Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Statistical Planning and Inference 130 (2005) 359–375
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Confidence intervals for two sample binomial distribution

Lawrence Brown[a,*], Xuefeng Li[b]

[a]*Statistics Department, Wharton School, University of Pennsylvania, 3730 Walnut Street., Philadelphia, PA 19104-6340, USA*
[b]*Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, 1350 Piccard drive, Rockville, MD 20850, USA*

Received 22 August 2003; accepted 5 September 2003
Available online 29 July 2004

Abstract

This paper considers confidence intervals for the difference of two binomial proportions. Some currently used approaches are discussed. A new approach is proposed. Under several generally used criteria, these approaches are thoroughly compared. The widely used Wald confidence interval (CI) is far from satisfactory, while the Newcombe's CI, new recentered CI and score CI have very good performance. Recommendations for which approach is applicable under different situations are given.

Keywords: Confidence intervals; Binomial distribution; Two-sample problem; Wald interval; Newcombe's interval; Jeffrey's prior

1. Introduction

Suppose that X and Y are two independent random variables drawn from two different populations that both have binomial distributions. The first is of size m and has success probability p_1. The second is of size n and has success probability p_2. We are interested in comparing the difference of the success probability between these two populations.

We let $X \sim \text{binomial}(m, p_1)$ and $Y \sim \text{binomial}(n, p_2)$ and let $\Delta = p_1 - p_2$. We want to find the confidence interval with approximate level $1 - \alpha$ for Δ. When $\Delta = 0$, the related

* Corresponding author.
E-mail address: lbrown@wharton.upenn.edu (L. Brown).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2003.09.039

testing problem is equivalent to the classical problem of testing independence in a 2×2 contingency table.

Because of its wide application in practice many approaches have been provided for this problem. However, most of them are concentrated on testing the independence hypothesis. Pearson (1947) proposed the χ^2 goodness-of-fit test, which is still widely used today. To improve its performance, Yates (1934) and Pearson (1947) gave different corrections or modifications to the χ^2 test. Fisher (1935) proposed the exact test. It is well known but less convenient for large sample sizes. See also Boschloo (1970) and Haber (1986). The likelihood ratio test was discussed by Wilks (1935). Freeman and Tukey (1950), Cox (1953) and Gart (1966) gave some other test statistics that have approximately χ^2 distribution with one degree of freedom. Barnard (1947) and Liddell (1976) suggested some tests in the spirit of ordering the sample space. Goodman (1964) explicitly gave a test statistic for the hypothesis $H_0 : p_1 - p_2 = \Delta$. Some other tests, including Bayesian tests, are discussed in the literature. For instance, see Howard (1998), Tango (1998) and Agresti and Caffo (2000). Chernoff (2002) provides another interesting procedure on testing $p_1 = p_2$.

The reason we mention the above tests is because of the dual relationship between statistical tests and confidence sets. We can always obtain a confidence set for the parameter we are interested by inverting the family of tests . But we cannot always get a clear and convenient form for the confidence interval of Δ from these tests.

The well-known Wald confidence interval (CI) can be derived from Goodman's test, in which the standard errors are evaluated at the maximum likelihood estimates. Because of its simplicity and convenience, it has gained nearly universal application in practice and in textbooks. An alternative procedure is provided by the score test approach. This is based on inverting the test with the standard errors evaluated at the null hypothesis. Wilson (1927) gave the score CI for the proportion of one binomial population. For the difference of the proportions of two populations, we need an estimate of the standard error to avoid the nuisance parameter. A heuristic idea is to use the constrained maximum likelihood estimate of the standard error in the test. For simplicity we refer to the resulting CI as the score CI in this paper. It does not have an easily explained form. When $m = n$, Yule and Kendall (1950) obtained a CI by inverting the χ^2 test. For the case $m \neq n$ we propose a modified Yule's CI by changing the variance estimate. Newcombe (1998) gave a hybrid score interval by using information from the single score intervals for p_1 and p_2. Combining informative Bayesian estimates and the general procedure of inverting tests yields what we refer to as the Jeffrey's CI as a pseudo Bayesian CI. Inspired by the score test, Agresti and Caffo (2000) gave another pseudo Bayesian CI. Real Bayesian CIs are not fully explored here because of their computing difficulty. Finally, we propose a CI that is similar to the Wald CI but has a recentered coefficient. We call it the recentered CI.

After some exploration, we chose six representatives for comparison, Wald CI, Newcombe's CI, Jeffrey's CI, Agresti's CI, score CI and recentered CI. We compare their performance under some plausible criteria in order to give a broad picture for this problem. Recommendation is given for practical application.

We first give a brief summary of existing CIs in Section 2. Since the Bayesian CI is different from others, we consider it separately in Section 3. In Section 4, we propose the recentered CI. All the criteria used to compare the performance of these CIs are listed in Section 5. Section 6 gives empirical results for the comparisons and describes the various