ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Dariusz Parys**

MULTIPLE ENDPOINTS

Abstract. In ANOVA we are mainly based on inter-treatment comparisons. Another common problems arising in biometric studies (especially in biomedical studies) is that of comparing two groups of patients (treatment and a control group) based on multiple response (called multiple endpoints).

In this paper we present the continuos and discrete approaches to multiple endpoints. In the case of continuous multiple endpoints we have common assumption in that the covariance matrices in group of the control and observation are equal. Let ρ be the correlation coefficient between Y_i and Y_j endpoints and ρ_i be the raw *p*-value obtained using some tests statistics for the *i*-th endpoints.

We can also proposed a general bootstrap approach which can be used to estimate the *p*-value without making any parametric and distributional or correctional assumptions.

Binary outcomes are common in medical studies. We present the modified Bonfferroni procedures and permutational procedures and we compare these procedures to each other.

Key words: multiple comparisons, multiple endpoints, bootstrap approach.

1. INTRODUCTION

In ANOVA we are mainly based on inter-treatment comparisons. Another common problems arising in biometric studies (especially in biomedical studies) is that of comparing two groups of patients (treatment and a control group) based on multiple response (called multiple endpoints).

Suppose there are $k \geqslant 2$ endpoints $Y_1, Y_2, \ldots, Y_k$. Denote by $\mathbf{Y}_0 = (Y_{01}, Y_{02}, \ldots, Y_{0k})$ and $\mathbf{Y}_1 = (Y_{11}, Y_{12}, \ldots, Y_{1k})$ the vectors of observations on a typical patient from a control group and the treatment group.

* Ph.D., Department of Statistical Methods, University of Łódź.

Let $\boldsymbol{\mu}_0 = (\mu_{01}, \mu_{02}, ..., \mu_{0k})$ and $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, ..., \mu_{1k})$ be the mean vectors of the two groups and let $\boldsymbol{\theta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ be the difference vector.

Two different types of questions are often posed:

1. Is there at least one endpoint for which the treatment is more effective than the control? Identify all such endpoints.

2. Do different endpoints point in the same direction with regard to the superiority of the treatment over the control? If, so, does the combined evidence support the treatment's superiority?

In this paper we present the continuous and discrete approaches to multiple endpoints (Hochberg, Tamhane 1987). In the case of continuous multiple endpoints we have a common assumption in that the covariance matrices in group of the control and observation are equal.

We also proposed a general modified bootstrap approach which can be used to estimate the *p*-value without making any parametric and distributional or correlational assumptions.

2. CONTINUOUS ENDPOINTS

Let Y_{0m}, $m = 1, 2, ..., n_0$, be n_0 i.i.d. observations from the control group and Y_{1m}, $m = 1, 2, ..., n_1$, be n_1 i.i.d. observations from the treatment group. A common assumption is that the covariance matrices of the Y_{lm} in each group $l = 0, 1$ are equal. Let ρ_{ij} be the correlation coefficient between Y_i and Y_j (the *i*-th and *j*-th endpoint) for $1 \leqslant i < j \leqslant k$.

Let p_i be the raw *p*-value obtained using some statistic for the *i*-th endpoint $1 \leqslant i \leqslant k$.

First we can mention the methods based only on the raw *p*-values for adjusting the p_i

$$p_{ai} = 1 - (1 - p_i)^{\sqrt{k}}, \quad 1 \leqslant i \leqslant k.$$

We can generalize this formula to depend on the ρ_{ij} as follows:

$$p_{ai} = 1 - (1 - p_i)^{k^{(1-\overline{\rho})}}, \quad 1 \leqslant i \leqslant k,$$

where $\overline{\rho}$ is the average of all the ρ_{ij}.

