Communications for Statistical Applications and Methods
2016, Vol. 23, No. 5, 433–444

http://dx.doi.org/10.5351/CSAM.2016.23.5.433
Print ISSN 2287-7843 / Online ISSN 2383-4757

Tests for homogeneity of proportions in clustered binomial data

Kwang Mo Jeong[1,a]

[a]Department of Statistics, Pusan National University, Korea

Abstract

When we observe binary responses in a cluster (such as rat lab-subjects), they are usually correlated to each other. In clustered binomial counts, the independence assumption is violated and we encounter an extra-variation. In the presence of extra-variation, the ordinary statistical analyses of binomial data are inappropriate to apply. In testing the homogeneity of proportions between several treatment groups, the classical Pearson chi-squared test has a severe flaw in the control of Type I error rates. We focus on modifying the chi-squared statistic by incorporating variance inflation factors. We suggest a method to adjust data in terms of dispersion estimate based on a quasi-likelihood model. We explain the testing procedure via an illustrative example as well as compare the performance of a modified chi-squared test with competitive statistics through a Monte Carlo study.

Keywords: clustered binomial data, quasi-likelihood, homogeneity of proportions, Pearson chi-squared statistic, intra-cluster correlation, variance inflation, likelihood ratio test

1. Introduction

Elementary units in a cluster often behave more similar alike than units from different clusters. There exists a correlation between observations when they belongs to the same cluster. For example, binary responses of dead (or not) female rat among foetuses in a female rat are correlated to each other. This fact violates the independence of binomial assumption that plays a crucial role in statistical methods. If we ignore this dependence structure, parameter estimates might be significantly underestimated, and thus we may have misleading results. In the presence of intra-cluster correlation (ICC) we expect a larger variance than that of binomial counts, the so called (overdispersion). Dichotomous outcomes are very common in cluster randomized trials. Many authors studied various statistical analyses for clustered binomial data and we may refer to Williams (1982), Paul (1982), Donner (1989), and Reed (2004). If we consider I groups of treatments consisting of clustered binomial data it can be tabulated by $I \times 2$ table in which columns represent two response categories of binomial counts (success or failure). Sometimes we routinely apply a classical Pearson chi-squared statistic even when it violates the independence assumption in binomial counts. In this case, Type I error of a Pearson chi-squared test is severely inflated and therefore it may lead to erroneous results.

There have been many studies on the various ways to modifying Pearson chi-squared statistics. The idea main point is to incorporate variance inflation into the chi-squared statistics by adjusting observed data. Among others, Donner (1989) suggested a method to modify the Pearson chi-squared statistic by ICC. The beta-binomial (BB) distribution is a reasonable one to model overdispersed

[1] Department of Statistics, Pusan National University, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea. E-mail: kmjeong@pusan.ac.kr

Published 30 September 2016 / journal homepage: http://csam.or.kr

binomial data, and the choice of ICC originates from a variance of BB that can be represented by ICC. On the other hand, Rao and Scott (1992) alternatively used the term of design effect that is obtained as the ratio of variances between clustered sample versus simple random binomial counts. This value is a kind of variance inflation factor (VIF) that is obtained by the asymptotic variance of a ratio estimator based on clustered binomial data. VIF is a measure for variance inflation of sample proportions under clustered binomial data against simple random binomial data. As an alternative to modifying chi-square statistics, Jeong (2015) suggested a testing procedure based on generalized linear models (that included random effects for clusters) called a generalized linear mixed model (GLMM). The likelihood ratio test (LRT) or Wald test are competitive ones to the chi-squared tests mentioned before. We may refer to Jeong and Lee (2013) for researches on clustered binomial data we may refer to Jeong and Lee (2013).

In Section 2, we briefly review approaches of modeling clustered binomial data that is commonly obtained from cluster randomized trials. The Quasi-likelihood model and the BB model are popular to analyze overdispersed binomial data. As measures of heterogeneity between clusters, we explain the terms ICC and VIF, that are used to modify chi-squared statistics. In Section 3, we propose a modified chi-squared statistic that can be applied to adjusted values by dispersion estimates. We also provide an illustrative example to explain the testing procedure. In Section 4, we compare the performance of competitive statistics through a Monte Carlo study to control nominal levels and higher powers. We summarize the results with concluding remarks in Section 5.

