 & -.2 \\ -.6 & .2 & -.2 & 1.0 & -.6 \\ .2 & .2 & -.2 & -.6 & 1.0 \end{pmatrix}.$$

The variances of each component are assumed to be equal, and the common value is assumed to be known. The univariate hypothesis tests of interest are

$$H_i : \mu_{1i} = \mu_{2i} \quad \text{versus} \quad K_i : \mu_{1i} \ne \mu_{2i},$$

and the resampling is done without replacement. Therefore, each resample is a permutation of the original data. Table 1 shows a comparison of the experimentwise type I error rate for the two methods. The nominal level was .05. Two sets of results are presented; in the first the SR method is based on 500 resamples per experiment, and in the second the SR method is based on 1,000 resamples. The SR method approximates the nominal significance level more closely than the HB method, and the accuracy improves with an increasing number of resamples. Table 2 provides a comparison of the power to reject individual hypotheses for a particular alternative. The effect size is the difference of the true means of the treatment and control groups. The SR method rejects the false null hypotheses (1 and 2) more often than the HB method and rejects the true null hypotheses at about the same rate as the HB method. Table 3 shows that with many false null hypotheses, a step-up test like HB has more power than a step-down test like SR.

5.2 SR versus HB, SU: Known and Equal Covariances

A second set of simulations were performed to compare the nominal level and power of the SR method with that of

Table 3. Proportion of Times That Each Hypothesis is Rejected

Hypothesis	Effect size	HB	SR
1	.2	.5177	.5115
2	.5	.9975	.9972
3	.6	.9999	.9999
4	.6	1.0000	1.0000
5	.5	.9976	.9973

NOTE: Results are with 500 resamples and are based on 200,000 replications of the test. The nominal level is .05.

Table 4. Proportion of Times That a Type I Error is Committed

Number of resamples	HB	SR	SU
1,000	.0416	.0485	.0496

NOTE: Results are based on 100,000 replications of the test. The nominal level is .05.

the SU method. We simulated multivariate normal data, with $k = 5$ and covariance matrix

$$\begin{pmatrix} 1.0 & .5 & .5 & .5 & .5 \\ .5 & 1.0 & .5 & .5 & .5 \\ .5 & .5 & 1.0 & .5 & .5 \\ .5 & .5 & .5 & 1.0 & .5 \\ .5 & .5 & .5 & .5 & 1.0 \end{pmatrix}.$$

The variances of each component are assumed to be equal, and the common value is assumed to be unknown. In addition, the correlation between any two components is assumed to be equal and known. Instead of attempting to program the SU method, we used the critical constants found in table 2 of Dunnett and Tamhane's paper with $\rho = .5$ and an infinite number of degrees of freedom. The univariate hypothesis tests of interest are

$$H_i : \mu_{1i} = \mu_{2i} \quad \text{versus} \quad K_i : \mu_{1i} \neq \mu_{2i},$$

and the resampling is done with replacement. A two-sample t test is used, where the estimate of the variance is pooled among the two groups to get test statistics for each univariate comparison. The nominal significance level in each case is .05. A comparison of the experimentwise type I error rates appears in Table 4, and the proportion of times each univariate hypothesis is rejected by the methods is shown for a particular alternative in Table 5.

The tables show that under conditions favorable to the SU method, the power of the SU method (as judged by each hypothesis separately) is slightly greater than that of the SR method; however, the power of SR is closer to the power of SU than to the power of HB. The SR method appears to provide a large portion of the benefit gained by the use of the correlation structure without the need for the structural assumptions of SU.

