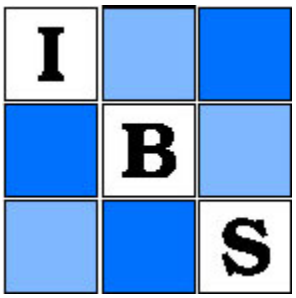

BIOMETRICS 48, 577–585
June 1992

SHORTER COMMUNICATIONS

EDITOR:
NIELS KEIDING

A Simple Method for the Analysis of Clustered Binary Data

J. N. K. Rao

Department of Mathematics and Statistics, Carleton University,
Ottawa, Ontario K1S 5B6, Canada

and

A. J. Scott

Department of Mathematics and Statistics, University of Auckland,
Private Bag, Auckland, New Zealand

SUMMARY

A simple method for comparing independent groups of clustered binary data with group-specific covariates is proposed. It is based on the concepts of design effect and effective sample size widely used in sample surveys, and assumes no specific models for the intracluster correlations. It can be implemented using any standard computer program for the analysis of independent binary data after a small amount of preprocessing. The method is applied to a variety of problems involving clustered binary data: testing homogeneity of proportions, estimating dose–response models and testing for trend in proportions, and performing the Mantel–Haenszel chi-squared test for independence in a series of 2×2 tables and estimating the common odds ratio and its variance. Illustrative applications of the method are also presented.

1. Introduction

Clustered binary data occur frequently in many fields of applications. For example, toxicological experiments designed to assess the teratogenic effects of chemicals often involve animal litters as experimental units. Another example involves the comparison of several groups of subjects in ophthalmologic studies, where each individual may contribute information on two eyes to the statistical analysis. Standard methods of analysis that ignore the cluster structure in such data tend to underestimate the true standard error of an estimated treatment difference. Similarly, the standard chi-squared tests may significantly inflate the Type I error rate (Haseman and Kupper, 1979).

Several alternative methods that take account of the within-cluster correlations have been proposed in the literature. Most of these methods, however, assume specific models for the intracluster correlations—for example, the beta–binomial model (Williams, 1975; Crowder, 1978; Paul, 1982), the correlated-binomial model (Kupper and Haseman, 1978), the multiplicative-binomial model (Altham, 1978), and the correlated probit regression model (Ochi and Prentice, 1984). Bonney (1987) modelled the dependence as a product of conditional probabilities, each of which is assumed to be logistic.

Zeger and Liang (1986) used a quasi-likelihood approach which specifies that a known function of the marginal expectation of a within-cluster observation is a linear function of

Key words: Binary response; Cochran–Armitage test; Dose–response models; Effective sample size; Homogeneity of proportions; Intracluster correlations; Mantel–Haenszel test.

known covariates, and assumes that the variance is a known function of the mean. In addition, a "working" correlation matrix for the observations for each cluster is specified; for example, the "independence working model" assumes that the observations for a cluster are independent. This setup leads to generalized estimating equations (GEE) that give consistent estimators of regression parameters and their variances (as the number of clusters goes to infinity) even when the within-cluster dependence is misspecified. Rotnitzky and Jewell (1990) extended the usual chi-squared tests of hypotheses on the regression parameters, under the Zeger–Liang setup.

In this paper, a simple method for comparing independent groups of clustered binary data with group-specific covariates is proposed. It is based on the concepts of "design effect" (deff) and "effective sample size" widely used in sample surveys (Kish, 1965, p. 259) and assumes no specific models for the intracluster correlations. Thus it is in the spirit of Zeger and Liang (1986) for more general problems involving within-cluster specific covariates. Our method, for the special case of comparing groups, is particularly simple, gives asymptotically correct results as the number of clusters in each group tends to infinity, and can be implemented using any standard computer program for the analysis of independent binary data after a small amount of preprocessing.

The method is applied to a variety of biometrical problems involving independent groups of clustered binary data—in particular, testing homogeneity of proportions, estimating dose–response models and testing for trends in proportions, and computing the Mantel–Haenszel chi-squared test statistic for independence in a series of 2×2 tables and estimating the common odds ratio and its variance when the independence hypothesis is rejected. Illustrative applications of the method are also given.