Now suppose that the Y_{0m} and Y_{1m} are multivariate normal. For testing $\mathrm{H}_{0i}: \theta_i = 0$ consider the usual test statistic

$$Z_i = \frac{\overline{Y}_{1i} - \overline{Y}_{0i}}{\sigma_i \sqrt{1/n_1 + 1/n_0}},$$

where $\overline{Y}_{1i}$ and $\overline{Y}_{0i}$ are the corresponding sample means and σ_i is the standard deviation of Y_i (usually estimated from data) $(1 \leqslant i \leqslant k)$. Note that $\mathrm{corr}(Z_i, Z_j) = \rho_{ij}$ $(1 \leqslant i < j \leqslant k)$. The raw p-values are given by

$$p_i = \mathrm{P}(Z_i \geqslant z_i \mid \theta = 0),$$

where z_i is the observed value of Z_i $(1 \leqslant i \leqslant k)$.

Recently most of authors have development the following *ad hoc* method, which is a hybrid of the multivariate normal and the p-value based methods. Let $z^{(\alpha)}$ be the upper α critical point of the univariate standard normal distribution. Then k' is found from

$$1-(1-\alpha)^{k'} = \mathrm{P}\left(\max_{1 \leqslant i \leqslant k} Z_i \geqslant z^{(\alpha)}\right).$$

Having found k', the adjustment p-values are calculated using

$$p_{ai} = 1-(1-p_i)^{k'} \quad (1 \leqslant i \leqslant k).$$

3. BOOTSTRAP APPROACH

The advantages of the bootstrap approach are that:

1) it is distribution free,

2) it accounts for the dependence structure automatically from the observed data,

3) it is very flexible in accommodating different tests for different endpoints.

We proposed a general bootstrap approach which can be used to estimate that ρ_{ai} without making any parametric distributional or correlational assumptions.

Let $y_{01}, y_{02}, \ldots, y_{0n_0}$ and $y_{11}, y_{12}, \ldots, y_{1n_1}$ be the observed data vectors from control and the treatment groups, respectively. Let $p_1, p_2, \ldots, p_k$ be the observed raw p-values obtained using appropriate two-sample tests for each endpoint. The bootstrap procedure operates as follows:

1) pool the two samples together,

2) draw bootstrap samples $y^*_{01}, y^*_{02}, \ldots, y^*_{0n_0}$ and $y^*_{11}, y^*_{12}, \ldots, y^*_{1n_1}$ with replacement from the pooled sample,

3) apply the appropriate two-sample tests to each of the k endpoints using, the bootstrap samples and calculate bootstrap p-values $p^*_1, p^*_2, \ldots, p^*_k$,

4) repeat steps 2 and 3 some large number (N) of times,

5) the bootstrap estimates of the adjustment p-values are then

$$\hat{p}_{ai} = \frac{\#(\min p_j^* \leqslant p_i)}{N} \quad (1 \leqslant i \leqslant k),$$

where $\#(\min p_j^* \leqslant p_i)$ is the number of simulations resulting in $p_j^* \leqslant p_i$.

3.1. Discrete endpoints

Binary outcomes are common in medical studies. Suppose that we divide randomly 100 patients into a control and a treatment group.

For each patient, k different sites (e.g. heart, skin) are examined for the occurrence of tumors. The k outcomes for each patient can be regarded as multiple endpoints.

Based on these data, it is of interest to determine if there is an increases incidence of tumors in the treatment group at certain sites. If π_{0i} and π_{1i} denote the tumor incidence rates at site i for the central and treatment groups, respectively, then this can be formulated as a multiple hypotheses testing problem

$$\mathrm{H}_i : \pi_{0i} = \pi_{1i} \quad \text{vs.} \quad A_i : \pi_{0i} < \pi_{1i} \quad (1 \leqslant i \leqslant k).$$

Let $\mathrm{H} = \bigcap_{i=1}^{k} \mathrm{H}_i$ and $A = \bigcap_{i=1}^{k} \mathrm{A}_i$.