5.3 SR versus HB, SU: Unequal Covariances

A final set of simulations were performed to compare the performance of the SR and SU methods when the basic assumptions on the data are not true. We simulated multivariate normal data, with $k = 5$ and covariance matrix

Table 6. Proportion of Times That a Type I Error is Committed

Number of resamples	HB	SR	SU
1,000	.0390	.0491	.0454

NOTE: Results are based on 50,000 replications of the test. The nominal level is .05.

$$\begin{pmatrix} 1.0 & .8 & .2 & .6 & .4 \\ .8 & 1.0 & .4 & .6 & .2 \\ .2 & .4 & 1.0 & .6 & .8 \\ .6 & .6 & .6 & 1.0 & .6 \\ .4 & .2 & .8 & .6 & 1.0 \end{pmatrix},$$

although the correlation between any two components was assumed to be known at .5. The variances of each component are assumed to be equal, and the common value is assumed to be unknown. Again, we used the critical constants found in table 2 of the Dunnett and Tamhane (1992) paper, with $\rho = .5$ and an infinite number of degrees of freedom to implement the SU method. The univariate hypothesis tests of interest are

$$H_i : \mu_{1i} = \mu_{2i} \quad \text{versus} \quad K_i : \mu_{1i} \neq \mu_{2i},$$

and the resampling is done with replacement. Again, a two-sample t-test is used, where the estimate of the variance is pooled among the two groups to get test statistics for each univariate comparison. The nominal significance level in each case is .05. Table 6 shows the comparison of experimentwise type I error rates. The proportion of times that each univariate hypothesis is rejected by the methods is shown for a particular alternative in Table 7. Here the SR method rejects the false hypotheses most often and approximates the nominal level most closely, whereas the SU method suffers a drop in the experimentwise type I error rate, indicating the SU method is sensitive to the assumption about the data correlation. In summary, the SR method is more generally applicable to multiple testing problems than the SU method, while providing a good approximation to the SU method when the assumptions of the SU method are met.

6. ANALYSIS OF MALFORMATION DATA

In this section the SR method is applied to Bernoulli data on the presence or absence of 55 different types of malformations in infants born of diabetic and nondiabetic women (treatment and control). (An article on this subject matter is under preparation, so the malformation names have been

Table 5. Proportion of Times That Each Hypothesis is Rejected

Hypothesis	Effect size	HB	SR	SU
1	0	.0154	.0172	.0172
2	.3	.6701	.6893	.6932
3	0	.0142	.0159	.0161
4	.2	.3059	.3245	.3261
5	0	.0148	.0169	.0169

NOTE: Results are with 1,000 resamples and are based on 100,000 replications of the test. The nominal level is .05.

Table 7. Proportion of Times That Each Hypothesis is Rejected

Hypothesis	Effect size	HB	SR	SU
1	0	.0205	.0243	.0224
2	.1	.0786	.0940	.0854
3	.5	.9941	.9955	.9952
4	.2	.0179	.0215	.0198
5	−.3	.6957	.7261	.7153

NOTE. Results are with 1,000 resamples and are based on 100,000 replications of the test. The nominal level is .05.

withheld.) The data, summarized in Table 8, come from the Diabetes in Early Pregnancy study (Mills et al. 1988). Fisher's exact test for 2×2 tables is used to give p values for each of the 55 comparisons between the treatment and control groups. The null hypothesis in each test is that the proportion of subjects with the condition is the same in the two groups, whereas the alternative is that the proportion is higher in the treatment group. Notice that for the SR method, we can either take T_i to be the test statistics from Fisher's exact test after normalization or we can use the p values from Fisher's exact test by noting that the event $T > t$ is equivalent to $P < p$, where P and p are the random and observed p values corresponding to T and t. We used the p values due to their availability in this case.