2. Basic Results

Suppose there are I conceptual populations of clusters receiving different treatments. Suppose also that independent random samples of clusters are drawn from the I populations. Let x_{ij} be the number of affected units among the n_{ij} units in the jth cluster ($j = 1, \ldots, m_i$) drawn from the ith population ($i = 1, 2, \ldots, I$). Our primary interest is in making inferences about the unknown overall proportions of affected units in the I populations, denoted by $p_1, p_2, \ldots, p_I$, utilizing the distributions created by random sampling of clusters.

A natural estimator of p_i is the overall sample proportion

$$\hat{p}_i = x_i/n_i, \tag{1}$$

where $x_i = \sum_j x_{ij}$ and $n_i = \sum_j n_{ij}$. Noting that $\hat{p}_i$ may be written as the ratio of two sample means, $\bar{x}_i = \sum_j x_{ij}/m_i$ and $\bar{n}_i = \sum_j n_{ij}/m_i$, we obtain an estimator of the variance of $\hat{p}_i$, for large m_i, as

$$v_i = m_i(m_i - 1)^{-1} n_i^{-2} \sum_{j=1}^{m_i} r_{ij}^2, \tag{2}$$

where $r_{ij} = x_{ij} - n_{ij}\hat{p}_i$; see, for example, Cochran (1977, p. 31). Under mild regularity conditions on the population variances of the n_{ij}'s and the r_{ij}'s, it follows that $(\hat{p}_i - p_i)/v_i^{1/2}$ is asymptotically $N(0, 1)$ as $m_i \to \infty$, for $i = 1, 2, \ldots, I$ [see Scott and Wu (1981) for details in the context of a finite population of clusters]. Also, v_i is a consistent estimator of $\text{var}(\hat{p}_i)$ in the sense that $m_i[v_i - \text{var}(\hat{p}_i)]$ converges in probability to 0 as $m_i \to \infty$, for each i.

The ratio of v_i to the estimated binomial variance $n_i^{-1}\hat{p}_i(1 - \hat{p}_i)$, denoted by

$$d_i = n_i v_i / [\hat{p}_i(1 - \hat{p}_i)], \tag{3}$$

represents the variance inflation due to clustering. The inflation factor d_i is called "design

effect" (deff) and $\tilde{n}_i = n_i/d_i$ "effective sample size" in survey sampling literature (Kish, 1965, p. 259).

We transform the aggregate data (x_i, n_i) to $(\tilde{x}_i, \tilde{n}_i)$, $i = 1, 2, \ldots, I$, where $\tilde{x}_i = x_i/d_i$, and treat $\tilde{x}_i$ as binomial$(\tilde{n}_i, p_i)$. This leads to asymptotically correct results (as $m_i \to \infty$ for each i) since the estimated binomial variance of $\tilde{p}_i = \tilde{x}_i/\tilde{n}_i = \hat{p}_i$ is given by $\tilde{p}_i(1 - \tilde{p}_i)/\tilde{n}_i = v_i$, which is the correct estimated variance of $\hat{p}_i$, and since

$$(\tilde{p}_i - p_i)/[\tilde{p}_i(1 - \tilde{p}_i)/\tilde{n}_i]^{1/2} = (\hat{p}_i - p_i)/v_i^{1/2}$$

is asymptotically $N(0, 1)$. Finally, noting that the covariance matrix of the $\hat{p}_i$'s is diagonal because of independent sampling from the I populations, we find that $[\sqrt{\tilde{n}_1}(\tilde{p}_1 - p_1), \ldots, \sqrt{\tilde{n}_I}(\tilde{p}_I - p_I)]'$ tends to multivariate normal with mean vector $\mathbf{0}$ and covariance matrix diag$[p_1(1 - p_1), \ldots, p_I(1 - p_I)]$. Thus, we may conclude that replacing $(n_i, \hat{p}_i)$ by $(\tilde{n}_i, \tilde{p}_i)$ or, equivalently, (x_i, n_i) by $(\tilde{x}_i, \tilde{n}_i)$, in any binomial-based procedure gives asymptotically correct results. Thus, the proposed method can be readily implemented using any standard computer program for the analysis of independent binary data after a small amount of preprocessing. We look at the application of this simple idea to a variety of problems in Section 3.

