The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data

Peter McCullagh

Journal of the American Statistical Association, Volume 81, Issue 393 (Mar., 1986), 104-107.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198603%2981%3A393%3C104%3ATCDOGS%3E2.0.CO%3B2-E

The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data

PETER McCULLAGH*

I consider the distribution of Pearson's statistic and of the likelihood-ratio goodness-of-fit statistic for discrete data in the important case where the data are extensive but sparse. It is argued that the appropriate reference distribution is conditional on the sufficient statistic for the unknown regression parameters, $\boldsymbol{\beta}$. The first three conditional asymptotic cumulants are derived by Edgeworth expansion, and these are used for the computation of tail probabilities. The principal advantage of the limit considered here, as opposed to the more usual χ^2 limit, is that the cell counts need not be large.

KEY WORDS: Conditional inference; Cumulants; Edgeworth approximation; Log-linear model; Linear logistic model; Sparse data.

1. INTRODUCTION

The object of this article is to derive, either exactly or approximately, the conditional distributions of certain commonly used goodness-of-fit statistics for discrete data in the important case where the data are sparse. For arbitrary log-linear or linear logistic models, exact specification of the conditional distribution is usually not feasible. For that reason, I concentrate on asymptotic approximations appropriate in the limit $n \to \infty$, as opposed to the more usual limit $\mu_i \to \infty$ (n fixed), where $\mu_i = E(Y_i)$ is the mean of the ith observation and $i = 1, \ldots, n$. In the context of contingency tables, n is the number of cells and μ_i is the mean count for cell i.

A number of authors have discussed the possibility of using normal rather than χ^2 approximations when the degrees of freedom are large (e.g., Morris 1975, Koehler and Larntz 1980, and Fienberg 1979). Although I do use normal and Edgeworth approximations, the main emphasis in this article is different in two important respects. First, I consider specifically the case in which there are unknown parameters to be estimated. Second, I argue that it is the conditional distribution of the statistic and not its marginal distribution that is relevant for assessing goodness of fit. Both of these are typically asymptotically normal but, as shown in Section 4, the difference between the two can be very large indeed. If there are no unknown parameters, the conditioning statistic disappears and we are left with the marginal distribution.

Out of the many possible goodness-of-fit statistics, I concentrate on the two most commonly used in applications—namely, the deviance

$$D = \sum \{2Y_i \log(Y_i/\hat{\mu}_i) - 2(Y_i - \hat{\mu}_i)\} \tag{1}$$

and Pearson's statistic

$$P = \sum (Y_i - \hat{\mu}_i)^2/\hat{\mu}_i. \tag{2}$$

In all calculations, I take $\mathbf{Y}$ to be a vector of length n, $\boldsymbol{\beta}$ to be an unknown parameter vector of length p, and μ to be a function of $\boldsymbol{\beta}$ specified by the model, either log-linear or linear logistic as appropriate. The appropriate conditioning variable to remove distributional dependence on $\boldsymbol{\beta}$ is the sufficient statistic for $\boldsymbol{\beta}$. More generally, if the sufficient statistic is not complete, it would be appropriate to condition on $\hat{\boldsymbol{\beta}}$, the maximum likelihood estimate of $\boldsymbol{\beta}$.

When there are no unknown parameters, or where the appropriate conditional distribution has a simple form, simulation can be used to determine the approximate null distributions of P and D. A number of authors have in fact made extensive Monte Carlo investigations of the marginal distributions of P and D for sparse data. When unknown parameters are present, however, it is usually not easy to simulate from the conditional distribution; unconditional simulation, on the other hand, is not relevant for significance testing and can give misleading answers, as shown in Section 3. The objective of this article is to give an approximate analytical solution, thereby avoiding the need for conditional simulation.

Section 2 is devoted to log-linear models for Poisson responses, and in Section 3 I consider binomial responses. Binary data are considered in Section 4.

