

Multiple Comparison Procedures with SAS

Chung-Kuei Chang, Prosoft Software Inc., Ambler, PA
Dror Rom, Prosoft Software Inc., Ambler, PA

Abstract

This paper discusses some commonly used multiple comparison procedures available in SAS. We focus on the following methods: modified Bonferroni Procedures, Dunnett's procedure for comparisons with a control treatment, with and without a covariate adjustment, step-down and closed testing procedures, adjusted p-values using Monte Carlo simulation, and calculation of exact adjusted p-values for discrete data. We will cover PROBMC, PROC GLM, PROC MULTTEST, and some macros for various methods. Various settings will be used to demonstrate the procedures, including multiple treatment comparisons, pairwise comparisons, dose response, and multiple endpoints in clinical trials.

Keywords: Multiple Comparisons, Multiplicity Adjustment, Closed Testing.

1. Introduction

Multiple comparison procedures (MCP) are frequently used in scientific studies. According to a survey, they are the second most frequently applied group of statistical methods, second only to the F -test (Hsu, 1996, page 175). This is mainly due to the fact that almost every study contains more than one response, more than one measurement on major response variables, or more than two groups. Taking pharmaceutical research as an example, multiple tests arise in several different types of studies including: multiple endpoints, comparison of more than two treatment groups, dose-response studies, interim analysis, sub-group analysis, meta-analysis, drug-screening, ..., etc.

Once the data are ready, with the power and convenience of today's statistical software, investigators tend to look at the problem from different angles and hence multiple tests are performed. Some of the tests are planned beforehand but some of them are just for "data snooping". A common practice is to report the p-value of each test performed. It is well known that when multiple tests are performed and a claim of significance is done by observing any significant individual p-value, then the chance of committing a false positive error can be inflated. Pocock, *et al* (1987) pointed out that clinical reports tend to exaggerate treatment difference due to overlooking various kinds of multiplicity. For instance, if an

unplanned sub-group analysis is carried out as suggested in "Munchausen's Statistical Grid" (Westfall and Young, 1993, Appendix I.)—keep partitioning each treatment group by all possible factors until a significant result is observed —, then any drug can almost always be declared as effective. This phenomenon can easily be illustrated by the following example in the analysis of multiple endpoints.

Suppose there are two treatment groups and k endpoints with continuous responses measured on each experimental unit. For endpoint i , $i = 1, 2, \dots, k$, a continuous test statistic T_i is used to test the null hypothesis that the mean responses of the two groups are the same for all endpoints, versus the alternative hypothesis that there is a difference between the two groups in at least one endpoint. Suppose each test is conducted at level α . Let us look at two extreme cases assuming that there is no treatment difference in any endpoint. First, assume that the k endpoints are perfectly correlated (which rarely ever happens in practice). Then, the probability of observing any significant result is still α . Secondly, assume that the k endpoints are independent. Then, the chance of observing at least one significant p-value can easily be seen to be $1 - (1 - \alpha)^k$. When $\alpha = 0.05$ and $k = 10$, for instance, the probability of committing a type I error at any endpoint is increased from 0.05 to 0.40. Therefore, one should use caution when interpreting results of multiple tests of significance. A way to address this problem is to use multiple comparison procedures that offer type I error control. Some procedures that are available in SAS are addressed in this paper.

2. Terminology and Notation

Suppose there is a set of k hypotheses of interest in an experiment: $\mathbf{H} = \{H_1, H_2, \dots, H_k\}$, each with a corresponding alternative hypothesis H_i' , $i = 1, \dots, k$. Each H_i is called an individual hypothesis, and the corresponding p-value, p_i , is called the individual p-value, raw p-value, or unadjusted p-value.

For any subset I of $A_k = \{1, 2, \dots, k\}$, define the subset intersection hypothesis $H_I = \cap_{j \in I} H_j$, and its alternative hypothesis $H_I' = \cup_{j \in I} H_j'$. When $I = A_k$, H_I is called the global null hypothesis, denoted as H , and the corresponding p-value is called the global p-value.

Denote the family of all non-empty subset intersection hypotheses as Ω . A decision rule for making inferences on Ω is called a multiple comparison procedure (MCP). An MCP is said to have a weak control of the familywise error rate (FWE) if under the global null hypothesis $H = H_1 \cap \dots \cap H_k$, the probability of falsely concluding any alternative hypotheses H_i , $i = 1, \dots, k$, is less than or equal to the pre-determined significance level α . An MCP is said to have a strong control of the FWE if it has a weak control of FWE under any H_i in Ω .

