Three Examples of Accurate Likelihood Inference

C. LOZADA-CAN and A. C. DAVISON

The modern theory of likelihood inference provides improved inferences in many parametric models, with little more effort than is required for application of standard first-order theory. We outline the relevant computations, and illustrate the calculations using a dilution assay, a zero-inflated Poisson regression model, and a short time series. In each case the effect of the higher order correction can be appreciable.

KEY WORDS: Autoregression; Bias reduction; Dilution assay; Higher order asymptotics; Likelihood; Zero-inflated Poisson distribution.

1. INTRODUCTION

Likelihood is a mainstay of statistical modeling. Inference in many applications is based on familiar large sample theory for the maximum likelihood estimator and the likelihood ratio statistic (Cox and Hinkley 1974, chap. 9). This theory involves approximations that may be justified when the sample size is large, but simulation is often needed to establish whether a given sample is sufficiently large in a new application. Thus the basic theory is often applied in cases where its performance is unclear, either because the sample is small but no better approach is readily available, or because although the sample appears large, the information content per observation is much smaller than is thought. The best-known situation where this second issue arises is in logistic regression, for which special approximations have been widely studied (Davison 1988; Strawderman and Wells 1998). It is not widely appreciated that standard likelihood theory can be readily improved and that the corresponding computations are relatively straightforward. The purpose of this article is to illustrate these ideas in three situations of increasing complexity: a single-parameter dilution assay; a zero-inflated Poisson regression model; and an autoregressive time series.

2. MODERN LIKELIHOOD THEORY

2.1 Basic Notions

We assume a parametric model with probability density function $f(y;\theta)$, with a d-dimensional parameter θ and a vector of continuous responses, $y = (y_1, \ldots, y_n)$. The log-likelihood, $\ell(\theta) = \log f(y;\theta)$, is maximized by the maximum likelihood estimator $\widehat{\theta}$. Under standard regularity conditions $\widehat{\theta}$ has an approximate normal distribution with mean θ and variance matrix $\jmath(\widehat{\theta})^{-1}$, where $\jmath(\theta) = -\partial^2\ell(\theta)/\partial\theta\,\partial\theta^{\mathrm{T}}$ is the observed information matrix, and the Wald pivot $\{\jmath(\widehat{\theta})\}^{1/2}(\widehat{\theta} - \theta)$ has a standard normal distribution. The approximate normal distribution of $\widehat{\theta}$ provides the most common basis for inference about elements of θ, but the implicit assumption that the estimator is symmetrically distributed around its target means that the resulting confidence intervals for components of θ tend to be centered wrongly, or equivalently that true significance levels for one-sided tests on such components may differ substantially from their nominal values, because asymmetry of the log-likelihood about its maximum is not accommodated. Moreover, these inferences are not invariant to transformations of the parameters.

Write $\theta = (\psi, \lambda)$, where ψ is a scalar component of θ for which inference is required and λ represents the remaining components, and let $\widehat{\theta} = (\widehat{\psi}, \widehat{\lambda})$ and $\widehat{\theta}_\psi = (\psi, \widehat{\lambda}_\psi)$ respectively denote the overall maximum likelihood estimator and the maximum likelihood estimator when ψ is held fixed. In this notation a preferable basis for inference on ψ is the likelihood root

$$r(\psi) = \mathrm{sign}(\widehat{\psi} - \psi)\big[2\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\}\big]^{1/2}, \qquad (1)$$

which takes potential asymmetry of the log-likelihood into account, and may be treated as an $\mathcal{N}(0,1)$ variable. The quantity $r(\psi)^2$ is the familiar likelihood ratio statistic. When testing the null hypothesis that $\psi = \psi_0$ against the one-sided hypothesis that $\psi > \psi_0$, we regard small values of the p-value $\Phi\{r(\psi_0)\}$ as casting doubt on the null hypothesis; here and below we use Φ to denote the cumulative probability function of the standard normal distribution. When $\theta = \psi$ is scalar, λ disappears from the expressions above and $\jmath(\theta)$ is scalar. It is easy to verify that apart from a possible sign change, $r(\psi)$ is invariant to reparameterizations of the form $(\psi, \lambda) \mapsto \{g(\psi), h(\psi, \lambda)\}$, in which g and h are bijective, so-called interest-respecting reparameterizations.

2.2 Higher Order Approximations

Improved likelihood inferences may be obtained through higher order asymptotics, on which there is a large literature summarized in books by Barndorff-Nielsen and Cox (1994), Pace and Salvan (1997), Severini (2000), and Brazzale, Davison, and Reid (2007), and in review articles such as Reid (2003). One basic formula is the so-called p^* approximation to the density of the maximum likelihood estimator conditioned on an ancillary statistic (Barndorff-Nielsen 1980, 1983, 1986), manipulation of which yields the modified likelihood root

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)}\log\left\{\frac{q(\psi)}{r(\psi)}\right\}, \qquad (2)$$

Claudia Lozada-Can is a Postdoctoral Researcher (E-mail: *Claudia_Lozada@live.com*) and Anthony Davison is Professor of Statistics (E-mail: *Anthony.Davison@epfl.ch*), Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, IMA-FSB-EPFL, Station 8, 1015 Lausanne, Switzerland. The work was supported by the Swiss National Science Foundation through the National Centre for Competence in Research on Plant Survival (*http://www2.unine.ch/nccr*). We are grateful to the referee, associate editor and editor, and to Alessandra Brazzale and Nancy Reid for their helpful comments on the work.

 DOI: 10.1198/tast.2010.09004

where $q(\psi)$ is a type of maximum likelihood or score statistic that we shall describe below. The quantity $r^*(\psi)$ provides confidence intervals and tests that improve those based on $r(\psi)$. For continuous responses, one-sided confidence intervals based on the significance function $\Phi\{r^*(\psi)\}$ have coverage error $\mathcal{O}(n^{-3/2})$ rather than the $\mathcal{O}(n^{-1/2})$ provided by $\Phi\{r(\psi)\}$. For discrete responses, the error increases to $\mathcal{O}(n^{-1})$. Thus $\Phi\{r^*(\psi)\}$ should provide much better inferences on ψ than does $\Phi\{r(\psi)\}$, and this has been confirmed in many empirical studies of the Barndorff-Nielsen approximation and its variants; see, for example, Severini (2000, sec. 7.6) and Brazzale, Davison, and Reid (2007).