It should be emphasized at the outset that I do not necessarily recommend either (1) or (2) as useful statistics for detecting lack of fit when the data are sparse. Indeed, it will be seen that D, in particular, often has very little diagnostic power and, in the case of binary data, D has no diagnostic power whatsoever. In this respect, the conclusions of this article contradict Cochran (1952, sec. 14), who suspected that in small samples, the likelihood-ratio test would be more powerful than Pearson's test.

The emphasis in this article is not on the details of regularity conditions but on obtaining usable results under reasonable assumptions when the data are extensive but sparse. By this I mean that the number of independent observations or "cells" is large but that, typically, each cell contributes only a small amount of information to the total. More precisely, I assume that p is fixed as $n \to \infty$ and that the maximum likelihood estimate of $\boldsymbol{\beta}$ is consistent; that is, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + O_p(n^{-1/2})$. For log-linear and linear logistic models, these conditions imply, in essence, that $\text{var}(Y_i)$ must be bounded away from zero as $n \to \infty$.

2. LOG-LINEAR MODELS

Suppose now that the components of $\mathbf{Y}$ are independent Poisson random variables with mean $E(Y_i) = \mu_i$. Writing $\eta_i = \log \mu_i$, the log-linear model becomes

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \tag{3}$$

* Peter McCullagh is Professor, Department of Statistics, University of Chicago, Chicago, IL 60637.

Journal of the American Statistical Association
March 1986, Vol. 81, No. 393, Theory and Methods

where $\boldsymbol{\eta}$ is a $n \times 1$ vector and $\mathbf{X}$ is a full-rank model matrix of known constants. The sufficient statistic $\mathbf{S} = \mathbf{X}^T\mathbf{Y}$, of dimension p, is complete, and the maximum likelihood estimate satisfies $\mathbf{X}^T\hat{\boldsymbol{\mu}} = \mathbf{S}$, where $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mu}_i = \exp(\hat{\eta}_i)$.

I consider first the conditional distribution of D given $\mathbf{S}$. Now, D may be decomposed in much the same way as a sum of squares, into the two quantities

$$\begin{aligned} D &= 2\sum\{Y_i\log(Y_i/\mu_i) - (Y_i - \mu_i)\} \\ &\quad - 2\sum\{\hat{\mu}_i\log(\hat{\mu}_i/\mu_i) - (\hat{\mu}_i - \mu_i)\} \\ &= \sum d(Y_i; \mu_i) - \sum d(\hat{\mu}_i; \mu_i). \end{aligned}$$

Under (3), the second term above is $O_p(1)$ and conditionally constant: its unconditional distribution is $\chi^2_p + O_p(n^{-1})$. Thus we require the conditional distribution of $\sum d(Y_i; \mu_i)$ given $\mathbf{S}$, at least approximately, for large n. By the central limit theorem, the joint distribution of $\sum d(Y_i; \mu_i)$ and $\mathbf{S}$ is $(p + 1)$-variate normal, the covariance matrix of $\mathbf{S}$ being $\mathbf{X}^T\mathbf{V}\mathbf{X}$, where $\mathbf{V} = \operatorname{diag}\{\mu_1, \ldots, \mu_n\}$. Writing $d_i = d(Y_i; \mu_i)$, $\kappa_1^{(i)} = E(d_i)$, $\kappa_2^{(i)} = \operatorname{var}(d_i)$, and $\kappa_{11}^{(i)} = \operatorname{cov}(d_i, Y_i) = d\kappa_1^{(i)}/d\eta_i$, the unconditional mean and variance of $\sum d_i$ may be written as $\kappa_1^{(\cdot)}$ and $\kappa_2^{(\cdot)}$, both involving sums over the components of the mean vector $\boldsymbol{\mu}$. The covariance of $\sum d_i$ and $\mathbf{S}$, expressed as a column vector, is $\mathbf{X}^T\boldsymbol{\kappa}_{11}$, where $\boldsymbol{\kappa}_{11}$ is a vector of length n with elements $\kappa_{11}^{(i)}$. To first order, it follows that, conditionally on $\mathbf{S}$, $\sum d_i$ is asymptotically normal with conditional mean

$$\kappa_1^{(\cdot)} + \boldsymbol{\kappa}_{11}^T\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}(\mathbf{S} - \mathbf{X}^T\boldsymbol{\mu}) + O(1) = \hat{\kappa}_1^{(\cdot)} + O(1) \quad (4)$$

and conditional variance

$$\kappa_2^{(\cdot)} - \boldsymbol{\kappa}_{11}^T\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\kappa}_{11} + O(n^{1/2}). \quad (5)$$