If in addition to performing the hypothesis tests, it is desired to report adjusted p values for the comparisons, a single-step resampling method (SS) is available (Westfall and Young 1989). By resampling for each comparison this method estimates the probability that when all null hypotheses are true, a p value as small as that observed will result. It is easy to see that the value of α_w^* of the SR method that corresponds to the smallest observed p value will equal the adjusted p value of SS, and that each subsequent α_w^* will be smaller than the corresponding p value of SS. The α_w^*'s generated by the SR method can be used to determine adjusted p values in the following way: $\text{Max}_{1 \le w \le r}\ \alpha_w^*$ is the adjusted p value for the hypothesis corresponding to $t_{(k-r+1)}$, and equals the smallest level at which the corresponding hypothesis could be rejected by the SR method when controlling the experimentwise error rate. In Table 9 we provide a comparison of the adjusted p values obtained by the SS method to those obtained by the SR method on the malformation data. The results reported are for the six comparisons that had observed p values less than .05, because these are the most interesting. Both methods are based on the same 10,000 with replacement samples of size 744 from the 744 subjects. Note that the Bonferroni, Holm, and Hochberg methods would all fail to reject all but the null hypothesis corresponding to malformation #32 when testing with $\alpha = .05$. It is thought that the improvement over SS that SR will attain should be more pronounced when the number of false null hypotheses is large.

Table 8. Number of Babies With Each Malformation Type

Malformation code	Group		Malformation code	Group	
	Diabetic	Nondiabetic		Diabetic	Nondiabetic
1	7	2	29	45	24
2	3	0	30	38	7
3	2	1	31	1	0
4	60	22	32	44	8
5	3	0	33	0	1
6	3	1	34	2	0
7	1	1	35	1	1
8	3	0	36	12	2
9	3	0	37	3	0
10	3	0	38	6	2
11	2	0	39	8	2
12	26	9	40	10	4
13	21	6	41	8	5
14	18	12	42	7	5
15	15	8	43	1	0
16	23	6	44	1	0
17	20	9	45	5	2
18	10	0	46	2	4
19	8	3	47	2	4
20	1	1	48	28	16
21	1	2	49	28	15
22	1	0	50	10	17
23	8	3	51	13	18
24	8	3	52	1	6
25	107	52	53	4	0
26	24	16	54	16	5
27	19	4	55	4	1
28	6	15			

NOTE: There were 467 babies born to diabetic women and 277 babies born to nondiabetic women.

Table 9. Unadjusted and Adjusted p Values for DIEP Data

Malformation number	p value	Adjusted p value	
		SS	SR
32	.00033	.0026	.0026
30	.00097	.0095	.0088
18	.00916	.1172	.1090
4	.02424	.3119	.2885
27	.03290	.3998	.3604
16	.04228	.4954	.4436

NOTE: Adjusted p values are based on 10,000 with replacement resamples.

7. CONCLUSIONS

After submission of the original version of this article, I was alerted to the work of Westfall and Young (1993). Westfall and Young considered the same step-down algorithm as that given here. But while we have presented the method in the case of multiple outcomes where the marginal distributions are all the same, Westfall and Young considered a more general hypothesis testing setup. The hypotheses do not have to be comparisons between the means of several outcome variables, but could come from any set of hypotheses under consideration. In such a case, the p values are used instead of the test statistics to perform the adjustments, so that a common scale is used. To show that the SR method is asymptotically consistent (i.e., that the experimentwise type I error rate actually approaches the nominal level), Westfall and Young assumed that the following six conditions are met (Westfall and Young 1993, p. 213: (8) is condition 2.1 on p. 42, (10) is in sec. 2.6, and (11)–(13) are eqs. 1–3 on p. 213).

Subset Pivotality (8)

Let $\mathbf{P}$ be the k-dimensional vector of random p values for the univariate tests. The distribution of $\mathbf{P}$ has the subset pivotality condition if the joint distribution of the subvector $\{P_i : i \in S\}$ is identical under the restrictions $\cap_{i \in S} H_i$ and $\cap_{i=1}^{k} H_i$, for all subsets $S = \{i_1, \ldots, i_j\}$ of true null hypotheses.

