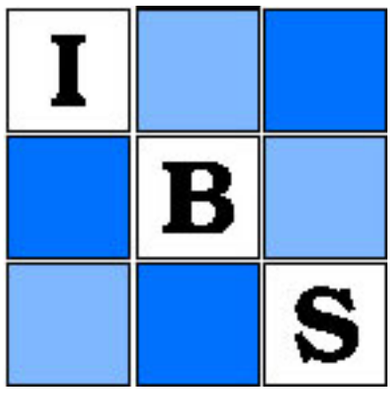

BIOMETRICS 45, 605–611
June 1989

Statistical Methods in Ophthalmology: An Adjusted Chi-Square Approach

Allan Donner

Department of Epidemiology and Biostatistics, The University of Western Ontario, Kresge Building, London N6A 5C1, Canada

SUMMARY

Ophthalmologic studies often compare several groups of subjects for the presence or absence of some ocular finding, where each subject may contribute two eyes to the analysis, the values from the two eyes being highly correlated. Rosner (1982, *Biometrics* **38,** 105–114) and Dallal (1988, *Biometrics* **44,** 253–257) proposed procedures for testing whether the proportion of affected eyes is the same among the different groups, while accounting for the intrasubject correlation. In this paper we propose an alternative approach, based on a simple adjustment of the standard Pearson chi-square test for the equality of proportions. The suggested approach utilizes information on subjects who supply only one eye to the analysis, and readily generalizes to studies in which more than two units of analysis are provided by each subject.

1. Introduction

Rosner (1982) pointed out that the fundamental unit for statistical analysis in ophthalmologic studies is often the eye rather than the person. If the purpose of the study is to compare $G \geq 2$ different groups of patients on some finding in an ocular examination, then an individual may contribute information on two eyes to the analysis, their value being, in general, highly correlated. In this case standard methods of analysis in which each eye is considered as an independent random variable are not valid. Dealing with the case of a binary outcome, Rosner (1982) proposed a model for testing whether the proportion of affected eyes is the same among the G groups of patients, while accounting for the intraperson dependence. Le (1988) extended this methodology to testing for a linear trend among the G proportions. The methodology has also been extended (Rosner, 1984) to situations where more than two units of analysis are provided by each individual, and which allow for covariate adjustment.

In this paper we deal with the case of a single binomially distributed eye-specific outcome variable, to which Rosner's (1982) results are applicable. Dallal (1988), criticizing the appropriateness of Rosner's model for this case, proposed an alternative approach based on compound multinomial sampling. In this paper we propose an alternative to these approaches, based on a simple adjustment of the standard Pearson chi-square test for homogeneity of proportions. The validity and power of the three proposed procedures are investigated using simulation.

2. Theory

Denote the number of subjects in the ith group by n_i ($i = 1, 2, \ldots, G$), where $N = \sum_{i=1}^{G} n_i$ is the total number of subjects in the study, each contributing one or two eyes to the

Key words: Intraclass correlation; Ophthalmology.

analysis. Let $Y_{ijk} = 1$ if the characteristic of interest is present at the kth eye of the jth subject in the ith group ($i = 1, 2, \ldots, G$), and $\theta_i = \Pr(Y_{ijk} = 1)$. Our primary aim is to test H_0: $\theta_1 = \theta_2 = \cdots = \theta_G$ vs H_1: $\theta_r \neq \theta_s$ for at least one pair (r, s). We first consider the approaches suggested by Rosner (1982) and Dallal (1988).

2.1 *Rosner's Approach*

Rosner (1982) assumed a "constant R" model, which states that $\Pr(Y_{ijk} = 1 \mid Y_{ij(3-k)} = 1) = R\theta_i$ for some positive constant R. If $R = 1$, then the eyes are completely independent, whereas if $R\theta_i = 1$, then the eyes are completely dependent. Let n_{il} denote the number of persons in the ith group with exactly l affected eyes ($i = 1, 2, \ldots, G$; $l = 0, 1, 2$), where $n_i = \sum_{l=0}^{2} n_{il}$. Rosner (1982) estimated the "effective number of eyes per person" under this model by

