Biometrika (2001), **88**, 1, *pp.* 13–52
© 2001 Biometrika Trust
Printed in Great Britain

Biometrika Centenary: Theory and general methodology

BY A. C. DAVISON

Department of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland

anthony.davison@epfl.ch

SUMMARY

Contributions to statistical theory and general methodology published in *Biometrika*, 1901–2000, are telegraphically reviewed.

Some key words: Bayesian inference; Estimating function; Foundations of statistics; Generalised regression model; Graphical method; Graphical model: Laplace approximation; Likelihood; Missing data; Model selection; Multivariate statistics; Non-regular model; Quasilikelihood; Saddlepoint: Simulation; Spatial statistics.

1. INTRODUCTION

Biometrika begins with a clear statement of purpose:

> It is intended that *Biometrika* shall serve as a means not only of collecting or publishing under one title biological data of a kind not systematically collected or published elsewhere in any other periodical, but also of spreading a knowledge of such statistical theory as may be requisite for their scientific treatment.

Its contents were to include

> (a) memoirs on variation, inheritance, and selection in Animals and Plants, based upon the examination of statistically large numbers of specimens … (b) those developments of statistical theory which are applicable to biological problems; (c) numerical tables and graphical solutions tending to reduce the labour of statistical arithmetic; (d) abstracts of memoirs, dealing with these subjects, which are published elsewhere; and (e) notes on current biometric work and unsolved problems.

Early volumes contained many memoirs on biological topics, but over the twentieth century *Biometrika* became a 'journal of statistics in which emphasis is placed on papers containing original theoretical contributions of direct or potential value in applications'. Thus, of the five types of contents envisaged by its founders, only (b) and to a lesser extent (c) remain, largely shorn of their biological roots. The purpose of this paper is to review *Biometrika*'s contributions to statistical theory and general methodology. To refer to every paper published in this area is evidently impossible: quite apart from the difficulty of staking out its boundaries, the result would be both unreadable and Pearsonian in length. Hence this review is selective, focusing on fairly recent developments which seem to me of probably lasting value but including some early work that has stood the test of time.

After some discussion of the foundations of statistics in § 2, an account is given in § 3 of likelihood and associated concepts, before prediction is touched on in § 4. Subsequent sections deal with estimating functions, model selection and Bayesian statistics, before contributions to simulation and asymptotic approximation are outlined in §§ 8 and 9.

Contributions in generalised regression are outlined in § 10, before the paper concludes with some miscellaneous topics.

2. Foundations

2·1. *Objective theories*

The foundations of statistical inference are discussed sporadically in early issues of *Biometrika.* Karl Pearson's views were apparently close to what was then called inverse probability and is now called a Bayesian approach, though he seems to have believed that the prior should have some frequency interpretation, and certainly his son Egon Sharpe Pearson believed that it should be subject to empirical assessment. Inverse probability had dominated discussions of inference for much of the 19th century, but for the first half of the 20th century it was largely eclipsed by the ideas of Fisher, Neyman and Egon Pearson. An important exception to this was the attempt to put objective Bayesian inference on a secure footing summarised by Jeffreys (1939). Wilk's (1941) verdict now seems ironic:

> From a scientific point of view it is doubtful that there will be many scholars thoroughly familiar with the system of statistical inference initiated by R. A. Fisher and extended by J. Neyman, E. S. Pearson, A. Wald and others who will abandon this system in favour of the one proposed by Jeffreys in which inverse probability plays the central role.

The rats have subsequently abandoned ship in numbers unthinkable in the 1940s. Bayesian contributions in *Biometrika* are briefly surveyed in § 7.

Inverse probability was initially strongly rejected by Fisher, though in later years his views seem to have softened, but his lifelong goal was the same as that of Jeffreys, namely the construction of an objective theory of inference, based on the data at hand. Fisher's fiducial inference transferred the initial uncertainty surrounding sample values into a density and hence measures of uncertainty for the parameter, given the data observed. In some cases this yields the frequentist confidence intervals proposed by Neyman in the 1930s, a coincidence that led early workers to believe that the difference between the approaches was merely a matter of presentation. This was indignantly rebutted by Neyman (1941), whose relations with Fisher had broken down well before. Kendall (1949) is a thoughtful and well-meaning attempt to synthesise positive aspects of the various theories. Fisher's lack of clarity over the basis of the fiducial method led to considerable unease about it, which seemed justified when counterexamples showed that it need not yield unique inferences (Anscombe, 1957; Lindley, 1958; Fraser, 1964). D. A. S. Fraser investigated the circumstances under which fiducial inferences behave sensibly, and in doing so developed his theory of structural inference based on group transformation models (Fraser, 1961, 1965, 1966), though its roots are much earlier (Pitman, 1938, 1939). Fraser established that, when fiducial inference for a scalar parameter is well defined, the model must be essentially a location family. A few subsequent attempts have been made to resurrect fiducialism, but it now seems largely of historical importance, particularly in view of its restricted range of applicability when set alongside models of current interest. Despite this, the group models introduced by Fraser are important in theories of conditional inference.

The foundations of frequentist inference were laid when Jerzy Neyman and Egon Pearson published in 1928 two massive and influential papers on hypothesis testing (Neyman & Pearson, 1928a,b). The eponymous lemma was published elsewhere five years later (Neyman & Pearson, 1933) but the familiar notions of simple and composite hypotheses and errors of the first and second kind appear, and the view of testing as decision-making lurks behind the arras. The first paper introduces the generalised likelihood ratio test for composite hypotheses, and applies it to normal, uniform and exponential models. The second shows the link between the likelihood ratio and chi-squared statistics, discusses the degrees of freedom of the latter and introduces minimum chi-squared estimation. A third discusses how small the sample may be before the large-sample approximation fails, by enumerating significance levels in samples of size ten (Neyman & Pearson, 1931). Welch (1933) applied the Neyman–Pearson approach to tests in the linear model.

2·2. *Principles of inference*

Most attempts to put statistical inference on a consistent footing rest on some subset of the sufficiency, conditionality, likelihood and invariance principles. The sufficiency principle simply says that two datasets from the same model that yield identical minimal sufficient statistics should give the same inferences on the model parameter θ. The conditionality principle rests on the notion of an ancillary statistic, namely a function A of the minimal sufficient statistic whose distribution does not depend on the parameter, and says that inference should be conducted using the relevant subset of the sample space, i.e. that portion of it in which A equals its observed value a. Following papers in the *Annals of Mathematical Statistics* by Welch (1958) and Cox (1958), Robinson (1975) gives examples where the conditionality principle conflicts with standard unconditional or Neyman confidence intervals. Both principles have been extended to situations where $\theta = (\psi, \lambda)$ consists of interest parameters ψ for which inference is required and nuisance parameters λ that are not of central concern (Bartlett, 1937; Cox, 1958; Barndorff-Nielsen, 1973, 1976; Sprott, 1975; Godambe, 1976, 1980; Basawa, 1981b; Liang, 1983). Approximate notions of sufficiency and ancillarity have also been discussed (Feigin & Reiser, 1979; Cox, 1980; Ryall, 1981; Liang, 1984; Skovgaard, 1986; Severini, 1993), with Efron & Hinkley (1978) particularly influential.

The strong likelihood principle states that, if the likelihoods under two possibly different models but with the same parameter are proportional, then inferences about θ should be the same in each case; its weak form is equivalent to the sufficiency principle. In particular, this implies that inference should not be influenced by elements of the sample space that were not in fact observed, appearing to rule out use of procedures such as significance tests and confidence intervals and paving a path towards some form of Bayesian inference.

Both the sufficiency and conditionality principles are accepted more widely than the likelihood principle, so Birnbaum's (1962) article in the *Journal of the American Statistical Association* caused consternation when he showed that acceptance of the first two entails acceptance of the third. Later work somewhat reducing the force of this result includes Kalbfleisch (1975), who distinguishes between experimental and mathematical ancillaries. The former are aspects of the data on which the inference would usually be conditioned whatever the precise form of the model, such as the design matrix in regression, while the latter are quantities that turn out to be distribution-constant once the full probability model has been specified. As experimental ancillaries typically determine the precision of the experiment, conditioning on them would normally be regarded as uncontroversial, while more questions surround mathematical ancillaries: in a particular model there may be several choices of these that give different inferences on θ. A striking example of this, the Cauchy location-scale model, is described by McCullagh (1992).

The invariance principle is closely related to the theory of structural inference mentioned above, and it too has its difficulties. One is that in any practical situation the use of the full group of transformations, however mathematically appealing, would not be sensible: it is more natural to measure distances between galaxies in parsecs than in nanometres. A second is that an invariant solution is often inadmissible, invariance considerations typically ruling out the use of shrinkage estimators.

3. Likelihood

3·1. *Primary notions*

Likelihood is central to much statistical theory and practice. The ideas of likelihood, sufficiency, conditioning, information and efficiency are due to R. A. Fisher, whose work established the excellent asymptotic properties of the maximum likelihood estimator and showed that likelihood gives a basis for exact conditional inference in location and scale models. The last 25 years have seen a second flowering of this theory, stimulated by influential work in the late 1970s and early 1980s.

To outline the basic theory, let $y_1, \ldots, y_n$ be the observed data, taken to be a realisation of a random sample $Y_1, \ldots, Y_n$ from a regular model whose density function $f(y; \theta)$ depends on a parameter θ. The corresponding loglikelihood $l(\theta)$ is maximised at the maximum likelihood estimator $\hat\theta$, and standard arguments tell us that in large samples $\hat\theta$ has an approximate normal distribution with mean θ and variance matrix $\mathscr{I}(\theta)^{-1}$, where

$$\mathscr{I}(\theta) = E\{J(\theta)\}, \quad J(\theta) = -\frac{\partial^2 l(\theta)}{\partial\theta\, \partial\theta^{\mathrm{T}}},$$

are respectively the expected or Fisher information and the observed information matrices. This gives a basis for confidence intervals for elements of θ. As stressed by Bartlett (1953a, b, 1955), inference may also be based on the asymptotically normal score statistic $U(\theta) = \partial l(\theta)/\partial\theta$, with mean zero and variance $\mathscr{I}(\theta)$. A third closely related basis for inference is the likelihood ratio statistic $W(\theta) = 2\{l(\hat\theta) - l(\theta)\}$, whose asymptotic distribution is chi-squared.

In practice θ is unknown, and so $\mathscr{I}(\theta)$ is unavailable. Hence confidence intervals based on normal approximation to $\hat\theta$ must be computed using $\mathscr{I}(\hat\theta)$, $J(\hat\theta)$ or another consistent estimator of $\mathscr{I}(\theta)$. Efron & Hinkley (1978) gave convincing evidence that confidence intervals that use $J(\hat\theta)$ are generally more appropriate than those based on expected information, because they obey the conditionality principle more closely. They are also simpler to compute, because no expectation is involved; there is no need to specify a censoring or missingness mechanism when data are incompletely observed. However, normal approximation to $\hat\theta$ may be inaccurate for small n, while lack of invariance to reparameterisation makes the resulting confidence intervals somewhat unsatisfactory.

In typical applications there are many parameters and we write $\theta = (\psi, \lambda)$, where only the interest parameter ψ is of central concern. For theoretical discussion it is convenient to focus on testing the null model $\psi = \psi_0$, tests of which may be inverted to give confidence intervals for ψ. Sometimes estimation is more easily performed for the full model, with no restriction on ψ; in such a case Cox & Wermuth (1990) show how to obtain approximate maximum likelihood estimates for the restricted model. In many models the score statistic enables assessment of fit by embedding the null model in a judiciously chosen wider class, so that $\psi \neq \psi_0$ corresponds to interesting departures. Tests may then be based on $U(\psi_0, \hat\lambda_{\psi_0})$, where $\hat\lambda_{\psi_0}$ is the null estimate of λ. This is often computationally convenient, because only the null model need be fitted (Bartlett, 1953a). Plots of contributions that different observations make to the test statistic can be used to diagnose the source of model failure; see, for example, Aranda-Ordaz (1981) or Hosking (1984). Mardia & Kent (1991) show how appealing to group structure can simplify construction of score tests. Peers (1971) compares score tests with those based on likelihood ratios and maximum likelihood estimators of $\hat{\psi}$, establishing that none is uniformly superior in the case of a simple null hypothesis.

In many models the maximum likelihood estimator is biased. In simple cases the bias, investigated by Haldane & Maynard Smith (1956), can be removed directly by subtracting an estimated correction, or indirectly by biasing the score statistic. Firth (1993) investigates the second option, showing it to be equivalent to use of a Jeffreys prior in linear exponential family models.

3·2. *Likelihood ratio statistic*

Large-sample inference for an interest parameter ψ is often based on the likelihood ratio statistic $W_p(\psi) = 2\{l(\hat{\psi}, \hat{\lambda}) - l(\psi, \hat{\lambda}_\psi)\}$ or equivalently the profile loglikelihood $l_p(\psi) = l(\psi, \hat{\lambda}_\psi)$, where $\hat{\lambda}_\psi$ is the maximum likelihood estimator of λ for fixed ψ. Although the large-sample distribution of $W_p(\psi)$ is chi-squared with degrees of freedom p equal to the dimension of ψ, this may provide a poor approximation if n is small or if λ has high dimension. One way to improve the approximation follows on writing $E\{W_p(\psi)\}$ as $p\{1 + b(\theta)/n\} + O(n^{-2})$. This implies that the modified likelihood ratio statistic $W_p(\psi)\{1 + b(\theta)/n\}^{-1}$ has expectation closer to the mean p of the approximating chi-squared distribution. The quantity $\{1 + b(\theta)/n\}$ is known as a Bartlett adjustment, following Bartlett (1937) who had investigated its performance in a special case; see also Box (1949). Lawley (1956) showed that in the continuous case this modification not only changes the mean of $W_p(\psi)$ but also improves the approximation to its entire distribution, so that $W_p(\psi)\{1 + b(\theta)/n\}^{-1}$ has a χ_p^2 distribution with relative error $O(n^{-2})$. In practice $b(\theta)$ must be replaced by $b(\hat{\theta})$, and as $\hat{\theta} - \theta = O_p(n^{-\frac{1}{2}})$ it appears that the error has order $n^{-3/2}$ (Harris, 1986), but Barndorff-Nielsen & Hall (1988) used properties of cumulant expansions to show that it remains $O(n^{-2})$. As this suggests, the approximation can be remarkably accurate even in small samples; see, for example, Porteous (1985) or Jensen (1986). G. M. Cordeiro and co-workers have derived Bartlett adjustments for classes of regression models (Cordeiro, 1987; Cordeiro & Paula, 1989) and investigated the extent to which the ideas extend to score statistics (Cordeiro & de Paula Ferrari, 1991; Cordeiro et al., 1994; Cribari-Neto & Ferrari, 1995). A puzzling point in the calculation of Bartlett adjustments is elucidated by Chesher & Smith (1995).

