SciMedCentral

Annals of Biometrics & Biostatistics

Research Article

Calculation of the Per Hypothesis Error Rate via Sums of Steck's Determinants

Alan D. Hutson*

Department of Biostatistics, University at Buffalo, USA

Abstract

In this note we provide a straightforward approach for calculation of the per hypothesis error rate in a multiple testing framework given a general stepwise testing procedure. This approach is based on a general result due to coined Steck's determinant. This result allows for a direct comparison across various testing procedures and illustrates some common misperceptions regarding the optimality of each method.

Corresponding author

Alan D. Hutson, Department of Biostatistics, University at Buffalo, 706 Kimball Tower, 3435 Main St., Buffalo, NY 14214-3000, USA, Email:

Submitted: 29 August 2013

Accepted: 14 October 2013

Published: 16 October 2013

Keywords

- Multiple outcomes
- Multiple testing
- Step-down and step-up procedures

INTRODUCTION

The classic stepwise multiple outcomes testing problem consists of developing procedures that control the familywise error (FWE) rate over a set of independent hypothesis tests at some overall prescribed level α, e.g. see Hochberg and Tamhane (1987) [1] for a detailed theoretical discussion. The FWE rate is basically a global concept defined as the probability of rejecting any null hypothesis or subset of null hypotheses conditional on the subset hypothesis or hypotheses being true, respectively. The FWE rate may be defined both weakly and strongly, e.g. see Hochberg [2]. The weak FWE rate is what is typically examined in the literature when comparing across stepwise procedures. For a detailed study of various stepwise approaches with respect to the weak FWE rate see Brown and Russell [3].

Note however that determination of the FWE rate is only the first step in terms of designing a study with respect to sample size or power calculations. In practice, once the FWE rate is fixed and the stepwise procedure is determined one then needs to determine the per hypothesis error rate, i.e. the probability of rejecting a specific hypothesis within the stepwise framework conditional on it being true. The most well-known application is the simple Bonferroni correction, used either in a stepwise fashion or in the more traditional sense, where by definition the per hypothesis error rate is α/k, where k is the number of independent hypotheses to be tested. In practical terms each of the k hypotheses then have differential levels of power at some fixed sample size and per hypothesis error rate or may have different sample size requirements at some fixed level of power and per hypothesis error rate. For clinical experiments, e.g. a biomarker study, studies are typically designed such that each specific hypothesis has at least a certain level of power at a fixed sample size and a fixed per hypothesis error rate.

Let us start with the general framework for stepwise a testing procedure, which is commonly based on ordered p-values. The key assumption is that p-values are i.i.d. uniformly distributed under their respective null hypotheses. Let $P_{(1)} < P_{(2)} < P_{(k)}$ denote the set of k ordered p-values corresponding to each of k independent p-values generated from null hypotheses, $H_{0(i)}$, i=1,2,⋯,k, where $H_{0(i)}$ is the ith "ordered" null hypothesis corresponding to the ordered p-value $P_{(i)}$. The general approach in stepwise testing is to either start with the smallest p-value or largest p-value and work your way successively up or down the sequence of ordered p-values, respectively, either rejecting $H_{0(i)}$ and continue testing down the sequence or stop testing successive $H_{0(i)}$'s after not rejecting the current hypothesis. A well-known and oft-cited procedure is the Bonferroni-Holm step-down procedure, e.g. see Holm (1979). This approach is given by the following algorithm:

1. Reject $H_{0(i)}$ if $P_{(1)} < \alpha / k$ and reject the global null hypothesis $H_0 = \bigcap_{i=1}^{k} H_{0i}$,
2. If $H_{0(i)}$ is rejected then reject each successive $H_{0(i)}$ if $P_{(1)} < \alpha/(k-i+1)$ and $P_{(i-1)} < \alpha/(k-i+2)$, else stop.