When the global null hypothesis is rejected, one concludes that at least one individual null hypothesis is false. To make the decision on the individual hypotheses, one can apply the closed testing procedure proposed by Marcus, et al (1976). A closed testing procedure rejects H_i if all hypotheses implying H_i are rejected by a corresponding α level test. This procedure has a strong control of the FWE. For instance, with $H = \{H_1, H_2, H_3\}$, H_1 is rejected by the closed testing procedure if H_1 , $H_1 \cap H_2$, and $H_1 \cap H_2 \cap H_3$ are all rejected by the global test procedure.

It is informative to report the adjusted p-value (for multiplicity) for each raw p-value, such that the decision can be made by comparing the adjusted p-value with the significance level. The p-value for an individual test is the probability of observing at least as extreme test statistic. It is also the minimum significance level such that the null hypothesis can be rejected. For instance, if the p-value is 0.051, then the null hypothesis is rejected if the significance level is greater than or equal to 0.051. Using the second interpretation, the adjusted p-value of a raw p-value is defined as the minimum significance level needed to reject the corresponding individual null hypothesis while controlling the FWE strongly. For example, if $H = \{H_1, H_2, H_3\}$, the adjusted p-value for testing H_1 is the maximum of the global p-values for testing H_1 , $H_1 \cap H_2$, and $H_1 \cap H_2 \cap H_3$.

3. Multiple Comparisons with the Control Group

Let $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(k)}$ be the ordered p-values. We use upper case to indicate that the p-values are random variable, and use lower case for the observed p-values. The global p-value is defined as $p = pr(P_{(1)} \leq p_{(1)})$.

3.1 One-Way ANOVA

In a one-way ANOVA, assume that the response of the j -th subject in the i -th group is $Y_{ij} = \mu_i + \varepsilon_{ij}$, where ε_{ij} are identically and independently distributed as a normal random variable with mean 0 and unknown variance σ^2 ; $i = 0, 1, \dots, k$, $j = 1, \dots, n_j$ ($i = 0$ represents the control

group). Let T_i be the t-statistic for testing the i -th null hypothesis $H_{0i}: \mu_i = \mu_0$, $i = 1, \dots, k$. Then,

$$T_i = \frac{(\bar{Y}_i - \bar{Y}_0) / \sqrt{1/n_i + 1/n_0}}{\sqrt{\hat{\sigma}^2}},$$

where $\bar{Y}_i$ is the sample mean of the i th group and

$\hat{\sigma}^2 = \sum_{i=0}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 / \sum_{i=0}^k (n_i - 1)$ is the pooled estimate of

σ^2 . Let $\nu = \sum_{i=0}^k (n_i - 1)$ be the error degrees of freedom

and $X = \nu \hat{\sigma}^2 / \sigma^2$. The T_i 's are called Dunnett's multivariate t-statistics. We want to calculate the p-value for the global null hypothesis $H = H_1 \cap \dots \cap H_k$ based on the joint distribution of the T_i 's.

Under the i -th null hypothesis,

$$\begin{aligned} T_i &= \frac{[(\bar{Y}_i - \mu_i) - (\bar{Y}_0 - \mu_0)] / (\sigma \sqrt{1/n_i + 1/n_0})}{\sqrt{\nu \hat{\sigma}^2 / \sigma^2}} \\ &= \frac{\frac{n_0}{\sqrt{n_0 + n_i}} \frac{\sqrt{n_i}(\bar{Y}_i - \mu_i)}{\sigma} - \frac{n_i}{\sqrt{n_0 + n_i}} \frac{\sqrt{n_0}(\bar{Y}_0 - \mu_0)}{\sigma}}{\sqrt{\nu \hat{\sigma}^2 / (\nu \sigma^2)}} \\ &= \frac{\sqrt{1 - \lambda_i^2} Z_i - \lambda_i Z_0}{\sqrt{X / \nu}}, \end{aligned} \quad (1)$$

where $\lambda_i = \sqrt{n_i / (n_0 + n_i)}$, Z_i and Z_0 follow the standard normal distribution, X has a chi-square distribution with ν degrees of freedom, and $(Z_0, Z_1, \dots, Z_k, X)$ are mutually independent. Since the T_i 's are independent conditional on X and Z_0 , the evaluation of the probability that all the T_i 's are less than or equal to a constant q can be simplified as:

$$\begin{aligned} &pr(T_i \leq q, i = 1, \dots, k) \\ &= E_{X, Z_0} [pr(\frac{\sqrt{1 - \lambda_i^2} Z_i - \lambda_i Z_0}{s} \leq q, i = 1, \dots, k | \sqrt{\frac{X}{\nu}} = s, Z_0 = z)] \\ &= \int_0^\infty \int_{-\infty}^\infty \phi(z) \prod_{i=1}^k \Phi(\frac{qs + \lambda_i z}{\sqrt{1 - \lambda_i^2}}) dz d\mu_\nu(s), \end{aligned} \quad (2)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively, the density and distribution functions of the standard normal random variable, and $d\mu_\nu(s)$ is the density function of $\sqrt{X / \nu}$. That is,

$$d\mu_\nu(s) = \frac{\nu^{\nu/2}}{\Gamma(\nu/2)\nu^{\nu/2-1}} s^{\nu-1} e^{-\nu s^2/2} ds.$$

