J. R. Statist. Soc. B (2006)
68, *Part* 3, *pp.* 495–508

Improved likelihood inference for discrete data

A. C. Davison

Ecole Polytechnique Fédérale de Lausanne, Switzerland

and D. A. S. Fraser and N. Reid

University of Toronto, Canada

[Received July 2004. Final revision December 2005]

Summary. Discrete data, particularly count and contingency table data, are typically analysed by using methods that are accurate to first order, such as normal approximations for maximum likelihood estimators. By contrast continuous data can quite generally be analysed by using third-order procedures, with major improvements in accuracy and with intrinsic separation of information concerning parameter components. The paper extends these higher order results to discrete data, yielding a methodology that is widely applicable and accurate to second order. The extension can be described in terms of an approximating exponential model that is expressed in terms of a score variable. The development is outlined and the flexibility of the approach is illustrated by examples.

Keywords: Binary regression; Categorical data; Conditional inference; Contingency tables; Likelihood; Negative binomial; Non-canonical link function

1. Introduction

In models with continuous response variables, recent developments in likelihood theory lead to p-value approximations for scalar parameters that are accurate to third order, and to marginal likelihoods for scalar or vector parameters that are accurate to the same order. By comparison the usual normal approximations for the distributions of quantities that are based on the maximum likelihood estimator or the likelihood root are accurate just to first order. In this paper we show how the higher order methods can be extended to the analysis of discrete data, following Davison and Wang (2002), who examined saddlepoint methods that give second-order approximations by embedding the discrete problem in a continuous model, and Pierce and Peters (1999), who argued that the continuous embedding model gives a more appropriate model for inference than the original discrete model; broadly similar conclusions were reached by Severini (2000a).

Recent likelihood theory shows that inference for a scalar component parameter $\psi(\theta)$ has a well defined p-value $p(\psi)$ for assessing $\psi(\theta) = \psi$, and that a scalar or vector parameter $\psi(\theta)$ has a marginal log-likelihood $l^*(\psi)$. The p-value is obtained from the observed log-likelihood $l(\theta)$ and a canonical parameterization $\varphi(\theta)$. In a full exponential family model $\varphi(\theta)$ is the canonical parameter; in more general models $\varphi(\theta)$ is constructed using sample space derivatives and approximate ancillarity. Furthermore with independent observations $y^1, \ldots, y^n$ we have

Address for correspondence: A. C. Davison, Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland.
E-mail: Anthony.Davison@epfl.ch

 1369–7412/06/68495

$$l(\theta) = \sum_{i=1}^{n} l^i(\theta) \tag{1}$$

where $l^i(\theta) = \log\{f^i(y^i;\theta)\}$ is the log-likelihood contribution from y^i. The canonical reparameterization $\varphi(\theta)$ can similarly be expressed as a sum

$$\varphi(\theta) = \sum_{i=1}^{n} \varphi^i(\theta) \tag{2}$$

where $\varphi^i(\theta)$ is a reparameterization contribution from the ith data component. The canonical reparameterization is defined only up to affine transformations, which have no effect on the inference results. To compute the p-value for inference on a scalar parameter of interest $\psi(\theta) = \psi$, we use the p-value function

$$p(\psi) = \Phi\{r + r^{-1} \log(q/r)\} \tag{3}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Expression (3) has been prominent in recent likelihood theory, accounts of which may be found in Barndorff-Nielsen and Cox (1994), Severini (2000b) and Reid (2003). In expression (3) both r and q are determined by the pair $\{l(\theta), \varphi(\theta)\}$ and the parameter of interest $\psi : r$ is the likelihood root

$$r(\psi) = \mathrm{sgn}(\hat{\psi} - \psi)[2\{l(\hat{\theta}) - l(\hat{\theta}_\psi)\}]^{1/2} \tag{4}$$

where $\hat{\theta}$ is the maximum likelihood estimator, $\hat{\theta}_\psi$ is the constrained maximum likelihood estimator for a given $\psi(\theta) = \psi$ and q is a maximum likelihood departure with a nuisance parameter adjustment. An expression for q is given in Appendix A.1.

Under moderate regularity conditions and assuming that the log-likelihood has the usual asymptotic properties as $n \to \infty$, the p-value and marginal log-likelihood are accurate to third order when the distribution of y is continuous.

In this paper we show how to use these results for the analysis of discrete data. The accuracy drops from third order to second order, and as in Davison and Wang (2002) the approximation involves a continuous embedding model, which is described in Appendix B. As we can achieve only $O(n^{-1})$ accuracy, we need only first-order approximate ancillary directions—whose construction we outline in Section 2—and these are more easily obtained than the second-order directions that are used for the continuous case. In Section 3 we show how these extended likelihood methods apply to inference in a general model for discrete data, and we illustrate this with 2×2 tables, binary regression with non-canonical link and Poisson regression with a non-linear component. In Section 4 we generalize to a model for overdispersion: the two-parameter negative binomial model.

As an initial example, suppose that we have n independent Bernoulli observations y_i, with probability of success p_i related to a covariate x_i by the logistic function

$$p_i = \exp(\beta_0 + \beta_1 x_i)/\{1 + \exp(\beta_0 + \beta_1 x_i)\}.$$

Exact inference for this binary regression model can be obtained by computing the conditional distribution of $\Sigma y_i x_i$ given Σy_i, and the conditional distribution can be very well approximated by the saddlepoint method; see for example Brazzale (2000). The arguments of Davison and Wang (2002) apply to this setting to show that the saddlepoint approximation with a continuity correction approximates the exact conditional distribution function with relative error $O(n^{-1})$, but more importantly that the saddlepoint approximation without continuity correction approximates a continuous embedding of the model, which in many ways is more useful for inference and approximates the mid-p-value. These arguments hold for infer-

ence about a linear function of the canonical parameter in full exponential families, including the odds ratio for 2×2 tables (Strawderman and Wells, 1998), matched pairs and multiple logistic regression. However, if the parameter of interest is a non-linear parameter of the exponential family, as for example if the binary regression model is based on a non-canonical link function, then the nuisance parameters cannot be exactly eliminated by conditioning, and the saddlepoint approach does not apply directly. In this setting the present approach effectively eliminates the nuisance parameters, by marginalization over a nuisance parameter distribution.