$$\hat{e} = \frac{2\hat{\lambda}(1 - \hat{\lambda})}{\hat{\lambda}(1 - \hat{\lambda}) + (\hat{R} - 1)\hat{\lambda}^2},$$

where $\hat{\lambda} = \frac{1}{2} \sum (n_{i1} + 2n_{i2})/N$, $\hat{R} = 4N \sum n_{i2}/(\sum n_{i1} + 2 \sum n_{i2})^2$ are the maximum likelihood estimators of λ, R, respectively, under H_0. An approximate test of H_0 is then given by referring

$$T = \left[\frac{\hat{e}}{\hat{\lambda}(1 - \hat{\lambda})}\right] \sum n_i(\hat{\lambda}_i - \hat{\lambda})^2 \tag{1}$$

to the chi-square distribution with $G - 1$ degrees of freedom, where $\hat{\lambda}_i = \frac{1}{2}(n_{i1} + 2n_{i2})/n_i$.

2.2 *Dallal's Approach*

A basic assumption behind Rosner's model is that the probability of success at one eye given a success at the other eye is proportional to θ_i. Dallal (1988) criticizes this assumption, pointing out that the constant R model will give a poor fit if the characteristic is almost certain to occur bilaterally with widely varying group-specific prevalence, because $R\theta_i$ cannot be close to 1 for all i unless the θ's themselves are nearly equal. Suppose instead that the probability of a success at one eye given a success at the other eye does not depend on θ_i, i.e., define $\Pr(Y_{ijk} = 1 \mid Y_{ij(3-k)} = 1) = \tau_i$ ($i = 1, 2, \ldots, G$). Dallal (1988) then proposed an analysis based on the following sequence of models:

Model 1: $\theta_1 = \theta_2 = \cdots = \theta_G$; $\tau_1 = \tau_2 = \cdots = \tau_G = \tau$

Model 2: $\theta_r \neq \theta_s$ for at least one pair (r, s); $\tau_1 = \tau_2 = \cdots = \tau_G = \tau$

Model 3: $\theta_r \neq \theta_s$ for at least one pair (r, s); $\tau_c \neq \tau_d$ for at least one pair (c, d).

Letting "$\cdot$" denote summation over the corresponding subscript, the expected values of the frequencies n_{il} have maximum likelihood estimates given by

Model 1: $$\mathrm{E}_{il} = \frac{n_{i\cdot}n_{\cdot l}}{n_{\cdot\cdot}}$$

Model 2: $$\mathrm{E}_{i0} = n_{i0},$$

$$\mathrm{E}_{i1} = \frac{(n_{i1} + n_{i2})(n_{\cdot 1})}{n_{\cdot 1} + n_{\cdot 2}},$$

$$\mathrm{E}_{i2} = \frac{(n_{i1} + n_{i2})(n_{\cdot 2})}{n_{\cdot 1} + n_{\cdot 2}},$$

Model 3: $E_{il} = n_{il}$

Testing H_0: $\theta_1 = \theta_2 = \cdots = \theta_G$ is equivalent to testing whether the proportion of unaffected individuals is the same in all groups, given a constant ratio of unilateral occurrence to bilateral occurrence, i.e., given a common τ. The desired test can therefore be obtained by comparing Model 1 vs Model 2 with the appropriate likelihood ratio statistic. Dallal (1988) shows that this statistic is given by

$$D = 2 \sum_{i=1}^{G} \sum_{l=0}^{2} n_{il} \log_e \left(\frac{E_{il}^{(1)}}{E_{il}^{(2)}} \right), \tag{2}$$

where D is referred to tables of the chi-square distribution with $G - 1$ degrees of freedom. Model 3 is not directly relevant to H_0, but is useful for testing other hypotheses of interest.

2.3 *An Adjusted Chi-Square Approach*

In this section we present an alternative approach to testing H_0 based on an adjustment of the usual Pearson chi-square statistic by an empirical estimate of the intraclass correlation between responses on two eyes of the same person. One advantage of this approach, aside from its intuitive attractiveness, is that it can be extended in a very simple fashion to handle the case in which an arbitrary number of units of analysis are provided by each person, as for example, in dental data. In Section 3, we compare the validity and power of all three approaches using Monte Carlo simulation.