Thus the standardized statistic

$$\{D - \hat{\kappa}_1^{(\cdot)}\}/\{\hat{\kappa}_2^{(\cdot)} - \hat{\boldsymbol{\kappa}}_{11}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\kappa}}_{11}\}^{1/2} \quad (6)$$

is conditionally $N(0, 1) + O_p(n^{-1/2})$ under H_o, large values being taken as evidence against H_o.

Table 1 gives values of the cumulants κ_1, κ_2, and κ_{11} for various small to moderate values of μ. For moderate to large values of μ we have $\kappa_{11} \to 0$, $\kappa_1 \to 1$, $\kappa_2 \to 2$ so that the mean and variance of D obtained from (4) and (5) are $n + O(1)$ and $2n + O(n^{1/2})$, respectively, and D is, to this order of approximation, independent of $\mathbf{S}$. Thus (4) and (5) are consistent with the usual chi-squared approximation based on the limit $\mu_i \to \infty$.

Table 1. Cumulants of d and Y for Poisson Distribution

$\mu = \kappa_{01}$	κ_1	κ_2	κ_3	κ_{11}	κ_{21}	κ_{12}
.1	.4741	.8604	3.6638	.2874	1.0408	.3135
.2	.6970	.7293	3.3595	.3481	1.1674	.4462
.5	1.0070	.7297	3.0025	.2912	1.5969	.8056
.8	1.1173	1.0913	2.6095	.1685	1.7746	1.2838
1.0	1.1468	1.3646	2.6567	.0950	1.7077	1.6633
1.5	1.1585	1.9185	3.8919	−.0328	1.1574	2.7298
2.0	1.1394	2.2330	5.8146	−.0940	.4733	3.8523
3.0	1.0946	2.4057	8.7875	−.1143	−.4404	6.0387
4.0	1.0645	2.3515	9.8608	−.0918	−.7063	8.1045
5.0	1.0467	2.2668	9.8813	−.0677	−.6620	10.1054
10.0	1.0188	2.0877	8.6387	−.0219	−.2123	20.0310
20.0	1.0088	2.0373	8.2380	−.0093	−.0793	40.0105
$\mu \to \infty$	$1 + (6\mu)^{-1}$	$\sim 2\kappa_1^2$	$\sim 8\kappa_1^3$	$\sim -(6\mu)^{-1}$	$\sim -4/(3\mu)$	$\sim 2\mu$

3. BINOMIAL DATA

Suppose now that the observations are independent and binomially distributed so that Y_i has parameter π_i and index m_i, which need not be large. It is required to test the adequacy of the linear logistic model $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is taken as unknown and $\eta_i = \log\{\pi_i/(1 - \pi_i)\}$. The test statistics D and P may be written as in (1) and (2) but where the sum extends over all $2n$ "cells"; that is, the sum is over "successes" as well as "failures." It is more convenient, however, to use the expressions

$$\begin{aligned} D = 2\sum_i [&Y_i\log(Y_i/\hat{\mu}_i) \\ &+ (m_i - Y_i)\log\{(m_i - Y_i)/(m_i - \hat{\mu}_i)\}] \end{aligned}$$

and $P = \sum_i(Y_i - \hat{\mu}_i)^2/\{m_i\hat{\pi}_i(1 - \hat{\pi}_i)\}$, with summation from 1 to n.

It is well known that for the purposes of fitting, testing, and estimation, linear logistic models can be considered as a special case of log-linear models. It is not possible to adapt the asymptotic results of Section 2 to linear logistic models, however, because the limit considered here would, if formulated in the log-linear framework, involve models with large numbers of parameters. The regularity conditions assumed in Section 1 would then be violated.