In the discrete case exact or approximate conditioning creates a distribution with a possibly complex lattice structure, and the maximum likelihood estimates may be on the boundary of the parameter space. For some discussion of these points see Albert and Anderson (1984) and Frydenberg and Jensen (1989). Our development does not address this complication; the results require the maximum likelihood estimate to be in the interior of the parameter space.

2. Canonical reparameterization

Consider independent variables $y^1, \ldots, y^n$ where y^i is a $d^i \times 1$ vector with model $f^i(y^i; \theta)$ and the common parameter θ is $1 \times k$. If y^i is continuous, the reparameterization component from the ith observation is a $1 \times k$ vector

$$\varphi^i(\theta) = \left. \frac{\partial l^i(\theta; y^i)}{\partial y^{i\mathrm{T}}} \right|_{y^{i0}} \times V^i \tag{5}$$

where T denotes matrix transpose and V^i is a $d^i \times k$ weight matrix that describes how parameter change near the maximum likelihood value $\hat{\theta}^0$ influences the ith data component. The first factor, the gradient of the log-likelihood at the observed data point y^0, gives the canonical parameter if the model is a curved exponential family. The second factor V^i gives a linear adjustment to the log-likelihood gradient; these weights V^i implicitly implement conditioning on an approximately ancillary statistic, reducing the dimension of the problem from that of the data $d^1 + \ldots + d^n$ to that of the parameter k. The numerical arrays V^i provide all the information that is needed concerning this conditioning (Fraser and Reid, 2001). For the continuous case the arrays V^i are defined in Appendix A at expression (23) and lead to third-order inference. We now address the modifications that are needed for inference in the discrete case.

Suppose first that a component y^i is a canonical variable in the curved exponential family model

$$f^i(y^i; \theta) = \exp\{\alpha^i(\theta) y^i - k^i(\theta)\} h(y^i), \tag{6}$$

and let $\mu^i(\theta) = E(y^i; \theta)$ be its mean. It is shown in Fraser and Reid (2001), section 7, that vectors V^i tangent to an approximate ancillary can be derived by describing the effect of θ on the variable y^i through its mean $\mu^i(\theta)$:

$$V^i = \frac{\partial}{\partial \theta} E(y^i; \theta) \bigg|_{\hat{\theta}^0} = \frac{\partial}{\partial \theta} \mu^i(\theta) \bigg|_{\hat{\theta}^0} = \mu^i_\theta(\hat{\theta}^0), \tag{7}$$

say, which is a $d^i \times k$ matrix. Then from expressions (5) and (6) we have

$$\varphi^i(\theta) = \alpha^i(\theta) V^i. \tag{8}$$

The co-ordinates of y^i need not be linearly independent: any alternative co-ordinates will lead to an equivalent $\varphi(\theta)$ owing to linearity and the use of the mean of the co-ordinates.

If the co-ordinates of y^i are actual score variables for the parameter θ then $\mu^i_\theta(\theta)$ is the expected information $i^i_{\theta\theta}(\theta)$ for that co-ordinate model.

In equation (6) y^i is the score variable for α^i in the full exponential family model. We extend this to more general settings by constructing $\varphi(\theta)$ and V^i with y^i replaced by the locally defined score variable

$$s^i = \frac{\partial}{\partial\theta}\log\{f(y^i;\theta)\}\bigg|_{\theta=\hat\theta^0}$$

and computing V^i as

$$V^i = \frac{\partial}{\partial\theta}E(s^i;\theta)\bigg|_{\theta=\hat\theta^0}. \tag{9}$$

We then have that the contribution of the ith observation to the local reparameterization is

$$\varphi^i(\theta) = \frac{\partial l^i(\theta;y^i)}{\partial s^i}\bigg|_{y^{i0}} V^i. \tag{10}$$

We then sum over i as at equations (1) and (2) and use the pair $\{l(\theta),\varphi(\theta)\}$ to obtain the p-value (3) as before. Although not illustrated in the examples here, we could also use $\{l(\theta),\varphi(\theta)\}$ to construct the marginal log-likelihood $l^*(\psi)$ as described in Fraser (2003). Fraser and Reid (2001), section 7, justified this construction through the tangent exponential model approximation to the original model.

A possible reparameterization of the model would give a new score variable that is a linear transformation of the initial score and would give a compensating linear transformation of the co-ordinates of the V^i; this has no effect on the resulting inference. In some models the calculations are easier if the score s^i is replaced by an affine transformation of it; we shall see this in the example in Section 4. As shown in section 7 of Fraser and Reid (2001), V^i is tangent to a first-order ancillary statistic; in curved exponential families it is tangent to the likelihood root for comparing the curved model with the full model. Inference based on a first-order ancillary is sufficient to obtain a second-order approximation. In the discrete case we can only obtain a second-order approximation in any case, as described in Appendix B and in Davison and Wang (2002), and the use of expected values to define V^i and hence $\varphi(\theta)$ avoids the need to specify a pivotal function, which would usually not be available for discrete models with a continuous parameter.

