The effect of positive dependence on chi-squared tests for categorical data

by

Leon J. Gleser and David S. Moore
Purdue University, USA

Technical Report #84-20

Department of Statistics
Purdue University

July 1984

The effect of positive dependence on chi-squared tests for categorical data

by

Leon J. Gleser and David S. Moore
Purdue University, USA

SUMMARY

We introduce a general definition of positive dependence for finite-state processes, and show that if successive observations are positively dependent, all tests that are asymptotically equivalent to standard Pearson chi-squared tests have asymptotic null distributions stochastically larger than those obtained under the usual independence assumptions. Ignoring positive dependence therefore leads to too frequent rejection of null hypotheses. This qualitative conclusion applies in particular to certain cases of Markov dependence, and of dependence induced by cluster sampling, that have been studied by previous authors.

Keywords: CATEGORICAL DATA; CONTINGENCY TABLES; CLUSTER SAMPLING; MARKOV CHAINS; POSITIVE DEPENDENCE

1. INTRODUCTION

Suppose that $X_1, X_2, \ldots, X_n$ are identically distributed categorical variables taking values $1,2,\ldots,M$. A statistician wishes to test a model that specifies the probabilities $P(X_t = i) = p_i(\theta)$ as functions of an m-dimensional parameter θ. (Alternatively, the model may state restrictions on the p_i, but we will use the equivalent parametric formulation.) A common example is the model of independence in a r×c contingency table, where $M = rc$ and θ consists of $(r-1) + (c-1)$ functionally free marginal probabilities. Such models are commonly tested by the Pearson statistic, estimating θ by an asymptotically efficient estimator based on the M observed cell frequencies. Under mild regularity conditions, the asymptotic null distribution of this statistic is χ^2_{M-m-1}. This theory assumes that the X_t are independent. What will be the effect of serial dependence among the X_t?

A number of authors have considered this question for various models, obtaining explicit results only in rather restricted cases. Altham (1979), Tavaré and Altham (1983) and Tavaré (1983) study the Pearson chi-squared test for independence in two-way contingency tables when the marginal variables Y_{row}, Y_{col} are in fact independent, but each follows a stationary ergodic Markov chain. When the r-state chain Y_{row} and the c-state chain Y_{col} are reversible, the asymptotic null distribution of the Pearson statistic is that of

$$\sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \left(\frac{1+\lambda_i \mu_j}{1-\lambda_i \mu_j} \right) Z_{ij}^2 \qquad (1.1)$$

where Z_{ij} are iid $N(0,1)$ and λ_i, μ_j are the non-unit characteristic roots of the transition matrices of Y_{row} and Y_{col}, respectively.

Another type of dependence is introduced by cluster sampling, as in Cohen (1976), Altham (1976) and Section 5 of Rao and Scott (1981). Rao and Scott (1984) describe the asymptotic behavior of chi-squared tests in loglinear models under general survey designs. Within this context, their asymptotic theory is more general than ours. Rao and Scott study statistics that assume knowledge of the sampling design, while we are concerned with the fate of a naive statistician who ignores serial dependence in analyzing categorical data.

Our goal is to give a general qualitative result: positive dependence among successive observations causes all Pearson-type tests for categorical data to reject a null hypothesis too often. More precisely, the asymptotic null distribution of the statistic is stochastically larger under positive dependence than in the iid case. This result applies not only to tests of models for the cell probabilities, but also to the full versus reduced model tests of nested hypotheses employed in loglinear or other generalized linear models. Though we discuss only chi-squared statistics, all conclusions apply equally to log likelihood ratio and other asymptotically equivalent statistics, a large class of which is described by Cressie and Read (1984). Statistics not in this class are generally used only when (a) model parameters are estimated other than by efficient estimators based on cell frequencies; or (b) the dependence structure is known and the statistic is adjusted to it. Case (a) is common in tests of fit for quantitative variables when estimators based on ungrouped data are available, but is rare in analysis of categorical data. The adjustments required in case (b) in some particular settings are discussed in the literature cited above.

We will employ the following general notion of positive dependence among identically distributed X_t.