I consider first the asymptotic distribution of D, which, for convenience, I decompose into the two components

$$D = \sum d(Y_i; \mu_i) - \sum d(\hat{\mu}_i; \mu_i), \quad (7)$$

where $d(Y_i; \mu_i) = 2Y_i\log(Y_i/\mu_i) + 2(m_i - Y_i)\log\{(m_i - Y_i)/(m_i - \mu_i)\} = d_i$. The second component in (7) is the log-likelihood ratio statistic for testing a simple null hypothesis against the linear logistic alternative. In other words, the second component is unconditionally approximately χ^2_p, but conditionally it is a constant that is $O(1)$. The algebra of Section 2 may now be applied with the newly defined cumulants $\kappa_1^{(i)} = E(d_i)$, $\kappa_2^{(i)} = \operatorname{var}(d_i)$, $\kappa_{11}^{(i)} = \operatorname{cov}(Y_i, d_i) = d\kappa_1^{(i)}/d\eta_i$, and so on. Evidently, these newly defined cumulants depend on both m_i and π_i. The standardized statistic

$$\{D - \hat{\kappa}_1^{(\cdot)}\}/\{\hat{\kappa}_2^{(\cdot)} - \hat{\boldsymbol{\kappa}}_{11}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\kappa}}_{11}\}^{1/2}, \quad (8)$$

where $\mathbf{V} = \operatorname{diag}\{m_i\pi_i(1 - \pi_i)\}$, is again conditionally approximately standard normal, provided only that the denominator is not zero (see Sec. 4). Large values of (8) would be taken as evidence against the linear logistic model.

4. BINARY DATA

I now take the problem considered in Section 3 and assume that $m_i = 1$ for all i. Then we find

$$\kappa_1^{(i)} = -2\{\pi_i\log\pi_i + (1 - \pi_i)\log(1 - \pi_i)\}$$

$$\kappa_2^{(i)} = 4\pi_i(1 - \pi_i)\log^2\{\pi_i/(1 - \pi_i)\} = 4\eta_i^2V_i$$

$$\kappa_{11}^{(i)} = -2\pi_i(1 - \pi_i)\log\{\pi_i/(1 - \pi_i)\} = -2\eta_iV_i.$$

The conditional variance (5) becomes

$$4\boldsymbol{\eta}^T\mathbf{V}\boldsymbol{\eta} - 4\boldsymbol{\eta}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\boldsymbol{\eta},$$

and since $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, this is identically zero. In other words, D

is conditionally degenerate at the point $-2\sum\{\hat{\pi}_i \log \hat{\pi}_i + (1-\hat{\pi}_i)\log(1-\hat{\pi}_i)\}$ and contains no information regarding lack of fit of the linear logistic model (Williams 1983). A similar result applies as an approximation in the case of log-linear models if all of the means are small and the constant vector, $\mathbf{1}$, lies in the column space of $\mathbf{X}$.

The preceding result shows that the total deviance is a function of the sufficient statistic and is therefore uninformative regarding lack of fit. The same is true of certain subtotals of the deviance in certain circumstances. Consider, for example, the statistic $T = \sum_i c_i d_i$. A straightforward calculation shows that if the vector with elements $c_i\eta_i$ lies in the column space of $\mathbf{X}$, then T is conditionally degenerate; that is, T is a function of the sufficient statistic. For example, if $\mathbf{X}$ is the incidence matrix for a one-way layout with k blocks, the k subtotals of D, one for each block, are all conditionally degenerate.

Note that if we were to use the unconditional asymptotic distribution of D, this would be based on $D \sim N(\kappa_1^{(\cdot)}, 4\boldsymbol{\eta}^T\mathbf{V}\boldsymbol{\eta})$, where both of the asymptotic moments are functions of the unknown $\boldsymbol{\beta}$. If we substitute $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and use $N(\hat{\kappa}_1^{(\cdot)}, 4\hat{\boldsymbol{\eta}}^T\hat{\mathbf{V}}\hat{\boldsymbol{\eta}})$ as an approximate reference distribution, we have identically that $D \equiv \hat{\kappa}_1^{(\cdot)}$, giving a tail probability of $\frac{1}{2}$ independently of the data.