Using the SAS PROBMC function, this probability can be calculated as:

probmc('dunnett1', q, ., v, k, of lam(1) ... lam(k)),

where lam(i) is the variable with value λ_i , $i = 1, \dots, k$. (You will receive an error message if lam(i) is replaced with the real value λ_i .) To obtain the value of q for a given p-value p , it is the value of:

probmc('dunnett1', ., p, v, k, of lam(1) ... lam(k)).

Let $t_{(1)}, \dots, t_{(k)}$ be the observed ordered statistics of the T_i 's (where $t_{(1)}$ is the minimum and $t_{(k)}$ is the maximum). For testing that at least one mean of the treated groups is less than that of the control, the *exact* global p-value using Dunnett's Test is:

$$\begin{aligned} & \Pr(\bigcup_{i=1}^k \{T_i \leq t_{(1)}\}) \\ &= 1 - \Pr(\bigcap_{i=1}^k \{T_i \geq t_{(1)}\}) \\ &= 1 - \Pr(\bigcap_{i=1}^k \{T_i \leq -t_{(1)}\}) \\ &= 1 - \text{probmc}(\text{'dunnett1'}, -t_{(1)}, ., v, k, \text{of lam(1) ... lam(k)}). \end{aligned}$$

Note that if $-t_{(1)}$ is too small, the value of the PROBMC function will be missing, instead of zero. To prevent from an error, you can first calculate the boundary value

$$t_b = \text{probmc}(\text{'dunnett1'}, ., 0.001, v, k, \text{of lam(1) ... lam(k)}).$$

If $-t_{(1)}$ is above t_b , call the function; otherwise output a flag indicating that the p-value is greater than 0.999.

With similar argument, the probability that all the absolute values of the t-statistics are not greater than q (≥ 0) is:

$$\begin{aligned} & \Pr(|T_i| \leq q, i = 1, \dots, k) \\ &= \int_0^\infty \int_0^\infty \phi(z) \prod_{i=1}^k \left[\Phi\left(\frac{\lambda_i z + qs}{\sqrt{1-\lambda_i^2}}\right) - \Phi\left(\frac{\lambda_i z - qs}{\sqrt{1-\lambda_i^2}}\right) \right] d\mu_v(s) dz. \end{aligned}$$

Using SAS PROBMC function, it can be calculated as:

probmc('dunnett2', q, ., v, k, of lam(1) ... lam(k)).

Let $q_{(1)} \leq \dots \leq q_{(k)}$ be the ordered absolute values of the t-statistics. Then the two-sided p-value is:

$$\begin{aligned} & \Pr(\bigcup_{i=1}^k \{|T_i| \geq q_{(k)}\}) \\ &= 1 - \Pr(\bigcap_{i=1}^k \{|T_i| \leq q_{(k)}\}) \\ &= 1 - \text{probmc}(\text{'dunnett2'}, q_{(k)}, ., v, k, \text{of lam(1) ... lam(k)}). \end{aligned}$$

3.2 One-Way ANCOVA

In Section 3.1, the numerical evaluation of the *exact* Dunnett's p-value is through a double integral. This is due to the fact that in a one-way ANOVA, the T_i 's are independent conditional on X and Z_0 [see equations (1)]. This is no longer true if covariates are in the model. Thus the PROBMC function might produce misleading p-values. Here we illustrate this with one covariate in the model.

Assume that the response of the j -th subject in the i -th group is $Y_{ij} = \mu_i + bZ_{ij} + \varepsilon_{ij}$, where ε_{ij} are identically and independently distributed as a normal random variable with mean 0 and unknown variance σ^2 ; $i = 0, 1, \dots, k$, $j = 1, \dots, n_{ij}$. From the theory of linear models (see, for example, Searle, 1987, pp. 172-173), we have:

$$\hat{\mu}_i = \bar{Y}_i - \hat{b}\bar{Z}_i,$$

$$\hat{b} = \frac{\sum_i \sum_j (Y_{ij} - \bar{Y}_i)(Z_{ij} - \bar{Z}_i)}{\sum_i \sum_j (Z_{ij} - \bar{Z}_i)^2},$$