3. Categorical data model

3.1. The log-likelihood and canonical parameter

Suppose that we have a response variable that can fall in one of d categories with corresponding indicator variables $y_1,\ldots,y_d$. Then $y=(y_1,\ldots,y_d)^{\mathrm{T}}$ has a multivariate Bernoulli distribution with probabilities $p_1(\theta),\ldots,p_d(\theta)$ proportional to say $\exp\{\alpha_1(\theta)\},\ldots,\exp\{\alpha_d(\theta)\}$. The log-likelihood contribution from this observation is

$$\alpha(\theta)y - \log\{A(\theta)\},$$

where $A(\theta)=\Sigma_j\exp\{\alpha_j(\theta)\}$ and $\alpha(\theta)=\{\alpha_1(\theta),\ldots,\alpha_d(\theta)\}$.

For a single observation from this multivariate Bernoulli model the derivative of the mean of the score variable y is

$$\frac{\partial}{\partial\theta}E(y;\theta)=\mu_\theta(\theta)$$
$$=[\mathrm{diag}\{p(\theta)\}-p(\theta)\,p^{\mathrm{T}}(\theta)]\frac{\partial\alpha^{\mathrm{T}}}{\partial\theta}.$$

The array in braces is the expected Fisher information for the canonical parameter of the full dimensional Bernoulli model and has dimension $d\times d$; the derivative of α has dimension $d\times k$.

Now consider independent multivariate Bernoulli responses y^i, with dimensions d^i, and with log-odds parameters $\alpha_1^i(\theta),\ldots,\alpha_d^i(\theta)$. For each observation y^i we have from expression (7) the ith weighting matrix

$$V^i=[\mathrm{diag}\{p^i(\hat\theta^0)\}-p^i(\hat\theta^0)\,p^{i\mathrm{T}}(\hat\theta^0)]\frac{\partial\alpha^{i\mathrm{T}}(\hat\theta^0)}{\partial\theta}$$

where $\hat\theta^0$ is the overall maximum likelihood estimate; then substituting in expression (5) we obtain the reparameterization contribution $\varphi^i(\theta)$.

We then compute $l(\theta)$ and $\varphi(\theta)$ by summing over i, as at equations (1) and (2), and use these to compute the p-value approximation (3) from expressions (4) and (21). This gives a p-value in this discrete case which is accurate to second order, $O(n^{-1})$, as a mid-p-value. If the parameters $\alpha^i(\theta)$ do not depend on i then $l(\theta)$ is the likelihood function for a d-category multinomial distribution and φ is linear in the canonical parameter for that exponential family model.

We illustrate this on some versions of the 2×2 table, and then on some regression models for discrete data.

3.2. 2×2 *tables*

Consider first a single multivariate Bernoulli variable y arrayed as a matrix

$$y=\begin{pmatrix}y_{11}&y_{12}\\y_{21}&y_{22}\end{pmatrix},\qquad\qquad \sum y_{jj'}=1,$$

with corresponding probabilities

$$p(\theta)=\begin{pmatrix}p_{11}(\theta)&p_{12}(\theta)\\p_{21}(\theta)&p_{22}(\theta)\end{pmatrix},\qquad\qquad \sum p_{jj'}=1, \tag{11}$$

so that

$$l(\theta)=y_{11}\log(p_{11})+y_{12}\log(p_{12})+y_{21}\log(p_{21})+y_{22}\log(p_{22});$$

the largest possible dimension for θ is $k=3$. Writing the mean as a column vector $\mu(\theta)=p(\theta)=(p_{11},p_{12},p_{21},p_{22})^{\mathrm{T}}$, we find that $\mu_\theta(\theta)$ consists of three column vectors that are orthogonal to a vector of 1s and the reparameterization can then be taken as

$$\varphi(\theta)=\left\{\log\left(\frac{p_{12}}{p_{11}}\right),\log\left(\frac{p_{21}}{p_{11}}\right),\log\left(\frac{p_{22}}{p_{11}}\right)\right\}.$$

With n independent observations $y^1,\ldots,y^n$ of this multivariate Bernoulli variable we have

$$l(\theta)=\sum y_{11}^i\log(p_{11})+\sum y_{12}^i\log(p_{12})+\sum y_{21}^i\log(p_{21})+\sum y_{22}^i\log(p_{22}),$$
$$\varphi(\theta)=\{n\log(p_{12}/p_{11}),n\log(p_{21}/p_{11}),n\log(p_{22}/p_{11})\},$$

although the constant multiple n in the second expression is unnecessary as the inferences are invariant to linear transformations of φ.

Now suppose that the row probabilities in equation (11) are known, thus reducing the dimension of θ to 2. We write these probabilities as

$$\begin{pmatrix} q_1 & p_1 \\ q_2 & p_2 \end{pmatrix},$$

and let $\theta = (p_1, p_2)$. The log-likelihood contribution from y^i is then

$$l^i(\theta) = y^i_{11}\log(q_1) + y^i_{12}\log(p_1) + y^i_{21}\log(q_2) + y^i_{22}\log(p_2).$$

Writing the probability array in the vector form $(q_1, p_1, q_2, p_2)^{\mathrm{T}}$, we obtain

$$\frac{\mathrm{d}\mu(\theta)}{\mathrm{d}\theta} = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{pmatrix},$$

and thus $\varphi^i(\theta) = \{\log(p_1) - \log(q_1), \log(p_2) - \log(q_2)\} = \{\log(p_1/q_1), \log(p_2/q_2)\}$. For n such observations we have

$$l(\theta) = n_{11}\log(q_1) + n_{12}\log(p_1) + n_{21}\log(q_2) + n_{22}\log(p_2),$$
$$\varphi(\theta) = \{\log(p_1/q_1), \log(p_2/q_2)\},$$

where we have omitted unneeded factors n_1 and n_2. This is the same log-likelihood and reparameterization as is obtained in the modelling of the 2×2 table as a comparison of two binomials; the familiar conditioning on row totals is here obtained automatically.

