Available online at www.sciencedirect.com
ScienceDirect
Journal of the Korean Statistical Society 37 (2008) 175–183
JOURNAL OF THE KOREAN STATISTICAL SOCIETY
www.elsevier.com/locate/jkss

New confidence intervals for the difference between two proportions in two-sample correlated binary data[☆]

Seung-Ho Kang

Department of Statistics, Ewha Womans University, Seoul 120-750, Republic of Korea

Received 1 May 2007; accepted 1 November 2007
Available online 21 April 2008

Abstract

Asymptotic confidence intervals for the difference between two proportions have been well developed in two-sample correlated binary data. But, the coverage probabilities of such asymptotic confidence intervals are much smaller than the nominal level in small samples, because the asymptotic confidence intervals rely on the large sample theory. The aim of this paper is to construct new confidence intervals whose performance is better than the existing confidence intervals in small samples. Assuming the beta-binomial model, we derive the Edgeworth expansion of the studentized test statistic. Then, we propose new confidence intervals by eliminating the skewness in the Edgeworth expansion. We conduct simulation studies to compare the new confidence intervals with the existing confidence intervals.

AMS 2000 Subject Classification: primary 62F25; secondary 62F05

Keywords: Beta-binomial distribution; Dentistry; Edgeworth expansion; Ophthalmology

1. Introduction

An important assumption that the outcomes are statistically independent is sometimes violated in biomedical areas. In this paper we are specially interested in correlated binary data (or clustered binary data). For example, in ophthalmologic studies observations from two eyes of the same patient are typically more similar than those from two eyes of different patients, and are thus correlated. We also often collect correlated binary data in periodontal studies. In such cases individual observations from each patient (called cluster) are correlated, even if those from different patients are independent.

In this paper we are interested in interval estimates for the difference between two probabilities of success in two-sample correlated binary data when the sample size is small. For example, Banting, Ellen, and Fillery (1985) conducted a longitudinal study of caries lesions on the exposed roots of teeth. Forty chronically ill subjects were followed for the development of root lesions over a one-year period. The data are given in Table 1.1. We would like to compute the 95% confidence interval for the difference of incidence of surfaces with root lesions between male and female.

☆ This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (KRF-2005-070-C00021).
E-mail address: seungho@ewha.ac.kr.

1226-3192/$ - see front matter
doi:10.1016/j.jkss.2007.11.002

Table 1.1
Longitudinal data of development of caries lesions over a one-year period

Male	0/4 1/1 2/2 0/2 2/4 0/3 0/3 0/3 0/1 1/2 0/2
Female	1/2 1/6 0/8 1/5 0/4 0/4 1/2 0/4 0/4 0/4 0/6 0/2 0/4 0/4 0/2 0/3 0/2 0/2 0/4 2/2 0/2 0/2 0/2 0/4 0/4 0/4 0/3 0/2 0/2

Number of development of root lesions/number of investigated teeth.

When the sample size is large, asymptotic confidence intervals have been well developed for this problem. For example, asymptotic confidence intervals based on the GEE estimator perform well when the sample size is large. However, the coverage probabilities of such methods are smaller than the nominal level in small samples. The details can be found in Section 4. In this paper we use the Edgeworth expansion to obtain more accurate asymptotic confidence intervals in small samples. Since we need the third and fourth moments to derive the Edgeworth expansion, we assume the beta-binomial distribution which is probably the most popular parametric model for correlated binary data (Williams, 1975).

The Edgeworth expansion has played an important role in modern statistics, but it has not been applied to discrete studentized test statistics until recently (Brown, Cai, & DasGupta, 2002; Zhou, Tsao, & Qin, 2004). One reason is that the distribution is a lattice, and the other reason is that the standard error in the denominator is replaced with its estimate. Zhou et al. (2004) obtained the Edgeworth expansion for the studentized difference between two binomial proportions in two-sample independent binary data. In this paper we employ the methods of Zhou et al. (2004) to construct new confidence intervals for the difference between two proportions in two-sample correlated binary data by assuming the beta-binomial distribution.

2. Notation and review

Suppose that we have independent clusters X_{ij} $(j = 1, \ldots, n_i, i = 1, 2)$ from a population with exposure $(i = 1)$ and a population with non-exposure $(i = 2)$ to a risk factor of interest, respectively. Each cluster is assumed to be a sum of m_{ij} units $(X_{ij} = \sum_{l=1}^{m_{ij}} X_{ijl}, j = 1, \ldots, n_i, i = 1, 2)$. We assume that $n = n_1 + n_2$ is large compared to m_{ij} so that asymptotic theory can be applied with respect to n. We define the random variable $X_{ijl} = 1$ with $P(X_{ijl} = 1) = \theta_i$ if the lth $(l = 1, \ldots, m_{ij})$ subject in the jth $(j = 1, \ldots, n_i)$ cluster from the ith $(i = 1, 2)$ population is a case, and $X_{ijl} = 0$ otherwise. For $l \neq l'$, X_{ijl} and $X_{ijl'}$ are correlated, because they belong to the same cluster. But, for $j \neq j'$, X_{ijl} and $X_{ij'l}$ are independent, because clusters are assumed to be independent. Note that X_{ij} denote the number of cases among m_{ij} units. We assume that X_{ij} follows the binomial distribution

$$P(X_{ij} = x_{ij}|\theta_i) = \frac{m_{ij}!}{x_{ij}!(m_{ij} - x_{ij})!}\theta_i^{x_{ij}}(1-\theta_i)^{m_{ij}-x_{ij}}, \quad x_{ij} = 0, 1, \ldots, m_{ij}$$

where a random variable θ_i follows the beta distribution with the parameters ζ_i and η_i. Then the marginal distribution of X_{ij} follows the beta-binomial

$$P(X_{ij} = x_{ij}) = \frac{m_{ij}!}{x_{ij}!(m_{ij} - x_{ij})!}\frac{B(x_{ij} + \zeta_i, m_{ij} - x_{ij} + \eta_i)}{B(\zeta_i, \eta_i)}, \quad x_{ij} = 0, 1, \ldots, m_{ij}$$

where $B(\zeta_1, \eta_1)$ is the beta function with the parameters ζ_1 and η_1. In this paper we use the following reparametrization. For $l \neq l'$

$$E(X_{ij}) = m_{ij}\zeta_i(\zeta_i + \eta_i)^{-1} \equiv m_{ij}p_i \qquad \mathrm{Corr}(X_{ijl}, X_{ijl'}) = (\zeta_i + \eta_i + 1)^{-1} \equiv \rho_i.$$

This reparametrization is useful, because the parameters $0 < p_i < 1$ and $\rho_i > 0$ are the mean and the intraclass correlation of X_{ijl}.

Let $X_{i1}, X_{i2}, \ldots, X_{in_i}$ be a random sample from a beta-binomial distribution with parameters p_i and ρ_i, $i = 1, 2$. The goal is to construct confidence intervals for $p_1 - p_2$. Let

$$\hat{p}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{M_{i1}}, \qquad M_{ij} = \sum_{l=1}^{n_i} m_{il}^j, \quad i = 1, 2.$$