Let us define $g_{(i)}(\alpha)$, where $g_{(i)}(\alpha) \le \alpha$ and $g_{(j)}(\alpha) \le g_{(j+1)}(\alpha)$, j=1,2,⋯,k−1, as the critical value for rejecting $H_{0(i)}$ at step i at some prescribed level α. Then the classic step-down algorithm may be written more generally as

1. Reject $H_{0(i)}$ if $p_{(1)} < g_{(1)}(\alpha)$ and reject the global null hypothesis $H_0 = \bigcap_{i=1}^{k} H_{0i}$.
2. If $H_{0(1)}$ is rejected then reject each successive $H_{0(i)}$ if $p_{(i)} < g_{(i)}(\alpha)$ and $P_{(i-1)} < g_{(i-1)}(\alpha)$, else stop.

As mentioned above, the most straightforward approach commonly used in practice is to choose all $g_{(i)}(\alpha) = \alpha/k$ based on a simple Bonferroni correction. Also used quite often in practice is the approach based on the work of Einot and Gabriel [4] where we define $g_{(i)}(\alpha) = 1 - (1-\alpha)^{1/k}$. These two commonly

used approaches don't necessarily need to be used in a stepwise manner. What is somewhat counterintuitive is that these two basic approaches may actually provide a testing procedure with superior properties in terms of the per hypothesis error rate than the Bonferroni-Holm method where $g_{(i)}(\alpha) = \alpha/(k-i-1)$, i.e. at first glance one would assume that employing a more complex stepwise procedure would yield an optimal testing procedure, when in fact it oftentimes does not.

The bounds where all $g_{(i)}(\alpha)$ are equal are obviously much easier to utilize and evaluate and do not necessarily have to be utilized in a stepwise fashion. Another distinct advantage of the Einot-Gabriel bound is that at the first step in the testing procedure and at each successive step the error rate is exactly α, whereas the Bonferroni and Bonferroni-Holm error rates are slightly less than α, i.e. the $Pr(\text{P(1)} \le 1 - (1-\alpha)^{1/k} \mid \text{all } H_{0(i)} \text{ true}) = \alpha$ and $Pr(\bigcup_{i=1}^{k} P_{(i)} \le 1-(1-\alpha)^{1/k} \mid \text{all } H_{0(i)} \text{ true}) = \alpha$. Another well-known stepwise procedures includes the approach of Simes (1986), where $g_{(i)}(\alpha) = i\alpha/k$, see also Hochberg (1988) and Hommel (1989). The approach of Simes also shares the property with the Einot-Gabriel method that $Pr(\bigcup_{i=1}^{k} P_{(i)} \le i\alpha / k \mid \text{all } H_{0(i)} \text{ true}) = \alpha$ and will be examined further in the next section.

One possibility for the popularity of the Bonferroni and Einot-Gabriel approaches over the various stepwise approaches is their ease of use in terms of study design such as a clinical trial with multiple endpoints and sample size considerations. This is primarily due to the fact that the per hypothesis error rate defined as Pr (rejecting any $H_{0i} \mid \text{all } H_{0(i)}$ true) $\le \alpha$ (1.1) is a straightforward calculation if $g_{(i)}(\alpha)$ equals either α / k or $1-(1-\alpha)^{1/k}$ for the Bonferroni and Einot-Gabriel approaches, respectively. This then allows for more straightforward examination of the statistical power for a given study. In this note we provide a straightforward approach for calculation of the per hypothesis error rate at (1.1) for the purpose of facilitating study design and to compare some commonly utilized stepwise approaches. This approach is based on a general result due to Steck (1971). This approach allows for a direct comparison across methodologies.

STECK'S DETERMINANT AND PER HYPOTHESIS ERROR RATE CALCULATIONS

Let $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$ denote the order statistics from an i.i.d. sample of size n from a uniform $U(0,1)$ distribution. Steck (1971) proved that

$$Pr(l_i \le U_{(i)} \le m_i, i = 1,2,\cdots,n) = det(\boldsymbol{S}), \tag{2.1}$$

for $l_i \le l_{i+1}$ and $m_i \le m_{i+1}$, where the elements $S_{ij} = \binom{j}{j-i+1}(m_i - l_j)_+^{j-i+1}$ or 0 according as $j-i+1$ is nonnegative or negative across $i=1,2,\cdots,n$ and $j=1,2,\cdots,n$, and $(x)_+ = \max(0,x)$. The matrix S is then seen to have the Hessenberg form with ones on the first subdiagonal and zeros below the first subdiagonal. Breth (1980) and Hutson (2002) [5,6] have utilized the main result of Steck [7] with respect to developing confidence bands for quantiles. Simes (1986) [8] utilized this result, but did not refer to it directly.