The p values are continuous random variables. (9)

Adjusted p values can be obtained without error. (10)

Let $P_i^{(N)}$ be the p values for a sample of size N. An arbitrary subset of $\{1, \ldots, k\}$ is denoted by S. Then under the null hypotheses H_i for $i \in S$,

$$\max_{i \in S'} P_i^{(N)} \xrightarrow{p} 0 \quad \text{as} \quad N \to \infty, \tag{11}$$

where $\xrightarrow{p}$ means convergence in probability and S' is the complement of S. Let $x_S^{(N)\alpha}$ be the α quantile of $\min_{i \in S} P_i^{(N)}$ when all the null hypotheses are true. Then there exists $\varepsilon_\alpha > 0$ such that

$$\min_S x_S^{(N)\alpha} \geq \varepsilon_\alpha, \quad \text{for all } N. \tag{12}$$

Also, under the null hypotheses H_i for $i \in S$,

$$\min_{i \in S} P_i^{(N)} \xrightarrow{w} P^\infty \quad \text{as} \quad N \to \infty, \tag{13}$$

where $\xrightarrow{w}$ means convergence in distribution and P^∞ is a continuous random variable.

For the special case of multiple outcomes considered here (Sec. 3), subset pivotality is always satisfied. Conditions (11)–(13) are difficult to verify with real data. They involve the joint distribution of p values, whereas in real applications the dependence structure of the p values is usually unknown. For Theorem 2, we have required only that

$$E[G(X_1, \ldots, X_{2N_0})] < \infty, \tag{14}$$

which is always satisfied if one considers the statistics T_i to be $1 - P_i$, where P_i is the random p value for component i. Therefore, in the case of testing the means of multiple outcomes, we have shown in general that the method is asymptotically conservative (i.e., the probability that any type I error is committed is asymptotically bounded above by α), which is clearly weaker than asymptotically consistent. This shows the robustness of the SR method with respect to the test used to perform the univariate comparisons, one of its fundamental strengths. For Theorem 1, we assume that adjusted p values can be evaluated without error [i.e., (10)], so an analogy can be made between the proof of Theorem 1 and the proof of consistency given by Westfall and Young. The proof of Theorem 2 is necessarily more involved, because it deals with the error present in adjusted p values. Theorem 2 also handles the discrete as well as the continuous case, which is not dealt with in general by Westfall and Young.

Because it is expected that step-up methods will generally have higher power than step-down methods when there are many false null hypotheses, a step-up resampling method is desirable. I am already pursuing this goal.

An advantage of the SR method is the ready availability of adjusted p values for the univariate tests. In addition to deciding whether to reject or accept each null hypothesis, it is simple to calculate adjusted p values for the comparisons by enforcing monotonicity on the α^* sequence.

Even if parametric tests are used to obtain the unadjusted p values (and test statistics), the adjusted p values obtained by the SR method are distribution free. Moreover, if the assumptions behind the parametric univariate tests are false, the method is still asymptotically conservative. It is not necessary for the univariate distributions of the data to be known even approximately. The method has been shown to provide excellent experimentwise type I error rate control while providing increased power to reject individual hypotheses when compared to the method of Hochberg (1989). It has also been shown to perform better than the step-up method of Dunnett and Tamhane (1992) when unequal correlations exist. Therefore, the SR method should be the method of choice for multiple comparisons when the distribution or correlation structure of the data is unknown.

APPENDIX: PROOFS

Proof of Theorem 1

Without loss of generality, we may assume that $H_1, \ldots, H_l$ are true, $H_{l+1}, \ldots, H_k$ are false, and $\max_{1 \leq j \leq l} t_j = t_{(r)}$ with $1 \leq l \leq r \leq k$. Let A be the event a type I error occurs,

$$A = \{T_{(k)} > \varphi_\alpha^{(k)}, T_{(k-1)} > \varphi_\alpha^{(k-1)}, \ldots, T_{(r)} > \varphi_\alpha^{(r)}\}.$$