2. Extra-variation in clustered binomial data

2.1. Homogeneity of proportions in binomial data

We introduce some notations to formulate a testing procedure for the homogeneity of proportions between several treatment groups. Let π_i denote a proportion of certain trait in group i, where $i = 1, 2, \ldots, I$, and I is the number of groups. The null hypothesis of homogeneity for the proportions between I groups can be expressed as

$$H_0 : \pi_1 = \pi_2 = \cdots = \pi_I \tag{2.1}$$

We now synonymously and interchangeably use the word subject and cluster as a synonym. Suppose that a subject j in group i consists of m_{ij} elementary units, each having a certain trait of interest, for example, diseased or not. Theses binary outcomes are correlated to each other because they are observed within the same subject. Let the binary responses take values 1 or 0 with probability π_i or $1 - \pi_i$, respectively. Let y_{ij} be the sum of binary responses having 1 for the units belonging to subject i. If we assume the independence between binary observations among m_{ij} units the y_{ij} follows a binomial distribution with probability π_i, denoted as $\text{Bin}(m_{ij}, \pi_i)$. If we let n_i be the number of subjects in group i, then $n = \sum_{i=1}^{I} n_i$ denotes the total number of subjects in a dataset. To simplify notations we denote $y_i = \sum_{j=1}^{n_i} y_{ij}$ and $m_i = \sum_{j=1}^{n_i} m_{ij}$. We introduce an illustrative example to clarify notations.

Example 1. Table 1 shows the seed counts that are germinated for the seed *Orobanche cernua*, cultivated in three kinds of dilutions 1/1, 1/25 and 1/625. The dataset had already been analyzed by Crowder (1978) using BB distribution. Columns of y_{ij} denote germinated counts among m_{ij} seeds in dilution group i.

Each dilution group contains various numbers of clusters; $n_1 = 6$, $n_2 = 5$, $n_3 = 5$. At the bottom line of Table 1, we also see that $m_1 = 240$, $y_1 = 34$, and so on. The binomial counts of Table 1 can

Table 1: Number of *Orobanche cernua* seeds germinated according to three dilution types

Dilution 1/1		Dilution 1/25		Dilution 1/625	
m_{1j}	y_{1j}	m_{2j}	y_{2j}	m_{3j}	y_{3j}
43	2	19	17	13	11
51	9	56	43	62	47
44	5	87	79	104	90
71	16	55	50	51	46
24	2	10	9	11	9
7	0				
240	34	227	198	241	203

Table 2: Cell counts aggregated according to dilution types and events

Dilution(i)	y_i	$m_i - y_i$	Total (m_i)
1/1 ($i = 1$)	34	206	240
1/25 ($i = 2$)	198	29	227
1/625 ($i = 3$)	203	38	241

be summarized as a 3×2 table (Table 2), where two columns denote total counts of seeds that are germinated versus non-germinated.

For testing H_0 we ordinarily use the classical Pearson chi-squared statistic applied to Table 2, that is defined by

$$X_P^2 = \sum_{i=1}^{I} \frac{(y_i - m_i\hat{\pi})^2}{m_i\hat{\pi}(1-\hat{\pi})}, \tag{2.2}$$

where $\hat{\pi} = \sum_i y_i / \sum_i m_i$ denotes a common estimator of π_i's under H_0. In applying the asymptotic theory of X_P^2 we implicitly assume the independence of binary responses that summed to y_{ij} within each cluster. However, Pearson test may be misleading because the dataset of Table 2 violates the binomial assumption. We note that the Pearson chi-squared test of (2.2) shows $X_P^2 = 342.94$ with a p-value of $3.40 * 10^{-75}$ that is extremely significant.

2.2. Modeling variance inflation

In this section, we introduce the quasi-likelihood model for binomial counts to represents extra-variation by a dispersion parameter. Let Y have a probability density $f(y;\theta)$ in exponential class given by

$$f(y;\theta) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(\phi, y)\right]. \tag{2.3}$$

From the form of (2.3) we routinely obtain $E(Y) = b'(\theta)$, and the following mean-variance relationship

$$\mathrm{Var}(Y) = b''(\theta)a(\phi). \tag{2.4}$$

If Y follows $\mathrm{Bin}(m, \pi)$ the density can be written in the form

$$\begin{aligned} f(y;\theta) &= \exp\left[y\log\left(\frac{\pi}{1-\pi}\right) + m\log(1-\pi) + \log\binom{m}{y}\right] \\ &= \exp\left[\frac{\frac{y}{m}\theta - \log\left(1+e^{\theta}\right)}{\frac{1}{m}} + \log\binom{m}{y}\right], \end{aligned} \tag{2.5}$$

where $\theta = \log(\pi/(1-\pi))$, and $\log(1+e^{\theta})$ corresponds to $b(\theta)$ in (2.3). In particular, $a(\phi) = 1/m$ does not depend an extra parameter ϕ, which is a main cause of overdispersion in binomial counts. We note that $E(Y/m) = \pi$, $\mathrm{Var}(Y/m) = \pi(1-\pi)/m$. In the presence of extra-variation that frequently occurs in clustered binomial data $\mathrm{Var}(Y/m)$ is larger than $\pi(1-\pi)/m$ that is the variance under binomial distribution. The derivative of $\log f(y;\theta)$ with respect to π is obtained from (2.5) using the chain rule and (2.4)

$$\begin{aligned}\frac{\partial \log f}{\partial \pi} &= \frac{\partial \log f}{\partial \theta}\frac{\partial \theta}{\partial \pi} \\ &= \frac{(y/m-\pi)}{1/m}\frac{1/m}{\mathrm{Var}(Y/m)}.\end{aligned} \tag{2.6}$$

We may generalize the variance function by allowing a dispersion parameter ϕ to overcome the overdispersion problem because (2.6) depends only on the distribution of Y through $E(Y/m) = \pi$ and $\mathrm{Var}(Y/m)$. Following the quasi-likelihood approach that was introduced by Wedderburn (1974), we take $v(\pi) = \phi\pi(1-\pi)/m$ as a variance function instead of $\mathrm{Var}(Y/m)$ in (2.6).