Definition 1. The process $\{X_t\}$ is positively dependent (PD) if all of the $M \times M$ joint probability matrices R_{ts} with (i,j)th entry $P(X_t = i, X_s = j)$ satisfy $a'R_{ts}\, a \geq 0$ for all vectors a.

Note that the matrices R_{ts} need not be symmetric. Any square matrix R such that $a'Ra \geq 0$ for all vectors a will be called generalized positive semidefinite (gpsd). Definition 1 is equivalent to $E\{h(X_t)h(X_s)\} \geq 0$ for all functions h, and to $\mathrm{Cov}\{h(X_t), h(X_s)\} \geq 0$ for all h. Gleser and Moore (1983) discuss the definition in these latter forms, and in the exchangeable case (R_{ts} symmetric) relate it to other notions of positive dependence. We shall see in Section 2 that PD arises quite naturally in a study of the behavior of tests for categorical data under serial dependence.

The previously mentioned special case of testing independence in a two-way table whose margins are generated by reversible Markov chains illustrates our results. PD is equivalent in the reversible Markov case to nonnegativity of the characteristic roots of the transition matrix. Under the null nypothesis of independent margins, the $\lambda_i\mu_j$ of (1.1) are among the roots of the transition matrix for the rc-state chain generating the table. Hence (1.1) shows how positive dependence makes the distribution of the Pearson statistic stochastically larger than $\chi^2_{(r-1)(c-1)}$. The size of this effect increases with the strength of the positive dependence, and can be arbitrarily large as $\lambda_i\mu_j \to 1$. Mixed signs among the $\lambda_i\mu_j$ (general dependence) have quite complicated effects on the distribution.

Our methods are based on those used by Moore (1982) and Gleser and Moore (1983) in the case of testing fit of quantitative variables to parametric families of distributions. Section 2 presents the basic asymptotic theory, while Section 3 gives the main result. Section 4 presents applications to Markov dependence and to cluster sampling, and gives some discussion of the practical consequences of our

results. Throughout, we are concerned with the behavior of tests when the model tested is in fact true, but the statistician ignores dependence among successive observations in conducting the test.

2. ASYMPTOTIC RESULTS

Observations $X_1,\ldots,X_n$ on an M-state process yield cell frequencies $N_1,\ldots,N_M$. A model specifies cell probabilities $p(\theta) = (p_1(\theta),\ldots,p_M(\theta))'$ in terms of θ in Ω, an open set in R^m. We assume that the model satisfies the usual conditions first stated by Birch (1964). In particular, when a specific θ_0 in Ω is the true value, it is assumed that the map $p(\theta)$ is totally differentiable at θ_0, that all $p_i(\theta_0) > 0$, and that the $M \times m$ matrix B with (i,j)th entry

$$p_i^{-\frac{1}{2}} \frac{\partial p_i}{\partial \theta_j} \tag{2.1}$$

has full rank $m < M$. (Here and throughout, $\theta = \theta_0$ is assumed when the argument θ is suppressed.) Birch's conditions apply to (in the iid case) multinomial sampling. All of our results hold for the Poisson and product multinomial sampling models as well. For example, the iid case of our (2.6) below appears for these sampling models as (4.174) of Haberman (1974), and the extension to dependent X_t is just as given here.

Birch's conditions imply that when X_t are iid, maximum likelihood estimators $\hat{\theta}_n$ eventually exist and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = (B'B)^{-1}B'V_n + o_p(1), \tag{2.2}$$

where V_n is the M-vector of standardized cell frequencies having ith component $(N_i - np_i)/(np_i)^{\frac{1}{2}}$. Birch also shows that the Pearson statistic $X^2(\hat{\theta}_n) = V_n'(\hat{\theta}_n)V_n(\hat{\theta}_n)$ for testing the model $p(\theta)$ satisfies

$$X^2(\hat{\theta}_n) = V_n'(I - P_B)V_n + o_p(1) \tag{2.3}$$

where $P_B = B(B'B)^{-1} B'$ is the orthogonal projection onto the range of B. Note that the asymptotic behavior of the right sides of (2.2) and (2.3) depends only on that of $V_n = V_n(\theta_0)$.