Computation of $q(\psi)$ through the p^* formula requires exact or approximate specification of the ancillary statistic a, so that the likelihood can be written as $\ell(\theta; \widehat{\theta}, a)$ and derivatives obtained with respect to $\widehat{\theta}$, with a fixed. This is straightforward in exponential family and group transformation models, but more generally it is awkward and approximations must be derived; see Reid and Fraser (2010), who discussed the relations between different versions of $q(\psi)$. These difficulties motivate the development of the tangent exponential model of Fraser and Reid (2001), which demands only knowledge of the tangent vectors to the surface on the sample space determined by fixed values of a, but not the ancillary statistic itself. When inference is required on a scalar parameter ψ in the presence of nuisance parameters λ, it is possible to write

$$q(\psi) = \frac{|\,\varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi) \quad \varphi_\lambda(\widehat{\theta}_\psi)\,|}{|\varphi_\theta(\widehat{\theta})|} \times \frac{|j(\widehat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\widehat{\theta}_\psi)|^{1/2}}, \quad (3)$$

in terms of a data-dependent reparameterization $\theta \mapsto \varphi(\theta)$ described below. The numerator of the first term on the right side of (3) is the determinant of a $d \times d$ matrix whose first column is $\varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi)$ and whose remaining columns are formed by the $d \times (d-1)$ matrix $\varphi_\lambda(\widehat{\theta}_\psi)$ whose (r, s) element is $\partial\varphi_r/\partial\lambda_s$, evaluated at $\widehat{\theta}_\psi$. The matrix appearing in the denominator is the $d \times d$ matrix with elements $\partial\varphi_r/\partial\theta_s$. The second term involves the determinants of the observed information matrix evaluated at the maximum likelihood estimator and of the $(d-1) \times (d-1)$ submatrix corresponding to λ, evaluated at $\widehat{\theta}_\psi$. When θ is scalar, comprising only ψ, then $\widehat{\theta}_\psi = \theta$ and expression (3) reduces to

$$q(\theta) = \frac{\varphi(\widehat{\theta}) - \varphi(\theta)}{|\varphi_\theta(\widehat{\theta})|} \times \{j(\widehat{\theta})\}^{1/2}. \quad (4)$$

It can be shown that $\varphi(\theta)$ is invariant to smooth interest-respecting reparameterizations, and thus so too is $r^*(\psi)$. Hence the data analyst may choose a nuisance parameter λ to simplify numerical and other aspects of the model, without altering the resulting inference.

Once $\varphi(\theta)$ and its derivative are available, the computation of (3) or of (4) involves three elements: the maximization of the log-likelihood with respect to λ for a grid of values of ψ and with respect to (ψ, λ); the computation of the observed information matrices; and the computation of $\varphi(\theta)$ and its derivatives. The observed information matrices and the derivatives $\varphi_\theta(\theta)$ can often be obtained numerically, so that analytical work is needed only to find $\varphi(\theta)$. Moreover, the same calculation of $\varphi(\theta)$ may be applied for different parameters ψ in turn, because all that changes in (3) is the columns of $\varphi_\theta(\theta)$ that are used, plus of course the maximization to obtain the grid $\widehat{\lambda}_\psi$ and its associated likelihood and observed information matrices.

2.3 Computation of $\varphi(\theta)$

A central role is played by the parameter $\varphi(\theta)$. In order to define it, we must temporarily distinguish the observed data y^0 and corresponding maximum likelihood estimate $\widehat{\theta}^0$ from the corresponding variables y and $\widehat{\theta}$. Suppose that the data consist of independent observations $y_1, \ldots, y_n$, possibly of different dimensions, and that the parameter is a vector $\theta = (\theta_1, \ldots, \theta_d)$. Then we may write (Fraser and Reid 2001; Davison, Fraser, and Reid 2006)

$$\varphi(\theta) = \sum_{k=1}^{n} i_k(\widehat{\theta}^0) \times \left.\frac{\partial \log f(y_k; \theta)}{\partial s_k}\right|_{y=y^0} \quad (5)$$

$$= \sum_{k=1}^{n} i_k(\widehat{\theta}^0) \times \left(\frac{\partial s_k}{\partial y_k}\right)^{-1} \left.\frac{\partial \log f(y_k; \theta)}{\partial y_k}\right|_{y=y^0} \quad (6)$$

$$= \sum_{k=1}^{n} V_k \left.\frac{\partial \ell(\theta; y)}{\partial y_k}\right|_{y=y^0}, \quad (7)$$

say, where $i_k(\theta)$ is the $d \times d$ expected information matrix corresponding to the density $f(y_k; \theta)$ of y_k, and

$$s_k = \left.\frac{\partial \log f(y_k; \theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}^0}$$

is the contribution made by y_k to the $d \times 1$ score vector. The inverses in $(\partial s_k/\partial y_k)^{-1}$ are taken componentwise, with the convention that a zero component in $\partial s_k/\partial y_k$ gives a zero component in $(\partial s_k/\partial y_k)^{-1}$. The $V_1, \ldots, V_n$ in (7) are $d \times 1$ vectors that are tangent to the ancillary statistic and that show how changing the parameter θ changes y (Fraser 2004); we discuss them further below.

Expression (6), which is appreciably easier to deal with than the somewhat forbidding expression (5), shows that in order to compute $\varphi(\theta)$ we require the expected information matrix and the score statistic, and the derivatives of the latter and of the log-likelihood with respect to the observation. These quantities are available for many parametric models. The derivatives may be obtained numerically if need be, but more troublesome is the Fisher information matrix, which involves an expectation. In some cases its computation can be avoided. For example, if the $y_1, \ldots, y_k$ form a random sample, so that the subscript k disappears from $i_k(\widehat{\theta}^0)$, then $\varphi(\theta) = i(\widehat{\theta}^0)\varphi'(\theta)$, say, and since the factor $|i(\widehat{\theta}^0)|$ cancels from the numerator and denominator in (3), there is no need to compute the expected information. Similar arguments apply in some regression models. If the $i_k(\theta)$ cannot be obtained analytically, then they may be computed numerically; see (A.1) below.