Given clustered binomial counts y_{ij} among m_{ij} units we discuss on the quasi-likelihood estimation. If we let L_i be the log-likelihood of n_i observations in group i the quasi-likelihood of (2.4) can generally be expressed as

$$\frac{\partial L_i}{\partial \pi_i} = \sum_{j}^{n_i} \frac{(y_{ij}/m_{ij} - \pi_i)}{\phi\pi_i(1-\pi_i)/m_{ij}}. \tag{2.7}$$

From (2.7) the quasi-MLE of π_i is given by $\hat{\pi}_i = y_i/m_i$, the same form under binomial distribution. But the asymptotic variance under quasi-likelihood, denoted by $\mathrm{var}_Q(\hat{\pi}_i)$, is

$$\mathrm{var}_Q(\hat{\pi}_i) = \frac{\hat{\phi}_i\hat{\pi}_i(1-\hat{\pi}_i)}{m_i} \tag{2.8}$$

the form that has an inflation factor ϕ_i multiplied to the $\pi_i(1-\pi_i)/m_i$ when $\hat{\phi}_i > 1$. Furthermore, the overdispersion parameter ϕ_i is routinely estimated as

$$\hat{\phi}_i = \frac{1}{n_i - 1}\sum_{j}^{n_i}\frac{\left(y_{ij} - m_{ij}\hat{\pi}_i\right)^2}{m_{ij}\hat{\pi}_i\left(1-\hat{\pi}_i\right)},$$

the value of Pearson chi-squared statistic for $n_i \times 2$ table divided by $n_i - 1$.

The BB distribution is an alternative one that models clustered binomial data. From Section 14.3 of Agresti (2013) we briefly introduce BB distribution in a hierarchical viewpoint; given p_i, we assume that y_{ij} follows $\mathrm{Bin}(m_{ij}, p_i)$, and p_i follows a beta distribution with two parameters a_i and b_i satisfying $a_i = c\pi_i$, and $b_i = c(1-\pi_i)$ for $c > 0$ and π_i. For simplification we assume a common value c across all groups. From the beta distribution we have $E(p_i) = \pi_i$ and $\mathrm{Var}(p_i) = \rho^2\pi_i(1-\pi_i)$ with $\rho^2 = 1/(c+1)$. If we integrate out p_i the marginal distribution of y_{ij} leads to BB distribution denoted as $\mathrm{BB}(m_{ij}, \pi_i; \rho)$. The marginal mean of y_{ij} can be computed as

$$E(y_{ij}) = E[E(y_{ij}|p_i)] = E(m_{ij}p_i) = m_{ij}\pi_i.$$

Similarly, the marginal variance of y_{ij} is given by

$$\mathrm{Var}(y_{ij}) = m_{ij}\pi_i(1-\pi_i)\left[1 + \rho^2(m_{ij}-1)\right]. \tag{2.9}$$

We note that $0 < \rho^2 < 1$, and $\mathrm{Var}(y_{ij})$ becomes larger as ρ^2 increases, as we seen in (2.9).

However, the variance inflation may be explained in terms of a design effect that is popularly used in sampling theory. A design effect of cluster sampling against simple random sampling is defined as the ratio of variances of estimators under each sampling design. We let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{m}_i = \sum_{j=1}^{n_i} m_{ij}/n_i$ be the group means of responses y_{ij} and cluster sizes m_{ij}, respectively. The estimate $\hat{\pi}_i = y_i/m_i$ of a group proportion can be written as a type of ratio estimator under cluster sampling

$$\hat{\pi}_i = \frac{y_i}{m_i} = \frac{\bar{y}_i}{\bar{m}_i}.$$

Following Cochran (1977) we obtain the asymptotic variance $\mathrm{var}_C(\hat{\pi}_i)$ of ratio estimator $\bar{y}_i/\bar{m}_i$, when n_i is large

$$\mathrm{var}_C(\hat{\pi}_i) = n_i(n_i - 1)^{-1} m_i^{-2} \sum_{j=1}^{n_i} (y_{ij} - m_{ij}\hat{\pi}_i)^2.$$

Rao and Scott (1992) tried a method adjusting observed data to modify Pearson chi-squared statistic in the clustered data. Design effect, denoted by f_i, is the ratio of $\mathrm{var}_C(\hat{\pi}_i)$ to the binomial variance $\hat{\pi}_i(1 - \hat{\pi}_i)/m_i$ given by

$$f_i = \frac{\mathrm{var}_C(\hat{\pi}_i) m_i}{\hat{\pi}_i(1 - \hat{\pi}_i)} \tag{2.10}$$

Design effect means an amount of variance inflation of $\hat{\pi}_i$ under cluster sampling against simple random sampling.