Now by noting that in general p-values are uniformly distributed conditional under the global null hypothesis being true we will then be able to directly calculate the per hypothesis error rate at (1.1) compactly via a sum of Steck's determinants. We will also be able to readily calculate the error rate $Pr(\bigcup_{i=1}^{k} P_{(i)} \le g_{(i)}(\alpha) \mid \text{all } H_{0(i)} \text{ true})$ for the global null hypothesis, $H_0 = \bigcap_{i=1}^{k} H_{0i}$. This provides a simple bound on the weak FWE rate for each test. Note that there are exceptions to this theoretical framework under the assumption of uniformly distributed p-values in the case where a nuisance parameter is embedded as part of the estimation and testing scheme, e.g. see Robins et al. [9]. We assume the uniform case for this note.

For a step-down procedure the per hypothesis error rate for the l^{th} hypothesis is defined as

$$Pr\left((P_l = P_{(1)}) \cap (P_{(1)} < g_{(1)}(\alpha)) \mid \text{all } H_{0(i)} \text{ true}\right) \tag{2.2}$$

where P_l is the unordered p-value corresponding to the l^{th} unordered hypothesis of interest. In general, the per hypothesis error rate is less than α. The function $g_{(j)}(\alpha) \in (0,1)$ denotes the adjusted critical value, α', for the j^{th} step of a given step-down procedure, e.g. for the Bonferroni-Holm procedure $g_{(j)}(\alpha) = \alpha/(k-i+1)$. Note that the events $P_l = P_{(j)}$ and $P_{(j)} < g_{(j)}(\alpha)$ are independent events so that (2.2) may be rewritten more compactly as

$$Pr\left(P_{(1)} < g_{(1)}(\alpha) \mid \text{all } H_{0(i)} \text{ true}\right)/k + \sum_{j=2}^{k} Pr\left(P_{(j)} < g_{(j)}(\alpha) \mid \bigcap_{i=1}^{j-1} P_{(i)} < g_{(i)}(\alpha) \cap \text{all } H_{0(i)} \text{ true}\right)/k. \tag{2.3}$$

In terms of the multiple testing problem we now see that the per hypothesis error rate given by equation (2.3) may now be written in terms of sums of Steck's determinant as

$$\sum_{j=1}^{k} det(\boldsymbol{S}(j)) \tag{2.4}$$

Where the l row and m column elements of S defined at (2.1) are given as a function of the index j by

$$S(j)_{lm} = \begin{cases} \binom{m}{m-l+1} g_{(l)}(\alpha)^{m-l+1}, & m-l+1 \ge 0 \text{ and } l \le j, \\ \binom{m}{m-l+1}, & m-l+1 \ge 0 \text{ and } l > j, \\ 0, & m-l+1 \ge 0 < 0, \end{cases} \tag{2.5}$$

where $l=1,2,\cdots,k$ and $m=1,2,\cdots,k$. Note that equation (2.5) holds for both one-sided and two-sided alternatives.

Similarly, the calculation of the weak FWE rate $Pr(\bigcup_{i=1}^{k} P_{(i)} \le g_{(i)}(\alpha) \mid \text{all } H_{0(i)} \text{ true})$ for the global null hypothesis, $H_0 = \bigcap_{i=1}^{k} H_{0i}$ is given by

$$1 - Pr(\bigcap_{i=1}^{k} g_{(i)}(\alpha) \le P_{(i)} \le 1 \mid \text{all } H_{0(i)} \text{ true}) = 1\text{-}det(\boldsymbol{S}), \tag{2.6}$$

where (2.6) the elements $S_{lm} = \binom{m}{m-l+1}(1 - g_{(j)}(\alpha))^{m-l+1}$ or 0 according as $m-l+1$ is nonnegative or negative across $l=1,2,\cdots,k$ and $m=1,2,\cdots,k$.