3. Testing Homogeneity

The homogeneity hypothesis is given by H_0: $p_1 = p_2 = \cdots = p_I$. If the standard chi-squared statistic

$$X^2 = \sum_{i=1}^{I} (x_i - n_i\hat{p})^2/[n_i\hat{p}(1 - \hat{p})], \tag{4}$$

with $\hat{p} = \sum x_i/\sum n_i$, is used without modification, then the asymptotic distribution of X^2 under H_0 would be a weighted sum of $I - 1$ independent χ^2 variables, each with 1 degree of freedom, with weights depending on the population inflation factors $D_i = n_i\text{var}(\hat{p}_i)/[p_i(1 - p_i)]$. Typically, these weights are larger than 1 because of positive intracluster correlations so that the actual Type I error rate is larger (sometimes much larger) than the nominal level (Scott and Rao, 1981).

Applying the method of Section 2 to X^2, i.e., replacing (x_i, n_i) by $(\tilde{x}_i, \tilde{n}_i)$ in (4), we get the adjusted chi-squared statistic

$$\tilde{X}^2 = \sum_{i=1}^{I} (\tilde{x}_i - \tilde{n}_i\tilde{p})^2/[\tilde{n}_i\tilde{p}(1 - \tilde{p})], \tag{5}$$

where $\tilde{p} = \sum \tilde{x}_i/\sum \tilde{n}_i$. Under H_0, $\tilde{X}^2$ is asymptotically distributed as a χ^2 variable with $I - 1$ degrees of freedom, unlike X^2. In the special case where the population inflation factors D_i are equal, say $D_i = D$ for $i = 1, 2, \ldots, I$, we may use $\tilde{x}_i = x_i/d$ and $\tilde{n}_i = n_i/d$, where d is a pooled estimate given by

$$(I - 1)d = \sum_{i=1}^{I} (1 - f_i)\frac{\hat{p}_i(1 - \hat{p}_i)}{\hat{p}(1 - \hat{p})}\, d_i, \tag{6}$$

with $f_i = n_i/n$ and $n = \sum n_i$ (see Scott and Rao, 1981). In this case, $\tilde{X}^2$ reduces to X^2/d.

Donner (1989) assumed a common intracluster correlation, ρ, across groups, in which case $D_i = 1 + (c_i - 1)\rho$ with $c_i = \sum_j n_{ij}^2/n_i$ under his model for the intracluster correlations. He proposed another adjusted chi-squared statistic

$$X_D^2 = \sum_{i=1}^{I} (x_i - n_i\hat{p})^2/[\hat{d}_i n_i\hat{p}(1 - \hat{p})], \tag{7}$$

where $\hat{d}_i = 1 + (c_i - 1)\hat{\rho}$ and $\hat{\rho}$ is the analysis of variance estimator of ρ. Unfortunately, the null distribution of X_D^2, under his setup, is χ^2 with $I - 1$ degrees of freedom only in

the special case of $D_i = D$ or $n_{ij} = \bar{n}$ for all (i, j). This follows from the asymptotic results of Scott and Rao (1981), but can be seen simply by considering the special case of $I = 2$ groups. In this case, X^2_D is asymptotically equivalent to

$$X^2_D(a) = \left[\frac{(\hat{p}_1 - \hat{p}_2)^2}{\text{var}(\hat{p}_1) + \text{var}(\hat{p}_2)}\right] \frac{\bar{D}_a \bar{D}_b}{D_1 D_2}, \tag{8}$$

with $\bar{D}_a = (n_1 D_1 + n_2 D_2)/(n_1 + n_2)$ and $\bar{D}_b = (n_2 D_1 + n_1 D_2)/(n_1 + n_2)$. Since $\bar{D}_a \bar{D}_b = D_1 D_2 + n_1 n_2 (n_1 + n_2)^{-2}(D_1 - D_2)^2 \geq D_1 D_2$ with equality if and only if $D_1 = D_2$, it follows from (8) that treating X^2_D as a χ^2 variable with $I - 1 = 1$ degree of freedom can lead to inflated Type I error rate. It may be noted, however, that X^2_D and X^2/d are asymptotically valid in the case of experimental comparisons involving random assignment of clusters to treatments.

Example 1. Table 1 gives the data from an experiment, taken from Williams (1975) and originally reported by Weil (1970). One group of 16 pregnant female rats was fed a control diet during pregnancy and lactation, while a second group of 16 pregnant female rats was fed a diet treated with a chemical. For each litter the number, n_{ij}, of pups alive at 4 days and the number, x_{ij}, of pups that survived the 21-day lactation period were recorded ($I = 2$, $m_1 = m_2 = 16$). Using the data (x_{ij}, n_{ij}) in Table 1 and formulae (2) and (3), we computed the inflation factors as $d_1 = 1.24$ for the control group and $d_2 = 3.95$ for the treated group.