Suppose there are n_0 patients in the control group and n_i on the treatment group. Let Y_{0i} and Y_{1i} be the numbers of patients in each group with tumors at site i $(1 \leqslant i \leqslant k)$. Then $Y_0 = (Y_{01}, Y_{02}, \ldots, Y_{0k})$ and $Y_1 = (Y_{11}, Y_{12}, \ldots, Y_{1k})$ are independent multivariate binomial vectors with correlated components. Let $y_0 = (y_{01}, y_{02}, \ldots, y_{0k})$ and $y_1 = (y_{11}, y_{12}, \ldots, y_{1k})$ be the corresponding observed data vectors. For each site i we have a 2×2 table

	Tumor	No tumor	Total
Control	y_{0i}	$n_0 - y_{0i}$	n_0
Treatment	y_{1i}	$n_1 - y_{1i}$	n_1
Total	m_i	$n - m_i$	n

where $n = n_0 + n_1$ is the total number of animals in the study.

The raw p_i can be obtained by conditioning on m_i and using Fisher's exact test

$$p_i = \sum_{y \leqslant y_{0i}} \frac{\binom{n_0}{y}\binom{n_1}{m_i - y}}{\binom{n}{m_i}} = \sum_{y \geqslant y_{1i}} \frac{\binom{n_0}{m_i - y}\binom{n_1}{y}}{\binom{n}{m_i}}, \quad i = 1, \ldots, k.$$

One may consider using the p_i to test the H_i and (by the UI method) $p_{\min}$ to test H. However, to account for the multiplicity of the tests, the adjusted p-values, $p_{a,i}$ and $p_{a,\min}$, must be used. For this purpose, the Bonferroni methods for continuous data are generally too conservative.

4. MODIFIED PROCEDURES

4.1. Tukey–Mantel procedure

The following formulas are easily generalized to calculate the $p_{a,i}$:

$$\mathrm{P}_{\mathrm{H}}(P_i \leqslant p_{\min} \,|\, m_i) \equiv p_i^* \leqslant p_{\min} \quad (1 \leqslant i \leqslant k),$$

$$p_{a,\min} = \min\left(\sum_{i=1}^{k} p_i^*, 1\right), \quad p_{a,\min} = 1 - \prod_{i=1}^{k}(1 - p_i^*).$$

4.2. Tarone's procedures

R. E. Tarone (1990) used this idea to sharpen the Bonferroni procedure as follows: Calculate the minimum value of p_i for each i if $m_i \leqslant n_1$ then

$$p_{i,\min} = \frac{\binom{n_1}{m_i}}{\binom{n}{m_i}} \quad (1 \leqslant i \leqslant k).$$

1. First check whether the Bonferroni procedure can be used with level α for each hypothesis. Since the FEW must be controlled at level α, this is possible only if these is at most one rejectable hypotheis, i.e., if

$$k_1 = \#(i : p_{i,\min} < \alpha) \leqslant 1.$$

If there are no rejectable hypotheses ($k_1 = 0$) then accept all H_i's. If $k_1 = 1$ then test that rejectable hypothesis at level α.

2. If $k_1 > 1$ then check whether the Bonferroni procedure can be used with level $\alpha/2$ for each hypothesis. Since the FEW must be controlled at level α, this is possible only if there are at most two rejectable hypotheses, i.e. if

$$k_2 = \#(i : p_{i,\min} < \alpha/2) \leqslant 2.$$

If $k_2 = 0$ then accept all H_i's. If $k_2 = 1$ or 2 then test those rejectable hypotheses each at level $\alpha/2$. If $k_2 > 1$ go to the next step.

3. In general, let

$$k_j = \#(i : p_{i,\min} < \alpha/j), \quad j = 1, 2, ..., k.$$

Note $k_1 \geqslant k_2 \geqslant ... \geqslant k_k$. Find the smallest $j = j^*$ such that $k_j \leqslant j$. Then test the rejectable H_i at level α/j^*.