Suppose now that the X_t are dependent, but that each has a common univariate marginal distribution. We require of the process $\{X_t\}$ only that under θ_0

$$\frac{N_i}{n} \longrightarrow p_i \quad \text{in probability, } i = 1,\ldots,M. \tag{2.4}$$

$$V_n \longrightarrow N(0,\Sigma) \text{ in law}, \qquad \Sigma = \lim_n \text{Cov}(V_n). \tag{2.5}$$

When $\{X_t\}$ is a stationary ergodic (i.e., aperiodic positive recurrent) Markov chain, (2.4) and (2.5) are always satisfied. See Cox and Miller (1965), p. 98 for (2.4) and Doob (1953), p. 228 for (2.5). More generally, if $\{X_t\}$ is a stationary process, the mean-square ergodic theorem states that (2.4) holds if and only if

$$n^{-1} \sum_{t=0}^{n-1} \{P(X_1 = i, X_{1+t} = i) - p_i^2\} \to 0.$$

This is a mixing condition on $\{X_t\}$. For such processes, central limit theorems

for stationary processes then imply (2.5) under a variety of stronger conditions. See e.g. Withers (1981) for references.

An examination of the proof in Birch (1964) shows that (2.4) and (2.5) are sufficient for (2.2) and (2.3). Consequently, the representations (2.2), (2.3) hold under our assumptions for both the iid and dependent cases. The resulting asymptotic distributions will differ because the limiting covariance matrix Σ of (2.5) will reflect dependence among the X_t. In fact, computation shows that

$$\text{Cov}\ (V_n) = I - qq' + Q_n$$

where $q = (p_1^{\frac{1}{2}}, \ldots, p_M^{\frac{1}{2}})'$, I is the identity matrix, and

$$Q_n = \frac{1}{n} \sum_{\substack{t,s=1 \\ t \neq s}}^{n} Q_{ts}$$

for Q_{ts} the $M \times M$ matrix with (i,j)th entry

$$\{P(X_t = i,\ X_s = j) - p_i p_j\} / (p_i p_j)^{\frac{1}{2}}.$$

Hence, $\Sigma = I - qq' + Q$, where $Q = \lim Q_n$. Note that $q'B = q'Q_{ts} = 0$, and that in the iid case all $Q_{ts} = 0$.

Consider next the testing of a reduced model specifying that $\theta = g(\tau)$ for τ in an open set Γ of R^q, $q < m$. If both the reduced model $p_R(\tau) = p\{g(\tau)\}$ and $p(\theta)$ satisfy Birch's conditions, and $g: \Gamma \to \Omega$ is totally differentiable at the τ_0 such that $\theta_0 = g(\tau_0)$, then the B-matrix (2.1) in terms of τ at τ_0 satisfies $B_\tau = B\Delta$, where Δ is the $m \times q$ matrix of derivatives $\partial\theta_i/\partial\tau_j$. Since the columns of B_τ are in the range of B, the projection $P_R = B_\tau (B_\tau' B_\tau)^{-1} B_\tau'$ satisfies $P_R P_B = P_B P_R = P_R$. Thus $P_B - P_R$ is an orthogonal projection of rank $m - q$.

The statistic for testing the reduced model $p_R(\tau)$ versus the full model $p(\theta)$ is (in one of several equivalent forms),

$$X_R^2 = X^2(\hat{\tau}_n) - X^2(\hat{\theta}_n)$$

where $\hat{\tau}_n$ is the mle or asymptotically equivalent estimator of τ from $N_1,\ldots,N_M$ under the reduced model. By (2.3) (note that V_n involves no estimation, so is the same in both models),

$$X_R^2 = V_n' (P_B - P_R)V_n + o_p(1) \tag{2.6}$$

when the reduced model is true. The expression (2.6) includes (2.3) as the special case in which the full model does not constrain the p_i, and $p(\theta)$ defines the reduced model. Another familiar example is the case of loglinear models, $\log p(\theta) = A\theta$ for an $M \times m$ matrix A of rank m. Reduced models set some components of θ (say the last $m - q$) to zero, so that $\theta = (\tau, \xi)$ and $g(\tau) = (\tau, 0)$, where τ has dimension q, and $\xi, 0$ have dimension $m - q$. Other generalized linear models $\phi\{p(\theta)\} = A\theta$ are also covered by our formulation.

