Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 710–722
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

The sizes of the three popular asymptotic tests for testing homogeneity of two binomial proportions

Seung-Ho Kang[a,*], Yonghee Lee[a], EunSug Park[b]

[a]Department of Statistics, Ewha Womans University, 11-1, DaeHyun-Dong, SeoDaeMun-Gu, Seoul 120-750, Korea
[b]Texas Transportation Institute, The Texas A&M University System, USA

Received 7 July 2005; received in revised form 20 March 2006; accepted 20 March 2006
Available online 24 April 2006

Abstract

In statistical hypothesis testing it is important to ensure that the type I error rate is preserved under the nominal level. This paper addresses the sizes and the type I errors rates of the three popular asymptotic tests for testing homogeneity of two binomial proportions: the chi-square test with and without continuity correction, the likelihood ratio test. Although it has been recognized that, based on limited simulation studies, the sizes of the tests are inflated in small samples, it has been thought that the sizes are well preserved under the nominal level when the sample size is sufficiently large. But, Loh [1989. Bounds on the size of the χ^2 test of independence in a contingency table. Ann. Statist. 17, 1709–1722], and Loh and Yu [1993. Bounds on the size of the likelihood ratio test of independence in a contingency table. J. Multivariate Anal. 45, 291–304] showed theoretically that the sizes are always greater than or equal to the nominal level when the sample size is infinite. In this paper, we confirm their results by computing the large-sample lower bounds of the sizes numerically. Applying complete enumeration which does not have any error, we confirm again the results by computing the sizes precisely on computer in moderate sample sizes. When the sample sizes are unbalanced, the peaks of the type I error rates occur at the extremes of the nuisance parameter. But, the type I error rates of the three tests are close to the nominal level in most values of the nuisance parameter except the extremes. We also find that, when the sample sizes are severely unbalanced and the value of the nuisance parameter is very small, the size of the chi-square test with continuity correction can exceed the nominal level excessively (for instance, the size could be at least 0.877 at 5% nominal level in some cases).

Keywords: Chi-square test; Likelihood ratio test; Continuity correction; Type I error rate; Nuisance parameter; Contingency table

1. Introduction

The analysis of 2×2 contingency tables for testing homogeneity of two binomial proportions is a ubiquitous problem in statistics. Although there has been a lot of research and debates for analyzing 2×2 contingency tables, in this paper, we focus on the sizes and the type I error rates of the three popular asymptotic tests: the Pearson chi-square test without continuity correction, the Pearson chi-square test with continuity correction and the likelihood ratio test. The three asymptotic tests are chosen, because they have been widely used if the sample sizes are not too small. For example, the FREQ procedure in SAS computes the p-values of the three tests. There are some other tests for testing homogeneity of two binomial proportions. For example, Storer and Kim (1990) examined the size and power of four exact tests

* Corresponding author. Tel.: +82 2 3277 3337; fax: +82 2 3277 3607.
E-mail address: seungho@ewha.ac.kr (S.-H. Kang).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2006.03.006

(unconditional exact test, approximate unconditional test, Fisher's exact test and Liddell's exact test) and three versions of the chi-square test. The unconditional exact test and the Fisher's exact test guarantee that the size will never exceed a nominal level and, Storer and Kim (1990) showed that the size of the approximate unconditional test rarely exceeds the nominal level. Therefore, exact tests are not considered in this paper.

The three asymptotic tests are based on the assumption that all the expected frequencies (e_{ij}) are reasonably large. Fisher (1925) suggested that the chi-square test without continuity correction should be used only when $\min(e_{ij}) > 5$, but subsequent authors have gradually relaxed this condition. For example, Cochran (1954) suggested $\min(e_{ij}) > 1$. Typically, textbook authors indicate that a satisfactory approximation is achieved when expected frequencies are restricted to values of five (for example, Rosner, 2000). Since their validity in small samples is open to question, there has been many works to investigate the validity of these methods in small samples with simulations (for example, Berry and Mielke, 1988; Bradley et al., 1979; Kraemer and Woolson, 1987; Lawal and Upton, 1990; Tate and Hyer, 1973). It was found that the sizes (the supremum of the type I error rate over the nuisance parameter) exceed the nominal level in either small or unbalanced data. However, these results are obtained by simulations when the sample sizes are small. What will happen to the sizes if the sample sizes converge to infinity? Does the sizes converge to the nominal level? A well known fact is $\lim_{n\longrightarrow\infty} h(p,n)=\alpha$ for any fixed value of p, where $h(p,n)$ represents the type I error rate depending on both n(sample size) and the nuisance parameter p(common success rate). Therefore, it might be expected that the sizes converge to the nominal level as the sample sizes go to infinity. Actually, it is not true.

This question can be restated as follows. If we take the supremum of the limits of the type I error rates over the nuisance parameter space, then we have

$$\sup_{0<p<1} \lim_{n\longrightarrow\infty} h(p,n) = \sup_{0<p<1} \alpha = \alpha.$$

Is the same result obtained if the order of $\sup_{0<p<1}$ and $\lim_{n\longrightarrow\infty}$ is exchanged? Note that, although the value of p is unknown, the value of p is fixed in reality. The size is defined as the supremum of the type I error rate over the nuisance parameter space. That is, the size represents the type I error rate at the worst possible case and has been frequently used in statistics theory to evaluate performance of tests. In reality, since the sample size is finite, we might be interested in the size at some finite sample size n. That is,

$$\sup_{0<p<1} h(p,n) \quad \text{for finite } n.$$

It is quite natural to investigate the behavior of the size as the sample size increases, because the asymptotic theory is one of main research tools in modern statistics. Then, we are interested in

$$\lim_{n\longrightarrow\infty} \sup_{0<p<1} h(p,n).$$

The question is whether or not $\lim_{n\longrightarrow\infty} \sup_{0<p<1} h(p,n) = \alpha$. The answer is no. Loh (1989), and Loh and Yu (1993), in large samples, proved that the sizes of Pearson's chi-square test with (and without) continuity correction and the likelihood ratio test in testing homogeneity of two binomial proportions are always greater than or equal to the nominal value α. Furthermore, they obtained the lower bounds of the limiting sizes and found that in some cases the lower bounds numerically are strictly greater than the nominal level. For example, when the ratio of column margins is 1:99 and the minimally required expected cell frequency is zero and $\alpha = 0.05$, the lower bound is 0.104. It implies $\lim_{n\longrightarrow\infty} \sup_{0<p<1} h(p,n) \geqslant 0.104$ and the size is twice as large as the nominal level. Does it contradict $\lim_{n\longrightarrow\infty} h(p,n)=\alpha$ for any fixed value of p? The answer is no. The lower bounds of limiting sizes are obtained when $\lim_{n\longrightarrow\infty} np \longrightarrow w$ for a fixed value of w (Loh, 1989; Loh and Yu, 1993). Therefore, as the sample size increases, the value of p at which the supremum occurs are likely to converge to zero.

In this paper, we investigate the following issues.

1. The numerical lower bounds obtained by Loh (1989), and Loh and Yu (1993) were computed in some limited cases. We will calculate the numerical lower bounds extensively by focusing on 2×2 contingency tables for testing homogeneity of two binomial proportions.
2. With complete enumeration we will compute the sizes in moderate sample sizes and compare them with the lower bounds obtained in 1.
3. Finally, we will compare the three tests in terms of preserving the nominal level.