Simple formulas for the vectors V_k in (7) are available in certain cases. If the response y_k is continuous, then one may take

$$V_k = \left.\frac{\partial y_k}{\partial \theta}\right|_{(y^0, \widehat{\theta}^0)} = -\left(\frac{\partial z_k}{\partial y_k}\right)^{-1} \left.\left(\frac{\partial z_k}{\partial \theta}\right)\right|_{\theta=\widehat{\theta}^0, y=y^0}, \quad (8)$$

where z_k is a pivotal quantity, available through the probability integral transform by taking $z_k = z(y_k, \theta) = F(y_k; \theta)$; these pivotal quantities must be defined componentwise. Equivalently, for location-scale and scale models with continuous responses one may take $z_k = (y_k - \mu_k)/\sigma_k$ and $z_k = y_k/\sigma_k$, respectively, where the location and scale parameters μ_k and σ_k may vary with k; this encompasses regression formulations in which $\mu_k = \mu(x^{\mathrm{T}}\beta)$ and $\sigma_k = \sigma(x_k^{\mathrm{T}}\gamma)$ are functions of covariate vectors x_k. For a discrete exponential family model, we may take (Davison, Fraser, and Reid 2006)

$$V_k = \left.\frac{\partial \mathrm{E}(y_k; \theta)}{\partial \theta}\right|_{\widehat{\theta}^0}; \tag{9}$$

this expression, which is readily derived from (7) on setting $\log f(y_k; \theta) = y_k^{\mathrm{T}} a(\theta_k) - \kappa_k(\theta)$, applies both for linear and curved exponential families. We illustrate these computations below. Although approximations to exact p-values have been proposed for use with discrete problems, in most cases the use of mid-p-values seems preferable, and corresponds to direct application of the modified likelihood root (2) with (3), (7), and (9). Pierce and Peters (1992) and Brazzale and Davison (2008) gave discussion and further references on discrete responses, and Brazzale, Davison, and Reid (2007) gave several examples of higher order methods applied to discrete response data.

3. LIMITING DILUTION ASSAY

Limiting dilution assays are used in areas like biology, public hygiene, and immunology to estimate quantities such as the relative frequency θ of a well-defined cell subtype in a population of cells. For instance, in microbiology they may be used to estimate the concentration of micro-organisms per unit volume of solution (Strijbosch et al. 1987; Strijbosch and Does 1988). The value of θ is estimated using data extracted from different dilutions and the single hit Poisson model described below.

In the notation of Mehrabi and Matthews (1995), m successive dilutions of an original preparation are made, the numbers of cells of all types at the different dilutions being $d_1 > \cdots > d_m$. At the jth dilution there are n_j replicates, but all that can be recorded is the absence or presence of the specified subtype in each replicate. To model this we let $y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\pi_j)$, where y_{ij} is an indicator of the presence of cells of the specified subtype in the ith replicate at dilution j. The number of cells of the subtype in each replicate at dilution j is taken to have a Poisson distribution with mean θd_j, so $\mathrm{E}(y_{ij}) = \Pr(y_{ij} = 1) = \pi_j = 1 - \exp(-\theta d_j)$. Apart from additive constants, the log-likelihood is therefore

$$\ell(\theta) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left[y_{ij} \log\{\exp(\theta d_j) - 1\} - \theta d_j \right].$$

This is a discrete exponential family model with a scalar parameter, so (7) and (9) are applicable, and yield

$$V_{ij} = \left.\frac{\partial \mathrm{E}(y_{ij})}{\partial \theta}\right|_{\theta=\widehat{\theta}} = d_j e^{-\widehat{\theta} d_j},$$

$$\varphi(\theta) = \sum_{j=1}^{m} n_j d_j e^{-\widehat{\theta} d_j} \log(e^{\theta d_j} - 1),$$

a result also found more laboriously through (6). Thus (4) equals

$$q(\theta) = \frac{\sum_{j=1}^{m} n_j d_j e^{-\widehat{\theta} d_j} \log\{(e^{\widehat{\theta} d_j} - 1)/(e^{\theta d_j} - 1)\}}{\sum_{j=1}^{m} n_j d_j^2 e^{-\widehat{\theta} d_j}/(1 - e^{-\widehat{\theta} d_j})} \times \left\{ \sum_{j=1}^{m} r_j d_j^2 \frac{e^{\widehat{\theta} d_j}}{(e^{\widehat{\theta} d_j} - 1)^2} \right\}^{1/2},$$

where $r_j = \sum_{i=1}^{n_j} y_{ij}$. Confidence intervals for θ may now be obtained simply by computing $r^*(\theta)$ from (2), with $\psi = \theta$ and $r(\theta) = \text{sign}(\widehat{\theta} - \theta)[2\{\ell(\widehat{\theta}) - \ell(\theta)\}]^{1/2}$. An improved point estimator may be found by numerical solution of the equation $r^*(\theta) = 0$, resulting in the so-called median unbiased estimator $\widehat{\theta}^*$. Below we discuss point and interval estimation for θ in turn.

Small-sample bias of the maximum likelihood estimator $\widehat{\theta}$ was investigated by Mehrabi and Matthews (1995), who applied the ideas of Firth (1993). The latter proposed that the usual score $U(\theta)$ be replaced by $U_F(\theta) = U(\theta) - i(\theta)b(\theta)$, where $i(\theta)$ is the Fisher information and $b(\theta)$ is the first term in the expansion of $\mathrm{E}(\widehat{\theta} - \theta)$ (Cox and Hinkley 1974, p. 309). The solution $\widehat{\theta}_F$ to the modified score equation is unbiased to order n^{-1}, where n is the total sample size. The detailed computations in the assay model are straightforward and may be found in the article by Mehrabi and Matthews (1995), who used simulation to show that $\widehat{\theta}_F$ has smaller bias and mean squared error than $\widehat{\theta}$.

Bias is widely used to measure the quality of point estimators, but it has the drawback of depending on the scale chosen, and it may be infinite in some discrete response models when the data fall on the boundary of the sample space. These problems are not shared by the median unbiased estimator, which satisfies $\Pr(\widehat{\theta}^* \le \theta) = 0.5$, at least approximately. Figure 1 shows the mean relative bias and the median bias $\Pr(\widehat{\theta} \le \theta)$ estimated from 10,000 datasets generated for different values of θ. The Firth estimator $\widehat{\theta}_F$ is very close to unbiased, but its median bias is appreciable, while $\widehat{\theta}^*$ has around a 5%–7% relative bias for all the θ considered, but is close to median unbiased. Particularly for $m = 4$, the median biases are unstable, owing to the very small variation in the discrete distributions for certain parameter values.

We now compare the properties of confidence intervals based on $\widehat{\theta}$, $\widehat{\theta}_F$, r, and r^*. We adopt the study design of Mehrabi and Matthews (1995) with $m = 4, 7$. The total numbers of cells per replicate in the dilutions were $\{49, 85, 49, 260, 454, 793, 1384\}$. We chose eleven values of θ equally spaced from 0.001 to 0.1, with $n_j \equiv n = 4, 6$, and performed a simulation with 10,000 replicates to compare coverage probabilities of the confidence intervals. If all responses are positive in all dilutions we follow Mehrabi and Matthews (1995) and replace the first observation by a zero to avoid having $\widehat{\theta} = +\infty$. The first observation is replaced in 1% of the datasets over all values of θ; the number of replacements decreases as m and n increase. The highest level of replacement, 7.5%, is observed when $m = n = 4$ and is due to the lack of variation mentioned above.