5. PERMUTATIONAL PROCEDURES

5.1. Brown and Fears procedure

To explain this method, introduce the notation $Y_0(S)$ and $Y_1(S)$ where $Y_0(S)$ (respectively, $Y_1(S)$) is the number of animals in the control group (respectively, treatment group) with at least one tumor at each site $i \in S \subseteq K = \{1, 2, ..., k\}$; if S is an empty set then the notation stands for patients with no tumors at any of the sites. Note

$$Y_{0i} = \sum_{S : i \in S} Y_0(S) \quad \text{and} \quad Y_{1i} = \sum_{S : i \in S} Y_1(S).$$

Let $Y_0(S) + Y_1(S) = m(S)$ be the total number of patients with at least one tumor at each site $i \in S$. The Brown and Fears method (Brown, Fears 1981) is based on the permulational (randomization) joint distribution on all $m(S)$, $S \subseteq \{1, 2, ..., k\}$ (not just the marginal totals m_i). Under

$$H : \pi_{0i} = \pi_{1i} \quad (1 \leqslant i \leqslant k),$$

this distribution is multivariate hypergeometric

$$P(Y_1 = y_1) = \sum \prod_{S \subseteq K} \binom{m(S)}{y_1(S)} \Big/ \binom{n}{n_1},$$

where $y_1 = (y_{11}, y_{12}, \ldots, y_{1k})$ and the sum is over all $y_1(S)$, $S \subseteq K$ such that

$$y_{1i} = \sum_{S : i \in S} y_1(S) \quad (1 \leqslant i \leqslant k).$$

Using this distribution, $p_{a,\min}$ is obtained from

$$p_{a,\min} = \mathrm{P_H}\left(\bigcup_{i=1}^{k} (Y_{1i} \geqslant c_i \mid m(S)\ \forall S \subseteq K) \right),$$

where c_i is the largest integer such that

$$\mathrm{P_H}(Y_{1i} \geqslant c_i \mid m_i) = p_i^* \leqslant p_{\min}.$$

5.2. Rom procedure

D. Rom (1992) proposed to test the overall null hypothesis H based on the adjusted p-value (denoted by p_a) that takes into account all the p-values instead of only the $p_{\min}$. Let $p_{(1)} \geqslant p_{(2)} \geqslant \ldots p_{(k)}$ be the ordered p-values and let $P_{(i)}$ be the r.v. corresponding to $p_{(i)}$. Then p_a is the probability of the event that

$$\{P_{(k)} < p_{(k)}\} \quad \text{or} \quad \{P_{(k)} = p_{(k)}\} \cap \{P_{(k-1)} < p_{(k-1)}\} \quad \text{or} \ \ldots \ \text{or}$$

$$\{P_{(k)} = p_{(k)}\} \cap \ldots \cap \{P_{(2)} = p_{(2)}\} \cap \{P_{(1)} < p_{(1)}\}.$$

Clearly, this probability is never larger (and often much smaller) than $p_{a,\min} = \mathrm{P}(P_{\min} \leqslant p_{\min})$. Therefore the test of H based on p_a is more powerful than the test based on $p_{a,\min}$.

6. EXAMPLE

In a hypothetical study 100 patients are randomly assigned with 50 each to the control and the treatment group. Only $k = 2$ tumor sites, A and B, are examined with the following results presented on table.

The marginal p-values using Fischer's exact test are: $p_1 = = P(Y_{11} \geqslant 5 | m_1 = 6) = 0.1022$ and $p_2 = P(Y_{12} \geqslant 8 | m_2 = 10) = 0.0457$. We shall now calculate $p_{a,\min}$ using the methods discussed above.

Site	Control	Treatment	Total
A only	0	3	3
B only	1	6	7
A and B	1	2	3
No Tumor	48	39	87
Total	50	50	100

First, for the Bonferroni procedure we have $p_{a,\min} = 2 \cdot 0.0457 = 0.0914$. Next, to apply the Tukey–Mantel procedure we need to calculate p_1^* and p_2^*. We have $P(Y_{11} \geqslant 6 | m_1 = 6) = 0.0133 < p_{\min}$ and $P(Y_{11} \geqslant \geqslant 5 | m_1 = 6) = 0.1022 > p_{\min}$; therefore $p_1^* = 0.0133$. Next, $p_2^* = 0.0457$. We have: $p_{a,\min} = 0.0133 + 0.0457 = 0.0590$. We get $p_{a,\min} = 1 - (1 - 0.0133)(1 - 0.0457) = 0.0584$.