These results are familiar in the iid case. We have remarked that they hold much more generally. Here is a summary statement.

Theorem 1. Suppose that $\{X_t\}$ is any process such that X_t are identically distributed and (2.4), (2.5) hold. Suppose also that a full model $p(\theta)$ and a reduced model $p_R(\tau) = p\{g(\tau)\}$ both satisfy the regularity conditions of Birch (1964). When θ, τ are estimated by estimators that are asymptotically efficient for $\{X_t\}$ iid, the Pearson statistic for testing p_R versus p satisfies

$$X_R^2 = V_n'(P_B - P_R)V_n + o_p(1)$$

when the reduced model is true and $\tau = \tau_0$.

3. POSITIVE DEPENDENCE

The framework of Section 2 can be used to study the effect of serial dependence on the asymptotic null distribution of tests for categorical data in considerable generality. Here we are interested in the effect of positive dependence in the sense of Definition 1.

Let $D = \text{diag}\,(p_1,\ldots,p_M)$. Then from

$$Q_{ts} = D^{-\frac{1}{2}} R_{ts} D^{-\frac{1}{2}} - qq'$$

and $q'Q_{ts} = 0$, it follows that $Q_{ts} + Q_{st}$ (which always has characteristic root 0 in the qq' direction) is gpsd if and only if R_{ts} is. Hence if $\{X_t\}$ is PD, Q_n and Q (which are symmetric) are psd.

It follows from (2.6) and the asymptotic normality of V_n that

$$X_R^2 \longrightarrow \sum_{i=1}^{M} \lambda_i Z_i^2 \quad \text{in law,}$$

where Z_i are iid $N(0,1)$ and $\lambda_i \equiv \lambda_i(W)$ are the characteristic roots of

$$W = (P_B - P_R)^{\frac{1}{2}} (I - qq' + Q) (P_B - P_R)^{\frac{1}{2}} . \tag{3.1}$$

Setting $Q = 0$ in (3.1) defines the iid case of this matrix, which we call W_{IID}. Since $W - W_{IID}$ is psd if Q is, $\{X_t\}$ PD implies that $\lambda_k(W) \geq \lambda_k(W_{IID})$, where $\lambda_k(E)$ is the kth largest characteristic root of the matrix E. (Of course, $W_{IID} = P_B - P_R$, so that its nonzero roots are $m - q$ 1's and X_R^2 is asymptotically χ^2_{m-q}.) We have proved our main result.

Theorem 2. If $\{X_t\}$ is PD and the conditions of Theorem 1 hold, the asymptotic null distribution of any Pearson-type statistic X_R^2 is stochastically larger

than when $\{X_t\}$ is iid.

4. EXAMPLES AND DISCUSSION

Markov dependence. Suppose $\{X_t\}$ to be a stationary ergodic Markov chain with transition matrix T and vector of stationary probabilities p. Then computation shows that

$$Q = D^{\frac{1}{2}}(Z - I)D^{-\frac{1}{2}} + D^{-\frac{1}{2}}(Z' - I)D^{\frac{1}{2}}, \tag{4.1}$$

where $Z^{-1} = I - (T-ep')$, e a vector of 1's, in agreement with Tavaré and Altham. The concept of PD for Markov chains is investigated in detail in Gleser and Moore (1985). If $r_{ij} = P(X_1 = i, X_{1+k} = j)$, a necessary condition for PD is that $r_{ii} \geq p_i^2$ for all i and k, and a sufficient condition is that $r_{ij} + r_{ji} \leq 2p_ip_j$ for all $i \neq j$ and all k. These conditions have obvious interpretations relating the joint distribution of (X_1, X_{1+k}) under PD to that under independence.