Therefore, the event that a type I error occurs is a subset of the event that the largest test statistic corresponding to a true null hypothesis exceeds $\varphi_\alpha^{(r)}$. Letting P_{H_l} stand for the probability under $H_1, \ldots, H_l$, we have

$$P_{H_l}\{A\} \leq P_{H_l}\{\max_{1 \leq j \leq l} T_j \geq \varphi_\alpha^{(r)}\}, \tag{A.1}$$

and we know from SR Algorithm 2 that

$$\alpha = P_{H_{(r)}}\{T_{(r)}^{(r)} \geq \varphi_\alpha^{(r)}\},$$

where $T_{(r)}^{(r)}$ is the maximal test statistic from the components corresponding to $H_{(1)}, \ldots, H_{(r)}$. Because this set contains $\{1, 2, \ldots, l\}$, we have

$$\alpha \geq P_{H_l}\{\max_{1 \leq j \leq l} T_j \geq \varphi_\alpha^{(r)}\}.$$

Substituting this into (A.1) gives

$$P_{H_l}\{\text{type I error}\} \leq \alpha.$$

Proof of Theorem 2

As in the proof of Theorem 1, we start by assuming that $H_1, \ldots, H_l$ are true, $H_{l+1}, \ldots, H_k$ are false, and $\max_{1 \leq j \leq l} t_j = t_{(r)}$ with $1 \leq l \leq r \leq k$. Let A denote the event a type I error occurs,

$$A = \{\alpha_1^* < \alpha, \alpha_2^* < \alpha, \ldots, \alpha_{k-r+1} < \alpha\}.$$

For notational simplicity, let $\underline{\mathbf{X}}_{2N} = (\mathbf{X}_1, \ldots, \mathbf{X}_{2N})$ and $\underline{\mathbf{x}}_{2N} = (\mathbf{x}_1, \ldots, \mathbf{x}_{2N})$. We have

$$\begin{aligned}
&P_{H_l}\{A \mid \underline{\mathbf{X}}_{2N} = \underline{\mathbf{x}}_{2N}\} \\
&\quad \leq P_{H_l}\{\alpha_{k-r+1}^* < \alpha \mid \underline{\mathbf{X}}_{2N} = \underline{\mathbf{x}}_{2N}\} \\
&\quad = P_{H_l}\left\{\frac{1}{M}\sum_{j=1}^{M} I[\max_i T_{ij}^* \geq t_{(r)}] < \alpha \,\middle|\, \underline{\mathbf{X}}_{2N} = \underline{\mathbf{x}}_{2N}\right\} \\
&\quad \leq P_{H_l}\left\{\frac{1}{M}\sum_{j=1}^{M} I[\max_{1 \leq i \leq l} T_{ij}^* \geq \max_{1 \leq i \leq l} t_i] < \alpha \,\middle|\, \underline{\mathbf{X}}_{2N} = \underline{\mathbf{x}}_{2N}\right\},
\end{aligned} \tag{A.2}$$

where the maximum in the third line extends over all i corresponding to $t_{(1)}, \ldots, t_{(r)}$ and the second inequality comes from the fact that the set of indices corresponding to $t_{(1)}, \ldots, t_{(r)}$ includes the set $\{1, \ldots, l\}$. Because the event in (A.2) depends only on the first l components, we may consider the truncated data $\underline{\mathbf{X}}_1^{(l)}, \ldots, \underline{\mathbf{X}}_{2N}^{(l)}$, which consist of only the first l components of $\underline{\mathbf{X}}_1, \ldots, \underline{\mathbf{X}}_{2N}$. Under $H_1, \ldots, H_l$, $\underline{\mathbf{X}}_1^{(l)}, \ldots, \underline{\mathbf{X}}_{2N}^{(l)}$ are independent and identically distributed with distribution function $F^{(l)}$, the joint distribution of the first l components of $\mathbf{X}_i$ obtained from F by integrating out the last $k - l$ components. Define the event

$$E_j = \{\max_{1\le i\le l} T_{ij}^* \ge \max_{1\le i\le l} t_i\}, \qquad j = 1, \ldots, M.$$

