ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 286, 2013

Dariusz Parys[*]

STEPWISE MULTIPLE TESTS PROCEDURES FOR DISCRETE DISTRIBUTIONS

Abstract. We presented some properties of new procedures of multiple hypotheses testing for discrete distributions. We choose the new procedure stepwise *TWW*$_k$, based on Tarone, Westfall and Welfinger ideas. We compare this procedure to multiple testing procedures like *T**, *TH** and others and we show the power advantage of this procedure.

Key words: multiple procedure, stepwise testing, discrete distribution.

I. INTRODUCTION

Multiple testing problem involves a family of hypotheses $H_{01},...,H_{0n}$ (alternative $H_{11},...,H_{1n}$). The hypotheses are tested simultaneously and a multiple level α has to be controlled. A ”valid” procedure to solve this problem will maintain strong control of the familywise error rate (FEW) at its nominal level α (i. e. the probability of rejecting at least one true H_{0i} $(i=1,...,n)$ is at most α, no matter which and how many H_{0i} are true (Hochberg and Tamhane, 1987). A simple way to solve the question is to use the Bonferroni method. This method rejects all hypotheses with p-values less than or equal to α/n. The Bonferroni method is conservative when the p-values are uniformly distributed, since it ignores correlation between the p-values. It does not make allowance for situations where one of the null hypotheses are clearly false. In addition, the Bonferroni method may become conservative due to discreteness of the sampling distribution, and this disadvantage is potentially worse than the two disadvantages mentioned earlier.

The existing and new multiple hypotheses procedures for discrete distribution are critically reviewed and compared to each other, for its power and average power.

Gart et al. (1979) note that for discrete data statistics, there actually exists the smallest attainable p-value α_i^* $(i=1,...,n)$ for each hypothesis. Thus, the number of significance tests could be reduced by eliminating those tests, for which the

[*] Ph.D., Chair of Statistical Methods, University of Lodz.

smallest p-values is higher than $\alpha\,(\alpha_i^* > \alpha)$. Tarone (1990) improved this idea by noting that even for hypotheses with $\alpha / n < \alpha_i^* < \alpha$ rejection may never be possible. For each integer k. define $R_k = (i : k\,\alpha_i^* < \alpha)$ and $m(k) =| R_K |$, where α is the nominal significance level and α_i^* is the minimum achievable level at site i. Thus $m(1)$ is the number of hypotheses that can be rejected at the nominal level α. If $m(1) > 1$, a correction for multiple comparisons should be considered.

For any integer $k < m(1)$, $m(k+1) \le m(k)$ and $m[m(1)] \le m(1)$ (Tarone, 1990), thus if the correction factor is $m(1)$, there may exist H_{0i} such that its $\alpha_i^* < \alpha / m(1)$, and one cannot reject those H_{0i}, whatever their p-value will be. By excluding those H_{0i}'s, the correction factor can be reduced until the smallest number k such that $m(k) \le k$ is reached. Define K to be the smallest value of k such that $m(k) \le k$. This reduction will only be effective for discrete data, since in continuous data $m(1) = m(2) = ... = m(n) = n$, so that, so that $K = n$ and the usual Bonferroni method is applied. The values K and R_k can be determined using only the information in the marginal total. Tarone's procedure (T) rejects H_{0i} if and only if H_{0i} is contained in R_k and $p_i < \alpha / K$, where p_i is the observed significance level at hypothesis i.

Unfortunately, T lacks alpha consistency (AC) (Roth, 1999). Hypothesis that is accepted at a gives α level may be rejected at a lower α level. Roth (1999) developed procedure T^* that modifies AC while simultaneously increasing the power. T^* maintains strong control of $FWE < \alpha$. The procedure rejects all H_{0i}'s such that $p_i < \alpha / K^*$ where $M = \{x \in [0,1] \,|\, m(x) \le x\}$ and $K^* = \inf\{x \in M\}$. A simple way to construct T^* in practice will be to arrange the smallest attainable p-value in an increasing manner $\alpha_{(1)}^* \le ... \le \alpha_{(n)}^*$. If $m(K) = K$ then $K^* = K$ else $K^* = \alpha / \alpha_{(K)}^*$. T^* does not stand the $FWE \le \alpha$ criterion is used, arises from those cases where $m(K^*) > K^*$.

Westfall and Wolfinger (W&W, 1997) suggested a different approach based on the full set of possible values for each P_i, rather than just on the minimum attainable p-values α_i^* for each P_i. They defined adjusted p-values $(p_j^{'})$ as $p_j^{'} = \Pr(\min P_i \le p_j)$ where P_i refers to the random p-values considered under their null hypotheses. If we define p_i $(i = 1, ..., n)$ as the observed p-values of given tests, given that the distribution of the test statistics is discrete, the observed values of the random p-values P_i will be $\{p_{it} : t = 1, ..., m_i\}$ (m_i is the maximum availed value for the ith test statistic) where $\Pr(P_i \le p_{it})$. The adjusted value, p'_j will be the probability that a p-value as small as p_j will be observed in the entire study when all null hypotheses are true.