Using a basic general approach, we assume that the clustering within an individual may be modelled by assuming that (i) for each pair of eyes the distribution of the number of successes is binomial with probability parameter P and (ii) the parameter P varies among individuals according to some (possibly unspecified) distribution on the interval (0, 1). If the expectation of P is p, then the variance of P is no greater than $p(1 - p)$, so that $\text{var}(P) = \rho p(1 - p)$, where ρ is a real number, $0 \leqslant \rho \leqslant 1$. We call ρ the intra-person correlation coefficient, which is assumed constant for all individuals in the sample. Now suppose individual j in group i contributes m_{ij} measurements to the analysis ($j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, G$), where $m_{ij} = 1$ or 2. Then the total number of measurements in the ith group is given by $M_i = \sum_{j=1}^{n_i} m_{ij}$, and the total number of successes by $A_i = \sum_{l=0}^{2} l n_{il}$. If $\rho = 0$, then H_0 may be tested by the standard Pearson chi-square statistic with $G - 1$ degrees of freedom, given by

$$X^2 = \sum_{i=1}^{G} X_i^2,$$

where

$$X_i^2 = \frac{(A_i - M_i\hat{\theta})^2}{M_i\hat{\theta}} + \frac{(M_i - A_i - M_i\hat{Q})^2}{M_i\hat{Q}},$$

and $\hat{\theta} = \sum A_i / \sum M_i$, $\hat{Q} = 1 - \hat{\theta}$. If $\rho > 0$, X^2 is no longer approximated by a chi-square distribution, but can be appropriately adjusted so that a modified form of the test can be applied. Let $\bar{C}_i = \sum_j m_{ij} C_{ij} / \sum_j m_{ij}$, where $C_{ij} = 1 + (m_{ij} - 1)\rho$. Then it follows from Brier (1980) that $X_A^2 = \sum_{i=1}^{G} X_i^2 / \bar{C}_i$ approximately follows a chi-square distribution with $G - 1$ degrees of freedom under H_0. To use this result in practice, however, an estimate of the unknown parameter ρ is required. We recommend an estimator that corresponds to the standard analysis of variance estimator of an intraclass correlation for a stratified cluster design (Donner, 1985). Let $\hat{\theta}_{ij} = a_{ij}/m_{ij}$ denote the proportion of successes for individual j in group i ($j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, G$), $\hat{\theta}_i = A_i/T_i$ the overall proportion of successes

in group i, and $M = \sum\sum_{ij} m_{ij}$ the total number of observations in the sample. Then the mean squared error among persons within groups and the mean squared error within individuals are given, respectively, by

$$\text{MSC} = \frac{\sum\sum_{ij} m_{ij}(\hat{\theta}_{ij} - \hat{\theta}_i)^2}{N - G},$$

$$\text{MSE} = \frac{\sum\sum_{ij} a_{ij}(1 - \hat{\theta}_{ij})}{M - N}.$$

An appropriate estimate of the intraperson correlation is then given by

$$\hat{\rho} = \frac{\text{MSC} - \text{MSE}}{\text{MSC} + (m_A - 1)\text{MSE}},$$

where

$$m_A = \frac{M - \sum_i (\sum_j m_{ij}^2/M_i)}{N - G}. \tag{3}$$

Thus, an approximate test of H_0 is obtained by substituting $\hat{\rho}$ for ρ in $\bar{C}_i$ and referring X_A^2 to tables of the chi-square distribution with $G - 1$ degrees of freedom. We refer to this procedure as the adjusted chi-square test.