When the chain is reversible (i.e., R_{ts} is symmetric), PD is equivalent to Q psd and to nonnegativity of all characteristic roots of T. Thus in the reversible case (which includes all 2-state chains), $\{X_t\}$ PD has a natural description in terms of T and is also the natural condition (Q psd) for all chi-squared statistics to be stochastically larger than for iid observations. The nonreversible case, which appears to be more common than reversible chains in scientific models with more than two states, is more complex. Counterexamples shows that neither of the equivalences mentioned above holds for nonreversible chains. But $\{X_t\}$ PD continues to imply that Q is psd and hence that Pearson-type tests can be misleading.

Cluster sampling. To fit cluster sampling into the serial dependence framework, we assume that X_t's from the same cluster are adjacent in the sequence. Suppose there are N clusters, of sizes M_ν for $\nu = 1,\ldots,N$. Setting $Z_{i\nu t} = 1$ if $X_t = i$ where X_t is the t-th observation from the νth cluster, we assume that $Z_{i\nu t}$ from different clusters are independent, that $E(Z_{i\nu t}) = p_i(\theta)$ for all ν,t, and that the matrix

$$\{E(Z_{i\nu t}Z_{j\nu s})\}_{i,j=1,\ldots,M} = R_\nu,\ t \neq s.$$

This is the model of Altham (1976), generalized to allow the matrices R_ν, which describe within-cluster dependence, to vary among clusters. Under this model, the joint probability matrices R_{ts} of Definition 1 are the same for all X_t, X_s in the same cluster (in particular, R_{ts} is symmetric), while X_t, X_s are independent if in different clusters.

In this model for cluster sampling, $\{X_t\}$ is PD if and only if all R_ν are psd. The models for positive dependence within clusters proposed by Altham and others, and reviewed in Section 5.2 of Rao and Scott (1981), are all special cases of this natural definition. It follows that Pearson statistics are stochastically larger under PD clustering than in the iid case whenever a central limit theorem for the N_i for the specified design is available. As in the Markov case, a usable null distribution is available only in restricted cases. Extending (5.3) of Rao and Scott (1981),

$$\text{Cov}\,(V_n) = I - qq' + \sum_{\nu=1}^{N} \left\{\frac{M_\nu(M_\nu - 1)}{n}\right\} D^{-\frac{1}{2}} R_\nu D^{-\frac{1}{2}}.$$

The limit of Q_n (the last term on the right) will rarely be tractable.

Rao and Scott (1984) study a more general class of statistics for survey data under regularity conditions similar to ours. Our naive statistician ignores

serial dependence; theirs is aware of the sampling design, and uses an estimator $\hat{p}$ of p based on this knowledge, rather than our N_i/n of (2.4). Rao and Scott give most attention to the use of such $\hat{p}$ in standard chi-squared statistics, identical to our X_R^2 if $\hat{p}_i = N_i/n$. These statistics are not properly standardized relative to the asymptotic covariance matrix of $\hat{p}$, and do not in general have χ^2 asymptotic null distributions. It is not surprising that the resulting distribution is stochastically larger than the iid-case distribution of X_R^2 when the limiting covariance matrix V of $\hat{p}$ is larger than the iid-case covariance matrix C of the N_i/n, in the sense that V-C is psd. This follows from Theorem 1 of Rao and Scott (1984). In principle, this result includes our cluster sampling example. But it does not lend itself to specifying general conditions such as PD that cause true null hypotheses to be rejected too often.

Discussion. In the light of the conclusions of this and earlier studies, statisticians who suspect serial dependence in their data should conduct a test for such dependence prior to applying standard procedures for categorical data analysis. When a finite Markov chain model is appropriate, Chatfield (1973) reviews tests for independence of successive events and Katz (1981) provides references to some more recent work. More general nonparametric tests for randomness include the classical runs tests and, in some circumstances, rank tests such as that proposed by Bartels (1982).

The analysis of data when serial dependence has been verified is a topic requiring further study. In the case of dependence induced by a sampling design, alternative analyses are available. The weighted least squares approach based on the Wald statistic (Koch, Freeman and Freeman, 1975) is well known, and Rao and Scott (1981, 1984) propose other statistics. Thomas and Rao (1984) have studied

the small-sample properties of a number of tests in this context. The development of procedures for dealing with more general dependence, e.g. in a time series, is at a more primitive stage. Distributions of classical statistics can be explicitly calculated only in a few special cases, such as that leading to (1.1).