A central role in such calculations is played by the Bartlett identities linking cumulants of the score statistic and its derivatives (Bartlett, 1953a,b), the first two of which are the standard formulae $E\{U(\theta)\} = 0$ and $\text{var}\{U(\theta)\} = I(\theta)$.

Although Bartlett adjustment can also improve the practical performance of the likelihood ratio statistic for discrete data, Frydenberg & Jensen (1989) showed that the theoretical improvement seen with continuous data does not carry over.

When the interest parameter ψ is scalar, a better basis for inference on ψ is a signed version of $W_p(\psi)$, namely $R(\psi) = \text{sign}(\hat{\psi} - \psi)W_p(\psi)^{\frac{1}{2}}$ (DiCiccio, 1984; McCullagh, 1984a), and a related quantity $R^*(\psi)$ which is analogous to a signed Bartlett-adjusted version of $W_p(\psi)$, though it is not obtained in the same way; see § 3·3. Jensen (1986) shows that confidence intervals based on $R^*(\psi)$ closely approximate those based on exact similar tests in exponential family models.

Bartlett-like adjustments for Bayesian quantities are given by DiCiccio & Stern (1993), while Cox (1984) investigates the relation between Bartlett adjustment and the notion of effective degrees of freedom.

Though often straightforward in principle, the calculations needed for asymptotic expansions of $\hat{\theta}$, $W_p(\psi)$ and related quantities such as $b(\theta)$ can be extremely tedious. One valuable way to simplify these calculations is through use of cumulants and index notation, systematised for statistical purposes by McCullagh (1984b). Symbolic computation can help conceal the sordid details (Stafford & Andrews, 1993).

3·3. *Conditional and marginal likelihood*

The profile likelihood $l_p(\psi)$ is often used to form confidence intervals for ψ, but it can perform badly if the distribution of $W_p(\psi)$ is not close to χ^2 and can fail entirely when the dimension of λ grows with n, as arises in so-called Neyman–Scott problems. Even in cases where large-sample results apparently work, the presence of nuisance parameters can cause poor small-sample performance, prompting the search for improvements (Lubin, 1981). One possibility is to use not the full likelihood but a related marginal or conditional density. Suppose that the original likelihood may be expressed as

$$f(y;\psi,\lambda)=f(\hat{\psi},\hat{\lambda}\,|\,a;\psi,\lambda)f(a)f(y\,|\,\hat{\psi},\hat{\lambda},a), \tag{1}$$

where $\hat{\psi}$ and $\hat{\lambda}$ are maximum likelihood estimators, $(\hat{\psi},\hat{\lambda},a)$ is minimal sufficient and a is ancillary. Factorisation (1) is implied in Bartlett (1937). Then all the information about the parameters is contained in the first term on the right-hand side of (1). In many important models this term can be expressed as one of

$$f(\hat{\lambda}\,|\,a;\psi,\lambda)f(\hat{\psi}\,|\,\hat{\lambda},a;\psi),\quad f(\hat{\lambda}\,|\,\hat{\psi},a;\psi,\lambda)f(\hat{\psi}\,|\,a;\psi),$$

where $f(\hat{\psi}\,|\,\hat{\lambda},a;\psi)$ and $f(\hat{\psi}\,|\,a;\psi)$ are respectively conditional and marginal densities that depend only on ψ. Regarded as functions of ψ, these are examples of conditional and marginal likelihoods. There may be a loss of information for ψ if inference is based on one of these, but it is often small in practice; see, for example, the discussion of the 2×2 table in Plackett (1977), a classic example earlier treated in detail by Barnard (1947a,b), Pearson (1947) and others. If we adhere to the conditionality principle, inference should be conducted conditional on the observed value of the ancillary, so we would wish to investigate conditional properties of the conditional or marginal likelihoods and quantities derived therefrom.

Conditioning and marginalisation may be used to eliminate nuisance parameters from linear exponential family and group transformation models, respectively. Almost the simplest transformation model is the location-scale model, in which the random sample $Y_1,\ldots,Y_n$ may be written as $Y_j=\eta+\tau Z_j$, where the density of Z_j is independent of $\theta=(\eta,\tau)$, expressed with no nuisance parameter. In this model the ordered values of the $A_j=(Y_j-\hat{\eta})/\hat{\tau}$ are ancillary, so the conditionality principle suggests that inference should be conditioned on A. We have exactly (Fisher, 1934)

$$f(\hat{\theta}\,|\,\alpha;\theta)=c(a)|J(\hat{\theta})|^{\frac{1}{2}}\exp\{l(\theta)-l(\hat{\theta})\}, \tag{2}$$

where $c(a)$ is a constant depending on a that makes the integral of (2) over $\hat{\eta}$ and $\hat{\tau}$ equal to unity for each fixed a and $|.|$ denotes determinant. More complex marginal likelihoods have been described by Levenbach (1972), Kalbfleisch & Prentice (1973), Pettitt (1983) and many others; see particularly Kalbfleisch & Sprott (1970). An important example is restricted likelihood, discussed in § 3·5.

In the linear exponential family $\exp\{t\psi + s\lambda - \kappa(\psi, \lambda) + c(t, s)\}$, the parameters λ can be eliminated by conditioning, giving a density $f(t \mid s; \psi)$ of the form $\exp\{t\psi - \kappa_s(\psi) + c(t, s)\}$ that can be treated as a conditional likelihood. Here the argument for conditioning stems from the desire to eliminate λ rather than from the conditionality principle. A key problem in this context is to obtain the normalising quantity $\kappa_s(\psi)$, which is typically not in a form useful for applications, though algorithms exist for some special cases (Gail et al., 1981).

Formula (2) applies in wider generality than merely location-scale models, as pointed out in a Royal Statistical Society discussion by Daniels (1958). It turns out to be exact for transformation models (Barndorff-Nielsen, 1980, 1983) and highly accurate much more widely, reproducing the saddlepoint approximation to the density of $\hat{\theta}$ in full exponential family models, for which it has error $O(n^{-3/2})$. Unlike the large-sample normal distribution of $\hat{\theta}$, it has the potential to give inferences that are independent both of the scale chosen for θ and of the particular form of likelihood function used. It generalises work of Cox (1980), Hinkley (1980), Durbin (1980a,b) and others and forms a useful basis for theoretical comparison of likelihood and Bayesian approaches (Efron, 1993). A practical difficulty is how it may be used for inference on components of θ.

Inference is typically required on one parameter at a time, based on a pivotal quantity. In the location-scale model suitable pivots are $Q_1 = (\hat{\eta} - \eta)/\hat{\tau}$ and $Q_2 = \hat{\tau}/\tau$, whose joint density may be derived from (2). Inference for η may be performed by integrating this joint density to give the marginal distribution of Q_1 given A, from which confidence intervals for η may be found by inversion. Similar arguments give the intervals for τ in the location-scale model and for other group transformation models. Exact inference is difficult because of the integration involved in obtaining $c(a)$ and the marginal distributions of the required pivots, and approximations are needed for use in practice; see § 9. In a continuous full exponential family a suitable pivot for scalar ψ is the probability integral transformation $\int_{-\infty}^{t} f(u \mid s; \psi)\, du$, but in discrete exponential families there is in principle no exact pivot. However, the quantity $R^*(\psi)$ mentioned in § 3·2 has a standard normal distribution to a high degree of approximation, typically to $O(n^{-3/2})$ in moderate deviation regions (Barndorff-Nielsen, 1986; Jensen, 1992) and to $O(n^{-1})$ in large-deviation regions, both conditionally on a and unconditionally. Its form is

$$R^*(\psi) = R(\psi) + R(\psi)^{-1} \log\{V(\psi)/R(\psi)\},$$

where $R(\psi)$ is the signed likelihood ratio statistic based on the profile loglikelihood and the form of $V(\psi)$ depends on the underlying model (Fraser, 1990; Barndorff-Nielsen, 1991). Sweeting (1995b) considers $R^*(\psi)$ from a Bayesian point of view. Unfortunately an explicit ancillary is often needed to compute $V(\psi)$, and hence Barndorff-Nielsen & Chamberlin (1994) and others have sought alternative forms.

3·4. *Modified profile likelihood*

As mentioned above, the profile loglikelihood can be a poor basis for inference in the presence of nuisance parameters. One explanation of this is that $\partial l_p(\psi)/\partial\psi$ does not have the usual properties of a score function; its mean is generally $O(1)$ rather than zero, as it would be for the score itself. This suggests modifying $l_p(\psi)$ by the addition of a function chosen to improve the properties of the profile score $\partial l_p(\psi)/\partial\psi$.

Barndorff-Nielsen (1983) gave a general form of modified profile loglikelihood, namely

$$l_m(\psi) = l_p(\psi) - \frac{1}{2}\log\left| -\frac{\partial^2 l(\psi, \hat{\lambda}_\psi)}{\partial\lambda\,\partial\lambda^{\mathrm{T}}} \right| - \log\left| \frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}^{\mathrm{T}}} \right|, \tag{3}$$

where the second determinant requires that the loglikelihood function be expressed in terms of the minimal sufficient statistic $(\hat{\psi}, \hat{\lambda}, a)$. Expression (3) recovers standard inferential quantities when these are known, as in Neyman–Scott cases, while in others it can greatly improve on $l_p(\psi)$. Calculation of its last term, however, entails writing the likelihood explicitly in terms of $\hat{\psi}$, $\hat{\lambda}$ and an ancillary statistic a. As mentioned in § 2·2, exact ancillaries may not exist, and may not be unique even when they do, so (3) is unavailable for many models of practical interest. One way to reduce the size of the last term, typically $O_p(1)$, is to parameterise the model in such a way that λ and ψ are orthogonal, that is to say, $E\{\partial^2 l(\psi, \lambda)/\partial\psi\,\partial\lambda^{\mathrm{T}}\} = 0$, for then the term is $O_p(n^{-1})$ and it can be ignored relative to the other terms, as pointed out in a Royal Statistical Society discussion paper of Cox & Reid (1987). Further work including comparisons of these modifications has been done by Cox & Reid (1992), Barndorff-Nielsen & McCullagh (1993), Fraser & Reid (1989), Liseo (1993) and Severini (1998a,b), while Cruddas et al. (1989) show their effect in a time series example. Mukerjee & Chandra (1991) derive a Bartlett adjustment for the Cox–Reid likelihood ratio statistic. A major effort has been put into such adjustments, in order to produce versions that work automatically in applications (Severini, 1999).

The idea of parameter orthogonality has ramifications beyond modified likelihoods; see, for example, Solomon & Taylor (1999).

Cox (1980) and McCullagh (1984a) discuss inference based on approximate ancillaries, the latter showing that the lack of uniqueness is not typically a problem provided inference is required only accurate to terms of size $O(n^{-\frac{1}{2}})$, while Barndorff-Nielsen (1995) describes approximations that do not require specification of an ancillary but which retain much of the accuracy of (3); see also Barndorff-Nielsen (1986).

3·5. *Restricted maximum likelihood*

An important application of marginal likelihood is to components of variance models. The simplest case is the normal linear model with a single additional level of random effects, in which the $n \times 1$ response y equals $X\beta + Z\eta + \varepsilon$, where X and Z are known $n \times p$ and $n \times k$ matrices and the elements of the $k \times 1$ vector η are independent normal variables independent of the $n \times 1$ vector of errors ε; β is an unknown parameter. The means of η and ε are zero, while the variance matrices are $I_k\gamma\sigma^2$ and $I_n\sigma^2$, γ and σ^2 being positive unknowns. Hence y is normal with mean $X\beta$ and variance matrix $\Omega = \sigma^2(\gamma ZZ^{\mathrm{T}} + I_n)$. The usual unbiased estimator of β is $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$, which is normal with mean β and variance matrix $X(X^{\mathrm{T}}X)^{-1}\Omega(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, independent of the residuals $r = y - X\hat{\beta}$ that contribute to the residual sum of squares, whose distribution depends only on σ^2 and γ. This suggests that inference for these parameters be based on the marginal density of r, or equivalently on $n - p$ linearly independent residuals. It turns out that the corresponding loglikelihood may be written as

$$-\tfrac{1}{2}\log|\Omega| - \tfrac{1}{2}\log|X^{\mathrm{T}}\Omega^{-1}X| - \tfrac{1}{2}r^{\mathrm{T}}\{\Omega^{-1} - \Omega^{-1}X(X^{\mathrm{T}}\Omega^{-1}X)^{-1}X^{\mathrm{T}}\Omega^{-1}\}r.$$

This result was derived by Patterson & Thompson (1971) in a different form, building on Hartley & Rao (1967) and on earlier ideas of Thompson (1962) and others. Inference based on such expressions is known as restricted maximum likelihood inference; a connection to Bayesian procedures was explored by Harville (1974). In the simplest case η is not present and hence $\Omega = \sigma^2 I_n$; the restricted loglikelihood then yields the marginal loglikelihood obtained from the chi-squared distribution of the residual sum of squares. Hence the resulting estimator of σ^2 is unbiased (Bartlett, 1937). The bias reduction due to use of restricted maximum likelihood can be substantial, especially if p and k are large. Similar bias reductions arise in other contexts where the parameter of interest appears in the variance structure. Examples are crossover designs (Matthews, 1989) and Gaussian random fields (Dietrich, 1991).