A special case of this test occurs when each individual contributes exactly $m = 2$ eyes to the analysis. The expression for X_A^2 then reduces to $X^2/(1 + \hat{\rho})$, where $1 + \hat{\rho}$ is the "variance inflation factor" associated with the intraperson dependence. A further simplification is that the mean squared errors MSC and MSE required for the computation of $\hat{\rho}$ reduce to

$$\text{MSC} = \frac{\sum_{i=1}^{G} \{\sum_{l=0}^{2} l^2 n_{il} - (\sum ln_{il})^2/(2n_i)\}}{N - G}$$

and

$$\text{MSE} = \frac{\{\sum_{i=1}^{G} \sum_{l=0}^{2} ln_{il} - \sum l^2 n_{il}/2\}}{M - N}.$$

Example We reanalyze the data presented by Rosner (1982) on 218 persons aged 20–39 with retinitis pigmentosa (RP), seen at the Massachusetts Eye and Ear Infirmary from 1970 to 1979. The sample is described in detail by Berson, Rosner, and Simonoff (1980). The patients, one from each of 218 separate families, were classified into the genetic types of autosomal dominant RP (DOM), autosomal recessive RP (AR), sex-linked RP (SL), and isolated RP (ISO). The distribution of patients by genetic type and number of affected eyes is given in Table 1.

Table 1
Distribution of the number of affected eyes for persons in each genetic type (Rosner, 1982)

	n_{i0}	n_{i1}	n_{i2}	n_i
DOM	15	6	7	28
AR	7	5	9	21
SL	3	2	14	19
ISO	67	24	57	148

The test statistic proposed by Rosner (1982) may be computed for these data from (1) as $T = 11.36$ ($P = .01$), while Dallal's statistic may be computed from (2) as $D = 8.86$ ($P = .03$). The adjusted chi-square statistic X_A^2 depends on the estimated intraperson correlation, which may be obtained from (3) as $\hat{\rho} = .647$. The value of X_A^2 is then given by $X_A^2 = 11.43$ ($P = .01$). Thus, the results from each test are in agreement that the prevalence rate of affected eyes varies significantly according to genetic type. Further analysis indicates that the significant variation can be largely attributed to differences between the SL group and the other groups, again consistent with Rosner's results.

3. Comparison of the Methods

In this section we report the results of a simulation study estimating the empirical levels of significance and powers associated with the statistics T, D, and X_A^2 under a reasonably wide range of parameter values. We assumed for the purposes of this study that the parameter P varies among individuals according to a beta distribution so that the number of successes a_{ij} contributed by $m = 2$ eyes of a given individual follows a beta–binomial distribution with density

$$\Pr(a_{ij}) = \binom{2}{a_{ij}} \frac{\Gamma(\alpha_{ij}+\beta_{ij})\Gamma(a_{ij}+\alpha_{ij})\Gamma(2-a_{ij}+\beta_{ij})}{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})\Gamma(2+\alpha_{ij}+\beta_{ij})}, \quad j=1,2,\ldots,n_i; \quad i=1,2,\ldots,G.$$

Under this model, which allows considerable flexibility and is widely used to model dependent data (e.g., Griffiths, 1973; Moore, 1987), the expected probability of a success in a randomly chosen individual is given by $\theta_{ij} = \alpha_{ij}/(\alpha_{ij} + \beta_{ij})$ and the intraperson correlation coefficient by $\rho = 1/(1 + \alpha_{ij} + \beta_{ij})$. For the purpose of the present investigation we set $\theta_{ij} = \theta_i$ ($j = 1, 2, \ldots, n_i$), so that each group of subjects was characterized by a single prevalence parameter θ_i and a common intraperson correlation parameter ρ.

Each experiment in the simulation included a test of H_0: $\theta_1 = \theta_2 = \cdots = \theta_G$ at the .05 level of significance, and was repeated over $I = 500$ independent iterations at each parameter combination. This choice of I assured that a deviation of .02 or more between the empirical significance level for a given procedure and a nominal level of .05 would be statistically significant ($\alpha = .05$, two-tailed). We performed the simulation on a CDC 1035 computer, using a FORTRAN 5 compiler and random number generator GGUBS from the International Mathematical and Statistical Library (IMSL).

One limitation of procedure D is that it cannot be computed if (i) either n_{i0} or $n_{i1} + n_{i2}$ equals 0 for any $i = 1, 2, \ldots, G$; or (ii) either $\sum_{i=1}^{G} n_{i1}$ or $\sum_{i=1}^{G} n_{i2}$ equals 0. We dealt with this problem in the simulation by setting $n_{il} = \frac{1}{2}$ for $n_{il} = 0$ ($i = 1, 2, \ldots, G$; $l = 0, 1, 2$).