3. Test statistics for homogeneity of proportions

3.1. Modified chi-squared statistic using quasi-likelihood

Classical Pearson chi-squared statistic has a tendency to inflate the Type I error severely for testing homogeneity of proportions in clustered binomial data. Many researchers have tried to adjust the Pearson chi-squared statistic by the VIF and ICC. In the quasi-likelihood model for binomial data the dispersion parameter ϕ_i plays the role of VIF compared to the binomial variance. We adjust a dataset in terms of ϕ_i to incorporate the inflated variance as follows. Denote the effective sample values as $\tilde{y}_i = y_i/\hat{\phi}_i$ and $\tilde{m}_i = m_i/\hat{\phi}_i$ that are adjusted by $\hat{\phi}_i$, and let $\tilde{\pi}_i = \tilde{y}_i/\tilde{m}_i$. Then we obtain the relationship between $\tilde{\pi}_i$ and $\hat{\pi}_i$ given by

$$\frac{\tilde{\pi}_i - \pi_i}{\sqrt{\tilde{\pi}_i(1 - \tilde{\pi}_i)/\tilde{m}_i}} = \frac{\hat{\pi}_i - \pi_i}{\sqrt{\mathrm{var}_Q(\hat{\pi}_i)}}, \tag{3.1}$$

where $\mathrm{var}_Q(\hat{\pi}_i)$ is given in (2.8). According to the asymptotic properties of MLE under distributional misspecification studied, among others, by White (1982) we have the asymptotic normality of (3.1). This fact shows that standard statistical test under binomial assumption can simply be applied to the adjusted values $\tilde{y}_i$, $\tilde{m}_i$, and it results in the same asymptotic distribution. We propose a modified chi-squared statistic applied to $\tilde{y}_i$, $\tilde{m}_i$ by

$$X_Q^2 = \sum_{i=1}^{I} \frac{(\tilde{y}_i - \tilde{m}_i\tilde{\pi})^2}{\tilde{m}_i\tilde{\pi}(1 - \tilde{\pi})}, \tag{3.2}$$

where $\tilde{\pi} = \sum_{i=1}^{I} \tilde{y}_i / \sum_{i=1}^{I} \tilde{m}_i$. In particular, if we use a common dispersion estimate $\hat{\phi}$, then $\tilde{\pi} = \hat{\pi}$ and X_Q^2 is simplified to

$$X_Q^2 = \frac{X_P^2}{\hat{\phi}},$$

where X_P^2 is the Pearson chi-squared statistic of (2.2), and $\hat{\phi}$ is an estimate using whole dataset across all groups. The asymptotic distribution of X_Q^2 is chi-squared with degrees of freedom (df) $I - 1$. In a later section of a simulation study we check the robustness of (3.2) when we use a common estimate $\hat{\phi}$ in the presence of moderate heterogeneity for ICC values between groups.

3.2. Other test statistics

Using adjusted values in chi-squared statistic seems to originate from Rao and Scott (1992). The Rao-Scott chi-squared statistic X_{RS}^2 by Rao and Scott (1992) is the Pearson chi-squared statistic applied to the adjusted values y_i/f_i, m_i/f_i by design effect f_i in (2.10). The Rao-Scott statistic X_{RS}^2 takes the form of (3.2), except that the adjusted values are computed differently from those of the quasi-likelihood approach. The asymptotic distribution of X_{RS}^2 is also asymptotically a chi-square with df $= I - 1$.

Donner (1989) alternatively suggested another kind of chi-squared statistic using ICC. We briefly explain the estimation of ICC according to Ridout *et al.* (1999), who suggested an ICC estimator and that is based on the analysis of variance (ANOVA). Let $ms_b^{(i)}$ and $ms_w^{(i)}$ be the mean squared error (MSE) for the between groups and within groups, respectively. The ANOVA type estimator of ICC for group i is defined by

$$\mathrm{ICC}_i = \frac{ms_b^{(i)} - ms_w^{(i)}}{ms_b^{(i)} + \left(n_0^{(i)} - 1\right) ms_w^{(i)}},$$

where $n_0^{(i)} = (n_i - 1)^{-1}\{m_i - \sum_{j=1}^{n_i} m_{ij}^2/m_i\}$. To modify Pearson chi-squared statistic Donner (1989) used a VIF of the form

$$d_i = 1 + \left(\sum_{j=1} \frac{m_{ij}^2}{m_i} - 1\right)\rho^2,$$

where ρ^2 denotes an ICC value. When we assume identical ICC for all groups the d_i has the form of variance inflation in (2.9) under BB model. The modified chi-squared statistic by Donner (1989) is

$$X_D^2 = \sum_{i=1}^{I} \frac{(y_i - m_i\hat{\pi})^2}{\hat{d}_i m_i \hat{\pi}(1 - \hat{\pi})}, \tag{3.3}$$

where $\hat{d}_i$ denotes an estimate of d_i. We are to estimate $\hat{\rho}^2$ by an ANOVA type ICC based on the whole dataset. The null distribution of X_D^2 in (3.3) is an asymptotically chi-square with df $= I - 1$ only when d_i's are common or m_{ij}'s are the same as commented by Rao and Scott (1992).