$$\text{var}(\hat{b}) = \frac{\sigma^2}{\sum_i \sum_j (Z_{ij} - \bar{Z}_i)^2},$$

$$\text{cov}(\bar{Y}_i, \hat{b}) = 0,$$

and

$$\begin{aligned} & \text{var}(\hat{\mu}_i - \hat{\mu}_0) \\ &= \sigma^2 \left[\frac{1}{n_i} + \frac{1}{n_0} + \frac{(\bar{Z}_i - \bar{Z}_0)^2}{\sum_i \sum_j (Z_{ij} - \bar{Z}_i)^2} \right], \end{aligned}$$

where $\hat{\mu}_i$ and $\hat{b}$ are the least square estimates of μ_i and b , respectively. Since

$$\bar{Y}_i \sim N(\mu_i + b\bar{Z}_i, \frac{\sigma^2}{n_i}),$$

and

$$\hat{b} \sim N(b, \frac{\sigma^2}{\sum_{i,j} (Z_{ij} - \bar{Z}_i)^2}),$$

the t statistic for testing $H_{0i}: \mu_i = \mu_0$ is:

$$\begin{aligned} T_i &= \frac{(\hat{\mu}_i - \hat{\mu}_0) / \sqrt{\sigma^2 c_i}}{\sqrt{V / (v \sigma^2)}} \\ &= \left\{ \frac{\bar{Y}_i - (\mu_i + b \bar{Z}_i)}{\sigma} - \frac{\bar{Y}_0 - (\mu_0 + b \bar{Z}_0)}{\sigma} \right. \\ &\quad \left. - \frac{(\hat{b} - b)(\bar{Z}_i - \bar{Z}_0)}{\sigma} \right\} / \sqrt{c_i (X / v)} \\ &= \left\{ \frac{W_i}{\sqrt{n_i c_i}} - \frac{W_0}{\sqrt{n_0 c_0}} \right. \\ &\quad \left. + \frac{\sqrt{(\bar{Z}_i - \bar{Z}_0)^2 B}}{\sqrt{c_i \sum_{i,j} (\bar{Z}_{ij} - \bar{Z}_i)^2}} \right\} / \sqrt{X / v}, \end{aligned}$$

where

$$W_i = [\bar{Y}_i - (\mu_i + b \bar{Z}_i)] / \sigma, i = 0, 1, \dots, k,$$

$$B = \frac{(\hat{b} - b) \sqrt{\sum_{i,j} (\bar{Z}_{ij} - \bar{Z}_i)^2}}{\sigma},$$

$$c_i = \frac{1}{n_i} + \frac{1}{n_0} + \frac{(\bar{Z}_i - \bar{Z}_0)^2}{\sum_{i,j} (\bar{Z}_{ij} - \bar{Z}_i)^2},$$

V/v is the mean square error with $v (= \sum_i n_i - k - 2)$ degrees of freedom, and $X = V/\sigma^2$. Note that $W_0, W_1, \dots, W_k$ and B follow a standard normal distribution, X follows a chi-square distribution, and $(W_0, W_1, \dots, W_k, B, X)$ are mutually independent.

Define:

$$\lambda_{i0} = \left[1 + \frac{n_0}{n_i} + \frac{n_0 (\bar{Z}_i - \bar{Z}_0)^2}{\sum_{i,j} (\bar{Z}_{ij} - \bar{Z}_i)^2} \right]^{-1/2},$$

and

$$\lambda_{i1} = \left[1 + \left(\frac{1}{n_i} + \frac{1}{n_0} \right) \frac{\sum_{i,j} (\bar{Z}_{ij} - \bar{Z}_i)^2}{(\bar{Z}_i - \bar{Z}_0)^2} \right]^{-1/2}, \quad (3)$$

then T_i can be expressed as:

$$T_i = \frac{\sqrt{1 - \lambda_{i0}^2 - \lambda_{i1}^2} W_i - \lambda_{i0} W_0 + \lambda_{i1} B}{\sqrt{X / v}}. \quad (4)$$

Since T_i 's are independent conditional on W_0, B , and X ,

$$\begin{aligned} &pr(T_i \leq q, i = 1, \dots, k) \\ &= \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \phi(w_0) \phi(b) \prod_{i=1}^k \Phi\left(\frac{qs + \lambda_{i0} w_0 - \lambda_{i1} b}{\sqrt{1 - \lambda_{i0}^2 - \lambda_{i1}^2}}\right) dw_0 db d\mu_v(s), \end{aligned} \quad (5)$$

where $d\mu_v(s)$ is the density function of $\sqrt{X/v}$. Equations (2) and (5) are quite different. Thus you should use the PROBMC function with caution if you have covariates in the model.

Note that if the covariate means of the treated groups are all very close to that of the control group, then λ_{i1} approaches zero [see equation (3)] and equation (4) is reduced to equation (1). Hence the PROBMC function will still produce satisfactory results.