4. Results

Table 2 shows the empirical significance levels associated with the statistics T, D, X_A^2, and X^2 for testing H_0: $\theta_1 = \theta_2 = \cdots = \theta_G$ at $\alpha = .05$ for various values of G, $n_i = n$ ($i = 1, 2, \ldots, G$), and ρ. The underlying prevalence rates θ_i are equal to .3 and .5. The significance levels associated with the unadjusted statistic X^2 are presented only to show the effect of ignoring the intraperson dependence on the observed Type I error rate.

It is seen from this table that Rosner's statistic T, Dallal's statistic D, and the adjusted chi-square statistic X_A^2 all give empirical significance levels reasonably close to nominal. As expected, the significance levels associated with the unadjusted chi-square statistic tend to be very much greater than nominal.

Table 3 shows the empirical powers associated with T, D, and X_A^2 for two different configurations of the θ_i at each value of $G = 4, 8$. It is interesting to note that the powers

Table 2
Empirical significance levels for testing H_0: $\theta_1 = \theta_2 = \cdots = \theta_G = \theta$ *corresponding to nominal .05 significance level based on 500 replications*

		$\theta = .3, G = 4$		$\theta = .3, G = 8$		$\theta = .5, G = 4$		$\theta = .5, G = 8$	
ρ		$n = 20$	$n = 40$	$n = 20$	$n = 40$	$n = 20$	$n = 40$	$n = 20$	$n = 40$
.5	T	.052	.046	.052	.052	.064	.054	.056	.070
	D	.064	.064	.054	.068	.056	.062	.058	.066
	X^2_A	.056	.046	.054	.060	.064	.056	.060	.070
	X^2	.160	.190	.232	.260	.144	.190	.224	.248
.7	T	.044	.060	.048	.048	.046	.050	.048	.056
	D	.052	.070	.054	.052	.050	.064	.054	.074
	X^2_A	.048	.062	.050	.048	.046	.050	.052	.056
	X^2	.236	.220	.362	.348	.204	.208	.342	.372
.9	T	.064	.064	.052	.050	.036	.062	.028	.050
	D	.062	.056	.046	.048	.038	.048	.030	.052
	X^2_A	.066	.064	.052	.050	.036	.062	.028	.050
	X^2	.274	.256	.464	.440	.262	.276	.416	.450
.95	T	.060	.052	.054	.048	.044	.054	.036	.038
	D	.054	.048	.050	.044	.048	.044	.046	.042
	X^2_A	.060	.054	.054	.050	.044	.054	.036	.038
	X^2	.290	.266	.496	.454	.258	.262	.448	.468

Table 3
Empirical powers for testing H_0: $\theta_1 = \theta_2 = \cdots = \theta_G = \theta$ *vs* H_{1A}, H_{1B} *corresponding to nominal .05 significance level based on 500 replications*

		H_{1A}*				H_{1B}**			
		$G = 4$		$G = 8$		$G = 4$		$G = 8$	
ρ		$n = 20$	$n = 40$	$n = 20$	$n = 40$	$n = 20$	$n = 40$	$n = 20$	$n = 40$
.5	T	.568	.900	.818	.988	.396	.740	.604	.924
	D	.560	.916	.794	.992	.402	.726	.576	.908
	X^2_A	.572	.902	.820	.988	.398	.742	.616	.926
.7	T	.554	.868	.730	.976	.386	.708	.572	.882
	D	.532	.868	.718	.986	.382	.690	.560	.874
	X^2_A	.562	.868	.736	.976	.392	.708	.578	.882
.9	T	.476	.832	.672	.972	.362	.628	.520	.842
	D	.454	.830	.640	.970	.356	.636	.514	.832
	X^2_A	.484	.834	.676	.972	.370	.628	.520	.846
.95	T	.496	.824	.670	.966	.360	.596	.488	.828
	D	.468	.810	.614	.968	.348	.600	.484	.822
	X^2_A	.498	.824	.670	.966	.360	.596	.488	.828