Table 1 shows the left and right noncoverage rates, and the sum of their absolute differences from the nominal one-sided

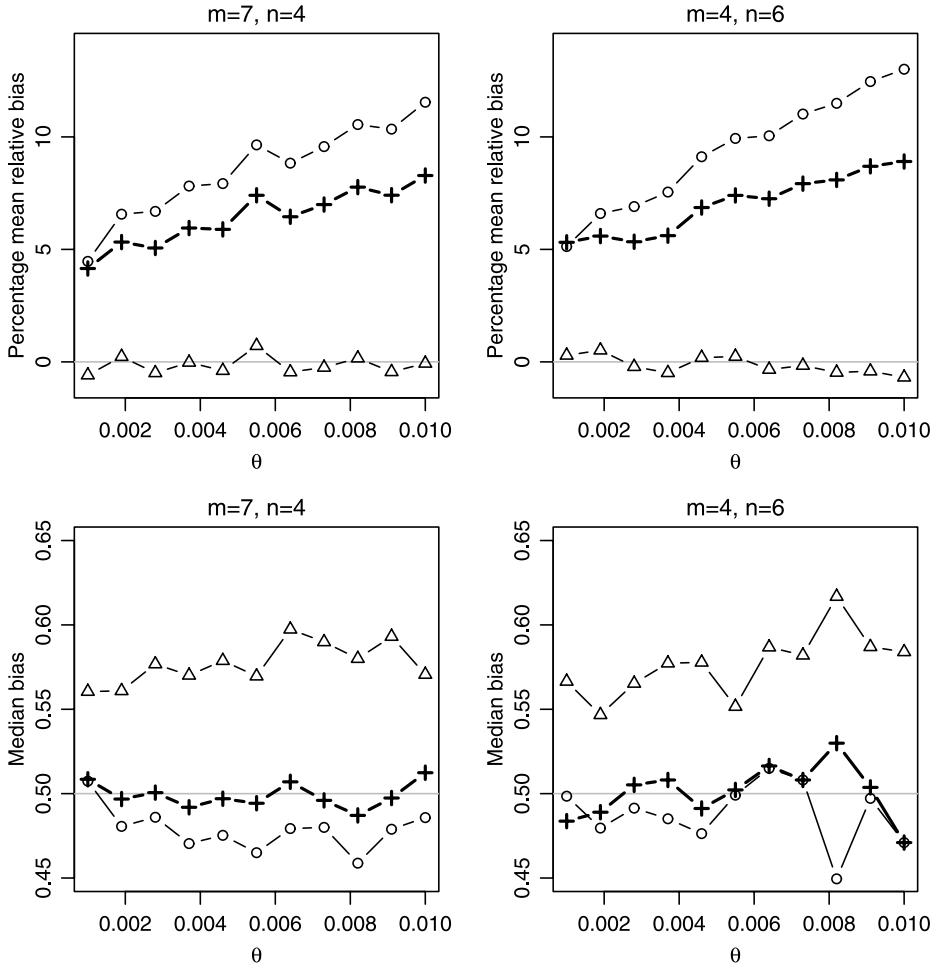

Figure 1. Mean relative bias (%) (upper panels) and median bias (%) (lower panels) for $\widehat{\theta}$ ($\circ$), $\widehat{\theta}_F$ ($\triangle$), and $\widehat{\theta}^*$ ($+$) as a function of θ, for two small-sample settings, based on 10,000 Monte Carlo simulations.

rate of 2.5%, for nominal 95% confidence intervals for $n = 6$ and $m = 7$, computed using the Wald pivot $\jmath(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta)$, $r(\theta)$, $r^*(\theta)$ and the Firth pivot $\jmath(\widehat{\theta})^{1/2}(\widehat{\theta}_F - \theta)$. The overall error rates, found by summing the left and right rates, are close to 5% for all θ, apart from the Wald and Firth pivots when $\theta = 0.001$. These pivots show strong asymmetry, however; the behavior of r and of r^* is much more symmetrical, with the error rates for r^* being better overall.

Table 1. Empirical left (L), right (R), and total (T) noncoverage, and sum of absolute differences (A) from nominal level (%) for nominal 95% confidence intervals based on the Wald and Firth pivots, and on r and r^*, as functions of θ, each with $n = 6$ replicates of size $m = 7$, based on 10,000 Monte Carlo simulations. Standard errors for each set of four columns are roughly 0.16, 0.16, 0.22, 0.22.

	Wald				r				r^*				Firth			
$\theta \times 10^4$	L	R	T	A	L	R	T	A	L	R	T	A	L	R	T	A
10	5.3	0.7	6.0	4.6	2.7	2.5	5.2	0.3	2.6	2.4	5.0	0.2	5.3	0.7	6.0	4.6
19	5.1	0.7	5.8	4.4	2.6	2.7	5.3	0.3	2.8	2.4	5.2	0.4	5.1	0.7	5.8	4.4
28	4.3	0.6	4.9	3.7	2.2	2.7	4.9	0.4	2.4	2.4	4.8	0.2	4.3	0.6	4.9	3.7
37	4.7	0.5	5.2	4.2	2.4	2.4	4.8	0.2	2.7	2.1	4.8	0.5	4.7	0.5	5.1	4.2
46	4.3	0.4	4.7	3.8	2.2	2.8	5.0	0.5	2.4	2.4	4.7	0.3	4.3	0.4	4.7	3.8
55	4.5	0.4	4.9	4.1	2.2	2.9	5.1	0.6	2.5	2.4	4.9	0.2	4.5	0.4	4.9	4.1
64	4.2	0.3	4.5	3.8	2.2	2.4	4.7	0.3	2.5	2.1	4.6	0.4	4.2	0.3	4.5	3.8
73	4.3	0.4	4.7	3.9	2.1	3.0	5.1	1.0	2.4	2.6	5.0	0.3	4.3	0.4	4.7	3.9
82	4.6	0.2	4.8	4.3	2.5	2.6	5.2	0.2	2.8	2.5	5.3	0.3	4.6	0.2	4.8	4.3
91	4.4	0.3	4.7	4.1	2.1	3.3	5.4	1.2	2.3	2.8	5.1	0.4	4.4	0.3	4.7	4.1
100	4.5	0.1	4.6	4.5	2.1	2.8	5.0	0.7	2.4	2.3	4.7	0.3	4.5	0.1	4.6	4.5

Our results thus suggest that although $\widehat{\theta}_F$ is almost an unbiased estimator, confidence limits should rather be based on r, or preferably, r^*. The effort needed to obtain $\widehat{\theta}_F$ and r^* is very similar.