Finally, we introduce the LRT for treatment effects based on GLMM. For the general theory of GLMM we may refer to Agresti (2013). The group effects can be represented by a categorical variable G taking i for group i, where $i = 1, \ldots, I$. By transforming G into indicator variables $x_1, \ldots, x_{I-1}$ for the first $I - 1$ groups we can formulate a logistic-normal GLMM with random intercept u_{ij} as

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha + \beta_1^G x_{1i} + \cdots + \beta_{I-1}^G x_{I-1,i} + \sigma u_{ij}, \tag{3.4}$$

where u_{ij} follows a normal distribution $N(0,1)$, and β_i^G denotes group i effect. The variance estimate $\hat{\sigma}^2$ provides a summary measure of heterogeneity between subjects. Testing $H_0 : \pi_1 = \cdots = \pi_I$ in (2.1) corresponds to the null model with $\beta_1^G = \cdots = \beta_I^G = 0$. If we simply denote the maximized log-likelihood under null and the model of (3.4) by L_0 and L_1, respectively, the likelihood ratio statistic can be expressed as

$$G^2 = -2(\log L_0 - \log L_1).$$

We may easily implement the LRT because statistical packages (such as R and SAS) provide log-likelihood values under the models. Furthermore, G^2 can be expressed as the difference between deviances under null and alternative models, that is,

$$G^2 = D_0 - D_1,$$

where D_0 and D_1 denote deviances under null and the assumed models, respectively. The null distribution of G^2 is asymptotically chi-square with df $= I - 1$.

Example 2. (Example 1 continued) We explain the chi-squared statistics discussed before through the data of Example 1. We find design effects for three groups as $f_1 = 2.45$, $f_2 = 2.38$, and $f_3 = 1.65$, respectively. The common ICC estimate and dispersion estimate are $\hat{\rho}^2 = 0.51$ and $\hat{\phi} = 23.62$, respectively. The chi-squared statistics and p-values (in parentheses) are: proposed statistic $X_Q^2 = 14.52$ ($p = 0.0007$), Donner statistic $X_D^2 = 12.0$ ($p = 0.0025$), and Rao-Scott statistic $X_{RS}^2 = 155.14$ ($p = 2.05 * 10^{-34}$), LRT statistic $G^2 = 40.23$ ($p = 1.84 * 10^{-9}$). However in, as we have computed in Example 1 of Section 2 the Pearson statistic shows $X_P^2 = 342.94$ ($p = 3.40 * 10^{-75}$). The proposed test and the Donner test show similar results but the Pearson test and the Rao-Scott test show very different results. The LRT locates between two extremes. Small group sizes as well as a relatively large ICC value may be the causes of the fact that X_P^2 and X_{RS}^2 work very differently from the others.

4. A Monte Carlo study

4.1. Design of experiments

In this section, a small scale simulation study has been designed to check the performance of test statistics: the proposed X_Q^2, Rao-Scott X_{RS}^2, Donner X_D^2, Pearson X_P^2, and LRT. We consider $I = 2$ and $I = 3$ with group sizes $(n_1, n_2) = (10, 10)$ and $(20, 20)$ for $I = 2$, and $(n_1, n_2, n_3) = (10, 10, 10)$ and $(15, 20, 20)$ for $I = 3$. A clustered binomial observation y_{ij} among m_{ij} trials is generated from BB distribution BB$(m_{ij}, \pi_i; \rho_i)$ introduced in Section 2.2. We assign π_i as $\pi_1 = 0.3$ and $\pi_2 = 0.3 + \delta$ for $I = 2$, where δ takes 0.0, 0.1, 0.2, and 0.3. Similarly, when $I = 3$ we let $\pi_1 = \pi_2 = 0.3$ and $\pi_3 = 0.3 + \delta$ with the same values of δ as previously defined. The value $\delta = 0.0$ corresponds to the null hypothesis and δ represents a measure of discrepancy between groups. We take $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_I)$ values as $\boldsymbol{\rho} = (0.3, 0.3)$, $(0.2, 0.4)$, and $(0.3, 0.5)$ when $I = 2$, and for $I = 3$ we take $\boldsymbol{\rho} = (0.3, 0.3, 0.3)$, $(0.3, 0.3, 0.4)$, and $(0.3, 0.4, 0.5)$. Given a dataset the cluster sizes m_{ij}'s are constants that may vary according to subjects. To imitate a real dataset we generate pseudo random counts from the Poisson distribution Poi(m) with mean $m = 7$ or 10. The generated counts may equal zeroes; therefore, we assign m_{ij} as the value of generated count plus one. The whole procedure of computation and model fits has been implemented via R. The library glmmML in R plays the role of fitting logistic-normal GLMM. The empirical proportions of rejecting H_0 among 1,000 replications are listed in Tables 3–6 according to the cases of designed experiments.