5. PEARSON'S STATISTIC

Using the summation convention, Pearson's statistic may be written as

$$P = (Y_i - \hat{\mu}_i)(Y_j - \hat{\mu}_j)\hat{V}^{ij},$$

where $\mathbf{V}^{-1} = \operatorname{diag}\{V^{-1}(\mu_i)\}$ is the inverse covariance matrix of $\mathbf{Y}$. A simple rearrangement leads to

$$P = (Y_i - \mu_i)(Y_j - \mu_j)\hat{V}^{ij} - 2(\hat{\mu}_i - \mu_i)(Y_j - \hat{\mu}_j)\hat{V}^{ij} - (\hat{\mu}_i - \mu_i)(\hat{\mu}_j - \mu_j)\hat{V}^{ij}.$$

The leading term is $O(n) + O_p(n^{1/2})$; the second term is $O_p(1)$ in view of the likelihood equations $\mathbf{X}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = 0$; the third term is $O_p(1)$ marginally and $O(1)$ conditionally. Now write

$$\hat{V}^{ij} = V^{ij} + V^{ijk}(\hat{\eta}_k - \eta_k) + O(n^{-1}) = V^{ij} + V^{ijk}a_k^l(Y_l - \mu_l) + O(n^{-1}),$$

where $V^{ijk} = dV^{ij}/d\eta_k$ is zero unless $i = j = k$ and $\mathbf{A} = \{a_j^i\} = \mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T$ is a symmetric $n \times n$ matrix. Thus, putting $Z_i = Y_i - \mu_i$, we have

$$P = Z_iZ_jV^{ij} + V^{ijl}a_l^kZ_iZ_jZ_k + O_p(1).$$

Using the results of McCullagh (1984), we find the unconditional moments of P to be

$$E(P) = n + O(1) \tag{9}$$

and

$$\operatorname{var}(P) = 2\sum_i\left(\frac{m_i - 1}{m_i}\right) + \mathbf{C}^T(\mathbf{V}^{-1} - \mathbf{A})\mathbf{C} + O(n^{1/2}), \tag{10}$$

where C_i is the ratio of the third to the second cumulant of Y_i. The limit $m_i \to \infty$ in (10) gives the variance for the Poisson case. A similar calculation for the covariance of S_j and P gives $\operatorname{cov}(S_j, P) = O(1)$ rather than $O(n)$ as was the case for the likelihood-ratio statistic. In other words, to first order in n, P and $\mathbf{S}$ are independent, the conditional variance being obtained by inserting consistent estimates in (10).

In the case of single samples and one-way layouts, we have $\mathbf{C}^T(\mathbf{V}^{-1} - \mathbf{A})\mathbf{C} = 0$ so that $\operatorname{var}(P) = 2\sum(m_i - 1)/m_i$. For Poisson observations this takes the value $2n$, whereas for binary data, $\operatorname{var}(P) = 0$ as might be expected. The quantity $2\sum(m_i - 1)/m_i$ is the leading term in the exact expression for $\operatorname{var}(P)$ derived by Haldane (1939). The error term is in fact $O(1)$ rather than $O(n^{1/2})$.

More precise calculations, correct conditionally to second order, were given by McCullagh (1985) for arbitrary linear exponential-family models. Moreover, a simple algorithm was given for computing the first three conditional cumulants, and this can be applied routinely regardless of the complexity of the model matrix.

6. SECOND-ORDER CALCULATIONS

To improve on the first-order normal-theory approximations to the distributions of D and P, it is necessary to compute the conditional mean up to and including terms that are $O(1)$ and the conditional variance up to and including terms that are $O(n^{1/2})$. A second-order correction involving the conditional skewness is applied using Edgeworth expansion. The formulas necessary for computing conditional cumulants were given by McCullagh (1984, sec. 6.1). Only the final result of such calculations is given here.