Another approach to obtaining Dunnett's adjusted p-value with covariates in the model is to use the LSMEANS statement in the GLM procedure. There are two options available, one for analytical approximation method and the other one for simulation method. Let y be the response variable and tg be the variable for treatment group. Suppose there are two covariates $x1$ and $x2$ and there is no interaction term in the model. The following code produces the lower-tailed adjusted p-values for many-to-one comparisons via simulation:

```
proc glm;
class tg;
model y = tg x1 x2;
lsmeans tg/adjust=simulate pdiff=control;
run;
```

In the above code, if "simulate" is replaced with "dunnett", the adjusted p-values are calculated using Hsu's (1992) factor-analytic covariance approximation. From our experience, the simulation method performs better than Hsu's approximation. However, if the covariate means are homogeneous, the PROBMC function provides even better results when used with the closed testing procedure to generate the adjusted p-values.

3.3 Discrete Response

Here we illustrate the calculation of the global p-value with ordinary response. We first review the calculation of the exact individual p-value for one-sided upper-tailed tests. Then extend the method to the calculation of the global p-value for multiple comparisons with a control.

The calculation for nominal response and one-sided lower-tailed or two-sided tests follows analogous method.

For a given $r \times c$ contingency table N_0 , let n_{ij} be the number of subjects in the i th treatment group with the j th response. Also, let r_i and c_j be the scores of the i th row and j th column, respectively. Since both the treatments and the response levels are assumed to be ordered, the row and column scores are supposed to be monotonically increasing. To test the null hypothesis of no treatment effect versus the alternative hypothesis of an increasing treatment effect on the response (also known as a positive correlation between the row and column scores), the Mantel's (1963) score test $T = \sum_{i,j} r_i c_j n_{ij}$ can be used. The exact one-sided p-value conditional on the row and column margins is:

$$p = \frac{\sum_{T(N) \geq T(N_0)} pr(N | N_0),}{\sum_{T(N) \geq T(N_0)} pr(N | N_0)},$$

where $T(N)$ is the test statistic of the $r \times c$ table N that has the same row and column margins as those of N_0 , $pr(MN_0)$ is the probability of N , conditional on the margins of N_0 . Under the null hypothesis, N follows a hypergeometric distribution and $pr(N) = \prod_i n_{i.}! \prod_j n_{.j}! / (n_{..}! \prod_{i,j} n_{ij}!)$, where $n_{i.}$ ($n_{.j}$) is the i th row sum (j th column sum) and $n_{..}$ is the grand total of N .

To calculate the exact global p-value for a set of k contrasts of the treatment groups, each contrast can be considered as a set of row scores and thus has a corresponding test statistic T for testing the individual hypothesis. For instance, if there are a control and three treated groups, the row scores for the pairwise comparisons with the control are: $(-1, 1, 0, 0)$, $(-1, 0, 1, 0)$, and $(-1, 0, 0, 1)$. The global p-value is then equal to:

$$\begin{aligned} p &= pr(P_{(1)} \leq p_{(1)} | N_0) \\ &= pr(T_{(1)} \geq t_{(1)} | N_0) \\ &= \sum_{T_{(1)} \geq t_{(1)}} pr(N | N_0), \end{aligned}$$

where $T_{(1)}$ is the maximum of $\{T = \sum_{i,j} r_{il} c_j n_{ij}, l = 1, \dots, k\}$, in which r_{il} is the i th row score of the l th contrast, and $t_{(1)}$ is the observed value of $T_{(1)}$.

For binary response data, SAS PROC MULTTEST generates adjusted p-values for multiple contrasts using bootstrap or permutation. The following code generates the adjusted p-values through bootstrap simulation for pairwise comparisons with the control when there are three treated groups:

```
proc multtest bootstrap nsample=10000 pvals;
  class tg;
  test ca(y);
```

```
contrast 'c vs d1' -1 1 0 0;
contrast 'c vs d2' -1 0 1 0;
contrast 'c vs d3' -1 0 0 1;
run;
```

In the above code, the PVALS option requests a summary table of raw and adjusted p-values, the CA option in the TEST statement indicates that the individual p-values are calculated with the Cochran-Armitage test (see Agresti, 1990), which has the same conditional exact p-value as Mantel's score test. For other options in the MULTTEST procedure, see: SAS/STAT® Software Changes and Enhancements through Release 6.11, 1996.

Currently there is no SAS procedure to adjust for multiplicity when the response is multinomial. We have developed a SAS macro to generate the global p-value for Mantel's (1963) test with multinomial response. Due to limited space, the macro is not attached. Our algorithm is summarized as follows:

1. Assign 0 to the global p-value.
2. Calculate the test statistic for each individual contrast and determine the maximum $t_{(1)}$.
3. Enumerate all the tables N with the same row and column margins as those of N_0 .
4. For each table in step 3, determine the corresponding $t_{(1)}$. If it is greater than or equal to the observed $t_{(1)}$, the global p-value is added by $pr(MN_0)$.