The standard chi-squared statistic for testing the equality of the two death rates, H_0: $p_1 = p_2$, has the observed value $X^2_{\text{obs}} = 8.94$. The associated incorrect P-value, treating X^2 as χ^2 with $I - 1 = 1$ degree of freedom, is less than .005. On the other hand, the adjusted statistic $\tilde{X}^2$ has the observed value $\tilde{X}^2_{\text{obs}} = 4.10$, and the associated correct P-value, treating $\tilde{X}^2$ as χ^2 with $I - 1 = 1$ degree of freedom, is about .047. Williams (1975) and Ochi and Prentice (1984) previously analysed this data set, using a beta–binomial model and correlated probit regression model, respectively. Observed values for the respective chi-squared statistics for testing H_0: $p_1 = p_2$ are 4.48 and 5.86, referred to χ^2 with 1 degree of freedom. These values suggest slight loss of power for the simple adjusted statistic $\tilde{X}^2$, although it is possible that their models do not fit the data well.

It is clear from the large difference between d_1 and d_2 that the statistic X^2/d, based on the pooled estimate d, and Donner's statistic X^2_D are not appropriate for this data set.

Table 1

Number, n_{ij}, of pups alive at 4 days and the number, x_{ij}, of pups that survived the 21-day lactation period for litter j in group i ($j = 1, 2, \ldots, 16$; $i = 1, 2$) (Weil, 1970; Williams, 1975)

Group	(x_{ij}, n_{ij})							
Control ($i = 1$)	(13, 13)	(12, 12)	(9, 9)	(9, 9)	(8, 8)	(8, 8)	(12, 13)	(11, 12)
	(9, 10)	(9, 10)	(8, 9)	(11, 13)	(4, 5)	(5, 7)	(7, 10)	(7, 10)
Treated ($i = 2$)	(12, 12)	(11, 11)	(10, 10)	(9, 9)	(10, 11)	(9, 10)	(9, 10)	(8, 9)
	(8, 9)	(4, 5)	(7, 9)	(4, 7)	(5, 10)	(3, 6)	(3, 10)	(0, 7)

4. Tests for Trend in Proportions

Suppose that a score z_i is associated with the ith group ($z_1 < z_2 < \cdots < z_I$) and that $p_i = K(\alpha + \beta z_i)$, where K is a monotone function that is twice differentiable. For example, the choice $K(a) = [1 + \exp(a)]^{-1}$ gives the familiar logistic linear regression model

$$\ln[p_i/(1 - p_i)] = \alpha + \beta z_i, \quad i = 1, \ldots, I. \tag{9}$$

We are interested in testing the null hypothesis of no trend, i.e., in testing the composite hypothesis H_0: $\beta = 0$ versus the one-sided alternative H_1: $\beta > 0$. A test statistic commonly used for this purpose is the Cochran–Armitage statistic (Cochran, 1954; Armitage, 1955)

$$t = (\sum x_i z_i - \hat{p} \sum n_i z_i)/[\hat{p}(1 - \hat{p}) \sum n_i (z_i - \bar{z})^2]^{1/2}, \tag{10}$$

provided $x_1, \ldots, x_I$ follow independent binomial distributions $B(n_1, p_1), \ldots, B(n_I, p_I)$, where $\bar{z} = \sum n_i z_i / \sum n_i$ and $\hat{p} = \sum x_i / \sum n_i$. Tarone and Gart (1980) showed that the statistic t gives the $C(\alpha)$ test of Neyman (1959) or the score test for any choice of the twice-differentiable function K. Under H_0, t has a limiting standard normal distribution, $N(0, 1)$, as $n_i \to \infty$ for each i, provided the x_i's are independent binomial variables.

Application of the Cochran–Armitage test to clustered data, without adjustment for intracluster correlations, however, leads to erroneous inferences. An asymptotically valid test is simply obtained by replacing (x_i, n_i) with $(\tilde{x}_i, \tilde{n}_i)$ in (10) and treating the resulting statistic, $\tilde{t}$, as a standard normal variable.