Finally we consider the more restricted model

$$p(\theta) = \frac{1}{4}\begin{pmatrix} 2+\theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}, \qquad 0 < \theta < 1.$$

This appears in discussions of ancillarity by Fisher (1956), Basu (1964) and Fraser (1979, 2004) and has the unusual feature that the row totals and column totals are each ancillary statistics for θ but the combination of them is not ancillary. The construction below, however, is conditional on an approximate ancillary statistic that is not needed explicitly.

The log-likelihood contribution for a single observation is

$$l(\theta) = y_{11}\log(2+\theta) + (y_{12} + y_{21})\log(1-\theta) + y_{22}\log(\theta).$$

The mean function $\mu(\theta)$ is $p(\theta)$, and we obtain

$$V = \mu_\theta(\hat{\theta}^0) = \frac{1}{4}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

yielding

$$\varphi(\theta) = \frac{1}{4}\{\log(2+\theta) - 2\log(1-\theta) + \log(\theta)\}.$$

For n independent observations of this form we have

$$l(\theta) = n_{11}\log(2+\theta) + (n_{12} + n_{21})\log(1-\theta) + n_{22}\log(\theta),$$
$$\varphi(\theta) = \frac{n}{4}\{\log(2+\theta) - 2\log(1-\theta) + \log(\theta)\},$$

where the factor $n/4$ can be ignored.

Note that V written in vector form is orthogonal to row differences and to column differences and thus in principle agrees with conditioning on both the row totals and the column totals.

The calculations that are needed for more general contingency tables are analogous.

3.3. Numerical examples

We now give two numerical illustrations of the accuracy of these higher order procedures in cases where an exact answer exists for comparison. We compute the ratio of the exact and approximate p-value as the sample size increases, choosing a fixed quantile for the comparison.

The first illustration concerns a Poisson variable with mean ψ. We take $y = \psi + \delta\psi^{1/2}$, for a specified value of δ, and consider $\psi \to \infty$; here ψ plays the role of n in independent sampling. The likelihood root r is readily computed and the maximum likelihood departure q equals $\sqrt{y}\log(y/\psi)$. We consider the behaviour of approximate significance probabilities as the sample size, or equivalently here ψ, increases with δ fixed. Let $p_\psi = \mathrm{pr}(Y \leqslant y; \psi)$ and $\tilde{p}_\psi$ denote exact and approximate significance probabilities, and suppose that

$$\tilde{p}_\psi = p_\psi(1 + b\psi^{-c}) + o(\psi^{-c})$$

for some c as $\psi \to \infty$. Then a log–log-graph of $|\tilde{p}_\psi / p_\psi - 1|$ against ψ will be linear with slope $-c$ and intercept $\log|b|$.

Fig. 1(a) shows such a graph (with $\delta = 2$), comparing $\tilde{p}_\psi$ given by $\Phi(r)$ and by $\Phi(r^*)$ with the mid-p-value $\mathrm{pr}(Y \leqslant y-1; \psi) + \frac{1}{2}\mathrm{pr}(Y = y; \psi)$, and comparing a continuity-corrected version of r^* with the exact value $\mathrm{pr}(Y \leqslant y; \psi)$. The relative errors show the expected dependence on sample size, and the good performance of the continuity-corrected version of r^* supports the arguments of Davison and Wang (2002). For $\psi = 1$ the relative errors of r and r^* are around 3% and less than 1%, whereas the continuity-corrected version of r^*, in which y is replaced by $y + \frac{1}{2}$, has relative error around 0.2% as an approximation to $\mathrm{pr}(Y \leqslant y; \psi)$.

For an example with a nuisance parameter consider the difference of log-odds for two binomial observations, with denominators $m_1 = m_2 = 2\psi$, $r_1 = \psi$ and $r_2 = \psi + \sqrt{\psi}$, for $\psi = 4, 9, 16, \ldots$,

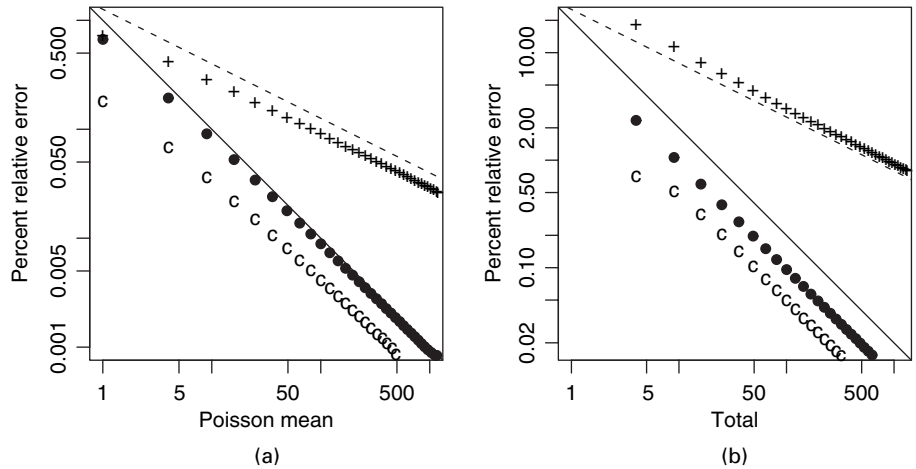

Fig. 1. (a) Comparison of exact and approximate probabilities for tests on the Poisson mean and (b) difference of log-odds for two binomial variables, for the likelihood root r (+) and the modified likelihood root r^* (•) (———, slope −1; -------, slope $-\frac{1}{2}$): here $\delta = 2$, corresponding to a significance level of around 0.025; (a) also shows the dependence on ψ of the significance probability $\phi(r^*)$ for the continuity-corrected observation $y + \frac{1}{2}$ (c)

with respective log-odds λ and $\lambda+\psi$. Again ψ plays the role of sample size, and calculations analogous to those which were outlined for the Poisson model yield the results that are shown in Fig. 1(b). Again we see the relative error of the approximation based on the likelihood root r decreasing as $\psi^{-1/2}$ and the relative error of the more refined approximation (3) decreasing as ψ^{-1}.