After rather lengthy calculations involving expansion of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ up to the quadratic term in $\mathbf{S} - \mathbf{X}^T\boldsymbol{\mu}$, the conditional expectation of D after linear and quadratic regression on $\mathbf{S}$ may be shown to be

$$E(D \mid \mathbf{S}) = \hat{\kappa}_1^{(\cdot)} - \tfrac{1}{2}\mathbf{1}^T\mathbf{X}^T\hat{\boldsymbol{\Sigma}}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{1} + O(n^{-1}), \tag{11}$$

where $\boldsymbol{\Sigma} = \operatorname{diag}\{\kappa_{12}^{(i)} - \kappa_{11}^{(i)}\kappa_{03}^{(i)}/\kappa_{02}^{(i)}\}$. In the case of the Poisson distribution, $\kappa_{03}^{(i)} = \kappa_{02}^{(i)} = \mu_i$ and the remaining cumulants are given in Table 1. For large μ_i we have $\kappa_{12}^{(i)} - \kappa_{11}^{(i)} \to 2V_i$ so that the $O(1)$ correction in (11) tends to p, as is to be expected on the basis of the χ^2 approximation. The conditional variance of D is $\operatorname{var}(D \mid \mathbf{S}) = \hat{\kappa}_2^{(\cdot)} - \hat{\boldsymbol{\kappa}}_{11}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\kappa}}_{11} + \varepsilon + O(n^{-1})$, where ε is a rather complicated expression involving $\mathbf{X}$ and the fourth-order cumulants κ_{22}, κ_{13}, and κ_{04}, as well as lower-order cumulants. In the simplest case, where the observations constitute a simple random sample, we find that

$$2\varepsilon = (\kappa_{22} - 2\gamma\kappa_{13} + \gamma^2\kappa_{04})/V + (\kappa_{12}^2 + \kappa_{21}\kappa_{03} - 4\gamma\kappa_{12}\kappa_{03} + 2\gamma^2\kappa_{03}^2)/V^2,$$

where $\gamma = \kappa_{11}/V$ and, in general $\varepsilon \to 2p$ as $\mu \to \infty$. For small degrees of freedom it is often advantageous to use the crude approximation

$$\operatorname{var}(D \mid \mathbf{S}) \simeq (1 - p/n)\{\hat{\kappa}_2^{(\cdot)} - \hat{\boldsymbol{\kappa}}_{11}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\kappa}}_{11}\}, \tag{12}$$

which has an error that is $O(1)$ as $n \to \infty$. For fixed n, the error in (12) is $O(\mu^{-1})$ as $\mu \to \infty$, and in this limit (11) and (12) both agree with the χ^2 approximation.

The conditional skewness of D may be written as

$$\hat{\kappa}_3^{(\cdot)} - 3\hat{\boldsymbol{\kappa}}_{11}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\kappa}}_{21} + 3\hat{\boldsymbol{\kappa}}_{11}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\kappa}_{12}\mathbf{X})(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\kappa}}_{11} - \hat{\gamma}_r\hat{\gamma}_s\hat{\gamma}_t x_i^r x_i^s x_i^t \hat{\kappa}_{03}^{(i)} + O(1), \quad (13)$$

where $\boldsymbol{\kappa}_{21}$ is a vector of length n, $\boldsymbol{\kappa}_{12} = \text{diag}\{\kappa_{12}^{(i)}\}$, and the summation convention has been employed in the final term with $i = 1, \ldots, n$ and $r, s, t = 1, \ldots, p$. The vector $\boldsymbol{\gamma}$ is given by $(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\kappa}_{11}$ and $\kappa_{03}^{(i)} = E(Y_i - \mu_i)^3$. In applications in which the degrees of freedom are few, it is best to multiply (13) by the factor $(1 - p/n)$, but this correction, though helpful, does not remove the $O(1)$ error term in (13).