If we integrate (A.2) over all $\underline{\mathbf{x}}_{2N}$, we get

$$P_{H_l}\{A\} \le \int \mathbf{P}\left\{\frac{1}{M}\sum_{j=1}^{M} I[E_j] < \alpha \middle| \underline{\mathbf{X}}_{2N}^{(l)} = \underline{\mathbf{x}}_{2N}^{(l)}\right\} dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots dF^{(l)}(\mathbf{x}_{2N}^{(l)}), \quad \text{(A.3)}$$

where the $\mathbf{P}$ inside the integral is the conditional probability with respect to resampling from the fixed original data values. Now consider taking the limit as M goes to infinity on both sides of (A.3). The bounded convergence theorem enables us to pass the limit inside the integral, and the law of large numbers results in a new right side for (A.3):

$$\lim_{M\to\infty} P_{H_l}\{A\} \le \int_{\underline{\mathbf{x}}_{2N}^{(l)}:\mathbf{P}\{E_s|\underline{\mathbf{X}}_{2N}^{(l)}=\underline{\mathbf{x}}_{2N}^{(l)}\}\le\alpha} dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots dF^{(l)}(\mathbf{x}_{2N}^{(l)}), \quad \text{(A.4)}$$

where s is arbitrary from $1, \ldots, M$. Recall that the resample $\mathbf{X}_1^*, \ldots, \mathbf{X}_{2N_0}^*$ is a with replacement sample from $\mathbf{x}_1, \ldots, \mathbf{x}_{2N}$. Similarly, let $\mathbf{Z}_1, \ldots, \mathbf{Z}_{2N_0}$ be the truncated resample consisting of the first l components of $\mathbf{X}_1^*, \ldots, \mathbf{X}_{2N_0}^*$. Therefore, $\mathbf{Z}_1, \ldots, \mathbf{Z}_{2N_0}$ are independent and identically distributed with distribution function $F_N^{(l)}(\mathbf{z}_w)$ given by

$$F_N^{(l)}(\mathbf{z}_w) = \frac{\sum_{i=1}^{2N} I[x_{ij} \le z_{wj}, j = 1, \ldots, l]}{2N},$$

the empirical distribution function of the original truncated sample. We can now express (A.4) as

$$\lim_{M\to\infty} P_{H_l}\{A\} \le \int_{\underline{\mathbf{x}}_{2N}^{(l)}:[\int_{E_s} dF_N^{(l)}(\mathbf{z}_1)\ldots dF_N^{(l)}(\mathbf{z}_{2N_0})]\le\alpha} dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots dF^{(l)}(\mathbf{x}_{2N}^{(l)}). \quad \text{(A.5)}$$

Event E_s occurs when a certain function of $\mathbf{z}_1, \ldots, \mathbf{z}_{2N_0}$ (i.e., the maximum of the first l test statistics) exceeds the same function of $\mathbf{x}_1, \ldots, \mathbf{x}_{2N_0}$. (Recall that the test statistics $t_1, \ldots, t_k$ are based on an initial sample of size $2N_0$ even though resampling occurs from the growing sample of size $2N$.) Therefore, event E_s is independent of N. If we now consider the limit superior of (A.5) as N goes to infinity, we get

$$\begin{aligned}\overline{\lim_{N\to\infty}} \lim_{M\to\infty} P_{H_l}\{A\} &\le \int I\left[\int_{E_s} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}) < \alpha\right] dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots \\ &+ \int I\left[\int_{E_s} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}) = \alpha\right] dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots, \end{aligned} \quad \text{(A.6)}$$

using the fact that for every $\underline{\mathbf{x}}^{(l)}$ in the space of sequences from $\mathcal{R}^l$,

$$\overline{\lim_{N\to\infty}} I\left[\int_{E_s} dF_N^{(l)}(\mathbf{z}_1) \ldots dF_N^{(l)}(\mathbf{z}_{2N_0}) \le \alpha\right] \le I\left[\int_{E_s} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}) \le \alpha\right].$$

The value of the second integral in (A.6) is either positive or zero.