Using discreteness

$$p'_j = 1 - \prod_{i=1}^{n} (1 - p_{it(j)}) \text{ where } p_{it(j)} = \begin{cases} \max_t \{p_{it} : p_{it} \le p_j\} & \text{if } \min_t \{p_{it}\} \le p_j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For each hypothesis, the procedure computes its adjusted p-value and compares it to $FWE = \alpha$. The latter procedure assumes independence between the tests, thus making the method rather conservative, although less than the Bonferroni method. In case of dependence, one way to bind the true values of p'_j, will be to use the Bonferroni inequality. (The discrete Bonferroni adjusted p-values are $p'_j = \min\{\sum_{i=1}^{n} P_{it(j)}, 1\}$. Another way, probably preferable, will be to calculate the exact $\min(P_i)$ distribution either exactly or using Monte Carlo (MC) resampling method.

We propose a new method, *TWW$_k$*, *FWE* and incorporates the discreteness of the distribution. This method will use W&W on the set defined by T_k,. *TWW$_k$*, rejects $\{H_{0i} \in R_k \mid p'_i \le \alpha\}$ where $p'_i = \Pr(\min P_i \le p_j)$, $\{j \mid H_{0j} \in R_k\}$. This method controls the $FWE \le \alpha$.

Some of the methods are universally more powerful than others, some are not universally so;

Claim 1. *T* is universally more powerful than T* (Roth, 1999).

Claim 2. *T_k is universally more powerful than T* (Roth, 1999).

Claim 3. *None of T_k* and *T* is universally more powerful than the other* (Roth, 1999).

Claim 4. *W&W method is universally more powerful than T*.*

Claim 5. *TWW$_k$, is universally more powerful than T_k.*

Claim 6. None of *W&W, and TWW$_k$/T_k, is universally more powerful than the others.*

II. STEPWISE PROCEDURES

Stepwise methods provide an increase of the power of multiple testing methods. These techniques are not unique to discrete distributions, but need to be discussed since they improve the power of the multiple hypotheses tests.

The procedure suggested by W&W can easily be adapted to stepwise analysis. The p-values are adjusted using the step down technique, by adjusting the smallest p-value according to min (P_i) distribution. The second smallest p-value is adjusted according to the min (P_i) distribution of all the variables excluding the variable whose unadjusted p-value was smallest, and so on.

Hommel and Krummenauer (1998) step down procedure, is similar to Holm's (1979) Bonferroni test, but incorporates T*. This procedure was named TH*:

(1) Set $I = \{1,..., n\}$,

(2) For j = 1, …, #I define $m_I(\alpha, j)$ =#$\{i \in I \mid \alpha_i^* \le \alpha / j\}$, number of hypotheses with indices $i \in I$ that can be rejected at level α / j. $K_I(\alpha) = \min(j = 1,..., \# I \mid m_I(\alpha, j \le j\}$ and $b_I(\alpha) = \alpha / K_I(\alpha)$.

(3) For $i \in I$ reject H_{0i} if $p_i \le b_I(\gamma)$ for some $0 < \gamma < \alpha$. (practically apply T* on I).

(4) Let J = index set of all hypotheses that have been rejected in step 3.

(5) If J is empty stop otherwise set $I = I – J$ and return to step 2.

Roth (1999) described a step up Procedure R based on Hochberg's procedure (H) (1998). Procedure R is composed of two procedures: Procedure L (that is closely related to H), and a component Procedure C. R rejects H_{0i} if it was rejected by either L or C.

Procedure *L*

(1) Accept the entire p_i's that are not in $R_1 = \{H_{0i} \mid \alpha_i^* < \alpha)$.

(2) Order the p_i's in R_1 from highest to lowest $p_{(1)} \ge, ..., \ge p_{(m(1))}$.

(3) Let $Q = \{j \mid p_{(j)} < \alpha / j, p_{(j)} \in R_1\}$ define $q = \min\{j \in Q\}$.

(4) Reject all of the $H_{0i} \in R_1$ such that $p_i < \alpha / q$.

Procedure *C*

(1) Consider only the $H_{0i} \in R_k$ order the p_i's from highest to lowest by $p_{(1)} \ge, ..., \ge p_{(m(k))}$ if $m(K) < K$ than $q_{(i)} = 0$ for $I = m(K), ..., K$ (K – as defined in 1 above).

(2) For $j = 1,..., K$ define $p_j^* = \max\big\{\{q(j)\} \cup \{p_i \mid H_{0i} \in R_j - R_k\}\big\}$.

(3) Let $W = \{j \mid p_j^* < \alpha / j\}$ define $w = \min\{j \in W\}$.

(4) Reject H_{0i} if $p_i < \alpha / w$.

R is valid if H is valid for all subsets of R_1 of size $\le q^*$ is defined as the larger of $m(K)$ and $\max\{\{0\} \cup \{i = 1,..., k-1 \mid R_i - R_{i+1}$ is not empty$\}\}$.

