Journal of Statistical Planning and Inference 142 (2012) 2059–2070

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Step-up and step-down methods for testing multiple hypotheses in sequential experiments

Shyamal K. De, Michael Baron *

Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080, United States

ARTICLE INFO

Article history:
Received 18 July 2011
Received in revised form
2 February 2012
Accepted 6 February 2012
Available online 14 February 2012

Keywords:
Bonferroni methods
Familywise error rate
Holm procedure
Sequential probability ratio test
Stopping boundaries
Wald approximation

ABSTRACT

Sequential methods are developed for testing multiple hypotheses, resulting in a statistical decision for each individual test and controlling the familywise error rate and the familywise power in the strong sense. Extending the ideas of step-up and step-down methods for multiple comparisons to sequential designs, the new techniques improve over the Bonferroni and closed testing methods proposed earlier by a substantial reduction of the expected sample size.

1. Introduction

1.1. Motivation

The problem of multiple inferences in sequential experiments arises in many fields. Typical applications are in sequential clinical trials with both efficacy and safety endpoints (Jennison and Turnbull, 1993) or several outcome measures of efficacy (O'Brien, 1984; Pocock et al., 1987), acceptance sampling with several different criteria of acceptance (Baillie, 1987; Hamilton and Lesperance, 1991), multichannel change-point detection (Tartakovsky and Veeravalli, 2004; Tartakovsky et al., 2003) and in microarray experiments (Dudoit et al., 2003). It is often necessary to find the statistical answer to each posed question by testing each individual hypothesis rather than giving one global answer by combining all the tests into one and testing a composite hypothesis.

Methods developed in this paper aim to test *multiple hypotheses* based on sequentially collected data, resulting in *individual decisions* for each individual test. They control the familywise error rate and the familywise power in the strong sense. That is, both probabilities of rejecting at least one true null hypothesis and accepting at least one false null hypothesis are kept within the chosen levels α and β under any set of true hypotheses. This condition is a multi-testing analogue of controlling both probabilities of Type I and Type II errors in sequential experiments. As a result, the *familywise power*, defined as the probability of detecting *all* significant differences at the specified alternative parameter values, is controlled at the level $(1-\beta)$ (see Shaffer, 1995, for three alternative definitions of familywise power).

* Corresponding author.
E-mail addresses: shyamal@utdallas.edu (S.K. De), mbaron@utdallas.edu (M. Baron).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2012.02.005

Under these conditions, proposed stopping rules and decision rules achieve substantial reduction of the expected sample size over all the existing (to the best of our knowledge) sequential multiple testing procedures.

1.2. Sequential multiple comparisons in the literature

The concept of multiple comparisons is not new in sequential analysis. Sequential methods exist for inferences about multivariate parameters (Ghosh et al., 1997, Sections 6.8 and 7.5). They are widely used in studies where inferences about individual parameters are not required.

Most of the research in sequential multiple testing is limited to two types of problems.

One type is the study of several ($k>2$) treatments comparing their effects. Sampled units are randomized to k groups where treatments are administered. Based on the observed responses, one typically tests a composite null hypothesis $H_0:\theta_1=\cdots=\theta_k$ against H_A: not H_0, where θ_j is the effect of treatment j for $j=1,\ldots,k$ (Betensky, 1996; Edwards, 1987; Edwards and Hsu, 1983; Hughes, 1993; Jennison and Turnbull, 2000, Chapter 16; O'Brien and Fleming, 1979; Siegmund, 1993; Wilcox, 2004; Zacks, 2009, Chapter 8). Sometimes each treatment is compared to the accepted standard (e.g., Paulson, 1962), and often the ultimate goal is selection of the best treatment (Jennison et al., 1982; Paulson, 1964).

The other type of studies involves a sequentially observed sequence of data that needs to be classified into one of the several available sets of models. In a parametric setting, a null hypothesis $H_0:\theta\in\Theta_0$ is tested against several alternatives, $H_1:\theta\in\Theta_1$ vs …vs $H_k:\theta\in\Theta_k$, where θ is the common parameter of the observed sequence (Armitage, 1950; Baum and Veeravalli, 1994; Novikov, 2009; Simons, 1967).

The optimal stopping rules for such tests are (naturally!) extensions of the classical Wald's sequential probability ratio tests (Govindarajulu, 2004; Wald, 1947; Wald and Wolfowitz, 1948; Siegmund, 1985). For the case of three alternative hypotheses, Sobel and Wald (1949) obtained a set of four stopping boundaries for the likelihood-ratio statistic. Their solution was generalized to a larger number of alternatives resulting in the *multi-hypothesis sequential probability ratio tests* (Dragalin et al., 1999; Lai, 2000).

1.3. Our goal—simultaneous testing of individual hypotheses

The focus of this paper is different and more general. We assume that the sequence of sampled units is observed to answer several questions about its parameters. Indeed, once the sampling cost is already spent on each sampled unit, it is natural to use it to answer more than just one question! Therefore, there are *d individual hypotheses* about parameters $\theta_1,\ldots,\theta_d$ of sequentially observed vectors $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots$

$$
\begin{aligned}
&H_0^{(1)}:\theta_1\in\Theta_{01} \text{ vs } H_A^{(1)}:\theta_1\in\Theta_{11},\\
&H_0^{(2)}:\theta_2\in\Theta_{02} \text{ vs } H_A^{(2)}:\theta_2\in\Theta_{12},\\
&\vdots\\
&H_0^{(d)}:\theta_k\in\Theta_{0d} \text{ vs } H_A^{(d)}:\theta_k\in\Theta_{1d}.
\end{aligned}
\tag{1}
$$

A few sequential procedures have been proposed for multiple tests similar to (1). One can conduct individual sequential tests of $H_0^{(1)},\ldots,H_0^{(d)}$ and stop after the first rejection or acceptance, as in Jennison and Turnbull (2000, Chapter 15). Hypotheses that are not rejected at this moment will be accepted, conservatively protecting the familywise Type I error rate (FWER-I).

Alternatively, one can assign level α_j and the corresponding Pocock or O'Brien–Fleming rejection boundary to the jth hypothesis. Then one conducts sequential or group sequential tests in a hierarchical manner, as proposed in Glimm et al. (2010), Tamhane et al. (2010), and Maurer et al. (2011) for testing primary, secondary, and possibly tertiary endpoints of a clinical trial. This procedure controls FWER-I at the level $\alpha=\sum\alpha_j$.

A different approach proposed in Tang and Geller (1999) and further developed in Bartroff and Lai (2010) allows to control FWER-I by testing *a closed set* of hypotheses. Along with the individual hypotheses $H_0^{(1)},\ldots,H_0^{(d)}$, this method requires to test all the composite hypotheses consisting of intersections $\cap H_0^{(j_k)}$, $1\le j_k\le d$, $1\le k\le d$. This results in mandatory testing of (2^d-1) instead of d hypotheses. As shown in Section 4, controlling the overall familywise Type I error rate will then require a rather large expected sample size.

While focusing on the Type I FWER, these procedures do not control the *familywise Type II error rate* and the familywise power. On the other hand, a Type II error, for example, on one of the tests of a safety clinical trial implies a failure to notice a side effect of a treatment, which is important to control.

Notice that sequential tests of single hypotheses are able to control probabilities of both the Type I and Type II errors. Extending this to multiple testing, our goal is to control *both familywise error rates* I and II and to do so at a *low sampling cost* by computing the optimal stopping boundaries and the optimal stopping rule followed by the optimal terminal decisions.