An important generalisation is to multilevel models, in which data are structured hierarchically. Such models are used in contexts as different as the social sciences and animal breeding experiments. For example, data on the attainments of individual pupils within classes within schools may be available, with random effects at each level of the hierarchy. Estimation that accounts for the variance structure imposed by these effects may be achieved using restricted maximum likelihood (Goldstein, 1986, 1987, 1989). An algorithm for estimation by standard maximum likelihood in these models is given by Longford (1987).

These ideas have been extended to nonlinear and generalised linear models by Gilmour et al. (1985), Goldstein (1991), Schall (1991) and others.

3·6. *Non-regular models*

The theory described above applies to settings where the usual regularity conditions hold. They can fail in such a variety of ways that general discussion is difficult. One type of failure occurs when parameters present under one model vanish under a contiguous model. One example is two-phase regression with an unknown changepoint θ, the simplest case of which has

$$Y_j = \beta_0 + \beta_1(x_j - \theta)I(x_j > \theta) + \sigma\varepsilon_j, \quad j = 1, \ldots, n,$$

where the ε_j are independent standard normal variables and $I(.)$ is an indicator function, so the regression slope is zero up to $x = \theta$ and subsequently β_1. A second example sets $Y_j = \xi_1 \sin(j\theta) + \xi_2 \cos(j\theta) + \sigma\varepsilon_j$. The parameter θ is meaningless if $\beta_1 = 0$ in the first case and $\xi_1 = \xi_2 = 0$ in the second, and standard likelihood ratio or other tests are hence inapplicable. One approach to such problems is detailed case-by-case analysis, as in Hinkley's (1969) discussion of two-phase regression, but the resulting intricate approximations are difficult to generalise. Simulation is now a natural alternative, but a useful and fairly satisfactory analytical approach was proposed by Davies (1977, 1987), based on upcrossings of stochastic processes. The chosen test statistic for comparison of the models is $W = \sup_\theta W(\theta)$, where $W(\theta)$ is the likelihood ratio statistic applicable if θ were fixed and known, and the supremum is over $\theta_L \leqslant \theta \leqslant \theta_U$. If $W(\theta)$ has an approximate chi-squared distribution with p degrees of freedom for each θ, then the process $W(\theta)$ as θ varies in $[\theta_L, \theta_U]$ should be related to a chi-squared process, and the distribution of W close to that of its supremum. Davies (1987) uses results on extrema of stochastic processes to show that the approximate significance level for W is $1 - F_p(W) + Vf_p(W)$, where F_p and f_p are the χ_p^2 distribution and density functions, and

$$\begin{aligned} V &= \int_{\theta_L}^{\theta_U} \left| \frac{\partial W^{\frac{1}{2}}(\theta)}{\partial\theta} \right| d\theta \\ &= |W^{\frac{1}{2}}(\theta_1) - W^{\frac{1}{2}}(\theta_L)| + |W^{\frac{1}{2}}(\theta_2) - W^{\frac{1}{2}}(\theta_1) + \ldots + |W^{\frac{1}{2}}(\theta_U) - W^{\frac{1}{2}}(\theta_m)|, \end{aligned}$$

where $\theta_1 < \ldots < \theta_m$ are the turning points of $W^{\frac{1}{2}}(\theta)$ in $[\theta_L, \theta_U]$. He also gives more accurate approximations, but examples suggest that this remarkably simple approach will be precise enough for many purposes.

Standard likelihood asymptotics break down if the true parameter value is on the boundary of the parameter space, and the maximum likelihood estimator then has a nonnormal limiting distribution. Chant (1974) investigates the relation between likelihood ratio and $C(\alpha)$ tests in this situation, and finds that they are asymptotically equivalent only when the true parameter value is not a boundary case. Likelihood inference in a particular case of such models was investigated by Crowder (1990). An important example of such models is in estimation and testing in standard models but under ordered alternatives (Bartholomew, 1959).

A different form of nonregularity arises because of failure of the Bartlett identities. These fail perhaps most commonly when θ contains an endpoint, as when $Y_j = \beta + \varepsilon_j$, where $\varepsilon_1, \ldots, \varepsilon_n$ are independent positive random variables from a density for which $f(\varepsilon) \sim \alpha\gamma\varepsilon^{\alpha-1}$ as $\varepsilon \downarrow 0$, and $\alpha, \gamma > 0$; here $\theta = (\beta, \alpha, \gamma)$. Densities in this framework include the three-parameter Weibull and gamma distributions, and the generalised extreme-value and Pareto distributions, which are useful in modelling extreme values in areas as diverse as strength of materials, athletic records, and environmental and financial applications. The degree of nonregularity depends on α. If $\alpha > 2$, the Fisher information for θ is finite and $\hat{\theta}$ has its usual large-sample properties of consistency, normality and efficiency. If $1 < \alpha \leqslant 2$, then there is a large-sample local maximum of the likelihood, but when $\alpha < 2$ the limiting distribution of $\hat{\beta}$ is nonnormal and $\hat{\beta} - \beta = O_p(n^{-\frac{1}{\alpha}})$, while when $\alpha = 2$ the limit is normal but $\hat{\beta} - \beta = O_p\{(n^{\frac{1}{2}} \log n)^{-1}\}$. When $0 < \alpha \leqslant 1$ the likelihood has no asymptotic local maximum but is maximised by taking $\hat{\beta} = \min Y_j$, giving $\hat{\beta} - \beta = O_p(n^{-1/\alpha})$, an order of convergence that cannot be improved. A scattered literature on such results, related particularly to the three-parameter Weibull distribution, was unified by Smith (1985), who discusses also the properties of maximum likelihood estimators for α, γ and other parameters, if any. When $1 < \alpha < 2$ their maximum likelihood estimators are asymptotically normal, independent of $\hat{\beta}$, and converge at the usual rate. He also proposes an alternative approach, replacing β by $\min Y_j$ and then estimating the remaining parameters from the likelihood for the other observations, acting as if β were known to equal $\min Y_j$. Smith (1994) extends these results to regression models $Y_j = \beta^{\mathrm{T}} x_j + \varepsilon_j$, showing that β can be estimated by a linear programming algorithm, and that the remaining parameters are amenable to the same treatment as in the location case.

The discussion above presupposes that the correct model has been fitted. In misspecified models this is not the case. Results analogous to those of standard likelihood theory then hold, but modified by the fact that asymptotically the best-fitting model is not the truth but that closest to the truth among the class fitted. Kent (1982) shows that in this case the likelihood ratio statistic behaves asymptotically like a weighted combination of χ_1^2 variables; the weights would all equal unity were the model correct. A related setting is robust estimation where the estimating function and loglikelihood derivative for the underlying model are different. Boente & Fraiman (1988) discuss nonregular likelihood inference in this context, and give resistant estimators with the same rates of convergence as maximum likelihood estimators for various nonregular contexts; some of their estimators are also efficient.

Rather different problems arise in inference for parameters of stochastic processes. Such problems are often called non-ergodic because the quantity analogous to the scaled information matrix that norms the maximum likelihood estimator typically converges not to a constant but to a random variable. The case perhaps most thoroughly investigated is the branching process, where the asymptotics depend sharply on the birth parameter. See Stigler (1970, 1971), Becker (1974), Basawa & Scott (1976), Sweeting (1978), Feigin & Reiser (1979) and Basawa (1981a). Sweeting (1992) gives a more general discussion of such problems.

Practical difficulties arise when the likelihood has multiple local maxima or its global maximum equals infinity. Some particular instances of this are investigated by Barnett (1966), Copas (1975) and Crisp & Burridge (1994), while Jensen et al. (1991) discuss global convergence of Newton-like algorithms for important cases.

A further nonstandard situation is the comparison of nonnested families of models, initiated by Cox (1961). This topic was subsequently developed by econometricians and in regression (McAleer, 1983). Kent (1986) clarifies the structure of likelihood ratio tests in this context.

4. Predictive inference

Prediction has long been a weak point of likelihood theory. Bayes's theorem gives a basis for inference not only on parameters but also on an unobserved random variable Z. Given the observed data $Y = y$, the posterior predictive density of Z may be written

$$f(z \mid y) = \int f(z \mid y;\, \theta)\pi(\theta \mid y)\, d\theta = \frac{\int f(z \mid y;\, \theta) f(y;\, \theta)\pi(\theta)\, d\theta}{\int f(y;\, \theta)\pi(\theta)\, d\theta}, \tag{4}$$

where $\pi(\theta)$ is a prior density for θ and $\pi(\theta \mid y)$ is the corresponding posterior density. Although Fisher proposed a form of likelihood for predictive inference, it was largely ignored until Hinkley (1979) attempted to put it on a firmer footing. The idea is to treat the joint density of Y and Z conditional on the value of a joint sufficient statistic for θ, $S = s(Y, Z)$, as a function of z, but this is feasible only when the model admits a reduction by sufficiency. When it can be calculated it is close to (4) with vague priors, suggesting that approximation to (4) might be profitable.

Butler (1989) gave several general forms of predictive likelihood, including the approximate conditional quantity

$$L_{\mathrm{AC}}(z \mid y) = \frac{f(y, z;\, \hat{\theta}_z)}{|I_z(\hat{\theta}_z)|^{\frac{1}{2}}} \frac{|I_z(\hat{\theta}_z)|}{|J(\hat{\theta}_z)J(\hat{\theta}_z)|^{\frac{1}{2}}},$$

where $\hat{\theta}_z$ is the maximum likelihood estimator of θ based on (y, z), $I_z(\theta)$ is the observed information computed from $f(y, z;\, \theta)$, and $J(\theta) = \partial^2 \log f(y, z;\, \theta)/\partial\theta\, \partial(y^{\mathrm{T}}, z^{\mathrm{T}})$ is a mixed partial derivative of the loglikelihood. The first ratio results from Laplace approximation to the numerator integral in (4) (Davison, 1986), while the second ensures that $L_{\mathrm{AC}}(z \mid y)$ has suitable invariance properties. Butler also discusses coverage properties of prediction intervals based on such likelihoods.

A natural alternative, the so-called estimative approach, bases inference for z on the conditional density $f(z \mid y;\, \hat{\theta})$, where $\hat{\theta}$ is typically the maximum likelihood estimate based on y, but by acting as if $\hat{\theta}$ equalled the true θ this understates the prediction uncertainty. A measure of this understatement based on the Kullback–Liebler divergence between the true density $p(z \mid y;\, \theta)$ and a predictive density, $q(z \mid y)$, is (Aitchison, 1975)

$$D_\theta(p, q) = \int dy\, f(y;\, \theta) \int dz\, p(z \mid y;\, \theta) \log\left\{\frac{p(z \mid y;\, \theta)}{q(z \mid y)}\right\}.$$

A candidate predictive density r is judged poorer than q if $D_\theta(p, r) > D_\theta(p, q)$. If the result is independent of θ, it turns out that the estimative density can be bettered by a predictive density of form (4), even if the prior used to construct it differs from that underlying the data generation mechanism. In such cases a 'smearing out' of the estimative density by averaging over values of $\hat{\theta}$ is always beneficial. Murray (1977), Ng (1980) and Komaki (1996) investigated the generality of this result, which holds for gamma and multivariate normal models, among others.

Harris (1989) suggested that the estimative density be averaged out by taking not $f(z \mid y; \hat{\theta})$ but $\int f(y \mid z; \hat{\theta}) f(\hat{\theta}; \theta)\, d\hat{\theta}$, and showed that with $\theta = \hat{\theta}$ this improves on the estimative density for the Poisson and binomial cases, and asymptotically in more general exponential families. He used parametric simulation to estimate the integral, but Vidoni (1995) suggested an analytical approach whereby (2) is used to replace the density of $\hat{\theta}$ and the integral is approximated by Laplace's method. This approach also yields prediction limits. More recent work in a *Bernoulli* article of Barndorff-Nielsen & Cox (1996) gives a general but somewhat complicated basis for approximate predictive intervals which are suitably invariant and satisfy the conditionality principle, while having good asymptotic properties under repeated sampling.

5. Estimating functions

Elementary discussion of maximum likelihood estimators usually stresses asymptotic efficiency rather than finite-sample properties, which are difficult to obtain because of the implicit nature of the score function. Starting in the 1960s finite-sample theory was developed from a different point of view, taking as basis score-like estimating functions that determine estimators rather than estimators themselves. If ψ is a scalar parameter to be estimated using data Y, the function $g(Y; \psi)$ is said to be an unbiased estimating function if $E\{g(Y; \psi)\} = 0$ for all ψ. With $g(y; \psi) = \partial \log f(y; \psi)/\partial\psi$ this determines the maximum likelihood estimator, but it also encompasses least squares, minimum chi-squared and robust estimators. In a paper in the *Annals of Mathematical Statistics* and under regularity conditions, Godambe (1960) defined g^* to be optimal in the class of all unbiased estimating functions if it minimised the ratio $E\{g^*(Y; \psi)^2\}/E\{\partial g^*(Y; \psi)/\partial\psi\}^2$ for all ψ. An asymptotic basis for this choice is that this ratio is the large-sample variance of $\hat{\psi}$ determined as the root of the equation $g(Y; \psi) = 0$. A Cramér–Rao argument then establishes that the estimating function $g(y; \psi) = \partial \log f(y; \psi)/\partial\psi$, the score function, is optimal in this finite-sample sense; this result extends to vector ψ. It was extended to give a finite-sample non-decision-theoretic justification for Bayesian inference on ψ by Ferreira (1982). Godambe & Thompson (1984) extend these ideas to robust estimation of a location parameter, giving theoretical tools to determine which among a variety of parameters can most effectively be estimated, while the extension to stochastic processes is given by Godambe (1985). Heyde & Morton (1993) discuss optimal estimating functions in the presence of constraints on ψ.