If H_0: $\beta = 0$ is rejected in favour of H_1: $\beta > 0$, we proceed to fit the logistic regression model (9) to the transformed data $(\tilde{x}_i, \tilde{n}_i)$, $i = 1, \ldots, I$, using standard methods for independent binomial data, leading to estimates $\tilde{\alpha}$ and $\tilde{\beta}$ with estimated standard errors $\tilde{\sigma}_\alpha$ and $\tilde{\sigma}_\beta$, respectively.

Goodness of fit of the logistic regression model (9) can be tested using the adjusted chi-squared statistic

$$\tilde{X}^2 = \sum (\tilde{x}_i - \tilde{n}_i p_i(\tilde{\alpha}, \tilde{\beta}))^2 / [\tilde{n}_i p_i(\tilde{\alpha}, \tilde{\beta}) q_i(\tilde{\alpha}, \tilde{\beta})], \tag{11}$$

where $q_i = 1 - p_i$. If the model is correct, then $\tilde{X}^2$ is asymptotically distributed as χ^2_{I-2}, a χ^2 variable with $I - 2$ degrees of freedom. Hence, the P-value is given by $\Pr(\chi^2_{I-2} > \tilde{X}^2_{\text{obs}})$, where $\tilde{X}^2_{\text{obs}}$ is the observed value of $\tilde{X}^2$. Treating $(x_1, n_1), \ldots, (x_I, n_I)$ as independent binomial data and then using the standard chi-squared goodness-of-fit statistic, X^2, leads to erroneous P-values.

Example 2. Paul (1982) analysed some experimental data on the number of live foetuses in a litter affected by treatment and the total number of live foetuses in a litter, for each of $I = 4$ dose groups: control (C), low dose (L), medium dose (M), and high dose (H). We omitted the high-dose group because of problems with high toxicity. The remaining data

Table 2

Data from Shell Toxicology Laboratory. (i) *Number of live foetuses in a litter affected by treatment.* (ii) *Total number of live foetuses in a litter* (*Paul*, 1982).

Group																		
Control, C	(i)	1	1	4	0	0	0	0	0	1	0	2	0	5	2	1	2	0
	(ii)	12	7	6	6	7	8	10	7	8	6	11	7	8	9	2	7	9
	(i)	0	1	0	0	0	0	3	2	4	0							
	(ii)	7	11	10	4	8	10	12	8	7	8							
Low dose, L	(i)	0	1	1	0	2	0	1	0	1	0	0	3	0	0	1	5	
	(ii)	5	11	7	9	12	8	6	7	6	4	6	9	6	7	5	9	
	(i)	0	0	3														
	(ii)	1	6	9														
Medium dose, M	(i)	2	3	2	1	2	3	0	4	0	0	4	0	0	6	6	5	
	(ii)	4	4	9	8	9	7	8	9	6	4	6	7	3	13	6	8	
	(i)	4	1	0	3	6												
	(ii)	11	7	6	10	6												

are reported in Table 2. Since the actual dose levels were not given, we assigned the scores $z_1 = 0$, $z_2 = 1$, and $z_3 = 2$ to the three dose groups, C, L, and M, respectively.

Using the data in Table 2, we calculated the variance inflation factors as $d_1 = 2.33$ for controls, $d_2 = 2.05$ for the low-dose group, and $d_3 = 2.49$ for the medium-dose group. The Cochran–Armitage statistic gives $t_{\text{obs}} = 4.54$, while the adjusted statistic gives $\tilde{t}_{\text{obs}} = 3.02$. Treating $\tilde{t}$ as $N(0, 1)$, we find the P-value to be less than .002. Thus, we may conclude that the data provide strong evidence of increasing trend in the response proportions.

We proceed to fit the logistic regression model (9) to the transformed data $(\tilde{x}_i, \tilde{n}_i)$, $i = 1, 2, 3$. The adjusted chi-squared statistic $\tilde{X}^2$ gives $\tilde{X}^2_{\text{obs}} = 2.15$. Treating $\tilde{X}^2$ as a χ^2 variable with $I - 2 = 1$ degree of freedom, we find the P-value to be .142. The logistic regression model (9) thus provides a good fit to the data. On the other hand, the standard chi-squared goodness-of-fit statistic, X^2, gives $X^2_{\text{obs}} = 4.72$, and treating X^2 as a χ^2 variable with 1 degree of freedom leads to an erroneous P-value of .03.