Case 1: The Second Integral of (A.6) Equals Zero

If this is the case, then (A.6) becomes

$$\overline{\lim_{N\to\infty}} \lim_{M\to\infty} P_{H_l}\{A\} \le \int I\left[\int_{E_s} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}) < \alpha\right] dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots \quad \text{(A.7)}$$

The inner integral in (A.7) can be written as

$$\int_{\underline{\mathbf{z}}_{2N_0}:G(\mathbf{z}_1,\ldots,\mathbf{z}_{2N_0})\ge G(\mathbf{x}_1^{(l)},\ldots,\mathbf{x}_{2N_0}^{(l)})} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}).$$

From (6), we assert that there exists at least one number U such that

$$\int_{\underline{\mathbf{z}}_{2N_0}:G(\mathbf{z}_1,\ldots,\mathbf{z}_{2N_0})\ge U} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}) < \alpha. \quad \text{(A.8)}$$

Let $U^* = \inf\{U : \text{(A.8) holds}\}$. Then we have

$$I\left[\int_{\underline{\mathbf{z}}_{2N_0}:G(\mathbf{z}_1,\ldots,\mathbf{z}_{2N_0})\ge G(\mathbf{x}_1^{(l)},\ldots,\mathbf{x}_{2N_0}^{(l)})} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}) < \alpha\right] = I[G(\mathbf{x}_1^{(l)}, \ldots, \mathbf{x}_{2N_0}^{(l)}) > U^*].$$

This makes (A.7) become

$$\overline{\lim_{N\to\infty}} \lim_{M\to\infty} P_{H_l}\{A\} \le \int I[G(\mathbf{x}_1^{(l)}, \ldots, \mathbf{x}_{2N_0}^{(l)}) > U^*]\, dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots dF^{(l)}(\mathbf{x}_{2N_0}^{(l)}). \quad \text{(A.9)}$$

If U^* satisfies (A.8), then we are done. If not, then there is a sequence of numbers $\{U_i\}$ that satisfy (A.8) and converge to U^*. The indicator functions $I[G(\underline{\mathbf{x}}_{2N_0}^{(l)}) \ge U_i]$ converge pointwise to the indicator function $I[G(\underline{\mathbf{x}}_{2N_0}^{(l)}) > U^*]$, so that the bounded convergence theorem gives

$$\int_{\underline{\mathbf{x}}_{2N_0}^{(l)}:G(\underline{\mathbf{x}}_{2N_0}^{(l)})>U^*} dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots dF^{(l)}(\mathbf{x}_{2N_0}^{(l)}) \le \alpha.$$

When combined with (A.9), we have the desired asymptotic bound on $P_{H_l}\{A\}$.

Case 2: The Second Integral of (A.6) Equals $\gamma > 0$.

In this case we may assume that there is at least one number U such that

$$\int_{\underline{\mathbf{z}}_{2N_0}:G(\underline{\mathbf{z}}_{2N_0})\ge U} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}) = \alpha. \quad \text{(A.10)}$$