4. ZERO-INFLATED POISSON MODEL

The Poisson distribution is often used for counts, but in practice data are commonly more dispersed than a Poisson model would suggest. In particular, biomedical data often display extra zeros, in which case the zero-inflated Poisson model (Lambert 1992; Böhning et al. 1999) may be used. Such a model presupposes that the response y has a Poisson distribution with probability π, and that $y = 0$ with probability $1 - \pi$. If the Poisson means are determined by a log-linear model, then apart from constants the log-likelihood $\ell(\pi, \beta)$ for a sample of independent responses y_j with corresponding covariates x_j may be written as

$$\sum_{\{j:y_j=0\}} \log\left[(1-\pi) + \pi \exp\{-\exp(x_j^{\mathrm{T}}\beta)\}\right] + \sum_{\{j:y_j>0\}} \{\log\pi + y_j x_j^{\mathrm{T}}\beta - \exp(x_j^{\mathrm{T}}\beta)\}. \qquad (10)$$

Expression (10) is readily generalized to allow π to depend on covariates.

Böhning et al. (1999) used this model to analyze dental data from a prospective study of schoolchildren from an urban area in Brazil. The children were all 7 years old at the beginning of the study, whose goal was to assess the effectiveness of six treatments to prevent dental caries. The treatments were allocated at random among 797 schoolchildren, with children from the same school receiving the same treatment. Table 2 shows the distribution of the number of decayed, missing, and filled teeth (DMFT) at the end of the study. The data display an excess of zeros, so Böhning et al. (1999) analyzed the data using a zero-inflated Poisson model with the following covariates: SCHOOL, which identifies the treatment; SEX, a binary covariate indicating the gender of the child; COLOUR, an ethnic group covariate with three categories; and $\log(\mathrm{DMFT}_b + 0.5)$, the number of DMFT at the beginning of the study. Thus the parameter vector β is of length ten, and $d = 11$ once π is included in θ.

Table 2. Number of children with decayed, missing, and filled teeth index (DMFT) at the end of the study.

DMFT	0	1	2	3	4	5	6
Counts	231	163	140	116	70	55	22

In this case the sample size is almost 800, so one would think that the usual likelihood approximations would be adequate for all parameters. However, the model involves a hidden binary response for each individual, corresponding to the response having a Poisson distribution, and these variables are confounded with the possibility that a Poisson variable is indeed observed but takes value zero. We therefore consider improved inference on the model parameters. In this more complex discrete response model we use (6), with the unenlightening expressions for computation of $\varphi(\theta)$ given in the Appendix.

Table 3 shows the parameter estimates and 95% confidence intervals computed using the Wald pivot, the likelihood root r, and the modified likelihood root r^*. For most parameters the differences are small; they are largest for β_1, β_3, and β_7, but in no case would inference be altered by using small-sample approximation.

5. AUTOREGRESSIVE MODEL

Biometric time series may be short and highly correlated, in which case inferences based on standard asymptotics may prove unreliable, because the equivalent number of independent observations may be appreciably smaller than the sample size. To illustrate this we take the most commonly used time series model, the Gaussian autoregressive process of order 1, under which the responses are related by

$$y_t - \mu = \rho(y_{t-1} - \mu) + \epsilon_t, \qquad t = 1, \ldots, T, \qquad (11)$$

where the ϵ_t are a random sample from the normal distribution with mean zero and variance σ^2, and we must have $|\rho| < 1$ for stationarity. An alternative representation of (11) is as the linear model $y = \mu 1_n + \sigma\epsilon$, $\epsilon \sim N_n(0, \Gamma)$, where 1_n is an $n \times 1$

Table 3. Effect estimates with 95% confidence interval for ZIP regression on the DMFT index for the covariates GENDER (β_2 corresponding to *boys*), ETHNIC (β_3 and β_4 corresponding to *white* and *black*), SCHOOL ($\beta_5, \ldots, \beta_9$ corresponding to schools 1, 2, 4, 5, and 6), and $\log(\mathrm{DMFT}_b + 0.5)$ computed using the Wald, r, and r^* pivots.

Parameter	Wald	r	r^*
π	0.955 (0.922, 0.987)	0.955 (0.920, 0.985)	0.953 (0.920, 0.986)
β_1	−0.148 (−0.336, 0.041)	−0.147 (−0.338, 0.040)	−0.151 (−0.341, 0.039)
β_2	0.007 (−0.101, 0.115)	0.007 (−0.101, 0.115)	0.005 (−0.102, 0.114)
β_3	0.050 (−0.066, 0.166)	0.050 (−0.066, 0.167)	0.053 (−0.063, 0.170)
β_4	−0.047 (−0.223, 0.130)	−0.047 (−0.225, 0.127)	−0.048 (−0.226, 0.128)
β_5	−0.237 (−0.415, −0.060)	−0.237 (−0.415, −0.060)	−0.235 (−0.413, −0.058)
β_6	−0.328 (−0.527, −0.130)	−0.328 (−0.528, −0.131)	−0.329 (−0.528, −0.131)
β_7	0.017 (−0.147, 0.181)	0.017 (−0.148, 0.181)	0.021 (−0.144, 0.185)
β_8	−0.241 (−0.412, −0.070)	−0.241 (−0.412, −0.071)	−0.239 (−0.410, −0.068)
β_9	−0.103 (−0.283, 0.077)	−0.103 (−0.283, 0.076)	−0.103 (−0.282, 0.078)
β_{10}	0.730 (0.651, 0.809)	0.730 (0.652, 0.810)	0.729 (0.653, 0.811)

vector of ones, $y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$, and the (i, j) element of the $n \times n$ matrix Γ equals $\rho^{|i-j|}/(1-\rho^2)$. More generally, the mean of y might be a linear combination of form $X\beta$, where X is an $n \times p$ matrix of explanatory variables and β is a $p \times 1$ vector of parameters. Rekkas, Sun, and Wong (2008) discussed higher order inference on the autocorrelation ρ in such a model. Here we treat ρ as a nuisance rather than as being of main interest, and investigate the extent to which accommodating serial dependence influences inference about μ. It is well known that ignoring such autocorrelation can lead to very seriously misleading inferences. For example, if $\rho = 1/3$, then the variance of the sample average $\bar{y}$ is inflated by a factor of roughly $1 + 2\rho/(1-\rho) = 2$, yet such autocorrelation would be detected only with probability 0.65 or so in a sample of size 50, based on the sample correlogram.