To apply the Tarone procedure (Tarone 1990), first calculate $p_{1,\min} = 0.0133$ and $p_{2,\min} = 0.0005$. Therefore $k_1 = 2$, $k_2 = 2$ and $j^* = 2$; thus no reduction in the number of rejectable hypotheses is achieved. Comparing the observed p_1 and p_2 with $\alpha / j^* = 0.025$, we find that neither site has a significant result at $\alpha = 0.05$.

To apply the Brown and Fears procedure (Brown, Fears 1981) we need the joint distribution of $Y_1 = (Y_{11}, Y_{12})$. From the marginal distributions of Y_{11} and Y_{12} we see that the largest values c_i such that $P(Y_{1i} \geqslant c_i | m_i) = p_{\min}$ are $c_1 = 6$ and $c_2 = 8$. Therefore

$$\begin{aligned} p_{a,\min} &= P\{(Y_{11} \geqslant 6) \cup (Y_{12} \geqslant 8)\} = \\ &= P\{Y_{11} \geqslant 6\} + P\{Y_{12} \geqslant 8\} - P\{(Y_{11} \geqslant 6) \cap (Y_{12} \geqslant 8)\} = \\ &= 0.0133 + 0.0457 - 0.022 = 0.0568. \end{aligned}$$

Notice that the Mantel–Tukey approximations, namely 0.0590 and 0.0584. are quite close to the exact $p_{a,\min}$. However, they are all greater than $\alpha = 0.05$ and so H cannot be rejected.

Finally we apply the Rom procedure (Rom 1992) to these data. Adding up the probabilities from joint distribution of Y_{11} and Y_{12} we find that $p_a = 0.0285$. Thus, in this example, only the Rom procedure yields a significant result.

REFERENCES

Brown C. C., Fears T. R (1981), *Exact Significance Levels for Multiple Binominal Testing with Application to Carcinogenicity Screens*, "Biometrics", 37, 763–774.

Hochberg Y., Tamhane A. C. (1987), *Multiple Comparisons Procedures*, Wiley, New York.

Rom D. (1992), *Strenghtening some common multiple test procedures for discrete data*, Statist. Medicine, 11, 511–514.

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

Dariusz Parys

WIELOKROTNE PUNKTY KRAŃCOWE

Większość procedur testowych, dotyczących porównań wielokrotnych, związanych jest z porównaniami między zabiegami medycznymi. W studiach biometrycznych często spotykamy się z problemem porównań między dwiema grupami pacjentów (grupą zabiegową i grupą kontrolną) opartymi na wielokrotnych wynikach (relacjach) zwanych punktami krańcowymi. Rozważamy $k \leqslant 2$ punktów końcowych $Y_1, Y_2, ..., Y_k$. Oznaczmy przez $\mathbf{Y}_0 = (Y_{01}, Y_{02}, ..., Y_{0k})$ oraz $\mathbf{Y}_1 = (Y_{11}, Y_{12}, ..., Y_{1k})$ wektory obserwacji typowego pacjenta z grupy kontrolnej i grupy zabiegowej.

Niech $\boldsymbol{\mu}_0 = (\mu_{01}, \mu_{01}, ..., \mu_{0k})$ oraz $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, ..., \mu_{1k})$ będą odpowiednio wektorami średnich z obu grup, natomiast $\boldsymbol{\theta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ będzie wektorem różnic. W artykule przedstawiono procedury testowe i ich modyfikacje dotyczące ciągłych i skokowych punktów krańcowych oraz zaproponowano podejście bootstrapowe do estymacji p-wartości.