Other approximations that may be applied with discrete data, although not developed specifically for them, have been suggested by Skovgaard (1996) and by Severini (1999). In the examples above straightforward calculations show that these yield r^* and r respectively. Severini's approximation involves moment estimators of expected values and large sample sizes may be needed to estimate these accurately.

3.4. Binary regression

Suppose that y^i follows a Bernoulli distribution with success probability p^i and a link function $g(p^i)=\lambda+\psi x_i$ with two parameters ψ and λ. The corresponding log-likelihood contribution is

$$l^i(\lambda,\psi)=y^i\,\mathrm{logit}\{p^i(\lambda,\psi)\}+\log\{1-p^i(\lambda,\psi)\} \tag{12}$$

where $\mathrm{logit}(u)=\log\{u/(1-u)\}$. Then since $E(y^i)=p^i(\lambda,\psi)$ we have

$$\begin{aligned} V^i &= \left.\frac{\partial p^i(\lambda,\psi)}{\partial(\lambda,\psi)}\right|_{\hat\theta^0} \\ &= \frac{1}{g'\{p^i(\hat\lambda^0,\hat\psi^0)\}}(1 \quad x_i) \end{aligned}$$

and

$$\varphi^i=\mathrm{logit}\{p^i(\lambda,\psi)\}V^i. \tag{13}$$

For the logistic regression model, the link function is $\mathrm{logit}\{p^i(\lambda,\psi)\}=\lambda+\psi x_i$, and equation (13) simplifies to

$$\varphi^i=(\lambda+\psi x_i)\frac{1}{p^i(\hat\lambda^0,\hat\psi^0)\{1-p^i(\hat\lambda^0,\hat\psi^0)\}}(1 \quad x_i),$$

and shows that $\varphi(\lambda,\psi)$ is just a linear transformation of (λ,ψ), the canonical parameter of the exponential family. As the inference is invariant to linear transformation, we can take $\varphi(\theta)=(\lambda,\psi)$, with

$$l(\theta)=\lambda\sum x_i+\psi\sum y^i x_i-\log\{1+\exp(\lambda+\psi x_i)\},$$

and so equation (21) in Appendix A.1 simplifies to

$$q=(\hat\psi-\psi)|j_{\theta\theta}(\hat\theta)|^{1/2}|j_{\lambda\lambda}(\hat\theta_\psi)|^{-1/2}.$$

Higher order approximations based on this have been implemented in the S language by Brazzale (2000).

The computations for vector covariates x_i extend those above in an obvious way. If we use a non-canonical link then $l(\theta)$ and $\varphi(\theta)$ are computed from expressions (12) and (13), although the explicit expression for q seems unenlightening.

For a numerical assessment of the effect of higher order adjustments we examine data from 53 people with prostate cancer (Brown, 1980). The binary response indicates the presence of

nodal involvement and depends on five dichotomous explanatory variables. We fit the model with all covariates and assess how a positive response depends on one of these covariates, the level of serum acid phosphatase. The corresponding regression parameter is denoted ψ; there are six parameters in all, including a constant. Fitting a logistic regression model, for which $g(p) = \text{logit}(p)$, gives $\hat{\psi} = 1.684$ with standard error 0.791; the signed likelihood ratio statistic for testing $\psi = 0$ is $r = 2.247$, with p-value 0.012. The corresponding value of r^* is 2.083, with p-value 0.019. The model with complementary log–log-link function $g(p) = \log\{-\log(1-p)\}$ gives $\hat{\psi} = 1.142$ with standard error 0.618, and the values of r and r^* for testing $\psi = 0$ are 1.968 and 1.843, with corresponding p-values 0.025 and 0.033. The higher order correction is in the same direction as with the logistic model but is more substantial. Examples 12.18 and 12.24 of Davison (2003) also described the use of higher order approximations for the logistic regression model, using the software of Brazzale (2000) which implements approximate conditional inference for logistic regression. As noted above the procedure that is outlined here recovers this conditioning when the parameter of interest is linear in the canonical parameter of the exponential family.

3.5. Extension to Poisson counts: smoking data

The method that was outlined for the multivariate Bernoulli model extends directly to the case of Poisson counts, which we illustrate on the data in Table 1 on the relationship between smoking and lung cancer in British male physicians. It shows the man-years at risk, T, and the number of individuals dying of lung cancer, y, cross-classified by the number of cigarettes smoked daily, x, and the number of years of smoking, t, taken to be age minus 20 years. Frome (1983) suggested that the mean deaths per man-year be modelled as

$$\lambda(x,t) = \exp(\theta_1) t^{\theta_2} \{1 + \exp(\theta_3)\, x^{\theta_4}\}, \qquad -\infty < \theta_1, \theta_3 < \infty, \quad \theta_2, \theta_4 > 0; \tag{14}$$

the death-rate in the absence of smoking is thus $\exp(\theta_1) t^{\theta_2}$. One aspect of interest is the value of θ_4, as $\theta_4 = 1$ would correspond to a linear increase in death-rate with x, and we shall investigate this.