In more general situations the model also contains a nuisance parameter λ. Godambe (1976) showed that, if there is a complete statistic $S = s(Y)$ such that $f(y; \psi, \lambda) = f(y \mid s; \psi) f(s; \psi, \lambda)$, then the estimating function optimal for ψ is $\partial \log f(y \mid s; \psi)/\partial\psi$, corresponding to the use of a conditional likelihood for ψ. Further papers considered related definitions of ancillarity and sufficiency (Godambe, 1980) and made the link between orthogonality of estimating functions and optimal estimating functions (Godambe, 1991). Lindsay (1982) generalised results of Godambe (1976) to more general likelihood factorisations such as partial likelihood (Cox, 1975), and examined the case where the conditioning statistic depends on λ. Globally optimal estimating equations are then unavailable, but a weaker optimality criterion yields a class of estimated conditional score functions that satisfy a similar information inequality.

A general approach to such problems was developed by Waterman & Lindsay (1996) using Bhattacharyya scores. The motivation stems from the observation that, if S is sufficient for λ, then the residual $U_\psi - E(U_\psi \mid S)$ from the projection of the score $U_\psi = \partial \log f(y;\, \psi, \lambda)/\partial\psi$ on to the space of S-measurable functions equals the conditional score $U_c(\psi) = \partial \log f(y \mid s;\, \psi)/\partial\psi$, which is the optimal estimating function. This suggests basing inference on residuals from other projections of U_ψ in cases where an exact conditional score does not exist. To do so, it is useful to treat U_ψ as an element of a Hilbert space of square-integrable functions, subspaces of which may be generated by the quantities $f(y;\, \psi, \lambda)^{-1}\partial^t \log f(y;\, \psi, \lambda)/\partial\lambda_{a_1} \ldots \partial\lambda_{a_t}$, the scaled tth-order derivatives of $f(y;\, \psi, \lambda)$ with respect to the elements of λ. Nested subspaces $\mathscr{B}_t$ of the Hilbert space are determined by taking $\mathscr{B}_0$ to be the subspace generated by the constant function, $\mathscr{B}_1$ to be that generated by the constant and first-order derivatives, $\mathscr{B}_2$ to be that generated by the constant, first- and second-order derivatives and so forth. We can now define a sequence of approximate scores by taking the residuals from projection of U_ψ on $\mathscr{B}_1, \mathscr{B}_2, \ldots$. If λ can be eliminated by conditioning on the statistic S independent of ψ, then the optimal conditional score can be approximated arbitrarily closely by taking successive elements of this sequence, without finding the conditioning statistic S or explicit calculation of U_c; λ is replaced by its maximum likelihood estimate. In problems where S depends on ψ, the nuisance parameter cannot be fully eliminated and properties of the score residual depend on the estimator of λ used. Examples suggest that $t = 2$ will often be adequate in practice.

A different approach is to seek a combination of data and ψ whose distribution depends only on ψ, and to base inference on this. Cox (1993) discusses the extent to which this is useful, and the modifications to the score equations needed for it to work. Morton (1981) takes a different approach, seeking a type of pivotal quantity $g(Y;\, \psi)$ whose distribution may depend on ψ but not on λ, and from which estimating functions may be formed; this turns out to be related to the residuals from projected scores discussed above.

6. Model selection

The labour of statistical arithmetic has been vastly reduced by the advent of modern computing. One effect of this is that many models are now routinely fitted to a single dataset, particularly but not exclusively in time series and regression analysis. The problem of multiple testing entailed by standard approaches to model comparison using likelihood ratio and related statistics was underscored when computers began to be widely used during the 1960s, and since the early 1970s a major effort has been devoted to methods for selection of the 'best' model or, more sensibly, a few 'good' models for a dataset. In time series this may boil down to selection of the order of an autoregressive process and in regression to the choice of covariates to be retained in a final model. Such automatic procedures treat all models on an equal footing, but in applications it is essential to appreciate that some models are more equal than others: substantive interpretation and consistency with established knowledge must also be taken into account.

A widely used approach is to choose the model that minimises Akaike's information criterion AIC $= 2(-\hat{l} + p)$, where $\hat{l}$ is the maximised loglikelihood when a p-parameter model is fitted (Akaike, 1973). This may be derived by an asymptotic argument as an estimate of minus the expected loglikelihood for the fitted model, with expectation taken with respect to the true model, supposed to be among those fitted, but there are important connections to prediction (Larimore, 1983). The reduction in $-\hat{l}$ when parameters are added is traded off against the penalty p for model complexity. Several variant criteria have been published, for example to allow for the true model not being among those fitted, or to change the multiplier of p (Bhansali & Downham, 1977; Kohn, 1977; Akaike, 1979), but their performance depends on the family of models in which the 'truth' is embedded (Atkinson, 1980). Unfortunately minimisation of AIC does not give consistent model selection, and in practice it often indicates models more complex than are warranted by the data. The difficulty is that complexity is insufficiently penalised. In a series of papers Hurvich & Tsai (1989, 1991, 1998) and Hurvich et al. (1990) have investigated a related approach for models with normal errors in which the penalty is $n(n+p)/(n-p-2)$ rather than p. The resulting criterion AIC$_c$ can appreciably increase the probability of choosing the correct model, particularly in conventional regression and time series model selection, but also in some wavelet applications and in multivariate regression (Fujikoshi & Satoh, 1997). Burman & Nolan (1995) discuss the generalisation of AIC to robust regression models.

Theoretical properties of AIC and related statistics in regression contexts have been investigated by Shibata (1981, 1984), Hurvich & Tsai (1995) and others, particularly in the case where the underlying model is not among those fitted, so that comparison with nonparametric regression procedures is appropriate.

There is a close connection to crossvalidation, also used for model selection. For *Biometrika*'s contributions to this, see Hall (2001).

7. Bayesian statistics

The Bayesian revival that began in the 1950s soon led to *Biometrika* publishing investigations of particular models important in applications, such as the linear model (Tiao & Zellner, 1964) and random effects models (Tiao & Tan, 1965, 1966), but also to broader methodological discussions, concerning particularly the robustness of Bayesian inferences. Examples are Box & Tiao (1962, 1964), who assess the sensitivity of posterior densities to distributional assumptions, replacing the normal with a heavier-tailed density in work prefiguring current practice, and the investigation of modelling of outliers by mixtures in Box & Tiao (1968).

The role and use of prior knowledge has always been a stumbling block to the universal acceptance of the Bayesian approach. Jeffreys attempted to circumvent this by constructing what he hoped could be regarded as objective priors, but these are often improper. The unease surrounding their use deepened when Stone & Dawid (1972) pointed out the marginalisation paradoxes that can arise with improper priors. Kass (1990) gives heuristic justification for the use of Jeffreys priors in cases where the likelihood boils down to a location family, though not necessarily in the data themselves. Reference priors provide an alternative to Jeffreys priors, particularly when parameters are ordered by their degree of interest for the analyst, and Sun & Berger (1998) discuss their construction in cases where there are different amounts of information about different parameters.

In empirical Bayes inference the sample is used to determine the parameters of the prior. The original procedure proposed by Robbins (1956) is unstable except in large samples, but in special cases smooth estimators of the marginal density of the data were shown by Maritz (1966, 1967, 1968) to have considerably lower risk than the original proposal. An interesting application of empirical Bayes is to the estimation of the size of Shakespeare's vocabulary (Efron & Thisted, 1976; Thisted & Efron, 1987), while Copas (1972) shows how empirical Bayes can improve precision when a standard treatment is used repeatedly in clinical trials. Efron & Morris (1972) discuss a vector generalisation of the celebrated James–Stein theorem on the inadmissibility of the sample average, and describe associated empirical Bayes estimators. Hill (1990) attempts to tie empirical Bayesian, frequentist and regular Bayesian inference into a single general framework.

A topic of perennial interest is that of when, how and by how much Bayes and frequentist inferences differ, discussed, for example, by Bartholomew (1965) and Datta & Ghosh (1995); the latter investigate the choice of prior that gives Bayesian credible and frequentist confidence intervals the same approximate coverage probability. Similar investigations are by Mukerjee & Dey (1993) and Datta (1996). Related ideas were exploited by Tibshirani (1989), who discusses the construction of priors that are noninformative for a single parameter of interest in problems where there are many nuisance parameters to which it is orthogonal; it turns out that this is not the Jeffreys prior, though it involves the Fisher information matrix. A more systematic treatment is by Liseo (1993), who compares the effect of using Jeffreys and reference priors in a variety of classical problems, concluding that reference priors generally give more acceptable inference than do classical methods or Jeffreys priors.

Bayesian modelling provides a simple way to avoid the sometimes awkward approximations needed for nonregular likelihood analysis. A particular case is in inference for changepoints, as in Smith (1975) and Raftery & Akman (1986).

As mentioned in § 9·1, Laplace and related approximations provide useful tools for dealing with the integrals that arise in Bayesian applications, and there are close links to classical ideas described in § 3·3. Sweeting (1995a, b) shows how likelihood-based approximations to conditional p-values, pivots and suchlike can be derived from their Bayesian counterparts using an unsmoothing argument, and investigates the relationships between these various approximations.

An important development around 1970 was the use of hierarchical models, in which parameters are treated as exchangeable. Smith (1973) discusses their application in simple analysis of variance models, making comparisons with least squares solutions, while Fearn (1975) treats growth curve modelling. Dawid (1977) discusses the implications of invariance properties weaker than exchangeability for Bayesian analysis of variance; see also Goldstein & Wooff (1998) for discussion of weak forms of exchangeability in linear regression. Hierarchical models are now widely used; see, for example, Raftery (1988) who gives an application to population size estimation, Malec & Sedransk (1992) and Mallick & Walker (1997) who describe their use in meta-analysis, and Roeder et al. (1998) who suggest how they can be used to account for variability in forensic science databases; for other Bayesian applications in forensic science, see Lindley (1977) and David & Mortera (1998). Difficulties that can arise when there are constraints on the parameter space of a hierarchical model are tackled by Chen & Shao (1998) using a version of importance sampling.

The Bayesian analogue of the significance test is the Bayes factor, discussed by Gûnel & Dickey (1974) in the context of contingency table modelling. McCulloch & Rossi (1992) describe an attempt to compute Bayes factors for hypotheses that impose nonlinear constraints on the parameter space, with applications to logistic regression. Raftery (1996) discusses their construction using Laplace approximation for families of generalised linear models and suggests how model uncertainty can be incorporated into inference in this setting, which is also discussed by Goutis & Robert (1998) using Kullback–Leibler divergence. Related work is by Dellaportas & Forster (1999), who use a reversible jump Markov chain Monte Carlo algorithm for model comparison in hierarchical and graphical log-linear models. There are close connections between approximations to the Bayes factor, of which the best-known, the Schwarz criterion, can be derived by Laplace approximations to the marginal density of the observed data, and classical analogues such as AIC. Pauler (1998) outlines how these approaches may be applied for mixed models with normal errors.

Goldstein (1976, 1980) considers the Bayesian use of a linear model as an approximation to an underlying nonlinear regression structure, choosing the regression coefficient to minimise the posterior expected sum of squares; this depends only on certain joint posterior moments.

A very important recent development has been the emergence of Markov chain Monte Carlo methods for use in Bayesian applications. This is discussed in § 8.

8. Simulation

Simulation has been used to guide theoretical work since "Student's" (1908) derivation of the t statistic, and for finite-sample comparison of statistical procedures since the 1960s, but only in the last 20 years has it become integral to routine data analysis. The potential of simulation for frequentist data analysis was highlighted by Efron's (1979) account of the bootstrap; see Hall (2001) for *Biometrika*'s contributions. Possibilities for Bayesian analysis were perceived somewhat later, but there is now a large literature primarily about Markov chain Monte Carlo simulation. The idea was first suggested by Metropolis et al. (1953) and became standard in fields such as statistical mechanics and physical chemistry. Its importance dawned on workers in spatial statistics and image analysis during the 1970s and 1980s, before the possibilities it offers for investigation of posterior distributions and likelihood functions were initially realised around 1990.

Suppose that we desire to sample from a distribution π on a large finite set, $\mathcal{S}$, and that π is known only up to a normalising constant, it being infeasible to visit every element of $\mathcal{S}$. Consider an irreducible aperiodic Markov chain X_t on $\mathcal{S}$, with q_{ij} denoting the probability of transition from state i to state j, for $i, j \in \mathcal{S}$; suppose that $q_{ij} = q_{ji}$, so the transition probability matrix is symmetric. Then Metropolis et al. (1953) suggest that a new state X_{t+1} be chosen as follows: given $X_t = i$, generate a proposal state $X'_{t+1} = j$ using the transition probability q_{ij}; then set $X_{t+1} = X'_{t+1}$ with probability $a_{ij} = \min\{\pi_j/\pi_i, 1\}$ and otherwise take $X_{t+1} = X_t$. If the resulting transition probabilities are p_{ij}, it is easy to verify that $\pi_i p_{ij} = \pi_j p_{ji}$ for all i and j, so the induced chain X_t is reversible with stationary distribution π. We can then invoke the ergodic theorem and estimate functions of π using the generated values $X_1, \ldots, X_T$, provided that T is large enough; the expected value of $f(X)$ with respect to the stationary distribution π is estimated by $T^{-1}\sum_t f(X_t)$.

This algorithm became widely used outside statistics, though the requirement that $q_{ij} = q_{ji}$ is highly restrictive. An important generalisation was the realisation of Hastings (1970) that, if a reversible chain is sought, and the acceptance probability is changed to $a_{ij} = \min\{(\pi_i q_{ij})/(\pi_j q_{ji}), 1\}$, then the transition probability matrix q_{ij} of the underlying chain need not be symmetric: the matrix corresponding to essentially any irreducible aperiodic chain can be used, though the choice of q_{ij} will strongly influence the practical value of the algorithm. In finite state spaces, Peskun (1973) showed that autocorrelation in the chain is reduced by making the acceptance probability as large as possible, subject to retaining the detailed balance condition $\pi_i p_{ij} = \pi_j p_{ji}$ for all i and j. The Hastings–Metropolis algorithm extends to continuous state spaces, and it and variants such as the Gibbs sampler are now used in a huge range of Bayesian models (George & Robert, 1992; Carlin & Gelfand, 1993; Glad & Sebastiani, 1995; Müller & Roeder, 1997; Chib & Greenberg, 1998).