Fitting the model (9) to the transformed data $(\tilde{x}_i, \tilde{n}_i)$ by standard methods, we obtain $\tilde{\alpha} = -2.08$ and $\tilde{\beta} = .64$ with estimated standard errors $\tilde{\sigma}_\alpha = .30$ and $\tilde{\sigma}_\beta = .21$.

5. Mantel–Haenszel Test

Consider a series of K 2×2 tables with fixed row totals (n_{1k}, n_{2k}) and number of "successes" (x_{1k}, x_{2k}) with associated success probabilities (p_{1k}, p_{2k}), $k = 1, \ldots, K$. It is of interest to test the null hypothesis H_0: $\psi = 1$, assuming a common odds ratio $\psi = (p_{1k}q_{2k})/(p_{2k}q_{1k})$, where $q_{tk} = 1 - p_{tk}$, $t = 1, 2$. Mantel and Haenszel (1959) proposed a chi-squared statistic for testing H_0, assuming that the x_{tk}'s are independent binomial variables $B(n_{tk}, p_{tk})$. It is given by

$$X^2_{\text{MH}} = \frac{[\sum_k (n_{1k}n_{2k}/n_k)(\hat{p}_{1k} - \hat{p}_{2k})]^2}{\sum_k [n_{1k}n_{2k}/(n_k - 1)]\hat{p}_k(1 - \hat{p}_k)}, \tag{12}$$

where

$$\hat{p}_{tk} = x_{tk}/n_{tk}, \quad \hat{p}_k = (x_{1k} + x_{2k})/(n_{1k} + n_{2k}), \quad n_k = n_{1k} + n_{2k}.$$

Under the null hypothesis, X^2_{MH} is asymptotically (as $n_{tk} \to \infty$ for each t, k) χ^2 with 1 degree of freedom.

With clustered data, we adjust X^2_{MH} to account for intracluster correlations. We compute the variance inflation factors d_{tk} using the cluster-level data (x_{tkj}, n_{tkj}) $(j = 1, \ldots, m_{tk}$; $t = 1, 2$; $k = 1, \ldots, K)$ in formulae (2) and (3), where x_{tkj} is the number of "successes" among the n_{tkj} units in the jth cluster in row t of table k. Note that the subscript i in (2) and (3) is now changed to the subscripts tk. An asymptotically (as $m_{tk} \to \infty$ for each t, k) valid test of H_0 is now simply obtained by replacing (x_{tk}, n_{tk}) with $(\tilde{x}_{tk}, \tilde{n}_{tk})$ in (12) and treating the resulting statistic, $\tilde{X}^2_{\text{MH}}$, as χ^2 with 1 degree of freedom, where $\tilde{x}_{tk} = x_{tk}/d_{tk}$ and $\tilde{n}_{tk} = n_{tk}/d_{tk}$.

Donald and Donner (1987) proposed another adjusted Mantel–Haenszel statistic, similar to $\tilde{X}^2_{\text{MH}}$, except that $\hat{p}_k$ is used instead of $\tilde{p}_k$ and the population inflation factors, D_{tk}, are estimated under the assumption of a common intracluster correlation.

If H_0 is rejected, we proceed to estimate the common odds ratio ψ and set confidence intervals on ψ. For independent binomial data, Mantel and Haenszel (1959) proposed the estimator

$$\hat{\psi}_{\text{MH}} = \frac{\sum_k (n_{1k}n_{2k}/n_k)\hat{p}_{1k}\hat{q}_{2k}}{\sum_k (n_{1k}n_{2k}/n_k)\hat{p}_{2k}\hat{q}_{1k}}, \tag{13}$$

where $\hat{q}_{tk} = 1 - \hat{p}_{tk}$. Hauck (1979) derived an estimator of its asymptotic variance as

$$\hat{V}_{\text{MH}} = \hat{\psi}^2_{\text{MH}}(\sum \hat{w}_k^2\hat{b}_k)/(\sum \hat{w}_k)^2, \tag{14}$$

where

$$\hat{w}_k = (n_{1k}^{-1} + n_{2k}^{-1})^{-1}\hat{p}_{2k}\hat{q}_{1k} \quad \text{and} \quad \hat{b}_k = (n_{1k}\hat{p}_{1k}\hat{q}_{1k})^{-1} + (n_{2k}\hat{p}_{2k}\hat{q}_{2k})^{-1}.$$

