Computational Statistics and Data Analysis 68 (2013) 202–212

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Multiple comparisons of k binomial proportions[☆]

Kane Nashimoto[a,∗,1], Kristin M. Haldeman[b], Christopher M. Tait[c]

[a] Department of Mathematics and Statistics, James Madison University, 800 South Main Street, MSC 1911, Harrisonburg, VA 22807, United States

[b] Department of Mathematics and Statistics, California State University, Long Beach, CA 90840, United States

[c] Mathematics and Computer Science Department, Hampden-Sydney College, Hampden-Sydney, VA 23943, United States

ARTICLE INFO

Article history:
Received 18 May 2012
Received in revised form 2 July 2013
Accepted 2 July 2013
Available online 10 July 2013

Keywords:
Analysis of variance
Binomial proportions
Familywise error
Multiple comparisons
Simultaneous inference

ABSTRACT

Comparisons of k independent binomial proportions are studied. Piegorsch (1991) compared the Studentized-range implementation of the Wald interval and the Bonferroni-adjusted interval, both of which performed poorly for small values of the true proportions. Agresti et al. (2008) showed that adding one pseudo observation of each type in forming the Wald interval, along with the Studentized-range implementation, greatly improved the performance. A new two-stage method of multiple comparisons (global test followed by pairwise tests) is proposed. For the pairwise tests, three procedures are proposed, which are the LSD type, modified LSD, and inverse-sine based. Simulation studies show that the new procedures have relatively high power and that the modified LSD and inverse-sine based procedures maintain the familywise error rate near the nominal level.

1. Introduction

We consider comparisons of k independent binomial proportions. Numerous studies have examined comparisons of several means, but relatively a small number of studies have devoted attention to comparisons of proportions. In this section, we provide a brief review of the existing procedures and describe the objective of our study.

Perhaps the most basic type of statistical inference involving a proportion is forming a confidence interval for the binomial parameter p. The widely known Wald interval, which is obtained by inverting the two-sided Wald test of $p = p_0$, has the form $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p} = X/n$, $X \sim \mathrm{BIN}(n,\ p)$, and $z_{\alpha/2}$ is the standard-normal value with $\alpha/2$ upper-tail probability. Many studies have shown (e.g., Agresti and Coull, 1998; Brown et al., 2001) that the Wald interval has poor coverage probabilities when p is close to 0 or 1. The Score interval, which is the inversion of the two-sided Score test of $p = p_0$, performs more favorably. This interval was first discussed in Wilson (1927). It can be shown that the midpoint of the Score interval is

$$\frac{X}{n + z_{\alpha/2}^2} + \frac{z_{\alpha/2}^2}{2(n + z_{\alpha/2}^2)}.$$

☆ This research was supported in part by the NSF Grant DMS-05-52577.

∗ Corresponding author. Tel.: +1 540 568 2285; fax: +1 540 568 6857.
E-mail address: nashimkx@jmu.edu (K. Nashimoto).

1 Additional summary of results pertinent to this research at http://educ.jmu.edu/~nashimkx/mcbinom.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.07.002

For the 0.95 confidence level, $z_{0.05/2} = 1.96$, which is close to 2. Substituting 2 for $z_{\alpha/2}$ simplifies the midpoint of the interval to $(X+2)/(n+4)$. Agresti and Coull (1998) showed that using $p^* = (X+2)/(n+4)$ and $n^* = n+4$ in place of $\hat{p}$ and n for the Wald interval greatly improves the coverage probabilities. Incidentally, $p^* = (X+2)/(n+4)$ is the Bayes estimate of the binomial parameter p under the square-error loss when the BETA(2, 2) prior is used. Another well-known interval is the Clopper–Pearson interval (Clopper and Pearson, 1934). It is the inversion of the equal-tail binomial test of $p = p_0$, and its coverage probability is guaranteed to be at least $1-\alpha$. However, the Clopper–Pearson interval can be overly conservative (i.e., unnecessarily wide) when n is not large.