A newly proposed stepwise method

Using the mechanisms described in Section 2, one can apply W&W stepwise method to the group of *p*-values with a hypothesis that belongs only to R_k. This method has properties similar to those of TWW_k (lack of AC, universally more powerful than T_k and T), but it has a higher power, since we use a stepwise method rather than a single step.

Comparison between stepwise multiple hypothesis methods

Comparison between single step methods and stepwise methods was not performed since matching stepwise method, which is more powerful. However, occasionally, one type of single step method is more powerful than a stepwise method.

Claim 7. *Stepwise W&W method is universally more powerful than TH*.*

Claim 8. *R/RMOD*, *W&W stepwise and stepwise TWW_k are not universally more powerful than each other.*

Claim 9. *R/RMOD* and *W&W stepwise are not universally more powerful than T_k / TWW_k and vice versa.*

Claim 10. *None of T_k / TWW_k and TH* are universally more powerful than each other.*

Claim 11. *Stepwise TWW_k is not universally more powerful than either TH* or T*.*

III. APPLICATIONS OF THE MULTIPLE TESTING PROCEDURES

Case 1 – animal carcinogenicity test

Several animal organs and tissues were examined for the presence of tumour caused by a test compound as in Tarone (1990). This was a three-arm study: control (0), low dose (1), and high dose (2). The groups consisted of equally spaces doses. The number of observed tumours was recorded for each type group (animal (mouse, rat), gender (male, female), and tumour site). A trend statistic of the following form was defined $T_j = X_{0j} \cdot 0 + X_{1j} \cdot 1 + X_{2j} \cdot 2$ were X_{ij} are the number of observed tumours at dose group *i*, and type group *j*. Upper-tailed *p*-values were computed for each type group, using Fisher's exact statistics.

All methods rejected both hypotheses {*male mouse liver, female mouse liver*} at the 0.01 level. At the 0.1 significance value, all methods rejected the {*male rat kidney, male mouse liver, female mouse liver*} hypotheses. None of the methods tested, including the new ones, was more powerful than the others testing the hypotheses in this case study.

Case 2 – relationship between DVT and three genetic factors

This case study tested the relationship between deep vein thrombosis (DVT) and three different genetic factors (Fact V, Fact II and MTHFR) (Salomon et al.,1999). The population was divided into healthy controls and those with DVT. Each subject was tested for the presence of one of the three genetic factors. The subjects were then divided into one of the eight available genetic groups (a genetic group is built of the combination of presence or absence of all the three factors).

All methods, except for RMOD and $R+C$ rejected the same hypotheses for all levels of significance (Fact V, Fact V + Fact II, and Fact V + MTHFR). RMOD and $R+C$ rejected these three hypotheses for the 0.01 and 0.05 significance levels, and rejected the "All 3 Factors" hypothesis in the 0.1 significance level. $R+C$ and RMOD was deemed more powerful than all other methods for this case.

This simulation is based on some samples of the animal experiment presented in Case 1. The samples differed by the weights given to different hypotheses using extended multinomial hyper-geometric distribution $c\prod_{j=1}^{s}\binom{n_j}{x_j}z_j^{x_i}$, $z_j = p_{1j}p_{2s}/p_{1s}p_{2j}$, z_j is the extension of the odds ratio to $2\times N$ tables, z_j stands for the ratio between each group and the control group. The number c is determined by the condition that the sum over its range is unity). The difference between samples was created by two-fold increase weight of the low dose, and the 2.5-fold increase weight of the high dose. Each sample consisted of 10,000 resample data sets, and was tested using four multiple comparison tests (R/RMOD, W&W, stepwise *TWW_k*), and at three different significance level (0.01, 0.05, 0.1).

The dependent case of extended multinomial hyper-geometric distribution

The data for this simulation was derived from Case 2.

The study looks at the odds ratio for developing DVT by each of the seven genetic groups vs. the group that carries no genetic risk factors. We applied five tests to this simulation: stepwise W&W, stepwise *TTW_k*, R, EMOD, and *R* + *C*. The *p*-values were calculated as one-sided test from the multinomial hyper-geometric distribution. The significance level was set at 0.05.

REFERENCES

Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple test of significance. Biometrika 75, 800-802.

Hochberg, Y., Tamhane, A., 1987. Multiple Comparison Procedures. Wiley, New York.

Roth, A. J., 1999. Multiple comparison procedures for discrete test statistics. J. Statist. Plann. Inference 82, 101-117.

Tarone, R. E., 1990. A modified Bonferroni method for discrete data. Biometrics 46,515-522.

Westfall, P. H., Wolfinger, R. D., 1997. Multiple tests with discrete distributions. Amer. Statist. 51, 3-7.

Dariusz Parys

KROCZĄCE PROCEDURY TESTÓW WIELOKROTNYCH DLA ROZKŁADÓW DYSKRETNYCH

Zaproponowano tutaj nowe kroczące procedury wielokrotnego testowania w przypadku danych pochodzących z populacji o rozkładzie dyskretnym. Wybierając procedurę TWW_k, opartą na badaniach Tarone'a, Westfalla i Welfingera porównano tę procedurę do innych procedur testowania wielokrotnego (m. in. T^*, TH^*) i pokazano większą moc tej procedury.