If we assume that the number of deaths in the ith cell is Poisson with mean $\mu^i(\theta) = T\,\lambda(x,t)$, then the log-likelihood can be expressed as

Table 1. Lung cancer deaths in British male physicians (Frome, 1983)†

Years of smoking, t	*Results for the following daily cigarette consumptions, x:*						
	Non-smokers	*1–9*	*10–14*	*15–19*	*20–24*	*25–34*	*⩾35*
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

†The table gives (man-years at risk)/(number of cases of lung cancer), T/y, cross-classified by years of smoking, taken to be age minus 20 years, and number of cigarettes smoked per day.

$$l(\theta) = \sum_{i=1}^{n} [y^i \log\{\mu^i(\theta)\} - \mu^i(\theta)],$$

where y^i is the number of deaths in the ith cell of the table. The maximum likelihood estimates and their standard errors based on the observed information matrix are $\hat\theta_1 = 2.94$ (0.57), $\hat\theta_2 = 4.46$ (0.33), $\hat\theta_3 = -1.12$ (1.00) and $\hat\theta_4 = 1.28$ (0.2). The value of the signed likelihood ratio statistic r for testing $\theta_4 = 1$ is 1.506, so the normal approximation to the distribution of r gives a one-sided significance level for testing linear dependence of the death-rate on x as 0.066.

The canonical parameterization is

$$\varphi(\theta) = \sum_{i=1}^{n} \log\{\mu^i(\theta)\} \times \left(\frac{\partial \mu^i(\theta)}{\partial \theta}\bigg|_{\hat\theta^0} \right);$$

from equation (14) we have

$$\frac{\partial \mu^i(\theta)}{\partial \theta} = T_i \exp(\theta_1) t_i^{\theta_2} (1 + \exp(\theta_3) x_i^{\theta_4}, \{1 + \exp(\theta_3) x_i^{\theta_4}\} \log(t_i), \exp(\theta_3) x_i^{\theta_4}, \exp(\theta_3) x_i^{\theta_4} \log(x_i))^{\mathrm{T}},$$

where terms involving x are understood to vanish for non-smokers. The elements of equation (3) are most simply obtained numerically. With $\theta_4 = 1$ we have $q = 1.47$, so $r^* = 1.491$, giving a significance level of 0.068, which is only a small change from the value based on r.

4. A more complex model

In this section we illustrate the computations on a model which is not a curved exponential family model, so the more general approach to computing $\varphi(\theta)$ that was described at the end of Section 2 is needed. We consider an overdispersed Poisson distribution, where the overdispersion is generated by assuming that the mean of the Poisson distribution follows a gamma distribution with mean μ and shape parameter λ. The resulting density is the negative binomial

$$f(y; \theta) = \frac{\Gamma(\nu + y)}{\Gamma(\nu)\, y!} \frac{\nu^\nu \mu^y}{(\nu + \mu)^{\nu + y}}, \qquad y = 0, 1, \ldots, \ \mu, \nu > 0. \tag{15}$$

The log-likelihood function based on a single observation is

$$l(\theta; y) = A(y + \nu) - A(\nu) + \nu \log(\nu) + y \log(\mu) - (\nu + y) \log(\nu + \mu), \tag{16}$$

where $A(\nu) = \log\{\Gamma(\nu)\}$ is the log-gamma function. We shall write $\zeta(\nu)$ for the digamma function $A'(\nu)$. Differentiating equation (16) with respect to μ and ν at $(\hat\mu^0, \hat\nu^0)$ gives $s = (s_1, s_2)$ where

$$\begin{aligned} s_1 &= \frac{\partial l}{\partial \mu}\bigg|_{\hat\theta^0} = \frac{y}{\hat\mu^0} - \frac{(\hat\nu^0 + y)}{(\hat\nu^0 + \hat\mu^0)}, \\ s_2 &= \frac{\partial l}{\partial \nu}\bigg|_{\hat\theta^0} = \zeta(\hat\nu^0 + y) - \zeta(\hat\nu^0) + 1 + \log(\hat\nu^0) - \log(\hat\nu^0 + \hat\mu^0) - \frac{(\hat\nu^0 + y)}{(\hat\nu^0 + \hat\mu^0)}. \end{aligned} \tag{17}$$

For the ith observation in a sample of size n from equation (15), we obtain V^i from expression (9) as

$$\begin{pmatrix} \dfrac{\hat{\nu}^0}{\hat{\mu}^0(\hat{\mu}^0+\hat{\nu}^0)} & 0 \\ 0 & \zeta'(\hat{\nu}^0)-\dfrac{1}{\hat{\nu}^0}-E\{\zeta'(\hat{\nu}^0+y^i)\}+\dfrac{1}{\hat{\nu}^0+\hat{\mu}^0} \end{pmatrix}. \tag{18}$$

The expectation in the expression for V^i_{22} in matrix (18) is for a single observation from density (15).

Note that s^i is affinely equivalent to the simpler form

$$\begin{pmatrix} y^i \\ \zeta(\hat{\nu}^0+y^i)-\dfrac{(y^i+\hat{\nu}^0)}{(\hat{\mu}^0+\hat{\nu}^0)} \end{pmatrix},$$

but by keeping the more complex form (18) we have that $V^i_{\alpha\beta}=(\partial/\partial\theta_\beta)\,E(s^i_\alpha)$ is the (α,β)th element of the expected information matrix in a single observation from model (16). In this example the parameters ν and μ are orthogonal so the array V^i is diagonal.

We use V^i to compute $\varphi(\theta)$ as

$$\begin{aligned} \varphi_1(\theta) &= \sum \frac{\partial l(\theta;y^i)}{\partial s^i_1} V^i_{11} \\ &= \sum_{i=1}^{n} \left\{ \zeta(\nu+y^{i0}) + \log\left(\frac{\mu}{\nu+\mu}\right)\right\} \end{aligned} \tag{19}$$

$$\begin{aligned} \varphi_2(\theta) &= \sum \frac{\partial l(\theta;y^i)}{\partial s^i_2} V^i_{22} \\ &= \sum_{i=1}^{n} \frac{\zeta(\nu+y^{i0})+\log\{\mu/(\nu+\mu)\}}{\zeta'(\hat{\nu}^0+y^{i0})-1/(\hat{\nu}^0+\hat{\mu}^0)} V^i_{22} \end{aligned} \tag{20}$$

where

$$\partial l(\theta)/\partial s^i_\alpha = \{\partial l^i(\theta)/\partial y^i\}(\partial s^i_\alpha/\partial y^i)^{-1}.$$

As usual we combine this with $l(\theta)$ to compute $r^*(\psi)$.