In such applications the distribution π is typically the posterior distribution of parameters and predictands, given the data. In frequentist statistics it may represent a distribution whose support or normalising constant is unknown. Besag & Clifford (1989, 1991) use such algorithms to construct null distributions for test statistics in complex problems, such as assessing the adequacy of models for contingency tables, of which an example is the Rasch model. In such cases the parameters may be eliminated by conditioning on suitable margins of the table, but the conditional distribution has sample space and normalising constant that are typically so awkward that the Hastings–Metropolis algorithm is the best means of estimating conditional p-values. The same ideas may be used to estimate conditional likelihoods, though importance sampling also has a role (Booth & Butler, 1999).

The use of Markov chains in this context is essentially constructive; subject to mild conditions, the analyst is free to build the algorithm in many possible ways. A natural idea is to implement it so as to maximise speed of convergence. One approach to this is to minimise posterior correlations among parameters, as discussed in normal linear mixed models by Gelfand et al. (1995), while another approach is post-simulation improvement to get the effect of a longer run of the chain (Liu et al., 1994; Casella & Robert, 1996). Roberts & Tweedie (1996) give sufficient conditions for geometric ergodicity of certain algorithms and discuss central limit theorems in this context. There are numerous approaches to assessing if a chain has converged; see, for example, Brooks & Roberts (1999) for a critical discussion of the use of quantiles of the output in this regard.

If the parameter space is continuous but its dimensions may vary, then the discussion above does not apply. Unfortunately, this situation is common in more complex applications, where, for example, a choice must be made between competing models or where a mixture distribution has an unknown number of components. Green (1995) generalised the Hastings–Metropolis algorithm to this situation by introducing reversible jump Markov chain Monte Carlo algorithms. Suppose that the parameter space for the kth of a countable collection of candidate models is $\mathscr{C}_k = \mathbb{R}^{n_k}$, with $n_1 < n_2 < \ldots$, and write the joint density of the model index k, its parameter $\theta^{(k)}$ and the data y as $\pi(k)\pi(\theta^{(k)} \mid k) f(y \mid k, \theta^{(k)})$. Bayesian inference in this context uses the posterior density $\pi(k, \theta^{(k)} \mid y)$, typically expressed as $\pi(k \mid y)\pi(\theta^{(k)} \mid y, k)$; often it makes little sense to average over different models. The goal of setting up a Markov chain with this posterior density as its stationary distribution is complicated by the varying dimension of $\theta^{(k)}$. As the chain is constructed so as to be reversible, then, if a transition is allowed between $\mathscr{C}_1$ and $\mathscr{C}_2$ say, the reverse jump must also be possible. The difficulty lies in setting up the jumps in such a way that the equilibrium joint distribution of the current and proposed next states has a density with respect to any symmetric dominating measure. This is achieved through an ingenious 'dimension-matching' condition, whereby the jump from $\mathscr{C}_1$ to $\mathscr{C}_2$ lands in a subspace of dimension n_1 determined by the point of origin and an auxiliary random variable of dimension $n_2 - n_1$, while the reverse jump treats the departure point in C_2 as an element of a subspace of dimension n_1. The resulting algorithm is widely applicable; see, for example, Denison et al. (1998) and Rue & Hurn (1999).

9. Saddlepoint and related approximations

9·1. *Laplace approximation*

A recurrent theme of statistical work is the importance of small-sample methods. Fisher, Student, Wishart and others derived exact results for many statistics derived from the normal model, for example in analysis of variance and multivariate statistics. One obvious choice of approximations for wider use is based on Edgeworth series and Fisher–Cornish inversion, but, though important in the theoretical study of density, distribution and quantile approximations, these can give negative density estimates and non-monotone approximations to distributions and quantile functions when applied in practice. A natural alternative toolkit uses asymptotic expansions for integrals, such as Laplace and saddlepoint approximations. The latter have been intensively studied over the past 30 or so years, building on the seminal *Annals of Mathematical Statistics* paper of Daniels (1954).

Suppose that we are confronted with an integral based on a sample of size n and written in the form

$$\int \exp\{-nh(u)\}\,du, \tag{5}$$

where $h(u)$ is $O(1)$, sufficiently well behaved, strictly convex in u and attains its minimum at $\tilde{u}$, at which point $\partial h(u)/\partial u = 0$. Here $nh(u)$ is typically a negative loglikelihood or some variant thereof. The key idea of Laplace approximation is Taylor series expansion of $h(u)$ about $\tilde{u}$ and neglect of the third- and higher-order terms, followed by integration of the resulting normal density. The result is

$$\left\{\frac{(2\pi)^p}{n^p|\partial^2 h(\tilde{u})/\partial u\,\partial u^{\mathrm{T}}|}\right\}^{\frac{1}{2}} \exp\{nh(\tilde{u})\}\{1+O(n^{-1})\}, \tag{6}$$

where p is the dimension of u and $|.|$ indicates the determinant; unlike an Edgeworth series expansion, (6) is guaranteed to be nonnegative under mild conditions on h. Moreover, the error term is relative, so this approach can yield approximations that work well for both small and large values of the integrand, unlike Edgeworth series, whose error is typically absolute and can therefore greatly exceed the approximation. Cox (1948) used this approach to derive approximate densities and distributions for the sample range, in work extended to differences of order statistics by Harvill & Newton (1995). Daniels (1956) applied the ideas to small-sample distributions for the serial correlation coefficient. Other applications are in mixed models (Wolfinger, 1993; Vonesh, 1996).

For distribution function approximations it is better to consider integrals of the form

$$\left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \int_{-\infty}^{u_0} a(u)e^{-ng(u)}\{1+O(n^{-1})\}\,du, \tag{7}$$

where u is scalar, $a(u) > 0$, and, in addition to possessing the properties of $h(u)$ above, $g(.)$ is such that $g(\tilde{u}) = 0$. Under regularity conditions and after several changes of variable, it turns out that (7) may be expressed as $\Phi(n^{\frac{1}{2}}r_0^*) + O(n^{-1})$, where

$$r_0 = \mathrm{sign}(u_0-\tilde{u})\{2g(u_0)\}^{\frac{1}{2}}, \quad r_0^* = r_0 - (r_0 n)^{-1}\log\left\{\frac{a(u_0)r_0}{g'(u_0)}\right\}. \tag{8}$$

The first of these expressions corresponds to a signed loglikelihood ratio statistic, and the second, which has a standard normal distribution to high order, is a modified version thereof. Fraser (1990), DiCiccio et al. (1990), DiCiccio & Martin (1991), DiCiccio & Field (1991) and Fraser et al. (1999) describe similar approaches for such integrals, for both Bayesian and frequentist problems, while Cheah et al. (1994) apply them to testing problems in exponential families.

Integrals such as (5) arise very commonly in Bayesian statistics, for example in calculating the marginal posterior density

$$\pi(\psi \mid y) = \frac{\int \pi(\psi, \lambda \mid y)\, d\lambda}{\int \pi(\psi, \lambda \mid y)\, d\lambda\, d\psi}$$

of a parameter ψ of interest in the presence of a nuisance parameter λ. The denominator integrand has form (5), with $h(u)$ replaced by $-n^{-1} \log \pi(\psi, \lambda \mid y)$, that is, minus the scaled loglikelihood plus log prior; then $\tilde{u}$ corresponds to the mode of the joint posterior density. The numerator integrand has a similar form, but depends on ψ. Tierney et al. (1989), Wong & Li (1992) and others have investigated the use of Laplace's method in this context, for posterior predictive densities and for posterior moments, while Kass et al. (1989) have considered how it can be used for assessment of influence and sensitivity in Bayesian analysis. DiCiccio et al. (1993), Sweeting (1995a) and Fraser et al. (1999) have considered related tail probability approximations.

9·2. *Saddlepoint approximation*

Saddlepoint approximation starts by considering the average $\bar{Y}$ of a sample of continuous independent and identically distributed random variables $Y_1, \ldots, Y_n$, whose cumulant-generating function, $K(u) = \log E(e^{uY})$, is assumed to exist in some open interval $a < u < b$ containing the origin. The Laplace inversion formula implies that $\bar{Y}$ has density given by the complex integral

$$f(\bar{y}) = \frac{n}{2\pi i} \int_{c-i\infty}^{c+i\infty} \exp[n\{K(u) - u\bar{y}\}]\, du,$$

where any path of integration for which $a < c < b$ may be used. The idea is to choose c to pass through a saddlepoint of the integrand, outside a neighbourhood of which the integrand is negligible. An argument like that leading to (6), but complicated by the necessity for complex integration, yields

$$f(\bar{y}) = \left\{\frac{n}{2\pi K''(\tilde{u})}\right\}^{\frac{1}{2}} \exp[n\{K(\tilde{u}) - \tilde{u}\bar{y}\}]\{1 + O(n^{-1})\}, \tag{9}$$

where $\tilde{u}$ satisfies the saddlepoint equation $K'(u) = \bar{y}$. Not only is the error in (9) relative rather than absolute, but it also turns out to be uniformly bounded in large-deviation regions, often giving an astonishingly accurate approximation far into the tail of the density. Generally the leading term on the right of (9) does not have unit integral, but when renormalised to do so the relative error improves to $O(n^{-3/2})$. The renormalised version is exact for the normal, gamma and inverse Gaussian densities (Daniels, 1980). Application of (7) to (9) gives an approximation to the probability that $\bar{Y} \leqslant y$, of form $\Phi(r^*)$, where r^* is given by

$$r^* = r + r^{-1} \log(v/r), \quad r = \text{sign}(u_y)[2n\{u_y y - K(u_y)\}]^{\frac{1}{2}}, \quad v = u_y / |K''(u_y)|^{\frac{1}{2}},$$

in which u_y satisfies the equation $K'(u) = y$. Here r is the signed likelihood ratio statistic for assessing whether or not the data have mean $E(Y)$, while v is the corresponding Wald statistic. Jensen (1992) showed that $\Phi(r^*)$ differs from the Lugannani–Rice approximation $\Phi(r) + \phi(r)(r^{-1} - v^{-1})$ by $O(n^{-3/2})$ in ordinary-deviation regions.

Although known in principle, the distribution of statistics used in multivariate analysis can be awkward to use in practice, making simple approximations desirable. Butler et al. (1992a,b) have advocated saddlepoint methods for this context. Similar comments apply to quadratic forms, used for comparison of parametric and nonparametric regression models (Kuonen, 1999).

The application of these approximations in small-sample likelihood inference was briefly discussed in § 3·3, and here we concentrate mainly on non- and semiparametric applications. One is to estimators defined through estimating functions $g(y; \psi)$, supposed monotonic decreasing in the scalar ψ for each y. Then, as pointed out by Daniels (1983), following Field & Hampel (1982), $\hat{\psi} \leqslant \psi$ if and only if $\sum g(Y_j; \psi) < 0$. Hence the probability that $\hat{\psi} \leqslant \psi$ may be read off from the distribution of the average $\sum g(Y_j; \psi)$, to which the results above apply directly. This idea applies to robust estimates of location and has found many uses elsewhere. Spady (1991) generalised it to regression problems, in which ψ is no longer scalar.

Such approximations are also useful in the context of finite-population sampling and resampling. Davison & Hinkley (1988) show how saddlepoint methods can be used to avoid Monte Carlo simulation in numerous bootstrap and randomisation contexts, while Booth & Butler (1990) set that work in a more general context and point out the link with exponential family distributions. Wang (1993) shows how the same ideas apply to finite-population sampling, while applications to time series autocorrelations are given by Phillips (1978) and Wang (1992). In the bootstrap context the approach is sometimes called the empirical saddlepoint; its accuracy has been investigated by Feuerverger (1989), Jing et al. (1994) and Monti & Ronchetti (1993). Daniels & Young (1991) use a mixture of saddlepoint and Laplace approximations to obtain the marginal distribution of the studentised average, as used in bootstrap resampling for confidence intervals for a mean; see also DiCiccio et al. (1992). Booth et al. (1999) describe the use of saddlepoint approximation and parametric asymptotics to avoid bootstrap calibration.

10. Generalised regression

10·1. *Generalised linear models*

One of the most important developments of the 1970s and 1980s was the unification of regression provided by the notion of a generalised linear model (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989) and its associated software, though the concept had appeared earlier (Cox, 1968). In such models the response Y is taken to have an exponential family distribution, most often normal, gamma, Poisson or binomial, with its mean μ related to a vector of regressor variables through a linear predictor $\eta = x^{\mathrm{T}}\beta$ and a link function g, where $g(\mu) = \eta$. The variance of Y depends on μ through the variance function $V(\mu)$, giving $\mathrm{var}(Y) = \phi V(\mu)$, where ϕ is a dispersion parameter. Special cases are: the normal linear model, for which $V(\mu) = 1$, ϕ is the error variance and the link is $g(\mu) = \mu$; models for gamma errors, $V(\mu) = \mu^2$ and link functions $g(\mu) = \mu^{-1}$ or $g(\mu) = \log \mu$; Poisson models for counts, with $V(\mu) = \mu$ and $g(\mu) = \log \mu$; and logistic models for proportions, with $V(\mu) = \mu(1 - \mu)/m$ and $g(\mu) = \log\{\mu/(1 - \mu)\}$. These last two models have $\phi = 1$. This brings under the same roof the linear model with normal errors, logistic and log-linear models and many less familiar regression models.