Table 3: Empirical sizes and powers of test statistics according to $\boldsymbol{\rho} = (\rho_1, \rho_2)$ when $I = 2$, $m_1 = m_2 = 10$, $n_1 = n_2 = 10$, and $\pi_1 = 0.3$, $\pi_2 = 0.3 + \delta$

$\boldsymbol{\rho}$	α	Test	δ			
			0.0	0.1	0.2	0.3
(0.3, 0.3)	0.05	Proposed	0.057	0.180	0.553	0.873
		Rao-Scott	0.068	0.224	0.598	0.892
		Donner	0.058	0.169	0.540	0.864
		Pearson	0.166	0.395	0.796	0.966
		LRT	0.074	0.215	0.608	0.897
	0.10	Proposed	0.104	0.302	0.693	0.934
		Rao-Scott	0.111	0.313	0.714	0.936
		Donner	0.101	0.290	0.679	0.932
		Pearson	0.246	0.487	0.849	0.982
		LRT	0.120	0.325	0.726	0.944
(0.2, 0.4)	0.05	Proposed	0.059	0.171	0.495	0.834
		Rao-Scott	0.083	0.210	0.546	0.865
		Donner	0.054	0.160	0.473	0.828
		Pearson	0.180	0.417	0.766	0.948
		LRT	0.079	0.203	0.539	0.848
	0.10	Proposed	0.124	0.273	0.636	0.906
		Rao-Scott	0.139	0.317	0.672	0.914
		Donner	0.120	0.261	0.623	0.903
		Pearson	0.270	0.513	0.828	0.968
		LRT	0.148	0.300	0.650	0.912
(0.3, 0.5)	0.05	Proposed	0.049	0.150	0.397	0.736
		Rao-Scott	0.069	0.194	0.455	0.762
		Donner	0.042	0.140	0.381	0.720
		Pearson	0.252	0.449	0.741	0.923
		LRT	0.065	0.159	0.428	0.760
	0.10	Proposed	0.110	0.238	0.555	0.828
		Rao-Scott	0.121	0.281	0.582	0.839
		Donner	0.101	0.232	0.542	0.819
		Pearson	0.338	0.514	0.806	0.950
		LRT	0.133	0.246	0.563	0.843

LRT = likelihood ratio test.

4.2. Simulation results

Table 3 shows the results when group sizes are relatively small as $n_1 = n_2 = 10$. We note that the column of $\delta = 0.0$ represents empirical test sizes. The proposed test shows slightly higher sizes than nominal levels 0.05 and 0.10 when two values of ρ_i's are different. The test by Donner also shows a similar pattern. Rao-Scott test and LRT are slightly more liberal in controlling nominal levels; however, a Pearson test does not control nominal levels at all. When groups have moderate sizes of $n_1 = n_2 = 20$ (Table 4), the proposed test has good performance in the control of levels and attaining high powers except the case $\rho = (0.3, 0.5)$. A Donner test also appears to be competitive with the proposed one, but has a slightly lower power than the proposed test.

Table 5 shows the results when group sizes are as small as $n_1 = n_2 = n_3 = 10$. The proposed test has higher empirical sizes than nominal levels when ρ_i's are different between three groups as $\boldsymbol{\rho} = (0.3, 0.4, 0.5)$. Donner test also show a similar pattern, and the other tests are poor for the control of nominal levels in almost cases of $\boldsymbol{\rho}$ values. We note that the Rao-Scott test has as large an empirical sizes as 0.111 at $\alpha = 0.05$ when $\boldsymbol{\rho} = (0.3, 0.4, 0.5)$. In particular, a Pearson test has an empirical size as large as 0.309 at $\alpha = 0.05$. It warns us that the classical Pearson chi-squared statistic should not

Table 4: Empirical sizes and powers of test statistics according to $\boldsymbol{\rho} = (\rho_1, \rho_2)$ when $I = 2$, $m_1 = m_2 = 10$, $n_1 = n_2 = 20$, and $\pi_1 = 0.3$, $\pi_2 = 0.3 + \delta$