An asymptotic $100(1-\alpha)\%$ confidence interval on ψ is given by $\hat{\psi}_{\text{MH}} \pm z_{\alpha/2}\hat{V}_{\text{MH}}^{1/2}$. Confidence intervals based on the log odds ratio, $\hat{\gamma}_{\text{MH}} = \ln \hat{\psi}_{\text{MH}}$, usually give better coverage than those based on $\hat{\psi}_{\text{MH}}$ since $\hat{\psi}_{\text{MH}}$ has a skewed distribution. Noting that the estimated variance of $\hat{\gamma}_{\text{MH}}$ is $\hat{\psi}_{\text{MH}}^{-2}\hat{V}_{\text{MH}}$ approximately, we get a $100(1-\alpha)\%$ confidence interval on $\gamma = \ln \psi$ as $\hat{\gamma}_{\text{MH}} \pm z_{\alpha/2}\hat{\psi}_{\text{MH}}^{-1}\hat{V}_{\text{MH}}^{1/2}$, which can be converted to an interval on ψ.

With clustered data, we simply replace (x_{tk}, n_{tk}) with $(\tilde{x}_{tk}, \tilde{n}_{tk})$ in (13) and (14) to get the estimate $\tilde{\psi}_{\text{MH}}$ with estimated variance $\tilde{V}_{\text{MH}}$. An asymptotic $100(1-\alpha)\%$ confidence interval on ψ is now obtained as $\tilde{\psi}_{\text{MH}} \pm z_{\alpha/2}\tilde{V}_{\text{MH}}^{1/2}$. For the log odds ratio, we have $\tilde{\gamma}_{\text{MH}} = \ln \tilde{\psi}_{\text{MH}}$ with estimated variance $\tilde{\psi}_{\text{MH}}^{-2}\tilde{V}_{\text{MH}}$, and hence a $100(1-\alpha)\%$ confidence interval on $\gamma = \ln \psi$ is given by $\tilde{\gamma}_{\text{MH}} \pm z_{\alpha/2}\tilde{\psi}_{\text{MH}}^{-1}\tilde{V}_{\text{MH}}^{1/2}$.

Donald and Donner (1987) proposed an alternative estimator of the variance of $\hat{\psi}_{\text{MH}}$, assuming a common intracluster correlation. Results of a simulation study, under a beta–binomial model for the intracluster correlations, are reported by Donald and Donner (1990).

Example 3. Donner and Banting (1989) analysed some data on gingivitis outcome among subjects (clusters) from a randomized clinical trial of a mouthwash. The data (aggregated over subjects) are reported in Table 3, classified by sex (males denoted by $t = 1$ and females by $t = 2$) and treatment group [control ($k = 1$), low concentration ($k = 2$), medium concentration ($k = 3$), high concentration ($k = 4$)]. The number of subjects, m_{tk}, varied from 42 to 128 (see Table 3). Using subject-level data (kindly supplied by Dr A. Donner), we computed the variance inflation factors as $d_{11} = 2.01$, $d_{12} = 3.16$, $d_{13} = 2.43$, $d_{14} = 2.82$ for males and $d_{21} = 2.05$, $d_{22} = 2.75$, $d_{23} = 2.57$, $d_{24} = 2.70$ for females.

The question of interest here is whether the proportion of disease-free sites among male subjects is higher than the corresponding proportion among female subjects. We obtain the observed values of X_{MH}^2 and $\tilde{X}_{\text{MH}}^2$ as 21.42 and 9.91, respectively. The P-value associated with the adjusted statistic $\tilde{X}_{\text{MH}}^2$, $\Pr(\chi_1^2 > 9.91)$, is less than .005, indicating significant overall gender difference. The observed value, 7.59, of the Donald–Donner statistic, assuming a common intrasubject correlation, is also highly significant.