The likelihood for (11) is straightforward to write down, but y_1 must be treated specially, either by giving it its marginal $N\{\mu, \sigma^2/(1-\rho^2)\}$ distribution under stationarity, or by computing the conditional likelihood for $y_2, \ldots, y_n$ given y_1. These approaches are asymptotically equivalent but can differ in small samples, especially for models that are almost nonstationary or noninvertible (Zivot and Wang 2006, pp. 76–77). We adopt the first approach, which yields the log-likelihood

$$\ell(\theta) = -\frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log(1-\rho^2) - \frac{1}{2\sigma^2}(y-\mu)^T\Gamma^{-1}(y-\mu). \tag{12}$$

The parameter of interest is $\psi = \mu$, and $\lambda = (\rho, \sigma^2)$ are treated as nuisance parameters. To compute the necessary higher order quantities we follow Rekkas, Sun, and Wong (2008) and take the pivots to be the elements of the $n \times 1$ vector of scaled martingale differences $z(y, \theta) = U(y-\mu)/\sigma$, where U is the $n \times n$ lower-triangular matrix

$$U = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -\rho & 1 & \ddots & \cdots & 0 \\ 0 & 0 & -\rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & -\rho & 1 \end{pmatrix}.$$

This choice ensures that the $z(y;\theta)$ have independent standard normal distributions. In this case (7) yields $\varphi^{\mathrm{T}}(\theta) = [\varphi_1(\theta), \varphi_2(\theta), \varphi_3(\theta)]$, where

$$\varphi_1(\theta) = \sigma^{-2}(y - \mu 1_n)^{\mathrm{T}}\Gamma^{-1}1_n,$$

$$\varphi_2(\theta) = -\sigma^{-2}(y - \mu 1_n)^{\mathrm{T}}\Gamma^{-1}\widehat{U}^{-1}\frac{\partial U}{\partial \rho}\bigg|_{\widehat{\theta}}(y - \widehat{\mu}1_n),$$

$$\varphi_3(\theta) = (\sigma^2\widehat{\sigma})^{-1}(y-\mu)^{\mathrm{T}}\Gamma(y - \widehat{\mu}1_n).$$

Figure 2 shows Gaussian QQ-plots of the Wald, r, and r^* pivots for the mean μ of simulated time series of length $n = 50$, with $\rho = 0, 0.5, 0.8$. The Wald statistic is clearly overdispersed even when $\rho = 0$, and strikingly so for $\rho \geq 0.5$, to an extent that would lead to overoptimistically narrow confidence intervals for μ at the usual levels of significance. Both r and r^* are much closer to normality, even for large ρ. This is confirmed by Table 4, which shows the right and left noncoverage proportions for two-sided nominal 50%, 25%, 10%, 5%, and 1% confidence intervals. The coverage of the 95% confidence interval for μ, based on the Wald statistic, is about 93% when $\rho = 0$, dropping to about 86% when $\rho = 0.8$. The corresponding values for r show less extreme undercoverage, while those for r^* are close to the nominal levels for all values of ρ considered. Very large values of the Wald pivot or of r^* arose for around 0.5% of the simulated samples, and these are excluded from the computations.

Rekkas, Sun, and Wong (2008) showed that these corrections have a similar effect when applied to confidence intervals and tests for the autocorrelation ρ, even when the mean is replaced by a linear model: the Wald statistic performs very poorly; the likelihood root is a considerable improvement; and the modified likelihood root seems to provide essentially exact inferences unless ρ is close to unity.

We illustrate the size of the higher order corrections using data from Diggle (1990) on the levels of luteinizing hormone in blood samples taken at 10-minute intervals from a healthy woman over an eight-hour period: 48 values taken from the early follicular phase of the women's menstrual cycle. The left panel of Figure 3 shows the time series, which is available as object `lh` in the `R` library `MASS` (Venables and Ripley 2002; R Development Core Team 2008). A first-order autoregressive model $Y_t - 2.41 = 0.57(Y_{t-1} - 2.41) + \epsilon_t$, with $\epsilon_t \overset{\text{iid}}{\sim} N(0, 0.2)$, seems reasonable for these data. The right panel of Figure 3 shows the behavior of the Wald pivot and of $r(\psi)$ and $r^*(\psi)$ for ψ taken to be the mean μ. The 95% confidence intervals are $(2.13, 2.70)$ based on the Wald pivot, $(2.08, 2.76)$ based on r, and $(2.03, 2.82)$ based on r^*, so there is a large higher order correction that widens the 95% confidence intervals based on the Wald pivot or on r.

6. DISCUSSION

Higher order approximation can produce appreciably better inferences when sample sizes are small, as in the first example. Perhaps more surprising is that even with a sample size of nearly 800, as in the second example, higher order correction may change parameter estimates by around 3%. More striking still is the improvement seen in the third example, in which standard confidence intervals can be quite poor despite the reasonable sample size of $n = 50$ with only three parameters. This suggests that higher order inference may be particularly useful in situations where dependence can radically reduce the effective number of observations.

APPENDIX

Higher Order Inference for Zero-Inflated Poisson Model

This appendix records the quantities needed to compute (5) for the zero-inflated Poisson model. Differentiation yields

$$\frac{\partial \log f(y;\mu)}{\partial y}\bigg|_{y=y^0} = \frac{\pi \exp(-\mu)\mu^y \log(\mu)}{(1-\pi)0^y + \pi\exp(-\mu)\mu^y}\bigg|_{y=y^0},$$

