Journal of Statistical Planning and Inference 143 (2013) 764–770

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A combined p-value test for multiple hypothesis testing

Shunpu Zhang [a,*], Huann-Sheng Chen [b], Ruth M. Pfeiffer [c]

[a] Department of Statistics, University of Nebraska Lincoln, Lincoln, NE 68583-0963, USA
[b] Statistical Methodology and Application Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD 20892-8317, USA
[c] Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892-7244, USA

ARTICLE INFO

Article history:
Received 5 October 2011
Received in revised form
5 October 2012
Accepted 8 October 2012
Available online 17 October 2012

Keywords:
Bonferoni test
Simes test
Combined p-value test
Family-wise error rate
Fisher's product test
Rank truncated product (RTP) test
Truncated product method

ABSTRACT

Tests that combine p-values, such as Fisher's product test, are popular to test the global null hypothesis H_0 that each of n component null hypotheses, $H_1,\ldots,H_n$, is true versus the alternative that at least one of $H_1,\ldots,H_n$ is false, since they are more powerful than classical multiple tests such as the Bonferroni test and the Simes tests. Recent modifications of Fisher's product test, popular in the analysis of large scale genetic studies include the truncated product method (TPM) of Zaykin et al. (2002), the rank truncated product (RTP) test of Dudbridge and Koeleman (2003) and more recently, a permutation based test—the adaptive rank truncated product (ARTP) method of Yu et al. (2009). The TPM and RTP methods require users' specification of a truncation point. The ARTP method improves the performance of the RTP method by optimizing selection of the truncation point over a set of pre-specified candidate points. In this paper we extend the ARTP by proposing to use all the possible truncation points $\{1,\ldots,n\}$ as the candidate truncation points. Furthermore, we derive the theoretical probability distribution of the test statistic under the global null hypothesis H_0. Simulations are conducted to compare the performance of the proposed test with the Bonferroni test, the Simes test, the RTP test, and Fisher's product test. The simulation results show that the proposed test has higher power than the Bonferroni test and the Simes test, as well as the RTP method. It is also significantly more powerful than Fisher's product test when the number of truly false hypotheses is small relative to the total number of hypotheses, and has comparable power to Fisher's product test otherwise.

1. Introduction

Assume that we have a collection of hypotheses H_i, $i=1,\ldots,n$. We are interested in testing the combined or global null hypothesis

$$H_0=(H_1\cap\ldots\cap H_n) \tag{1.1}$$

The global Type I error rate and the global power of a testing procedure are defined as P (reject $H_0|H_0$ is true) and P (reject $H_0|H_0$ is false), respectively.

Denote the p-values corresponding to $H_1,\ldots,H_n$ by $p_1,\ldots,p_n$, and their ordered values by $p_{(1)}\le\ldots\le p_{(n)}$. The best known multiple test is the Bonferroni test. Fixing the Type I error rate at α, the Bonferroni test rejects H_0 if $p_i\le\alpha/n$ for any

* Corresponding author.
E-mail address: szhang3@unl.edu (S. Zhang).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.10.004

$i=1,\ldots,n$; *i.e.*, if $p_{(1)} \le \alpha/n$. The Bonferroni test is a conservative test in the sense that the true Type I error rate is always less than or equal to the nominal Type I error rate α. Šidàk (1967) provided an improved version of the Bonferroni test. The Šidàk test has exact control of the Type I error rate provided all the hypotheses are independent, and is more powerful than the Bonferroni test. Further improvements over the Bonferroni test include the well-known Holm test (Holm, 1979) and the Simes test (Simes, 1986), among others.

When the global null hypothesis H_0 is rejected by a test, it is often desirable to provide statements on individual hypotheses (i.e., which specific hypotheses should be rejected). For such tests, the Type I error rate P (reject $H_0|H_0$ is true) is the probability of falsely rejecting at least one true null hypothesis–the family–wise error rate (FWER). It is well known that the Bonferroni test and the Šidàk test control the FWER under the global null hypothesis H_0, i.e., they have "weak" control of the FWER. In addition, they control the FWER regardless of the composition of the hypotheses, and thus also have "strong" control of the FWER. However, the original Simes test only has weak control of the FWER; neither does it provide statements on individual hypotheses when the global null hypothesis is rejected. In order to provide individual statements, Simes (1986) proposed a modified step-up version of the Simes test. Hochberg (1988) proposed a step-up version of the Simes test (the Simes–Hochberg test) which has strong control of the FWER. Hommel (1988) proposed a strictly more powerful modification (neither step-down nor step-up) of the Simes test that also has strong control of the FWER. A comprehensive review of these methods can be found in Shaffer (1995). It is worth mentioning that the step-up Simes test leads to the false discovery rate (FDR) control procedure of Benjamini and Hochberg (1995).