For a numerical illustration of these calculations, we take the data from Bissell (1972) on the numbers of faults, y, in lengths of cloth, x (m $\times 10^2$). We suppose that y^i follows the two-parameter negative binomial distribution (15), where now $\mu=\mu^i=\beta x_i$, and we take the common shape parameter ν to be the parameter of interest. Instead of $(\partial s^i_1/\partial y^i)^{-1}V^i_{11}=1$ as at equation (19) it is now x_i. The mean and variance of y^i are βx_i and $\beta x_i+(\beta x_i)^2/\nu$.

The overall maximum likelihood estimate of ν is $\hat{\nu}=8.694$ with standard error 4.207. The 95% confidence interval for ν based on the normal approximation to the distribution of the likelihood root $r(\nu)$ is (3.68, 28.41). Using the normal approximation to the distribution of $r^*(\nu)$ the 95% confidence interval is (3.35, 24.13). Here the second-order correction moves the interval towards the origin and so increases the fitted response variances. This move is in the same direction as with normally distributed responses, for which higher order corrections correspond to taking the appropriate denominator for a sum of squares, and hence tend to increase maximum likelihood variance estimates.

Acknowledgements

The work was supported by the Swiss National Science Foundation, Ecole Polytechnique Fédérale de Lausanne, and the Natural Sciences and Engineering Research Council of Canada.

The authors thank the referees and Associate Editor for very helpful comments on an earlier version.

Appendix A: Details on the *p*-value formula

A.1. Computing q from φ

We now present the formulae that are needed to convert $\{l(\theta), \varphi(\theta)\}$ to approximate p-values; for further details see Fraser *et al.* (1999) or Reid (2003). For inference on a scalar parameter of interest $\psi(\theta) = \psi$, we use the p-value function that is defined at equation (3), with r the likelihood root. The complementing function q is a nuisance parameter adjusted maximum likelihood departure,

$$q(\psi) = \mathrm{sgn}(\hat{\psi} - \psi)|\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)| \left\{ \frac{|\hat{j}_{\varphi\varphi}|}{|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|} \right\}^{1/2}. \tag{21}$$

In equation (21) $\chi(\theta)$ is a surrogate for $\psi(\theta)$ and is linear in $\varphi(\theta)$,

$$\chi(\theta) = \left. \frac{\partial\psi/\partial\varphi}{|\partial\psi/\partial\varphi|} \right|_{\hat{\theta}_\psi} \varphi^{\mathrm{T}}(\theta), \tag{22}$$

with $\partial\psi/\partial\varphi$ obtained as $(\partial\psi/\partial\theta)(\partial\varphi^{\mathrm{T}}/\partial\theta)^{-1}$ and both θ and φ taken as row vectors. This linear surrogate has a contour or level surface that is tangent to $\psi(\theta)$ at $\hat{\theta}_\psi$. Formulae for the specialized observed informations are recorded at Appendix A.3.

A.2. The maximum likelihood values

For the computation of r in equation (4) we need the profile log-likelihood. If $\theta = (\psi, \lambda)$, where λ is explicitly available, then this is obtained simply by substituting $\hat{\lambda}_\psi$ for λ in the full log-likelihood. If an explicit nuisance parameterization is not available, then we can typically compute the profile log-likelihood $l_{\mathrm{p}}(\psi) = l(\hat{\theta}_\psi)$ by maximizing $l(\theta) + \alpha\{\psi(\theta) - \psi\}$ over (θ, α), which gives $\hat{\theta}_\psi$ and the Lagrange multipler $\hat{\alpha}_\psi$. The corresponding tilted likelihood or Lagrangian

$$\tilde{l}(\theta) = l(\theta) + \hat{\alpha}_\psi\{\psi(\theta) - \psi\}$$

can be used for calculating q; see Fraser *et al.* (1999).

A.3. The informations and estimated variances

The expression in braces in equation (21) is the reciprocal of an estimate of the variance of $|\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)|$, and is a ratio of observed Fisher information matrices for the full parameter and for the nuisance parameter, both recalibrated in terms of φ. They can be computed by rescaling the usual information determinants:

$$|\hat{j}_{\varphi\varphi}| = |\hat{j}_{\theta\theta}||\partial\varphi^{\mathrm{T}}/\partial\theta|^{-2},$$
$$|j_{(\lambda\lambda)}(\hat{\theta}_\psi)| = |j_{\lambda\lambda}(\hat{\theta}_\psi)||\varphi_{\lambda^{\mathrm{T}}}(\hat{\theta}_\psi)\,\varphi^{\mathrm{T}}_{\lambda^{\mathrm{T}}}(\hat{\theta}_\psi)|^{-1}$$

where the parentheses enclosing λ are to indicate that the nuisance parameter has been calibrated locally in terms of $\varphi(\theta)$.

A.4. The weighting matrix V^i for the continuous case

In the continuous case a full dimensional pivotal quantity can describe how the ith co-ordinate is influenced by parameter change near the observed maximum likelihood value. For the ith co-ordinate let $z^i(y^i; \theta)$ be the ith pivotal quantity. The array V^i is obtained from the total derivative for that co-ordinate pivotal:

$$V^i = \left. \frac{\mathrm{d}y^i}{\mathrm{d}\theta} \right|_{y^{i0}, \hat{\theta}^0} = -\left(\frac{\partial z^i}{\partial y^i} \right)^{-1} \left. \frac{\partial z^i}{\partial \theta} \right|_{y^{i0}, \hat{\theta}^0}, \qquad i = 1, \ldots, n, \tag{23}$$

where $\hat{\theta}^0 = \hat{\theta}(y^0)$ is the maximum likelihood estimator obtained from the full data, and the leftmost derivative is calculated for a fixed value of the pivot.