For comparisons of two independent proportions, perhaps the most commonly used is the Wald interval, which is of the form $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$, where $\hat{p}_i = X_i/n_i$ and $X_i \overset{\text{indep}}{\sim} \text{BIN}(n_i,\ p_i)$. Like its one-sample version, this interval shows poor coverage probabilities when p_1 and p_2 are both close to 0 or 1 (see Agresti and Caffo, 2000). The Score interval can be obtained by inverting the two-sided Score test of $p_1 - p_2 = 0$. The advantage of this interval is that the inference based on the interval is consistent with that based on the Score test conducted at the significance level α. Newcombe (1998) proposed an interval that is a combination of the previously mentioned single-sample Score intervals for p_1 and p_2. Letting l_i and u_i respectively denote the lower and upper limits of the single-sample Score interval for p_i, Newcombe's interval is of the form

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}},\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}}\right).$$

Newcombe showed that this interval has superb coverage probabilities for the 0.90 and 0.95 confidence levels. Agresti and Caffo (2000) proposed a modification of the Wald interval by adding one pseudo observation of each type (event/nonevent) to each sample. Thus, their interval is $(\tilde{p}_1 - \tilde{p}_2) \pm z_{\alpha/2}\sqrt{\tilde{p}_1(1-\tilde{p}_1)/\tilde{n}_1 + \tilde{p}_2(1-\tilde{p}_2)/\tilde{n}_2}$, where $\tilde{p}_i = (X_i+1)/(n_i+2)$ and $\tilde{n}_i = n_i + 2$. This interval is less conservative than Newcombe's when p_1 and p_2 are both close to 0 or 1.

Only a limited number of studies have investigated comparisons of k independent proportions. For comparisons of k means, the Tukey–Kramer method (Kramer, 1956) is popularly used to form simultaneous confidence intervals. The Tukey–Kramer method uses a Studentized-range q critical value as the multiplier of the estimated standard error, as opposed to a Student t critical value. Hence, a natural generalization of this method to comparisons of k proportions is to replace $z_{\alpha/2}$ for the intervals described in the preceding paragraph with $q_{\alpha,k,\infty}/\sqrt{2}$, which is the Studentized-range value with upper-tail probability α for k populations with infinite degrees of freedom (divided by $\sqrt{2}$). Piegorsch (1991) applied this method to the Wald interval and compared its performance with the Bonferroni-adjusted Wald interval. He noted that both intervals suffer undercoverage, with the Bonferroni-adjusted one being slightly better due to its width. Agresti et al. (2008) extended the "add one pseudo observation" idea to comparisons of k populations. The resulting simultaneous confidence intervals,

$$(\tilde{p}_i - \tilde{p}_j) \pm \frac{q_{\alpha,k,\infty}}{\sqrt{2}}\sqrt{\frac{\tilde{p}_i(1-\tilde{p}_i)}{\tilde{n}_i} + \frac{\tilde{p}_j(1-\tilde{p}_j)}{\tilde{n}_j}}, \quad 1 \le i \ne j \le k,$$

perform well for pairwise comparisons.

In this study, we extend the study by Agresti et al. (2008) and develop new procedures for comparing k binomial proportions. We evaluate the performances of the proposed procedures, along with some existing ones, within the framework of hypothesis testing, rather than constructing simultaneous confidence intervals. The reason for this consideration is that it will allow us to implement a two-stage method (global test followed by pairwise tests). It is well known that such a method, in general, is more sensitive to true differences, if existent. Hayter (1986) investigated the familywise error rate (FWER) of the least significant difference (LSD) test, which led to the Fisher–Hayter multiple-comparison method. The Fisher–Hayter method remains powerful while controlling the FWER properly. Nashimoto and Wright considered the use of a two-stage method for comparing k means (2005a, 2005b) and k medians (2007) under a simple order, incorporating the Fisher–Hayter method. In Section 2, we describe the proposed method. In Section 3, we conduct simulation studies to evaluate the performance characteristics of the proposed procedures, followed by an application in Section 4. Concluding remarks are given in Section 5.

2. Proposed method

Let $X_i \overset{\text{indep}}{\sim} \text{BIN}(n_i,\ p_i)$, $i = 1, 2, \ldots, k$, $0 < p_i < 1$, and $\hat{p}_i = X_i/n_i$. To compare the k binomial proportions, we propose a two-stage method, in which a global test of homogeneity is conducted first and, only if the test is significant, pairwise tests are conducted.

2.1. Global test

For testing $H_0: p_1 = p_2 = \cdots = p_k$ vs. $H_1: p_i \ne p_j$ for some $1 \le i \ne j \le k$, the likelihood ratio test (LRT) is given in Robertson et al. (1988). The LRT statistic involves $\hat{p}_i$, which are approximately normal by the central limit theorem, but their