$\boldsymbol{\rho}$	α	Test	δ			
			0.0	0.1	0.2	0.3
(0.3, 0.3)	0.05	Proposed	0.044	0.322	0.856	0.998
		Rao-Scott	0.055	0.332	0.858	0.999
		Donner	0.041	0.312	0.844	0.998
		Pearson	0.148	0.542	0.950	0.999
		LRT	0.052	0.344	0.874	0.999
	0.10	Proposed	0.099	0.459	0.918	0.999
		Rao-Scott	0.104	0.462	0.917	0.999
		Donner	0.095	0.448	0.911	0.999
		Pearson	0.242	0.624	0.973	0.999
		LRT	0.095	0.470	0.926	0.999
(0.2, 0.4)	0.05	Proposed	0.049	0.336	0.843	0.993
		Rao-Scott	0.050	0.354	0.865	0.993
		Donner	0.044	0.323	0.832	0.989
		Pearson	0.157	0.594	0.943	0.999
		LRT	0.055	0.331	0.850	0.992
	0.10	Proposed	0.098	0.462	0.913	0.997
		Rao-Scott	0.103	0.488	0.916	0.997
		Donner	0.094	0.450	0.908	0.997
		Pearson	0.235	0.687	0.959	1.000
		LRT	0.105	0.449	0.910	0.997
(0.3, 0.5)	0.05	Proposed	0.060	0.261	0.713	0.964
		Rao-Scott	0.060	0.277	0.723	0.967
		Donner	0.053	0.241	0.696	0.961
		Pearson	0.241	0.572	0.916	0.999
		LRT	0.068	0.244	0.710	0.971
	0.10	Proposed	0.106	0.374	0.811	0.987
		Rao-Scott	0.107	0.386	0.820	0.988
		Donner	0.101	0.354	0.797	0.984
		Pearson	0.345	0.645	0.945	0.999
		LRT	0.119	0.348	0.801	0.989

LRT = likelihood ratio test.

be used in clustered binomial data. When group sizes are as large as $n_1 = 15$, and $n_2 = n_3 = 20$ in Table 6, the performance of the proposed test becomes better even better than when ρ_i's are different between three groups. The performance of the Donner test also seems to be parallel to the proposed one but with slightly lower powers than ours. Generally, when group sizes are small to moderate, the Rao-Scott test and LRT are not good at controlling nominal levels, and a Pearson test shows a severe inflation of empirical sizes in all cases of the simulation study.

5. Concluding remarks

Binary outcomes observed in the same cluster are correlated to each other. This dependence relationship violates the crucial assumption of independence in analyzing clustered binomial data. An ordinary Pearson chi-squared test for homogeneity of proportions may lead to erroneous results due to the dependence relationship in clustered binomial data. Many researchers have tried to correct the inflation of a Type I error in a Pearson chi-squared statistic by incorporating variance inflation of proportion estimates.

Table 5: Empirical sizes and powers of test statistics according to $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$ when $I = 3$, $m_1 = m_2 = m_3 = 7$, $n_1 = n_2 = n_3 = 10$, and $\pi_1 = \pi_2 = 0.3$, $\pi_3 = 0.3 + \delta$

$\boldsymbol{\rho}$	α	Test	δ			
			0.0	0.1	0.2	0.3
(0.3, 0.3, 0.3)	0.05	Proposed	0.058	0.158	0.420	0.851
		Rao-Scott	0.109	0.228	0.451	0.901
		Donner	0.054	0.147	0.417	0.825
		Pearson	0.223	0.378	0.688	0.965
		LRT	0.090	0.174	0.484	0.886
	0.10	Proposed	0.127	0.246	0.597	0.933
		Rao-Scott	0.172	0.313	0.623	0.937
		Donner	0.116	0.231	0.533	0.923
		Pearson	0.319	0.504	0.820	0.977
		LRT	0.151	0.300	0.599	0.947
(0.3, 0.3, 0.4)	0.05	Proposed	0.040	0.172	0.470	0.798
		Rao-Scott	0.086	0.207	0.517	0.810
		Donner	0.034	0.154	0.439	0.774
		Pearson	0.189	0.403	0.766	0.960
		LRT	0.059	0.203	0.513	0.830
	0.10	Proposed	0.107	0.265	0.612	0.886
		Rao-Scott	0.137	0.304	0.630	0.884
		Donner	0.096	0.247	0.593	0.870
		Pearson	0.292	0.496	0.821	0.976
		LRT	0.121	0.289	0.652	0.904
(0.3, 0.4, 0.5)	0.05	Proposed	0.076	0.119	0.378	0.710
		Rao-Scott	0.111	0.163	0.432	0.717
		Donner	0.067	0.101	0.361	0.679
		Pearson	0.309	0.405	0.748	0.940
		LRT	0.096	0.134	0.413	0.745
	0.10	Proposed	0.145	0.217	0.512	0.811
		Rao-Scott	0.187	0.239	0.535	0.813
		Donner	0.134	0.189	0.484	0.791
		Pearson	0.399	0.511	0.800	0.955
		LRT	0.157	0.225	0.547	0.832

LRT = likelihood ratio test.