We use the complete algorithm (see Verbeek and Kroonenberg, 1985) to exhaust all the tables satisfying the row and column margins. If the table is too large, we use the Monte Carlo simulation to generate random tables satisfying the row and column margins (see Boyett, 1979). The simulated global p-value is the proportion of the random tables with $t_{(1)}$ greater than or equal to that of the observed table.

3.4 Dose response studies

Suppose an experiment contains a control (zero dose) group and several treated groups with increasing doses, and that the response is a monotone function of the dose. Tukey, et al (1985) proposed a step-down trend test to determine the no-statistical-significance-of-trend (NOSTASOT) dose (the lowest dose that is not significantly different from the control). They showed that in the context of a monotonic dose response, the step-down trend test is more powerful than Dunnett's test with the same error rate control.

With Tukey's trend test, you first test the significance of the trend with all groups included. If it is significant, you exclude the highest dose group and perform the same test on the remaining doses; otherwise you stop. You continue until you find significance, or you reached the lowest dose

group. If you reach the lowest dose group and the p-value is significant, conclude that there is no NOSTASOT dose for this study.

The following code can be used to test the significance of the trend:

```
proc reg;
  model y = tg;
run;
```

Note that the treatment group "tg" is used as a continuous variable. You will get the same test result if "proc reg" is replaced with "proc glm". However, the OUTTEST option in PROC REG does not output the p-value of "tg" to a data set, and the OUTSTAT option in PROC GLM does not output the estimated value of "tg" to a data set. If you are developing a macro to perform the step-down trend test with the option of calculating one-sided p-value, you can use the OUTSTAT option in PROC GLM to obtain the two-sided p-value of the trend test, then use the OUTSTAT option in PROC REG to check the direction of the slope to determine the one-sided test. You can also use both PROC MIXED and PROC OUTPUT (see SAS/STAT® Software Changes and Enhancements through Release 6.11, 1996) to obtain the estimate of the slope and the two-sided p-value.

For discrete data, you can use PROC FREQ with CMH1 option in the TABLE statement to calculate the two-sided p-value for testing the significance of the trend. The described step-down trend test can be easily implemented to determine the NOSTASOT dose.

4. Multiplicity Adjustment for Multiple Endpoints

In drug efficacy or safety studies, there are often multiple endpoints of interest. For example, in clinical carcinogenicity studies, the survival time and quality of life are two primary responses. In animal reproductive toxicology studies, the responses of interest include litter size, loss of implantation sites, dams' body weight, and food consumption, etc. In this section, we discuss some multiplicity adjustment methods for multiple endpoints. The discussion includes an adjustment based on the multivariate distribution of the endpoints, and another, based on the raw p-values.

The SAS MULTTEST procedure is a powerful tool for multiplicity adjustment based on the simulated multivariate distribution of the raw p-values. The procedure can adjust continuous and discrete responses simultaneously, for one-sided and/or two-sided tests. For example, suppose there are a control and three treated groups with increasing dose. Assume that there are seven endpoints (y1-y7) of interest, of which y1 to y4 are continuous and y5 to y7 are binary. The following code

generates the two-sided adjusted p-values using the permutational resampling method for testing the significance of trend from group 1 to group 4:

```
proc multtest permutation nsample=10000 pvals;
  class tg;
  test mean(y1-y4);
  test ca(y5-y7);
  contrast 'trend' 0 1 2 3;
run;
```

Note that the MULTTEST procedure does not provide multiplicity adjustment for multinomial response variables. For the calculation of exact and simulated global p-values for multinomial response, see Rom and Chang (1995), and Chang and Rom (1994).

The MULTTEST procedure also provides multiplicity adjustment based on the raw p-values. The available options include the Bonferroni, Sidak (1967), Holm (1979), Hochberg (1988), and Benjamini and Hochberg's (1995) methods. The Sidak, Holm, and Hochberg methods are improvements to the Bonferroni procedure. The Benjamini and Hochberg's method is for the control of the false discovery rate. The procedure may not control the FWE.

The Bonferroni and Holm procedures do not require distributional assumption. The Sidak procedure requires the assumption of positively orthon dependence. This type of distributions includes: absolute statistics from various multivariate t distributions, certain multivariate statistics that are asymptotically chi-square, F statistics with common denominators and positively dependent numerators, absolute multivariate normal statistics with zero mean and arbitrary dispersion, ..., etc. (see Holland and Copenhaver, 1987). The Benjamini and Hochberg's method controls the false discovery rate if the p-values are independent.