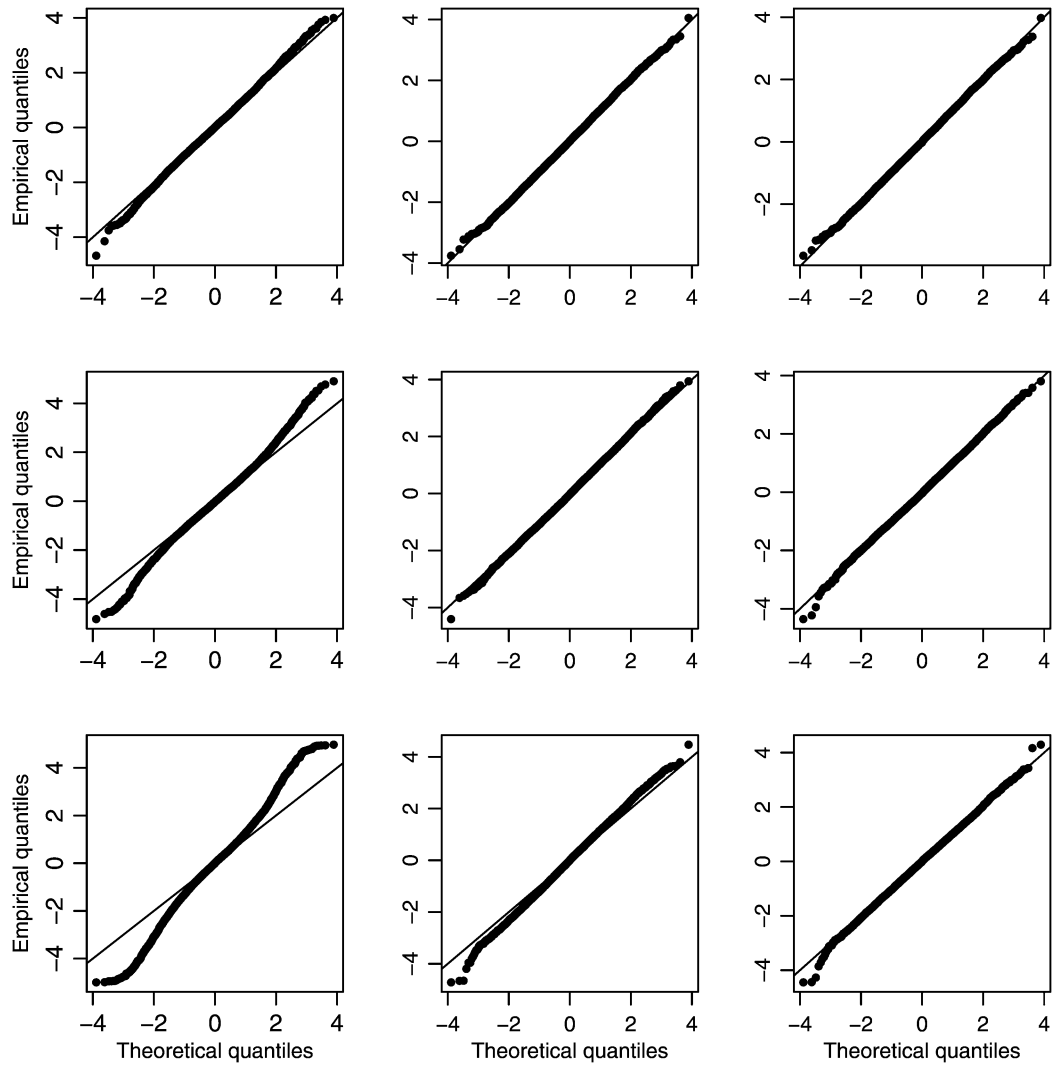

Figure 2. Gaussian QQ-plots of the Wald pivot (left column), r (middle column), and r^* (right column) for the mean of simulated autoregressive processes with $\rho = 0$ (top row), $\rho = 0.5$ (middle row), and $\rho = 0.8$ (bottom row), based on 10,000 simulated series of length $n = 50$.

where $0^y = 0$ if $y \in \mathbb{N}$ and equals 1 otherwise, and

$$s(y;\mu) = \left(\frac{\exp(-\mu)\mu^y - 0^y}{(1-\pi)0^y + \pi\exp(-\mu)\mu^y}, \right.$$
$$\left. \frac{\pi\exp(-\mu)\mu^y(y/\mu - 1)}{(1-\pi)0^y + \pi\exp(-\mu)\mu^y} \right)\Bigg|_{(\widehat{\pi}^0, \widehat{\mu}^0)}.$$

The term $i(\widehat{\theta}^0)$ is readily available by numerical computation of

$$i(\widehat{\theta}^0) = \sum_{y=0}^{\infty} s(\widehat{\theta}^0)s(\widehat{\theta}^0)^T f(y;\theta)\Big|_{(\widehat{\pi}^0, \widehat{\mu}^0)}, \qquad \text{(A.1)}$$

Table 4. Left and right noncoverage probabilities (%) for two-sided nominal 50%, 25%, 10%, 5%, and 1% confidence intervals for the mean parameter μ of simulated autoregressive time series of length $n = 50$, based on 10,000 Monte Carlo replications.

		Left tail %					Right tail %				
Nominal level		25	12.5	5	2.5	0.5	25	12.5	5	2.5	0.5
$\rho = 0$	Wald	25.08	12.99	5.71	3.35	0.84	25.97	13.68	6.32	3.48	1.06
	r	24.97	12.60	5.14	2.70	0.49	25.80	13.16	5.73	2.68	0.54
	r^*	24.42	12.06	4.83	2.42	0.45	25.26	12.64	5.30	2.38	0.44
$\rho = 0.5$	Wald	27.56	15.43	7.28	4.62	1.62	25.90	14.55	7.05	4.36	1.78
	r	27.25	14.32	5.89	3.16	0.66	25.66	13.71	5.77	3.15	0.75
	r^*	25.87	12.82	5.04	2.51	0.49	24.52	12.29	4.83	2.53	0.50
$\rho = 0.8$	Wald	29.20	18.60	11.12	7.93	4.07	28.75	17.88	10.40	7.21	3.53
	r	28.82	17.13	8.18	4.81	1.31	28.28	16.24	7.63	4.26	1.32
	r^*	25.63	13.44	5.63	2.92	0.66	25.28	12.70	5.06	2.63	0.60
Standard error		0.50	0.35	0.22	0.16	0.07	0.50	0.35	0.22	0.16	0.07

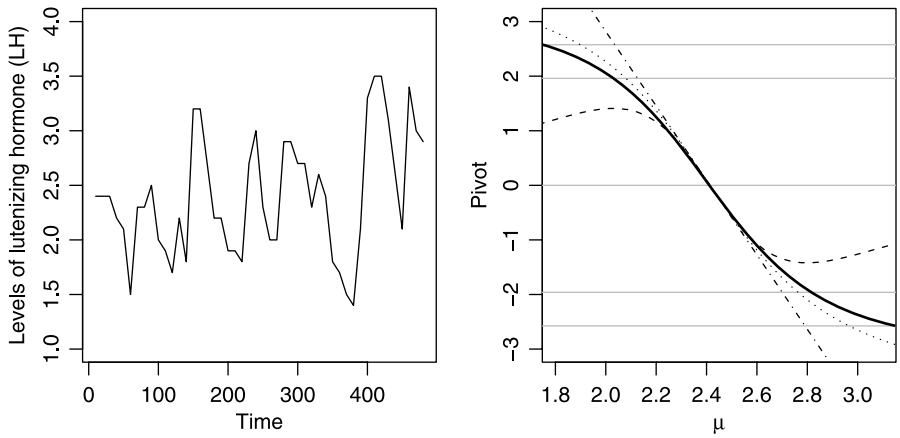

Figure 3. Concentration of luteinizing hormone in blood samples (Diggle 1990). Left panel: data. Right panel: plots of Wald pivot (dot–dash), r (dots), r^* (solid), and q (dashes) as functions of mean μ. Intersection of a pivot curve with the gray horizontal lines yields confidence limits at levels 0.005, 0.025, 0.5, 0.975, and 0.995.