Let $U^* = \sup\{U : \text{(A.10) holds}\}$. As in the previous case, we can show that (A.10) holds for U^* as well. Thus $G(\underline{\mathbf{X}}_{2N_0})$ has an atom at U^* but not at any other U satisfying (A.10). Therefore,

$$\begin{aligned}&\int I\left[\int_{\underline{\mathbf{z}}_{2N_0}:G(\underline{\mathbf{z}}_{2N_0})\ge G(\underline{\mathbf{x}}_{2N_0}^{(l)})} dF^{(l)}(\mathbf{z}_1) \ldots \right. \\ &\qquad \left. dF^{(l)}(\mathbf{z}_{2N_0}) = \alpha\right] dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots dF^{(l)}(\mathbf{x}_{2N_0}^{(l)}) \\ &= \int I[G(\underline{\mathbf{X}}_{2N_0}^{(l)}) = U^*]\, dF^{(l)}(\mathbf{x}_1^{(l)}) \ldots dF^{(l)}(\mathbf{x}_{2N_0}^{(l)}) = \gamma, \end{aligned} \quad \text{(A.11)}$$

while at the same time we have

$$I\left[\int_{\mathbf{z}_{2N_0}: G(\mathbf{z}_{2N_0}) \geq G(\underline{\mathbf{x}}^{(l)}_{2N_0})} dF^{(l)}(\mathbf{z}_1) \ldots dF^{(l)}(\mathbf{z}_{2N_0}) < \alpha\right]$$

$$= I[G(\underline{\mathbf{x}}^{(l)}_{2N_0})]. \quad \text{(A.12)}$$

Plugging (A.11) and (A.12) into (A.6), we get

$$\begin{aligned}
&\overline{\lim_{N\to\infty}} \lim_{M\to\infty} P_{H_l}\{A\} \\
&\leq \int I[G(\underline{\mathbf{X}}^{(l)}_{2N_0}) > U^*]\, dF^{(l)}(\mathbf{x}^{(l)}_1) \ldots dF^{(l)}(\mathbf{x}^{(l)}_{2N_0}) + \gamma \\
&= \int I[G(\underline{\mathbf{X}}^{(l)}_{2N_0}) \geq U^*]\, dF^{(l)}(\mathbf{x}^{(l)}_1) \ldots dF^{(l)}(\mathbf{x}^{(l)}_{2N_0}) \\
&\quad - \int I[G(\underline{\mathbf{X}}^{(l)}_{2N_0}) = U^*]\, dF^{(l)}(\mathbf{x}^{(l)}_1) \ldots dF^{(l)}(\mathbf{x}^{(l)}_{2N_0}) + \gamma \\
&= \alpha.
\end{aligned}$$

The result follows by realizing that this is an asymptotic bound on $P_{H_l}\{A\}$.

[Received April 1993. Revised March 1994.]

REFERENCES

Dunn, O. J. (1959), "Confidence Intervals for the Means of Dependent Normally Distributed Variables," *Journal of the American Statistical Association,* 54, 613–621.

Dunn, O. J., and Massey, F. J. (1965), "Estimation of Multiple Contrasts Using *t* Distributions," *Journal of the American Statistical Association,* 60, 573–583.

Dunnett, C. W., and Tamhane, A. C. (1992), "A Step-Up Multiple Test Procedure," *Journal of the American Statistical Association,* 87, 162–170.

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.

Mills, J. L., Knopp, R. H., Simpson, J. L., Jovanovic-Peterson, L., Metzger, B. E., Holmes, L. B., Aarons, J. H., Brown, Z., Reed, G. F., Bieber, F. R., Van Allen, M., Holtzman, I., Ober, C., Peterson, C. M., Withiam, M. J., Duckles, A., Mueller-Heubach, E., Polk, B. F., and the NICHD-Diabetes in Early Pregnancy Study (1988), "Lack of Relation of Increased Malformation Rates in Infants of Diabetic Mothers to Glycemic Control During Organogenesis," *New England Journal of Medicine,* 318, 671–676.

Simes, R. J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika,* 73, 751–754.

Sidak, Z. (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," *Journal of the American Statistical Association,* 62, 626–633.

——— (1971), "On Probabilities of Rectangles in Multivariate Student Distributions: Their Dependence on Correlations," *The Annals of Mathematical Statistics,* 42, 169–175.

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.

——— (1993), *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment* (Vol. 1), New York: John Wiley.