The Hochberg (1988) procedure controls the FWE strongly if the Simes (1986) global test procedure controls the type I error. The Simes procedure rejects the global null hypothesis if $p_{(i)} \leq i\alpha/k$, for any $i = 1, \dots, k$, where $p_{(1)} \leq \dots \leq p_{(k)}$ are the ordered p-values. The Simes procedure controls the type I error at the designated level α if the p-values are independent. Sarkar and Change (1997) proved that the Simes procedure controls the type I error for certain type of multivariate distributions with positive dependence.

By applying the closed testing procedure to the Simes global test procedure, Hommel (1988) proposed the following procedure that has a strong control of the FWE: First determine $j = \max \{i \in \{1, \dots, k\} : p_{(k-i+1)} > i'\alpha/i, i' = 1, \dots, i\}$. If the maximum does not exist, reject all the null hypotheses; otherwise reject $H_{(j)}$ with $p_{(j)} \leq \alpha/j$.

Hommel (1989) proved that his procedure is more powerful than the Hochberg (1988) procedure.

The Hommel procedure is not available in the MULTTEST procedure. We developed a SAS macro to calculate the adjusted p-value based on the Hommel (1988) procedure (see Appendix).

5. Conclusion

We discussed several multiplicity adjustment methods for different settings that are commonly encountered in our practice. Although you should adjust for multiple tests performed, it is also inappropriate to adjust for every measured response and every test performed. Otherwise, you can hardly conclude any significant result. Multiplicity adjustment should be well planned beforehand. For instance, in clinical efficacy studies, multiplicity adjustment should be applied to primary endpoints, which are identified through pilot study or experience.

The other issue you should put into consideration is the selection of the type I error to be controlled. In the aforementioned clinical efficacy studies, after identifying the primary endpoints, a strong control of FWE should be desired. However, at the first stage of drug screening studies, in which twenty to thirty chemical compounds are investigated for the potential of treating disease, the control of false discovery rate (Benjamini and Hochberg, 1995) or per-comparison error rate (Hochberg and Tamhane, 1987) might be appropriate.

Among the multiplicity adjustment methods and SAS function/procedures discussed, some provide adjusted p-values. You can make decision directly from the adjusted p-values. Some provide the global p-values only. If the strong control of the FWE is necessary, you can apply the closed testing procedure discussed in Section 2 to make the decision on the individual null hypotheses.

References

- Agresti A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons.
- Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57, pp. 289-300.
- Chang, C-K., and Rom, D. M. (1994), "On the Analysis of Multiple Correlated Binary Endpoints in Medical Studies," *Proceedings of the 26th Symposium on the Interface*, pp. 182-186.
- Hochberg, Y. (1988), "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 75, pp. 800-802.
- Hochberg, Y. and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, New York: John Wiley & Sons.
- Holland, B. S. and Copenhaver, M. D. (1987), "An Improved Sequentially Rejective Bonferroni Test Procedure," *Biometrics*, 43, pp. 417-423.
- Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, pp. 65-70.
- Hommel, G. (1988), "A Stagewise Rejective Multiple Test Procedure Based on A Modified Bonferroni Test," *Biometrika*, 75, pp. 383-386.
- Hommel, G. (1989), "A Comparison of Two Modified Bonferroni Procedures," *Biometrika*, 76, pp. 624-625.
- Hsu, J. C. (1996), *Multiple Comparisons*, New York: Chapman and Hall.
- Hsu, J. C. (1992), "The Factor Analytic Approach to Simultaneous Confidence Intervals for Multiple Comparisons with the Best," *Journal of Computational Statistics and Graphics*, 1, 151-168.
- Mantel, N. (1963), "Chi-square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure," *Journal of the American Statistical Association*, Vol. 58, pp. 690-700.
- Marcus, R., Peritz, E., and Gabriel, K. R. (1976), "On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance," *Biometrika*, 63, pp. 655-660.
- Rom, D. M. and Chang, C-K. (1995), *Multcomp 3.0--A Statistical Package for Multiple Comparison Procedures*, Ambler, PA: ProSoft Software Inc.
- Sarkar, K. Sanat and Chang, C-K., (1997), "The Simes Method for Multiple Hypothesis Testing with Positively Dependent Test Statistics," *Journal of the American Statistical Association*, 92, pp. 1601-1608.
- SAS Institute Inc. (1997), *SAS/STAT® Software: Changes and Enhancements through Release 6.12*, Cary, NC: SAS Institute Inc., 1167 pp.
- SAS Institute Inc. (1989), *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute Inc., 846 pp.

Searle, R. S., 1987, *Linear Models for Unbalanced Data*, New York: John Wiley & Sons.

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

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

Tukey, J. W., Ciminera, J. L., and Heyse, J. F. (1985), "Testing the Statistical Certainty of a Response to Increasing Doses of a Drug," *Biometrics*, 41, pp. 295-301.