Since H_0: $\psi = 1$ is rejected, we proceed to estimate the common odds ratio ψ and set confidence intervals on ψ. We obtain $\tilde{\psi}_{\text{MH}} = 1.52$ and $\tilde{V}_{\text{MH}} = .0414$ so that an asymptotically correct 95% confidence interval on ψ is given by [1.12, 1.92]. On the other hand, the

Table 3
Gingivitis outcome (by gender) at 3 months on the lingual surface of mandibular molars and bicuspids (Donner and Banting, 1989)

Treatment group		No. of patients	No. of surfaces with no gingivitis	Total no. of surfaces	Prop. free of gingivitis	Odds ratio
Control	Male	78	103	404	.25	2.34
	Female	123	84	658	.13	
Low	Male	45	81	229	.35	1.47
	Female	57	91	336	.27	
Intermediate	Male	68	96	378	.25	1.19
	Female	128	149	669	.22	
High	Male	42	102	200	.51	1.16
	Female	60	154	325	.47	

independent binomial assumption gives $\hat{\psi}_{MH} = 1.47$, $\hat{V}_{MH} = .0156$, and an "incorrect" 95% confidence interval [1.23, 1.72] whose length is .49 compared to .80 for the "correct" interval.

6. Remarks

The simple method presented here for analysing cluster correlated binary data makes no assumption on the dependence structure among binary observations within each cluster. The price for this robustness will be some loss in power compared to optimal tests under a specified model for the intracluster correlations, provided the assumed model fits the data well. It would be useful to study the relative powers under different intracluster models and deviations from the assumed models.

Acknowledgements

Our thanks are due to three referees, the associate editor, and Dr A. Donner for several constructive comments and suggestions. This research was supported by a grant from the National Sciences and Engineering Research Council of Canada.

Résumé

On propose une méthode simple permettant la comparaison de groupes indépendants de données binaires groupées, avec des valeurs de covariables spécifiques du groupe. Cette méthode est fondée sur des notions bien connues dans les sondages, notamment celles de taille effective de l'échantillon et d'effet du plan de son recueil; elle ne suppose aucun modèle particulier pour les corrélations intra-groupes. Après une brève étape de préparation des données, la méthode peut être mise en oeuvre à l'aide de n'importe quel logiciel standard traitant des données binaires. Elle s'applique à une variété de problèmes traitant des groupes de données binaires: test de l'égalité des fréquences, estimation de modèles de doses–réponses et test de tendance, ainsi qu'au test du chi deux de Mantel–Haenszel d'indépendance pour une série de tables 2×2 avec estimation du odds-ratio commun et de sa variance. A titre d'illustration, on présente quelques applications.

References

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statistics* **27,** 162–167.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11,** 375–386.

Bonney, G. W. (1987). Logistic regression for dependent binary observations. *Biometrics* **43,** 951–973.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.

Cochran, W. G. (1977). *Sampling Techniques,* 3rd edition. New York: Wiley.

Crowder, M. J. (1978). Beta–binomial ANOVA for proportions. *Applied Statistics* **27,** 34–37.

Donald, A. and Donner, A. (1987). Adjustments to the Mantel–Haenszel chi-squared statistic and odds ratio estimator when the data are clustered. *Statistics in Medicine* **6,** 491–499.

Donald, A. and Donner, A. (1990). A simulation study of the analysis of sets of 2×2 contingency tables under cluster sampling: Estimation of a common odds ratio. *Journal of the American Statistical Association* **85,** 537–543.

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

Donner, A. and Banting, D. (1989). Adjustment of frequently used chi-square procedures for the effect of site-to-site dependencies in the analysis of dental data. *Journal of Dental Research* **68,** 1350–1354.

Haseman, J. K. and Kupper, L. L. (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics* **34,** 69–76.

Hauck, W. W. (1979). The large-sample variance of the Mantel–Haenszel estimator of a common odds ratio. *Biometrics* **29,** 817–819.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34,** 69–76.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22,** 719–748.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. In *Probability and Statistics*, U. Grenander (ed.), 213–234. New York: Wiley.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71,** 531–543.

Paul, S. R. (1982). Analysis of proportions of affected foetuses in teratological experiments. *Biometrics* **38,** 361–370.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77,** 485–497.

Scott, A. J. and Rao, J. N. K. (1981). Chi-squared tests for contingency tables with proportions estimated from survey data. In *Current Topics in Survey Sampling*, D. Krewski, R. Platek, and J. N. K. Rao (eds), 247–266. New York: Academic Press.

Scott, A. J. and Wu, C. F. J. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association* **76,** 98–102.

Tarone, R. E. and Gart, J. J. (1980). On the robustness of combined tests for trends in proportions. *Journal of the American Statistical Association* **75,** 110–116.

Weil, C. S. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology* **8,** 177–182.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31,** 949–952.

Zeger, S. L. and Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.

Received August 1990; *revised May and September* 1991; *accepted September* 1991.