One aspect of this unification was the more widespread use of likelihood concepts, with different models compared through a likelihood ratio statistic called the deviance, thereby extending familiar notions from analysis of variance. A second was the realisation that iterative use of weighted least squares estimation can form the basis of a general algorithm for fitting regression models. The estimating equations for a generalised linear model for independent responses $Y_1, \ldots, Y_n$ and corresponding covariate vectors $x_1, \ldots, x_n$ may be expressed as

$$\sum_{j=1}^{n} x_j \frac{\partial \mu_j}{\partial \eta_j} \frac{Y_j - \mu_j}{V(\mu_j)} = 0, \qquad (10)$$

or in matrix form

$$D^{\mathrm{T}} V^{-1}(Y - \mu) = 0, \qquad (11)$$

where D is the $n \times p$ matrix of derivatives $\partial \mu_j / \partial \beta_r$, and the $n \times n$ covariance matrix V is diagonal if the responses are independent but not in general. Taylor expansion of (11) shows that the solution $\hat{\beta}$ may be found by repeated linear regression of a constructed response variable on the columns of D with weight matrix V^{-1}. The existence and uniqueness of $\hat{\beta}$ has been studied by Wedderburn (1976) and others, while Jørgensen (1983) discusses large-sample inference for such models and their nonlinear extensions.

The idea of decoupling the systematic and random structure of the model, that is, covariate matrix and response distribution, has obvious wider implications. One can envisage building models by fitting different covariate structures on to a chosen response distribution using an appropriate link, as in generalised additive models (Hastie & Tibshirani, 1990), in which smoothers are added to the linear predictor η to give a semiparametric model. Link functions themselves can be chosen on subject-matter or statistical grounds, or estimated parametrically or nonparametrically if the data are extensive enough.

10·2. *Quasilikelihood*

Data are often overdispersed relative to a textbook model. For example, although the variance of count data is often proportional to their mean, the constant of proportionality ϕ may exceed the value anticipated under a Poisson model, so $\mathrm{var}(Y) = \phi\mu$ for $\phi > 1$. One way to deal with this is to model explicitly the source of overdispersion by the incorporation of random effects; see § 3·5. The resulting integrals can considerably complicate computation of the likelihood, however, and a simpler approach is through quasilikelihood (Wedderburn, 1974).

Quasilikelihood is perhaps best seen as an extension of generalised least squares. To see why, note that (11) is equivalent to $U(\beta) = 0$, where $U(\beta) = \phi^{-1} D V^{-1}(Y - \mu)$. Asymptotic properties of $\hat{\beta}$ stem from the relations $E(U) = 0$ and $\mathrm{cov}(U) = -E(\partial U / \partial \beta)$, corresponding to standard results for a loglikelihood derivative. However, these properties do not depend on a particular probability model, requiring merely that $E(Y) = \mu$ and $\mathrm{cov}(Y) = \phi V(\mu)$, subject also to some regularity conditions. Hence $\hat{\beta}$ has the key properties of a maximum likelihood estimator, namely consistency and asymptotic normality, despite not being based on a fully-specified probability model. Moreover, it may be computed simply by solving (11), that is, behaving as if the exponential family model with variance function $V(\mu)$ were correct. The scale parameter ϕ is estimated by $\hat{\phi} = (n - p)^{-1}(Y - \hat{\mu})^{\mathrm{T}} V(\hat{\mu})^{-1}(Y - \hat{\mu})$, and the asymptotic covariance matrix of $\hat{\beta}$ is $\hat{\phi}(D^{\mathrm{T}} V D)^{-1}$ evaluated at $\hat{\beta}$. A unified asymptotic treatment of such estimators from overdispersed count and proportion data is provided by Moore (1986); see also Morton (1991), who discusses the choice of variance functions in the binomial case.

Despite being consistent and asymptotically normal, the quasilikelihood estimator $\hat{\beta}$ need not be asymptotically fully efficient. Firth (1987) showed that under modest overdispersion its efficiency is generally high, confirming an earlier conclusion of Cox (1983), who had established that overdispersion on the borderline of detectability has little effect on estimation of a suitably chosen parameter. It turns out that moment parameters are suitable, so that high efficiency can be expected if, for example, a log-linear model determines the expected value μ of an observed count; this is precisely the case with quasilikelihood analysis. If the correct variance function is used, then McCullagh (1983) has shown that quasilikelihood estimators have the smallest asymptotic variance among estimators derived from estimating functions that are linear in Y; this is a large-sample generalisation of the Gauss–Markov theorem. Crowder (1987) gives examples in which quasilikelihood estimators are inferior to those based on quadratic estimating equations, but these are mainly of theoretical interest, as the resulting procedures depend on the third and fourth cumulants of Y, which are typically unspecifiable and can be estimated rather poorly at best; see also Firth (1987). These cumulants also enter into the resulting standard errors, which therefore tend to be unreliable. A further difficulty is that, unlike in the linear case, estimators derived from quadratic estimating functions are consistent only if the variance function is correctly specified, and it may be hard to be sure of this in applications. One effect of misspecifying the variance function is to invalidate the usual chi-squared theory underlying score- and estimate-based tests and confidence intervals. One remedy for this is to use an 'information sandwich' (Rotnitzky & Jewell, 1990), replacing $(D^{\mathrm{T}}VD)^{-1}$ with $(D^{\mathrm{T}}VD)^{-1}D^{\mathrm{T}}V^{-1}\operatorname{cov}(Y)V^{-1}D(D^{\mathrm{T}}VD)^{-1}$, where $\operatorname{cov}(Y)$ is a residual-based estimator of the true variance function. Such estimators can be unstable, however, and in practice it can be better to sacrifice generality on the altar of stability.

Although its efficiency relative to maximum likelihood estimation will often be high, the relationship of quasilikelihood with generalised least squares implies that its resistance to outliers is low. Morgenthaler (1992) investigates the extent to which this may be overcome by least absolute deviation estimation of β, in which $\sum|(Y_j-\mu_j)/V(\mu_j)^{\frac{1}{2}}|^q$ is minimised. For discrete data this introduces a bias which can only be removed by assuming that the underlying distribution is fully known. This is not required for continuous data, and when $q=1$ this approach amounts to modelling the median rather than mean response.

As mentioned above, the equation $U(\beta)=0$ is an analogue to the usual likelihood equation, and it is natural to ask under what circumstances a log-quasilikelihood can be obtained by integrating $U(\beta)$, just as a loglikelihood is obtained by integrating a score. When the responses are independent, we can define a log-quasilikelihood as

$$Q(\beta;\, y)=\sum\int_{y_j}^{\mu_j}\frac{y_j-u}{\phi V_j(u)}\,du,$$

where $V_j(u)$ is the jth diagonal element of the covariance matrix $V(\mu)$, which gives $\partial Q(\beta;\, y)/\partial\beta=U(\beta)$. If $V(\mu)$ is not diagonal, however, it is typically impossible to produce a unique function Q, because the integral depends on the path along which it is evaluated; the corresponding Q is not a conservative vector field. Hanfelt & Liang (1995), following Li (1993), have investigated the extent to which this non-uniqueness is important, and conclude that in many circumstances approximate likelihood ratios can be defined that are sufficiently well behaved for inference in large samples; see McLeish & Small (1992) for related discussions and further examples.

In practice, the elements of $\mathrm{cov}(Y)$ may involve other parameters, giving

$$\mathrm{var}(Y_j) = \phi(x_j;\ \gamma)V_j(\mu_j;\ \lambda),$$

say. With independent responses, examples of this are $\phi = 1$, $V_j(\mu_j;\ \lambda) = \mu_j + \lambda\mu_j^2$, which arises on supposing that the mean of a Poisson variable is itself distributed as gamma with mean μ_j and shape parameter λ^{-1}, and

$$\phi(x;\ \gamma) = \exp(x^{\mathrm{T}}\gamma), \quad V_j(\mu_j;\ \lambda) = 1,$$

which enables joint modelling of mean and dispersion as dependent on covariates, useful in industrial quality experiments. Although this idea can be generalised to other settings, the normalising constant required for comparison of values of λ and γ is missing from the quasilikelihood. This led Nelder & Pregibon (1987) to extend quasilikelihood by adding extra terms to $Q(\beta;\ y)$ to enable λ and γ to be estimated; one interpretation of the resulting log extended quasilikelihood is through saddlepoint approximation to a linear exponential family with the chosen variance function, if one exists. Unfortunately, the resulting estimators of λ and γ are inconsistent, though they can behave adequately in small and moderate samples.

10·3. *Longitudinal, clustered and other correlated data models*

Applications of generalised linear models may involve numerous individuals, on each of which correlated responses are obtained. As with models for independent data, the responses may be continuous, binary or counts; sometimes more complicated observations are available, as for example with rainfall data, which typically comprises positive values interspersed with zeros. A medical example might involve binary repeated measures on individuals collected on several occasions, perhaps with missing data. This falls into the framework above, with the variance matrix block diagonal, the blocks corresponding to individuals. A simple structure for the jth block is $A_j^{1/2}R_j(\alpha)A_j^{1/2}$. Here A_j is a diagonal matrix containing the posited variance for the n_j responses on the jth individual, of form $\mathrm{diag}\{\mu_{j1}(1-\mu_{j1}), \ldots, \mu_{jn_j}(1-\mu_{jn_j})\}$ if the responses are binary, for example; and $R_j(\alpha)$ is a correlation matrix whose structure depends on the supposed underlying dependence. The simplest specification, $R_j(\alpha) = I$, ignores correlation among the responses for each individual, but other possibilities include modelling serial correlation among them, or taking an equicorrelation matrix in which all the off-diagonal elements equal a single intra-cluster correlation parameter α. In an influential paper, Liang & Zeger (1986) develop a modelling strategy using estimators based on such models, calling the resulting analogue of (11) a generalised estimating equation. They use information sandwich variances like those in § 10·2.

This approach has a number of potential advantages over fully parametric competitors such as log-linear and logistic models. In applications interest typically focuses on how mean parameters change in response to varying covariates, that is, on marginal features of the model, but in complex applications it can be hard to express this simply in terms of the natural parameters of the corresponding exponential family, which express conditional dependencies among variables. This difficulty is overcome by use of generalised estimating equations, in which marginal changes in means may be expressed in a natural way. Other advantages are robustness of the estimating equation approach to incorrect choice of variance function and the possibility of allowing for overdispersion. A good deal of subsequent work has refined the initial approach taken by Liang & Zeger (1986); see, for example, Lipsitz et al. (1991), who show that use of odds ratios rather than correlations has advantages for binary data, and Liang et al. (1992) and its discussion. Crowder (1995), however, cautions that a bad choice of working correlation matrix $R_j(\alpha)$ can lead to the disaster of inconsistent estimators of β and α; thought is needed in specifying the estimating functions to be fitted.

These and related ideas have been extended to models for count data with hierarchical covariance structure (Morton, 1987) and to continuous data with multiplicative errors (Firth & Harris, 1991), where there is a rather simple analogue of standard analysis of variance procedures for split-plot-like experiments. A rather different approach is taken by Zeger (1988), who develops quasilikelihood estimation for time series of counts. Paik (1996) discusses quasilikelihood estimation for models with missing covariates and Buonaccorsi (1996) discusses an estimating function approach to measurement error for general regression models, while Preisser & Qaqish (1996) extend deletion diagnostics to estimators determined by generalised estimating equations.

10·4. *Local models*

One major change during the last two decades has been the development, implementation and now widespread use of smoothing procedures. A wide range of methods for local density and curve estimation, each with its advantages and disadvantages, is now available to the data analyst. Contributions in *Biometrika* to this area are reviewed in Hall (2001), and here we simply note some connections with the regression models discussed above. One approach to local estimation of the mean $\mu(x)$ of a response Y as a function of the scalar covariate x is through weighting the contribution to a system of estimating equations according to their distance from the point at which local estimation is required. Then (10) becomes

$$\sum_{j=1}^{n} h^{-1} w\left(\frac{x_j - t}{h}\right) \frac{\partial \mu_j}{\partial \beta} \frac{Y_j - \mu_j}{V(\mu_j)} = 0, \qquad (12)$$

where t is the value of x at which it is required to estimate μ, $w(.)$ is a weighting function such as the normal density, and h is a bandwidth. As $h \rightarrow \infty$ the system reduces to (10), while as $h \rightarrow 0$ the estimation is based entirely on the observations closest to t. One interpretation of (12) is that each observation is given a dispersion parameter ϕ_j inversely proportional to its weight $h^{-1}w\{(x_j - t)/h\}$, so data for which x is close to t are observed more precisely than those further away. Evidently the ideas underlying our earlier discussion are also applicable here: the data may be overdispersed relative to likely models, with consistent estimation of $\mu(t)$ now possible as $n \rightarrow \infty$ and $h \rightarrow 0$ with $nh \rightarrow \infty$, even if the chosen variance function is wrong. A local model such as this is particularly useful when assessing the adequacy of a parametric form for $\mu(t)$, and this is discussed by Azzalini et al. (1989) and Firth et al. (1991).

11. Miscellanea

11·1. *Graphical methods*

Graphical representations of data play a crucial role at all stages of applied work, and has become increasingly important with the rise of the computer. One useful tool is the quantile–quantile plot, widely used both to compare different samples and to compare a single sample with quantiles of a probability distribution, perhaps even more common when applied to regression residuals. Wilk & Gnanadesikan (1968) discuss these and related probability plotting techniques, with many examples, while Michael (1983) and Coles (1989) describe an improved probability plot whose points are variance-stabilised, and related test statistics.

There is a close connection between probability plots and goodness-of-fit statistics that use the empirical distribution function, which have been extensively investigated; see, for example, Durbin (1961, 1975) and Stephens (1977, 1979).

Graphical methods are important in multivariate statistics. The simplest is the scatterplot. Fisher & Switzer (1985) describe the chi-plot, a transformation of the variables in a scatterplot intended to emphasise different forms of dependence. A more sophisticated tool is the biplot (Gabriel, 1971), which has proved useful as a dimension-reducing tool in principal components analysis and elsewhere, extended from the familiar Pythagorean metric to nonlinear metrics by Gower & Harding (1988) and further generalised by Gower (1990, 1992).