* H_{1A} prevalence rates ($G = 4$): .1, .2, .3, .4; ($G = 8$): .1, .2, .3, .4, .1, .2, .3, .4.
** H_{1B} prevalence rates ($G = 4$): .2, .2, .4, .4; ($G = 8$): .2, .2, .4, .4, .2, .2, .4, .4.

of the three procedures are generally very similar. However, the powers associated with X^2_A are usually slightly higher than those associated with the other two test statistics. It is also interesting to note that the powers associated with the statistic T tend to be slightly but consistently higher than those associated with D.

5. Discussion

The basic clustering model discussed in Section 2.3 is well known, where alternative distributions for P are the beta, the logistic-normal, and the probit-normal [see Williams

(1988) for a recent review]. The results of the investigation above show that it leads to an effective procedure for comparing several different groups of patients with respect to the prevalence of a specified ocular finding. It tends to be at least as powerful as other specialized procedures proposed for this problem and, as a natural extension of the standard Pearson chi-square test, is perhaps more intuitively attractive.

A very practical advantage of the adjusted chi-square test is that it utilizes information on individuals who supply only one eye to the analysis. Application of the statistic T or D requires that such individuals be excluded from the analysis, which is clearly inefficient unless there is some systematic reason for the missing information. A related advantage of the adjusted chi-square test is that it can be applied to studies in which more than two units of analysis are provided by each individual, as in the case of dental data. Suppose that individual j in group i contributes m_{ij} measurements to the analysis ($j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, G$). Then the total number of observations in the ith group is given by $M_i = \sum_{j=1}^{n_i} m_{ij}$ and the total number of successes by $A_i = \sum_{j=1}^{n_i} a_{ij}$. The expressions for X_A^2 and $\hat{\rho}$ remain as given in Section 2.3, with X_A^2 again referred to tables of the chi-square distribution with $G - 1$ degrees of freedom.

Résumé

Dans les études ophthalmologiques, on compare souvent plusieurs groupes de sujets sur la présence ou l'absence d'une caractéristique oculaire; quand les deux yeux d'un sujet sont examinés, les valeurs de ceux-ci sont très correlées. Rosner (1982, *Biometrics* **38,** 105–114) et Dallal (1988, *Biometrics* **44,** 253–257) ont proposé des procédures pour tester si la proportion d'yeux affectés est la même dans les différents groupes en tenant compte de la corrélation intra-sujet. Dans ce papier, nous proposons une autre approche basée sur un simple ajustement du test du chi-deux standard de Pearson pour l'égalité des proportions. L'approche suggérée utilise l'information des sujets qui contribuent pour un oeil seulement à l'analyse, et généralise facilement à des études où on analyse plus de deux résultats par sujet.

References

Berson, E. L., Rosner, B., and Simonoff, E. (1980). An outpatient population of retinitis pigmentosa and their normal relatives: Risk factors for genetic typing and detection derived from their ocular examinations. *American Journal of Ophthalmology* **89,** 763–765.

Brier, S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika* **67,** 591–596.

Dallal, G. E. (1988). Paired Bernoulli trials. *Biometrics* **44,** 253–257.

Donner, A. (1985). The analysis of intraclass correlation in multiple samples. *Annals of Human Genetics* **49,** 75–82.

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta–binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics* **29,** 637–648.

Le, C. T. (1988). Testing for linear trends in proportions using correlated otolaryngology or ophthalmology data. *Biometrics* **44,** 299–303.

Moore, D. F. (1987). Modelling the extraneous variance in the presence of extra-binomial variation. *Applied Statistics* **36,** 8–14.

Rosner, B. (1982). Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes. *Biometrics* **38,** 105–114.

Rosner, B. (1984). Multivariate methods in ophthalmology with applications to other paired-data situations. *Biometrics* **40,** 1025–1035.

Williams, D. A. (1988). Extra-binomial variation in toxicology. In *Proceedings of the* 14*th International Biometric Conference*, Namur, Belgium, July 1988.

Received April 1988; *revised October* 1988.