The formulae that were given in Fraser (2003) can be used to convert $\{l(\theta), \varphi(\theta)\}$ to a marginal log-likelihood for ψ, whether ψ is scalar or vector.

Appendix B: A continuous approximation to the discrete model

We use a continuous model that can be made arbitrarily close to the discrete model, and we apply the asymptotic methods to the continuous model; this gives the reduction of dimension and the separation of component parameters. Also, as mentioned in Section 2 we determine the influence of the parameter on a data point in terms of the mean of a locally defined score variable (Fraser and Reid, 2001). As the simplest discrete model we examine in detail the component Bernoulli model

$$f(y;\theta) = \frac{\exp\{y\,\varphi(\theta)\}}{\exp\{\varphi(\theta)\} + \exp\{-\varphi(\theta)\}}, \qquad y = -1, 1, \tag{24}$$

and construct a corresponding continuous model with the same score parameter and score variable

$$f_c(y;\theta) = k_c^{-1}\{\varphi(\theta)\}\frac{\exp\{y\,\varphi(\theta)\}}{\exp\{\varphi(\theta)\} + \exp\{-\varphi(\theta)\}}, \qquad y \in S_c, \tag{25}$$

where

$$S_c = (-1 \pm c) \bigcup (1 \pm c),$$

and c is an auxiliary parameter. The normalizing constant is

$$k_c(\varphi) = 2c\frac{\sinh(c\varphi)}{c\varphi} = 2c\left\{1 + \frac{(c\varphi)^2}{3!} + \frac{(c\varphi)^4}{5!} + \ldots\right\}.$$

The multivariate Bernoulli and more general discrete models can be obtained by compounding the simple Bernoulli model, possibly with appropriate conditioning.

The gradient of the log-likelihood of the continuous model with respect to y is $\varphi(\theta)$, which is the canonical parameter of the discrete model. The likelihood function and the distribution function for the continuous model approximate that of the discrete model with error $O(c^2)$ as $c \to 0$. For the asymptotic analysis we assume that we have n independent components from models of the type (25), with possibly different φ-functions but a common parameter θ, and consider $n \to \infty$; the effect of c can then be made arbitrarily small. The p-value approximation is obtained from Cakmak *et al.* (1998), using Taylor expansions in $n^{-1/2}$ neighbourhoods of the observed data point.

The discrete model can be viewed as obtained by rounding to the nearest integer; the round-off is of order $O(n^{-1/2})$ but by interpreting the p-value as a mid-p-value the effect is of order $O(n^{-1})$. The use of the mid-p-value interpretation avoids the concern for continuity corrections.

References

Albert, A. and Anderson, J. A. (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and Asymptotics*. London: Chapman and Hall.

Basu, D. (1964) Recovery of ancillary information. *Sankhya* A, **26**, 3–16.

Bissell, A. F. (1972) A negative binomial model with varying element sizes. *Biometrika*, **59**, 435–441.

Brazzale, A. R. (2000) Practical small-sample parametric inference. *PhD Thesis*. Department of Mathematics, Swiss Federal Institute of Technology, Lausanne.

Brown, B. W. (1980) Prediction analysis for binary data. In *Biostatistics Casebook* (eds R. G. Miller, B. Efron, B. W. Brown and L. E. Moses), pp. 3–18. New York: Wiley.

Cakmak, S., Fraser, D. A. S., McDunnough, P., Reid, N. and Yuan, X. (1998) Likelihood centered asymptotic model exponential and location model versions. *J. Statist. Plann Inf.*, **66**, 211–222.

Davison, A. C. (2003) *Statistical Models*. Cambridge: Cambridge University Press.

Davison, A. C. and Wang, S. (2002) Saddlepoint approximations as smoothers. *Biometrika*, **89**, 933–938.

Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Fraser, D. A. S. (1979) *Inference and Linear Models*. New York: McGraw-Hill.

Fraser, D. A. S. (2003) Likelihood for component parameters. *Biometrika*, **90**, 327–339.
Fraser, D. A. S. (2004) Ancillaries and conditional inference (with discussion). *Statist. Sci.*, **19**, 333–369.
Fraser, D. A. S. and Reid, N. (2001) Ancillary information for statistical inference. In *Empirical Bayes and Likelihood Inference* (eds S. E. Ahmed and N. Reid), pp. 185–207. New York: Springer.
Fraser, D. A. S., Reid, N. and Wu, J. (1999) A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, **86**, 249–264.
Frome, E. L. (1983) The analysis of rates using Poisson regression models. *Biometrics*, **39**, 665–674.
Frydenberg, M. and Jensen, J. L. (1989) Is the 'improved likelihood ratio statistic' really improved in the discrete case? *Biometrika*, **76**, 655–661.
Pierce, D. A. and Peters, D. (1999) Improving on exact tests by approximate conditioning. *Biometrika*, **86**, 265–277.
Reid, N. (2003) Asymptotics and the theory of inference. *Ann. Statist.*, **31**, 1695–1731.
Severini, T. A. (1999) An empirical adjustment to the likelihood ratio statistic. *Biometrika*, **86**, 235–247.
Severini, T. A. (2000a) The likelihood ratio approximation to the conditional distribution of the maximum likelihood estimator in the discrete case. *Biometrika*, **87**, 939–945.
Severini, T. A. (2000b) *Likelihood Methods in Statistics*. Oxford: Clarendon.
Skovgaard, I. M. (1996) An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, **2**, 145–166.
Strawderman, R. L. and Wells, M. T. (1998) Approximately exact inference for the common odds ratio in several 2×2 tables (with discussion). *J. Am. Statist. Ass.*, **93**, 1294–1306.