COMPARING FOUR COMMON APPROACHES

In this section we compare the per hypothesis error rate and weak FWE rate for four commonly used procedures [4,8,10,11]. We compared each approach using an overall FWE rate of α = 0.05. Table 1 provides the calculation of the per hypothesis

Table 1: Per hypothesis error rates for four common procedures.

k	Bonferroni	Einot-Gabriel	Bonferroni-Holm	Simes
2	0.02500	0.02532	0.02563	0.02563
3	0.01667	0.01695	0.01696	0.01723
4	0.01250	0.01274	0.01266	0.01298
5	0.01000	0.01021	0.01010	0.01041
6	0.00833	0.00851	0.00840	0.00869
10	0.00500	0.00512	0.00503	0.00523

Table 2: Weak FWE rates for four common procedures.

k	Bonferroni	Einot-Gabriel (exact)	Bonferroni-Holm	Simes
2	0.04938	0.05000	0.05000	0.05000
3	0.04917	0.05000	0.04941	0.05000
4	0.04907	0.05000	0.04918	0.05000
5	0.04901	0.05000	0.04907	0.05000
6	0.04897	0.05000	0.04901	0.05000
10	0.04889	0.05000	0.04890	0.05000

error rate at (2.2) based on the sums of Steck's determinant for four common procedures used in practice for k=2, 3,4,5,6,10. Similarly, in Table 2 we calculated the weak FWE rate bound $Pr(\bigcup_{i=1}^{k} P_{(i)} \le g_{(i)}(\alpha))$ for the same four procedures.

Interestingly, we see that the Einot-Gabriel method is superior to the Bonferroni-Holm method in terms of the per hypothesis error rate for $k > 3$ and in terms of the weak FWE rate. The Simes approach is only slightly better than the other three methods. Note that even though the procedure due to Simes has a weak FWE rate shown to be equal to α overall, the error rates at intermediate steps may be less than α. The same is true for the Bonferroni approach (when used in a stepwise manner) and the Bonferroni-Holm approach, i.e. the Pr ($P_{(1)} < \alpha/k$ | all $H_{0(i)}$ true) < α. In contrast, the Einot-Gabriel correction provides an exact α level test at each step, if utilized in a stepwise fashion. In terms of practical considerations the Einot-Gabriel correction is straightforward to implement. In terms of theoretical considerations it compares well when considered against other approaches in terms of specific and overall error control.

ACKNOWLEDGEMENTS

This research is supported in part by NIH grant 1R03DE02085101A1.

REFERENCES

1. Hochberg Y, Tamhane AC. Multiple Comparison Procedures. New York: Wiley. 1987.
2. Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika. 1988; 75: 800-802.
3. Brown BW, Russell K. Methods of correcting for multiple testing: operating characteristics. Stat Med. 1997; 16: 2511-2528.
4. Einot I, Gabriel KR. A Study of the Powers of Several Methods of Multiple Comparisons. Journal of the American Statistical Association. 1975; 70: 574-583.
5. Breth M. Quantile estimation and probability coverages. Australian Journal of Statistics. 1980; 22: 207-211.
6. Hutson AD. Exact Bootstrap Confidence Bands for the Quantile Function via Steck's Determinant. Journal of Computational and Graphical Statistics. 2002; 11: 471-482.
7. Steck GP. Rectangle Probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions. The Annals of Mathematical Statistics. 1971; 42: 1-11.
8. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986; 73: 751-754.
9. Robins JM, van der Vaart A, Ventura V. Asymptotic distribution of P values in composite null models. Journal of the American Statistical Association. 2000; 95: 1143-1156.
10. Holm S. A Simple Sequentially Rejective Multiple test Procedure. Scandinavian Journal of Statistics. 1979; 6: 65-70.
11. Hommel G. A Comparison of Two Modified Bonferroni Procedures. 1989; 76: 624-625.

Cite this article

Hutson AD (2013) Calculation of the Per Hypothesis Error Rate via Sums of Steck's Determinants. Ann Biom Biostat 1(2): 1006.