11·2. *Robustness and diagnostics*

Although the term 'robust' seems to have been introduced by Box (1953), the notion that inferences should be stable across a variety of models had been expressed much earlier; for example, Rider (1929) investigated the properties of the t statistic in nonnormal data. Analysis of variance was rapidly seen to be a major advance whose simplest basis is the assumption of normal errors, but there was concern about its sensitivity to this. Egon Pearson considered this analytically and by simulation, concluding that moderate nonnormality posed no serious problem (Pearson, 1928, 1929, 1931). Randomisation lends validity to analysis of variance regardless of the assumed error distribution, and was shown by Pitman (1937), Welch (1937) and others to give results very close to those obtained under normal theory. The alphabet soup of modern robustness developed from the 1960s onwards; tidbits floating in *Biometrika* are Devlin et al. (1975), McKean & Hettmansperger (1978) and Stefanski et al. (1986).

A counterpart to the idea of robustifying statistical methods is the use of diagnostic techniques, intended to detect departures from assumptions, which are widely used particularly in regression modelling; the associated ideas of goodness of fit, residuals and influence measures are commonplace. A key notion is that tests for model failure be accompanied by plots intended to highlight possible sources of difficulty and to suggest remedies. Failures can be general or caused by a few unusual observations, and the literature for detecting them splits along these lines. Examples of the first include the score test for heteroscedasticity in regression and the nonlinearity plot proposed by Cook & Weisberg (1983, 1994), while the second are represented by the simulation envelopes and approach to the detection of masking suggested by Atkinson (1981, 1986) and the discussion of influence in generalised linear models in Thomas & Cook (1989).

An approach to Bayesian outlier detection that avoids modelling outliers is described by Chaloner & Brant (1988), who compute the posterior probability that the unseen error is large relative to the observed standard deviation. Weiss & Cook (1992) describe an approach to Bayesian influence analysis.

11·3. *Multivariate analysis*

Although Karl Pearson developed principal component analysis as an exploratory tool from a geometric viewpoint at the very beginning of the 20th century (Pearson, 1901), the formal methods of multivariate analysis using the normal model began to develop around 1930, spurred by Wishart's (1928) derivation of the joint distribution of the covariance matrix for a multivariate normal sample. Wilks (1932) applied the ideas of Neyman and Pearson to hypothesis tests on mean vectors and variance matrices for multivariate normal populations, and hence obtained the statistic that now bears his name; this work was amplified by Lawley (1938) and Hsu (1940). Pearson & Wilks (1933) described a systematic approach to multivariate analysis of variance, giving examples of the application of the earlier results. Hotelling (1936) introduced canonical correlation analysis in a massive and influential paper that was followed up by Bartlett (1941), who made connections to likelihood ratio and other testing criteria. Other authors who applied likelihood methods to multivariate models were Plackett (1947) and Rao (1948), the latter author introducing the famous cork data, while Pillai (1956) discusses the distributions of extreme eigenvalues in multivariate analysis.

In a brief note Welch (1939) discussed the foundations of discriminant analysis, advocating general use of the likelihood ratio and pointing out that it yields Fisher's linear discriminant rule in the spatial case of normal populations. Subsequent workers have greatly extended discrimination, in particular to discrete data (Anderson, 1972; Aitchison & Aitken, 1976; Titterington, 1977) and to mixed discrete-continuous data (Krzanowski, 1979; Vlachonikolis, 1990).

A major unification was that of Jöreskog (1970), who developed maximisation algorithms for likelihood estimation for the multivariate normal model with structural covariance matrices of the type arising in factor analysis, educational testing, variance components and linear structural relationships.

An important development since 1980 has been the expression of multivariate dependence in terms of graphical models, an idea that has its roots in path analysis and which is useful in many areas of the social and medical sciences. For example, Wermuth & Lauritzen (1983), in work related to Goodman (1973), show how restrictions on the conditional independence structure of the variable in a contingency table yield two classes of models, one graphical and one recursive, respectively appropriate for studying associations among variables and for expressing relationships between explanatory and response variables. The intersection of these classes is the important class of decomposable models, which admits simple sufficient statistics and maximum likelihood estimators (Frydenberg & Lauritzen, 1989). Practical issues such as model selection, fitting and interpretation are considered by Edwards & Kreiner (1983) and Edwards & Havranek (1985). Tests on graphical Gaussian models involve constraints on the inverse covariance matrix, setting an element of which to zero amounts to asserting that the corresponding variables are conditionally independent given the rest. Roverato & Whittaker (1998) use Isserlis matrices to give a general discussion of properties of these models. Asmussen & Edwards (1983) discuss the types of model that may be appropriate when response variables are involved and the relation to collapsible tables; see also Madigan & Mosurski (1990). Cox & Wermuth (1992) investigate the construction of joint models for binary and continuous responses, with emphasis on those derived from the normal distribution.

Notions of causality lie at the root of much statistical analysis, particularly in the medical and social sciences, but have mostly been discussed informally until fairly recently. Rosenbaum & Rubin (1983) give an account of notions of causality in observational studies that centres on the role of the propensity score, the conditional probability of assignment to a treatment given the observed covariates, adjustment for which removes bias due to such covariates. Aspects of the extent to which graphical models can aid causal inference are dissected in Pearl (1995) and its useful discussion.

11·4. *Missing data*

Missing data arise in many contexts and had been dealt with informally for many years before the first general account of their potential effect on likelihood inference (Rubin, 1976). Let Y denote the full set of observations that might be made, and let M be a collection of indicators of whether or not the corresponding elements of Y have been observed. Thus the observed value (Y, M) might be $(y, 1)$ or $(?, 0)$ if Y was scalar. More generally, let y_{obs} and y_{mis} denote the observed and missing parts of y. Suppose that we wish to base inference for θ on the observed data, (y_{obs}, m), the likelihood for which may be written

$$f(y_{\text{obs}}, m;\, \theta, \gamma) = \int \text{pr}(m \mid y_{\text{obs}}, y_{\text{mis}};\, \gamma) f(y_{\text{mis}};\, \theta)\, dy_{\text{mis}} f(y_{\text{obs}};\, \theta),$$

where γ is a nuisance parameter which may play a role in the missingness mechanism $\text{pr}(m \mid y;\, \gamma)$. It is assumed that θ and γ are distinct parameters, in the sense that there are no a priori ties in the form of parameter space restrictions or links between prior information on them.

Then there are three cases: data are missing completely at random if $\text{pr}(m \mid y;\, \gamma) = \text{pr}(m;\, \gamma)$, so the pattern of missingness is unaffected by the observations, seen or unseen; data are missing at random if $\text{pr}(m \mid y;\, \gamma) = \text{pr}(m \mid y_{\text{obs}};\, \gamma)$, so the pattern of missingness is affected only by the observations actually seen; and nonignorable nonresponse, in which the missingness mechanism depends on not only the observed but also the missing data. The original formulation was in terms of data observed at random, meaning that $\text{pr}(m \mid y;\, \gamma) = \text{pr}(m \mid y_{\text{mis}};\, \gamma)$; the conjunction of this and missing at random entails missing completely at random. If data are missing at random or completely at random, the integral plays no role in the likelihood, but in the third case the missingness mechanism is inextricably tied into it. Bayesian inferences, or frequentist inferences that use only the observed likelihood, will be unaffected by the mechanism for data missing at random, but it will affect modes of inference requiring sample-space expectations, such as confidence intervals based on Fisher information. The strong conditions above can be weakened in some cases (Goffinet, 1987).

These ideas have been widely applied, for example in sampling and in longitudinal data analysis (Sugden & Smith, 1984; Molenberghs et al., 1997). A related development is the study of data coarsening, which encompasses missingness and other forms of incompleteness (Heitjan, 1994). Tests have been developed to distinguish whether data are missing at random or missing completely at random (Chen & Little, 1999), but little can generally be done when there is nonignorable nonresponse and sensitivity analysis seems the best way forward (Molenberghs et al., 1997).

A related development has been the use of imputation, various forms of which are widely used to replace missing data in sample surveys and medical statistics (Vach & Schumacher, 1993). The idea is to replace missing data with values generated randomly from a suitable distribution, so that standard full-data methods can be applied, with uncertainty augmented appropriately by repeating the simulation procedure. The coverage accuracies of confidence intervals and efficiencies of variance estimators from different imputation procedures are compared by Wang & Robins (1998) and Robins & Wang (2000), who describe estimators that improve considerably over those previously in use

Another important development was the realisation that many awkward problems simplify when the observed data are regarded as merely part of an ideal full dataset for which analysis would be easy, and the identification of the EM algorithm in a Royal Statistical Society discussion paper by Dempster et al. (1977). Some subsequent work concerns the many possible applications of the algorithm, to areas as diverse as contingency table analysis (Laird, 1978), data analysis using the multivariate normal model (Rubin & Szatrowski, 1982; Titterington & Jiang, 1983; Didelez & Pigeot, 1998), the analysis of data with mixed discrete and continuous responses (Little & Schluchter, 1985), measurement error in regression (Schafer, 1987) and electron microscope autoradiography (Aykroyd & Anderson, 1994).

It was realised early on that, though rather stable the EM algorithm can be slow, and much effort has been devoted to accelerating its linear rate of convergence. Its steps involve an expectation or E-step in which the conditional expectation of the loglikelihood given the observed data is computed, followed by a maximisation or M-step in which a derived loglikelihood is maximised. Among the approaches used to speed up the EM algorithm have been replacement of the potentially complicated M-step with a number of simpler maximisations (Meng & Rubin, 1993; Liu & Rubin, 1994; Kowalski et al., 1997) and expansion of the parameter space (Liu et al. 1998).

11·5. *Spatial statistics*

Spatial statistics developed rapidly in the 1970s after the relation between Gibbs distributions and Markov random fields as generalisations of Markov chains was clarified by the Hammersley–Clifford theorem. Brook (1964) is a precursor to this, giving consistency conditions that must be obeyed by Markov random fields. Earlier pioneering work on spatial autoregressions had been done by Whittle (1954), who proposed models and methods of estimation for lattices of data, analogous to the autoregressive process widely used in time series, and applied them to two datasets, one showing a clear nugget effect. Besag (1972) took up these ideas, studying the autocorrelation properties of such processes and considering some of Whittle's examples in detail, while Besag & Moran (1975) considered estimation and testing in lattice models in which the responses are Gaussian. An important technique for estimation in such models was based on maximising the pseudo-likelihood, the product of conditional densities for each observation on the lattice given its neighbours, whose efficiency was found to be generally acceptable by Besag (1977) in cases where the spatial correlation is not too strong. Künsch (1987) generalised such models to cases where only the increments and not the observations themselves are stationary, thereby overcoming one of their chief drawbacks; he described an extended version of the maximum likelihood estimator proposed by Whittle (1954). Such models are further investigated by Besag & Kooperberg (1995), who use ideas from graphical Gaussian modelling to circumvent some of the associated difficulties.

An apparently quite different approach is that of Green (1985), who fits least squares smoothing models by generalised least squares. There is a close connection to restricted maximum likelihood estimation; see also Williams (1986). Relationships among different estimators that allow for spatial effects in field experiments are discussed by Draper & Faraggi (1985) and Zimmerman & Harville (1989). Further approaches to spatial smoothing and noise removal use the excellent local adaptivity properties of wavelet expansions (Donoho & Johnstone, 1994; Abramovich & Silverman, 1998).

A closely related topic is image analysis. Markov random fields provide pixel-level models for images, while object recognition demands higher-level models. These have been studied, for example, by Rue & Hurn (1999), whose idea is to use reversible jump Markov chain Monte Carlo with suitable priors for the numbers and shapes of objects in the image; an unknown object type is used to accelerate convergence of the algorithm.

Models for spatial point processes are considered by Strauss (1975), who suggests a Markovian model for clustering and inhibition. Kelly & Ripley (1976) point out difficulties with the clustering model, and establish the spatial Markov structure in the inhibitory case, while Wolpert & Ickstadt (1998) discuss a wide class of Bayesian hierarchical models for doubly stochastic Poisson processes whose underlying intensity is a gamma random field; fitting is by Markov chain Monte Carlo simulation. Lund & Rudemo (2000) give a likelihood method for estimation of an underlying scene when a point process is observed with noise.

Mardia & Marshall (1984) establish consistency of maximum likelihood estimation for spatial Gaussian processes, though Warnes & Ripley (1987) point out that maximum likelihood can give nonsensical estimates of covariance functions in this context.

12. The future

What would Karl Pearson make of a current issue of *Biometrika*? Statistical theory and methods have developed so much and in so many unexpected ways over the past century that detailed prediction would be foolhardy. One broad trend has been the mathematisation of the subject, which has greatly clarified key notions. It has also enabled ready transfer of ideas from fields such as probability, stochastic processes, algorithmics, optimisation and so forth, despite occasional complaints that journals have become unreadable by non-specialists, comments that were already being made in the early years of the twentieth century!

Developments in statistics have in their turn stimulated research in other fields, an example being how modern simulation has renewed interest in Markov chain theory among probabilists. Such interaction seems likely to be increased by the ever more complex stochastic models needed in applications.

Perhaps the dominant trend is the effect of the astonishing advances in computation without which much of modern statistics would not have developed. A consequence of this is the increasingly detailed modelling of phenomena in ways unthinkable only 15 years ago, based on data whose form and quantity would then have seemed a dream, or perhaps a nightmare! One result is increasing diversity, as researchers become more immersed in particular areas of application. This brings with it the potential for further fragmentation of the discipline of statistics, so a continuing and increasingly important role for journals such as *Biometrika* is to be a medium of transfer for new theory and methods among sub-fields.

A related development is the rise of the Internet, with its vast possibilities for acquiring data and rapidly making them available for analysis. Although it is important to underscore that not all data are equally reliable, the simple availability of so much information will certainly have a profound impact, and interaction with computer science and algorithmics seems bound to increase as we look for ways to cope with it.

The century ends as it began, with a wave of enthusiasm for the biological sciences. Historically these have been much more dependent on statistical thinking than the physical sciences, so the opportunities for statistics seem bright, if we seize them.

Acknowledgement

I am grateful to Sir David Cox and the editor for comments on a draft and to numerous other persons for suggesting articles for inclusion. I thank particularly Marc-Olivier Boldi, Alessandra Brazzale, Gorana Capkun, Valérie Chavez-Demoulin and Armin Roehrl. The work was supported by the Swiss National Science Foundation.