It is also sometimes possible to explicitly compute the "true" statistic, i.e. the quadratic form in the standardized cell frequencies $V_n(\hat{\theta}_n)$ whose centering matrix is a generalized inverse of the limiting covariance matrix of the $V_n(\hat{\theta}_n)$. When testing fit to a completely specified model (θ_0 known), Σ in (2.5) is $I-qq' + Q$. If $I + Q$ is nonsingular, which is generally the case, $(I+Q)^{-1}$ is a generalized inverse of Σ. Hence when $\hat{Q}$ is a consistent estimator of Q, $V_n' (I+\hat{Q})^{-1} V_n$ has the χ^2_{M-1} limiting null distribution. In the Markov case, Q is given by (4.1) and can in principle be consistently estimated by employing the obvious count estimators of stationary and transition probabilities. Altham (1979) has given upper and lower bounds on the "true" statistic that are quite generally true and will often be useful.

ACKNOWLEDGEMENTS

This work was supported by NSF Grant DMS-8121948.

REFERENCES

Altham, P.M.E. (1976) Discrete variable analysis for individuals grouped into families. Biometrika, 63, 263-269.

Altham, P.M.E. (1979) Detecting relationships between categorical data observed over time: A problem of deflating a χ^2 statistic. Appl. Statist., 28, 115-125.

Bartels, R. (1982) The rank version of von Neumann's ratio test for randomness. J. Amer. Statist. Assoc., 77, 40-46.

Birch, M.W. (1964) A new proof of the Pearson-Fisher theorem. Ann. Math. Statist., 35, 817-824.

Chatfield, C. (1973) Statistical inference regarding Markov chain models. Appl. Statist., 22, 7-20.

Cohen, J.E. (1976) The distribution of the chi-squared statistic under cluster sampling from contingency tables. J. Amer. Statist. Assoc., 71, 665-670.

Cox, D.R. and Miller, H.D. (1965) The Theory of Stochastic Processes. New York: Wiley.

Cressie, N. and Read, T.R.C. (1984) Multinomial goodness-of-fit tests. J.R. Statist. Soc. B, 46.

Doob, J.L. (1953) Stochastic Processes. New York: Wiley.

Gleser, L.J. and Moore, D.S. (1983) The effect of dependence on chi-squared and empiric distribution tests of fit. Ann. Statist., 11, 1100-1108.

Gleser, L.J. and Moore, D.S. (1985) Positive dependence in Markov chains. J. Lin. Alg. Appl., to appear.

Haberman, S.J. (1974) The Analysis of Frequency Data. Chicago: University of Chicago Press.

Katz, R.W. (1981) On some criteria for estimating the order of a Markov chain. Technometrics, 23, 243-249.

Koch, G.G., Freeman, D.H. Jr. and Freeman, J.L. (1975) Strategies in the multivariate analysis of data from complex surveys. Intl. Statist. Rev., 43, 59-78.

Moore, D.S. (1982) The effect of dependence on chi-squared tests of fit. Ann. Statist., 10, 1163-1171.

Rao, J.N.K. and Scott, A.J. (1981) The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. J. Amer. Statist. Assoc., 76, 221-230.

Rao, N.N.K. and Scott, A.J. (1984) On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. Ann. Statist., 12, 46-60.

Tavaré, S. (1983) Serial dependence in contingency tables. J. R. Statist. Soc. B, 45, 100-106.

Tavaré, S. and Altham, P.M.E. (1983) Dependence in goodness of fit and contingency tables. Biometrika, 70, 139-144.

Thomas, D.R. and Rao, J.N.K. (1984) A monte carlo study of exact levels for chi-squared goodness-of-fit statistics under cluster sampling. Technical Report No. 35, Carleton University Laboratory for Research in Statistics and Probability.

Withers, C.S. (1981) Central limit theorems for dependent variables. Z. Wahrscheinlichkeitstheorie verw. Geb., 57, 509-534.