Westfall, P. H. and Young, S. S., 1993, *Resampling-Based Multiple Testing*, New York: John Wiley & Sons.

Appendix: SAS Macro to Calculate the Adjusted P-Values Based on Hommel's Procedure

For a given set of unadjusted p-values and a significance level, this macro generates adjusted p-values based on the closed testing procedure using the Simes (1986) global test procedure, and make the local inference based on Hommel's (1988) procedure. In fact, we can prove that both methods are equivalent; that is, the decision based on the generated adjusted p-values are the same as the one based on Hommel's procedure. After running the macro for a given set of raw p-values, you can verify that if you run the macro again with alpha less than or equal to a generated adjusted p-value, then the individual null hypothesis is rejected by Hommel's procedure; otherwise it is retained. Note that Hommel's procedure has a strong control of the FWE.

```
%macro Hommel(
  data = /* the input data set */,
  rawp = /* the variable name for the raw p-values */,
  alpha = /* the significance level */);

***Sort the raw p-values from large to small and exclude
the missing p-value.;
proc sort data=&data out=pdata;
  where (&Rawp ^= .);
  by descending &rawp;
run;

*** Count the # of non-missing p-values (= &np);
data _null_;
  set pdata nobs=nobs;
  if _N_=1 then call symput("np", trim(left(nobs)));
  stop;
run;

proc transpose data=pdata out=pdata prefix=rawp;
  var &rawp;
```

```
run;

data adjustp;
  set pdata;
  array rawp(&np);
  keep raw_p adjp;
  *** Calculate the adjusted p-value for each raw p-value;;
  do i=1 to &np;
    raw_p=rawp(i);
    adjp=0;
    *** For each # of individual hypotheses included in the
    global null hypotheses, determine the minimum
    significance level such that the Simes global test
    can reject the global null hypotheses;;
    do nh=1 to &np;
      *** nh=# of hypotheses in the global test.;
      if nh<=i then
        do;
          *** x=the minimum significance level to reject
          the i-th largest p-value. If the number of
          hypotheses in the global test is less than or
          equal to i, i can be the smallest p-value in the
          global test. Thus rawp(i) <= x/nh. That is, the
          minimum of x is nh*rawp(i).;
          x=nh*rawp(i);
          *** If there are more than one individual hypothesis
          in the global test, the least possible situation to
          reject the test is when rawp(i) and the largest p-
          values are included in the test. In this case, the
          null hypothesis can be rejected if rawp(k) i.e.
          ((nh-k+1)/k)*x, for any k=1 to (nh-1).;
          if nh>1 then
            do;
              do k=1 to (nh-1);
                x=min(x, rawp(k)*nh/(nh-k+1));
              end;
            end;
          else if nh>=i then
            do;
              *** Using the same reasoning;;
              x=1;
              do k=1 to nh;
                x=min(x, rawp(k)*nh/(nh-K+1));
              end;
            end;
          adjp=max(x, adjp);
        end;
      output;
    end;
  proc print;
    var raw_p adjp;
    title "HOMMEL'S ADJUSTED P-VALUE";
  run;

  *** Determine whether to reject or retain the individual
  hypotheses for the given significance level alpha,
  using the Hommel procedure.;
```

```

data decision;
set pdata;
array rawp(&np);
*** Decide  $j = \max\{\dots\}$ .;
j=0;
do i=1 to &np;
  k=0;
  pass=1;
  do until(k=i or pass=0);
    *** pass=0: The condition  $p_{(j)} > \alpha/j$  &
       $p_{(j-1)} > 2*\alpha/j, \dots, p_{(n-1)} > (j-1)*\alpha/j$  &
       $p_{(n)} > \alpha$  does not hold.;

    k=k+1;
    if rawp(k) <= (i-k+1)*&alpha/i then pass=0;
    if k=i and pass=1 then j=i;
  end;
end;
*** output the decision.;
if j=0 then
do;
  decision="REJECT";
  do i=1 to &np;
    raw_p=rawp(i);
    output;
  end;
end;
else if j>0 then
do;
  critical = &alpha/j;
  do i=1 to &np;
    raw_p=rawp(i);
    if raw_p<=critical then decision="REJECT";
    else decision="RETAIN";
  end;
end;
proc print;
var raw_p decision j;
title "THE RESULT FROM HOMMEL'S PROCEDURE,
ALPHA = &alpha";
run;

%mend Hommel;

```

Authors

Chung-Kuei Chang, 517 Barbara Drive, Norristown, PA 19403, Tel.: (610) 239-9459, e-mail: chung_kuei@yahoo.com.

Dror Rom, 12 East Butler Pike, Ambler, PA 19002, Tel.: (215) 591-9010, e-mail: rom@prosof.com.

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.