truncated for small enough densities. If $y \neq 0$, then

$$\left(\frac{\partial s}{\partial y}\right)^{-1} = \Big(0, \{(1-\pi)0^y + \pi\exp(-\mu)\mu^y\}^2 / \big(\pi(1-\pi)\exp(-\mu)\mu^{y-1}0^y\{y\log(\mu)+1-\mu\log(\mu)\} + \pi^2\exp(-2\mu)\mu^{2y-1}\big)\Big)\Big|_{(\widehat{\pi}^0,\widehat{\mu}^0)},$$

and otherwise

$$\left(\frac{\partial s}{\partial y}\right)^{-1} = \left(\frac{\{(1-\pi)0^y + \pi\exp(-\mu)\mu^y\}^2}{0^y\exp(-\mu)\mu^y\log(\mu)}, \{(1-\pi)0^y + \pi\exp(-\mu)\mu^y\}^2 / \big(\pi(1-\pi)\exp(-\mu)\mu^{y-1}0^y\{y\log(\mu)+1-\mu\log(\mu)\} + \pi^2\exp(-2\mu)\mu^{2y-1}\big)\right)\Big|_{(\widehat{\pi}^0,\widehat{\mu}^0)}.$$

For each component of the vector β we then apply the chain rule, writing $\partial\ell/\partial\beta_i = \partial\ell/\partial\mu\,\partial\mu/\partial\beta_i$ and $\partial\mu/\partial\beta_i = \exp(x^{\mathrm{T}}\beta)x_i$.

[Received January 2009. Revised January 2010.]

REFERENCES

Barndorff-Nielsen, O. E. (1980), "Conditionality Resolutions," *Biometrika*, 67, 293–310. [131]

——— (1983), "On a Formula for the Distribution of the Maximum Likelihood Estimator," *Biometrika*, 70, 343–365. [131]

——— (1986), "Inference on Full or Partial Parameters Based on the Standardized Signed Log Likelihood Ratio," *Biometrika*, 73, 307–322. [131]

Barndorff-Nielsen, O. E., and Cox, D. R. (1994), *Inference and Asymptotics*, London: Chapman & Hall. [131]

Böhning, D., Dietz, E., Schlattman, P., Mendona, L., and Kirchner, U. (1999), "The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology," *Journal of the Royal Statistical Society, Ser. A*, 162, 195–209. [135]

Brazzale, A. R., and Davison, A. C. (2008), "Accurate Parametric Inference for Small Samples," *Statistical Science*, 23, 465–484. [133]

Brazzale, A. R., Davison, A. C., and Reid, N. (2007), *Applied Asymptotics: Case Studies in Small Sample Statistics*, New York: Cambridge University Press. [131-133]

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall. [131,133]

Davison, A. C. (1988), "Approximate Conditional Inference in Generalized Linear Models," *Journal of the Royal Statistical Society, Ser. B*, 50, 445–461. [131]

Davison, A. C., Fraser, D. A. S., and Reid, N. (2006), "Improved Likelihood Inference for Discrete Data," *Journal of the Royal Statistical Society, Ser. B*, 68, 495–508. [132,133]

Diggle, P. J. (1990), *Time Series: A Biostatistical Introduction*, Oxford: Clarendon Press. [136,138]

Firth, D. (1993), "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, 80, 27–38. [133]

Fraser, D. A. S. (2004), "Ancillaries and Conditional Inference," *Statistical Science*, 19, 333–369. [132]

Fraser, D. A. S., and Reid, N. (2001), "Ancillary Information for Statistical Inference," in *Empirical Bayes and Likelihood Inference*, eds. E. Ahmed and N. Reid, New York: Springer, pp. 185–210. [132]

Lambert, D. (1992), "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14. [135]

Mehrabi, Y., and Matthews, J. N. S. (1995), "Likelihood-Based Methods for Bias Reduction in Limiting Dilution Assays," *Biometrics*, 51, 543–549. [133]

Pace, L., and Salvan, A. (1997), *Principles of Statistical Inference From a Neo-Fisherian Perspective*, Singapore: World Scientific. [131]

Pierce, D. A., and Peters, D. (1992), "Practical Use of Higher Order Asymptotics for Multiparameter Exponential Families" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 54, 701–737. [133]

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [136]

Reid, N. (2003), "Asymptotics and the Theory of Inference," *The Annals of Statistics*, 31, 1695–1731. [131]

Reid, N., and Fraser, D. A. S. (2010), "Mean Loglikelihood and Higher Order Approximations," *Biometrika*, 97, 159–170. [132]

Rekkas, M., Sun, Y., and Wong, A. (2008), "Improved Inference for First-Order Autocorrelation Using Likelihood Analysis," *Journal of Time Series Analysis*, 29, 513–532. [136]

Severini, T. A. (2000), *Likelihood Methods in Statistics*, Oxford: Oxford University Press. [131,132]

Strawderman, R. L., and Wells, M. T. (1998), "Approximately Exact Inference for the Common Odds Ratio in Several 2×2 Tables" (with discussion), *Journal of the American Statistical Association*, 93, 1294–1306. [131]

Strijbosch, L. W. G., and Does, R. J. M. M. (1988), "Comparison of Bias-Reducing Methods for Estimating the Parameter in Dilution Series," *Communication in Statistics—Simulation and Computation*, 17, 1173–1190. [133]

Strijbosch, L. W. G., Buurman, W. A., Does, R. J. M. M., Zinken, P. H., and Groenewegen, G. (1987), "Limiting Dilution Assays. Experimental Design and Statistical Analysis," *Journal of Immunological Methods*, 97, 133–140. [133]

Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics With S* (4th ed.), New York: Springer. [136]

Zivot, E., and Wang, J. (2006), *Modeling Financial Time Series With S-PLUS* (2nd ed.), New York: Springer. [136]