The approximate significance level can now be obtained via the standardized statistic

$$Z = \{D - E(D \mid \mathbf{S})\}/\{\text{var}(D \mid \mathbf{S})\}^{1/2},$$

using the Edgeworth expansion,

$$\Pr(Z \geqslant z \mid \mathbf{S}) = 1 - \Phi(z) + \phi(z)(z^2 - 1)\rho_3/6 + O(n^{-1}), \quad (14)$$

where ρ_3 is the standardized conditional skewness of Z obtained from (12) and (13), the latter formula with the correction factor $1 - p/n$.

7. A SIMPLE EXAMPLE

I consider a particularly simple example with independent and identically distributed Poisson observations with $n = 3$, $p = 1$. Such a small value of n is unfavorable for the limits considered here. The exact and approximate first three conditional cumulants of D are given in Table 2 for $\bar{y} = 1, 2, 5$. These are to be compared with the cumulants of χ_2^2, which are 2, 4, 16. Bearing in mind that errors in the higher-order cumulants have a progressively decreasing effect on probability calculations, the approximations (11), (12), and (13) are surprisingly good. In the cases $\bar{y} = 1$ and $\bar{y} = 2$, D has support on 3 and 7 points, respectively. Thus no continuous approximation is likely to be adequate in these cases. For $\bar{y} = 5$, D is supported on 27 points and the Edgeworth approximation seems to be preferable to the χ_2^2 approximation, at least in the region of most interest. In this particular case the limiting normal distribution for D is the same as that given by Morris (1975), because the conditional distribution, given $\bar{y}$, is multinomial.

Table 2. Exact and Approximate Conditional Cumulants of D

	κ_1		κ_2		κ_3	
S	*Approx.*	*Exact*	*Approx.*	*Exact*	*Approx.*	*Exact*
$\bar{y} = 1$	2.66	2.58	2.71	3.29	4.43	3.35
$\bar{y} = 2$	2.43	2.38	4.46	5.26	11.81	15.26
$\bar{y} = 5$	2.12	2.11	4.53	4.61	19.72	20.79
χ_2^2	2.0		4.0		16.0	

Routine computation of the conditional cumulants (11), (12), and (13) is feasible for log-linear models. For linear logistic models, on the other hand, the computations are excessively heavy because the required cumulants depend both on the binomial probability and on the binomial index. The corresponding computations for Pearson's statistic given in Section 4 (and with greater accuracy in McCullagh 1985) are simpler, and routine computation is quite feasible. The results of a small-scale simulation study were reported by McCullagh (1985) for binary data with $n = 50, p = 3$, giving very sparse data. These simulation results confirm (9) and (10) but demonstrate the inadequacy of the normal approximation. The Edgeworth approximation with skewness correction gives better results, and I would expect the same to be true for the deviance statistic.

[*Received February 1984. Revised July 1985.*]

REFERENCES

Cochran, W. G. (1952), "The χ^2 Test of Goodness of Fit," *Annals of Mathematical Statistics*, 23, 315–345.

Fienberg, S. E. (1979), "The Use of Chi-squared Statistics for Categorical Data Problems," *Journal of the Royal Statistical Society*, Ser. B, 41, 54–64.

Haldane, J. B. S. (1939), "The Mean and Variance of χ^2 When Used as a Test of Homogeneity When Expectations Are Small," *Biometrika*, 31, 346–355.

Koehler, K. J., and Larntz, K. (1980), "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials," *Journal of the American Statistical Association*, 75, 336–344.

McCullagh, P. (1984), "Tensor Notation and Cumulants of Polynomials," *Biometrika*, 71, 461–476.

——— (1985), "On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential-Family Models," *International Statistical Review*, 53, 61–67.

Morris, C. L. (1975), "Central Limit Theorem for Multinomial Sums," *The Annals of Statistics*, 3, 165–188.

Williams, D. A. (1983), "The Use of the Deviance to Test the Goodness of Fit of a Logistic-Linear Model to Binary Data," *GLIM Newsletter*, 6, 60–63.