The chi-squared statistic by Donner (1989) has the form scaled by a VIF based on the BB model for clustered binomial data. The ICC is used to compute variance inflation. However in, Rao and Scott (1992) introduced design effects to adjust observed values to which the ordinary Pearson statistic is applied. In this paper, we have proposed the quasi-likelihood approach to adjust observed data by dispersion estimates. The asymptotic distribution of the proposed chi-square statistic has been discussed by following the asymptotic theory of a quasi-MLE under model misspecification. The proposed chi-squared test is simple to use and similar to the test by Rao and Scott (1992) because the ordinary chi-squared statistic can be applied to the adjusted values that result in the same asymptotic chi-square distribution.

According to a simulation study the proposed statistic appears to perform well in empirical sizes and powers even when group sizes are small. The modified statistic by Donner (1989) also shows good performances, and Jeong (2015) showed that the LRT and Wald test perform well in the control of levels and attaining higher powers. However, the LRT shows a tendency to exceed nominal levels in this study of small group sizes. In conclusion, we recommend the proposed test based on quasi-likelihood and the one by Donner (1989) when testing the homogeneity of proportions in clustered binomial data.

Table 6: Empirical sizes and powers of test statistics according to $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$ when $I = 3$, $m_1 = m_2 = m_3 = 10$, $n_1 = 15$, $n_2 = 20$, $n_3 = 20$, and $\pi_1 = \pi_2 = 0.3$, $\pi_3 = 0.3 + \delta$

$\boldsymbol{\rho}$	α	Test	δ			
			0.0	0.1	0.2	0.3
(0.3, 0.3, 0.3)	0.05	Proposed	0.051	0.307	0.859	0.998
		Rao-Scott	0.064	0.334	0.874	0.998
		Donner	0.047	0.299	0.842	0.998
		Pearson	0.220	0.632	0.966	1.000
		LRT	0.052	0.335	0.864	0.998
	0.10	Proposed	0.092	0.447	0.922	0.999
		Rao-Scott	0.118	0.468	0.924	0.999
		Donner	0.087	0.422	0.915	0.999
		Pearson	0.337	0.716	0.979	1.000
		LRT	0.104	0.461	0.926	1.000
(0.3, 0.3, 0.4)	0.05	Proposed	0.050	0.258	0.805	0.992
		Rao-Scott	0.067	0.292	0.801	0.990
		Donner	0.043	0.248	0.786	0.990
		Pearson	0.257	0.599	0.958	1.000
		LRT	0.063	0.270	0.809	0.994
	0.10	Proposed	0.115	0.384	0.886	0.997
		Rao-Scott	0.126	0.387	0.877	0.996
		Donner	0.106	0.366	0.875	0.996
		Pearson	0.363	0.685	0.977	1.000
		LRT	0.123	0.378	0.889	0.997
(0.3, 0.4, 0.5)	0.05	Proposed	0.063	0.229	0.689	0.961
		Rao-Scott	0.092	0.247	0.691	0.960
		Donner	0.059	0.195	0.662	0.948
		Pearson	0.358	0.629	0.946	1.000
		LRT	0.068	0.220	0.691	0.971
	0.10	Proposed	0.114	0.344	0.801	0.984
		Rao-Scott	0.142	0.352	0.793	0.982
		Donner	0.100	0.325	0.786	0.983
		Pearson	0.451	0.693	0.963	1.000
		LRT	0.126	0.326	0.803	0.982

LRT = likelihood ratio test.

Acknowledgement

This work was supported by a 2-Year Research Grant of Pusan National University.

References

Agresti A (2013). *Categorical Data Analysis* (3rd ed), John Wiley & Sons, Hoboken, NJ.

Cochran WG (1977). *Sampling Techniques* (3rd ed), John Wiley & Sons, Hoboken, NJ.

Crowder MJ (1978). Beta-binomial ANOVA for proportions, *Journal of the Royal Statistical Society, Series C: Applied Statistics*, **27**, 34–37.

Donner A (1989). Statistical methods in ophthalmology: an adjusted chi-squared approach, *Biometrics*, **45**, 605–611.

Jeong KM (2015). Goodness-of-fit for the clustered binomial models, *Journal of the Korean Data Analysis Society*, **17**, 1725–1737.

Jeong KM and Lee HY (2013). Modeling overdispersion for clustered binomial data, *Journal of the Korean Data Analysis Society*, **15**, 2343–2356.

Paul SR (1982). Analysis of proportions of affected foetuses in teratological experiments, *Biometrics*,

38, 361–370.
Rao JNK and Scott AJ (1992). A simple method for the analysis of clustered binary data, *Biometrics*, **48**, 577–585.
Reed JF (2004). Adjusted chi-square statistics: application to clustered binary data in primary care, *Annals of Family Medicine*, **2**, 201–203.
Ridout MS, Demétrio CGB, and Firth D (1999). Estimating intraclass correlation for binary data, *Biometrics*, **55**, 137–148.
Wedderburn RWM (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika*, **61**, 439–447.
White H (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1–25.
Williams DA (1982). Extra-binomial variation in logistic linear models, *Journal of the Royal Statistical Society, Series C: Applied Statistics*, **31**, 144–148.

Received August 17, 2016; Revised September 22, 2016; Accepted September 23, 2016