Despite the popularity of multiple testing procedures which have strong control of the FWER, testing procedures with only weak control of the FWER remain powerful tools for exploratory analysis in which the researchers are more interested in finding whether a group of hypotheses as a whole is true or false rather than the individual hypotheses. For example, in genome-wide case-control association studies (GWAS), the association of individual single-nucleotide polymorphisms (SNPs) with a certain disease is tested. However, often investigators are interested whether particular genes or genetic regions are associated with disease, rather than individual SNPs. This has motivated the development of more powerful testing procedures for testing the global null hypothesis H_0 and also our current research. After briefly reviewing existing combined p-value tests in Section 2, we propose an extended version of the ARTP which uses all the possible truncation points $\{1,\ldots,n\}$ as the set of candidate truncation points in Section 3. The advantage of the proposed method compared to the ARTP method is that it does not require selection of the set of candidate truncation points. The theoretical probability distribution of the test statistic under the global null hypothesis is also derived in this section. Simulation results reported in Section 4 show that the proposed test has higher power than the classical multiple tests and several existing combined p-value tests in many situations. The conclusions and some discussion are provided in Section 5.

2. A brief review of the existing combined *p*-value methods for testing the global null hypothesis H_0

The combined p-value tests are based on the notion that several non-significant results together may suggest significance and hence detect departures from H_0. A popular combined p-value test for testing the global hypothesis H_0 is Fisher's product test (Fisher, 1932), based on

$$\prod_{i=1}^{n} p_i, \quad F = -2\sum_{i=1}^{n} \ln(p_i), \tag{2.1}$$

that follows a χ^2_{2n} distribution under H_0 and the assumption that the tests are independent.

Another popular combined p-value test for testing the global null hypothesis H_0 is the Simes test that rejects H_0 if $p_{(i)} \le i\alpha/n$, for any $i=1,\ldots,n$. It thus uses combined evidence through the order statistics. For $n=2$, the Simes test rejects $H_0=\{H_1 \cap H_2\}$ as long as $p_{(2)} \le 0.05$ while the Bonferroni test requires that $p_{(2)} \le 0.025$. Nevertheless, it is advantageous to combine the p-values through multiplication rather than using individual order statistics. By considering the rejection regions of the Simes test and Fisher's product test for $n=2$ it can easily be seen that Fisher's product test is in most situations much more relaxed in declaring pairs of p-values significant than the Simes test, and thus more powerful than the Simes test. For example, let $n=2$, $H_0=\{H_1 \cap H_2\}$ and assume $p_1=p_2=0.06$. Fisher's product test rejects H_0 since $F=11.25$ is larger than the critical value $\chi^2_4(1-0.05)=9.48$, while both the Bonferroni test and the Šidàk test fail to reject H_0. However, while Fisher's product test is generally more powerful than the Simes test, it has lower power for rejecting the global null hypothesis H_0 when only a small proportion of p-values is small. This restrictiveness becomes more extreme as n increases. For example, assume that $n=10$ and all but one p-value are close to 1. Then Fisher's product test requires the one significant p-value to be less than 1.51×10^{-7} to reject H_0. As the critical value of the Simes test for declaring significance of individual p-values for $n=10$ is 0.005 it is much more powerful than Fisher's product test in this case.

Other approaches to combining p-values include the sum of the p-values (Edgington, 1972), or sums of normal-transformed p-values (Stouffer et al., 1949). A systematic comparison of combining p-values methods for testing independent tests is provided in Loughin (2004). In this paper, we will only focus on the combined p-value method through the product. Several recent methods that combine p-values through their product are the TPM of Zaykin et al. (2002), the RTP test of Dudbridge and Koeleman (2003), and the ARTP method of Yu et al. (2009). The TPM uses the