Biometrika references

Abramovich, F. & Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115–29.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547–54.

Aitchison, J. & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413–20.

Akaike, H. (1979). A Bayesian extension of the minimum aic procedure of autoregressive model fitting. *Biometrika* **66**, 237–42.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Anscombe, F. J. (1957). Dependence of the fiducial argument on the sampling rule. *Biometrika* **44**, 464–9.

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68**, 357–63.

Asmussen, S. & Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70**, 567–78.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika* **67**, 413–8.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 13–20.

Atkinson, A. C. (1986). Masking unmasked. *Biometrika* **73**, 533–41.

Aykroyd, R. G. & Anderson, C. W. (1994). Use of the EM algorithm for maximum likelihood estimation in electron microscope autoradiography. *Biometrika* **81**, 41–52.

Azzalini, A., Bowman, A. W. & Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

Barnard, G. A. (1947a). Significance tests for 2 × 2 tables. *Biometrika* **34**, 123–38.

Barnard, G. A. (1947b). 2 × 2 tables. A note on E. S. Pearson's paper. *Biometrika* **34**, 168–9.

Barndorff-Nielsen, O. (1973). On M-ancillarity. *Biometrika* **60**, 447–55.

Barndorff-Nielsen, O. (1976). Nonformation. *Biometrika* **63**, 567–72.

Barndorff-Nielsen, O. (1980). Conditionality resolutions. *Biometrika* **67**, 293–310.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–65.

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–22.

Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557–63.

Barndorff-Nielsen, O. E. (1995). Stable and invariant adjusted profile likelihood and directed likelihood for curved exponential models. *Biometrika* **82**, 489–99.

Barndorff-Nielsen, O. E. & Chamberlin, S. R. (1994). Stable and invariant adjusted directed likelihoods. *Biometrika* **81**, 485–99.

Barndorff-Nielsen, O. E. & Hall, P. (1988). On the level-error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika* **75**, 374–8.

Barndorff-Nielsen, O. E. & McCullagh, P. (1993). A note on the relation between modified profile likelihood and the Cox–Reid adjusted profile likelihood. *Biometrika* **80**, 321–8.

Barnett, V. D. (1966). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika* **53**, 151–65.

Bartholomew, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika* **46**, 36–48.

Bartholomew, D. J. (1965). A comparison of some Bayesian and frequentist inferences. *Biometrika* **52**, 19–35.

Bartlett, M. S. (1941). The statistical significance of canonical correlations. *Biometrika* **32**, 29–37.

Bartlett, M. S. (1953a). Approximate confidence intervals. *Biometrika* **40**, 12–9.

Bartlett, M. S. (1953b). Approximate confidence intervals: II. More than one parameter. *Biometrika* **40**, 306–17.

Bartlett, M. S. (1955). Approximate confidence intervals. III. A bias correction. *Biometrika* **42**, 201–4.

Basawa, I. V. (1981a). Efficient conditional tests for mixture experiments with applications to the birth and branching processes. *Biometrika* **68**, 153–64.

Basawa, I. V. (1981b). Efficiency of conditional maximum likelihood estimators and confidence limits for mixtures of exponential families. *Biometrika* **68**, 515–23.

Basawa, I. V. & Scott, D. J. (1976). Efficient tests for branching processes. *Biometrika* **63**, 531–6.

Becker, N. (1974). On parametric estimation for mortal branching processes. *Biometrika* **61**, 393–9.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64**, 616–8.

Besag, J. & Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76**, 633–42.

Besag, J. & Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika* **78**, 301–4.

Besag, J. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 733–46.

Besag, J. E. (1972). On the correlation structure of some two-dimensional stationary processes. *Biometrika* **59**, 43–8.

Besag, J. E. & Moran, P. A. P. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* **62**, 555–62.

Bhansali, R. J. & Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's fpe criterion. *Biometrika* **64**, 547–52.

Boente, G. & Fraiman, R. (1988). On the asymptotic behaviour of general maximum likelihood estimates for the nonregular case under nonstandard conditions. *Biometrika* **75**, 45–56.

Booth, J. G. & Butler, R. W. (1990). Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* **77**, 787–96.

Booth, J. G. & Butler, R. W. (1999). An importance sampling algorithm for exact conditional tests in log-linear models. *Biometrika* **86**, 321–32.

Booth, J. G., Hobert, J. P. & Ohman, P. A. (1999). On the probable error of the ratio of two gamma means. *Biometrika* **86**, 439–52.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika* **36**, 317–46.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40**, 318–35.

Box, G. E. P. & Tiao, G. C. (1962). A further look at robustness via Bayes' theorem. *Biometrika* **49**, 419–32.

Box, G. E. P. & Tiao, G. C. (1964). A Bayesian approach to the importance of assumptions applied in the comparison of variances. *Biometrika* **51**, 153–67.

Box, G. E. P. & Tiao, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika* **55**, 119–29.

Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* **51**, 481–3.

Brooks, S. P. & Roberts, G. O. (1999). On quantile estimation and Markov chain Monte Carlo convergence. *Biometrika* **86**, 710–7.

Buonaccorsi, J. P. (1996). A modified estimating equation approach to correcting for measurement error in regression. *Biometrika* **83**, 433–40.

Burman, P. & Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika* **82**, 877–86.

Butler, R. W. (1989). Approximate predictive pivots and densities. *Biometrika* **76**, 489–501.

Butler, R. W., Huzurbazar, S. & Booth, J. G. (1992a). Saddlepoint approximations for the generalized variance and Wilks' statistic. *Biometrika* **79**, 157–69.

Butler, R. W., Huzurbazar, S. & Booth, J. G. (1992b). Saddlepoint approximations for the Bartlett–Nanda–Pillai trace statistic in multivariate analysis. *Biometrika* **79**, 705–15.

Carlin, B. P. & Gelfand, A. E. (1993). Parametric likelihood inference for record breaking problems. *Biometrika* **80**, 507–15.

Casella, G. & Robert, C. P. (1996). Rao–Blackwellisation of sampling schemes. *Biometrika* **83**, 81–94.

Chaloner, K. & Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651–9.

Chant, D. (1974). On asymptotic tests of complete hypotheses in nonstandard conditions. *Biometrika* **61**, 291–8.

Cheah, P. K., Fraser, D. A. S. & Reid, N. (1994). Multiparameter testing in exponential models: Third order approximations from likelihood. *Biometrika* **81**, 271–8.

Chen, H. Y. & Little, R. J. A. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika* **86**, 1–13.

Chen, M.-H. & Shao, Q.-M. (1998). Monte Carlo methods for Bayesian analysis of constrained parameter problems. *Biometrika* **85**, 73–87.

Chesher, A. & Smith, R. J. (1995). Bartlett corrections to likelihood ratio tests. *Biometrika* **82**, 433–6.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–61.

Coles, S. G. (1989). On goodness-of-fit tests for the two-parameter Weibull distribution derived from the stabilized probability plot. *Biometrika* **76**, 593–8.

Cook, R. D. & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1–10.

Cook, R. D. & Weisberg, S. (1994). Transforming a response variable for linearity. *Biometrika* **81**, 731–7.

Copas, J. B. (1972). Empirical Bayes methods and the repeated use of a standard. *Biometrika* **59**, 349–60.

Copas, J. B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika* **62**, 701–4.

Cordeiro, G. M. (1987). On the corrections to the likelihood ratio statistic. *Biometrika* **74**, 265–74.

Cordeiro, G. M., Botter, D. A. & de Paula Ferrari, S. L. (1994). Nonnull asymptotic distributions of three classic criteria in generalised linear models. *Biometrika* **81**, 709–20.

Cordeiro, G. M. & de Paula Ferrari, S. L. (1991). A modified score test statistic having chi-squared distribution to order n^{-1}. *Biometrika* **78**, 573–82.

Cordeiro, G. M. & Paula, G. A. (1989). Improved likelihood ratio statistics for exponential family nonlinear models. *Biometrika* **76**, 93–100.

Cox, D. R. (1948). A note on the asymptotic distribution of range. *Biometrika* **35**, 310–5.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.

Cox, D. R. (1980). Local ancillarity. *Biometrika* **67**, 279–86.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika* **70**, 269–74.

Cox, D. R. (1984). Effective degrees of freedom and the likelihood ratio test. *Biometrika* **71**, 487–93.

Cox, D. R. (1993). Unbiased estimating equation derived from statistics that are functions of a parameter. *Biometrika* **80**, 905–9.

Cox, D. R. & Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* **79**, 408–11.

Cox, D. R. & Wermuth, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika* **77**, 747–61.

Cox, D. R. & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79**, 441–61.

Cribari-Neto, F. & Ferrari, S. L. P. (1995). Second order asymptotics for score tests in generalised linear models. *Biometrika* **82**, 426–32.

Crisp, A. & Burridge, J. (1994). A note on nonregular likelihood functions in heteroscedastic regression models. *Biometrika* **81**, 585–7.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika* **74**, 591–7.

Crowder, M. (1990). On some nonregular tests for a modified Weibull model. *Biometrika* **77**, 499–506.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82**, 407–10.

Cruddas, A. M., Reid, N. & Cox, D. R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika* **76**, 231–7.

Daniels, H. E. (1956). The asymptotic distribution of serial correlation coefficients. *Biometrika* **43**, 169–85.

Daniels, H. E. (1980). Exact saddlepoint approximations. *Biometrika* **67**, 59–63.

Daniels, H. E. (1983). Saddlepoint approximations for estimating equations. *Biometrika* **70**, 89–96.

Daniels, H. E. & Young, G. A. (1991). Saddlepoint approximation for the studentized mean, with an application to the bootstrap. *Biometrika* **78**, 169–79.

Datta, G. S. (1996). On priors providing frequentist validity of Bayesian inference for multiple parametric functions. *Biometrika* **83**, 287–98.

Datta, G. S. & Ghosh, J. K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika* **82**, 37–45.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–54.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–32.

Davison, A. C. & Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75**, 417–31.

Dawid, A. P. (1977). Invariant distributions and analysis of variance models. *Biometrika* **64**, 291–8.

Dawid, A. P. & Mortera, J. (1998). Forensic identification with imperfect evidence. *Biometrika* **85**, 835–49.

Dellaportas, P. & Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–33.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85**, 363–77.

Devlin, S. J., Gnanadesikan, R. & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531–46.

DiCiccio, T. J. (1984). On parameter transformations and interval estimation. *Biometrika* **71**, 477–85.

DiCiccio, T. J. & Field, C. A. (1991). An accurate method for approximate conditional and Bayesian inference about linear regression models from censored data. *Biometrika* **78**, 903–10.

DiCiccio, T. J., Field, C. A. & Fraser, D. A. S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77**, 77–95.

DiCiccio, T. J. & Martin, M. A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference. *Biometrika* **78**, 891–902.

DiCiccio, T. J., Martin, M. A. & Young, G. A. (1992). Fast and accurate approximate double bootstrap confidence intervals. *Biometrika* **79**, 285–95.

DiCiccio, T. J., Martin, M. A. & Young, G. A. (1993). Analytical approximations to conditional distribution functions. *Biometrika* **80**, 781–90.

DiCiccio, T. J. & Stern, S. E. (1993). On Bartlett adjustments for approximate Bayesian inference. *Biometrika* **80**, 731–40.

Didelez, V. & Pigeot, I. (1998). Maximum likelihood estimation in graphical models with missing values. *Biometrika* **85**, 960–6.

Dietrich, C. R. (1991). Modality of the restricted likelihood for spatial Gaussian random fields. *Biometrika* **78**, 833–9.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.

Draper, N. R. & Faraggi, D. (1985). Role of the Papadakis estimator in one- and two-dimensional field trials. *Biometrika* **72**, 223–6.

Durbin, J. (1961). Some methods of constructing exact tests. *Biometrika* **48**, 41–55.

Durbin, J. (1975). Kolmogorov–Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika* **62**, 5–22.

Durbin, J. (1980a). Approximations for densities of sufficient estimators. *Biometrika* **67**, 311–33.

Durbin, J. (1980b). The approximate distribution of partial serial correlation coefficients calculated from residuals from regression on Fourier series. *Biometrika* **67**, 335–49.

Edwards, D. & Havranek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72**, 339–51.

Edwards, D. & Kreiner, S. (1983). The analysis of contingency tables by graphical models. *Biometrika* **70**, 553–65.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26.

Efron, B. & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457–81.

Efron, B. & Morris, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59**, 335–47.

Efron, B. & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63**, 435–48.

Fearn, T. (1975). A Bayesian approach to growth curves. *Biometrika* **62**, 89–100.

Feigin, P. D. & Reiser, B. (1979). On asymptotic ancillarity and inference for Yule and regular nonergodic processes. *Biometrika* **66**, 279–84.

Ferreira, P. E. (1982). Estimating equations in the presence of prior knowledge. *Biometrika* **69**, 667–9.

Feuerverger, A. (1989). On the empirical saddlepoint approximation. *Biometrika* **76**, 457–64.

Field, C. A. & Hampel, F. R. (1982). Small-sample asymptotic distributions of M estimators of location. *Biometrika* **69**, 29–46.

Firth, D. (1987). On the efficiency of quasi-likelihood estimation. *Biometrika* **74**, 233–45.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Firth, D., Glosup, J. & Hinkley, D. V. (1991). Model checking with nonparametric curves. *Biometrika* **78**, 245–52.

Firth, D. & Harris, I. R. (1991). Quasi-likelihood for multiplicative random effects. *Biometrika* **78**, 545–55.

Fisher, N. I. & Switzer, P. (1985). Chi-plots for assessing dependence. *Biometrika* **72**, 253–65.

Fraser, D. A. S. (1961). The fiducial method and invariance. *Biometrika* **48**, 261–80.

Fraser, D. A. S. (1964). Fiducial inference for location and scale parameters. *Biometrika* **51**, 17–24.

Fraser, D. A. S. (1965). Fiducial consistency and group structure. *Biometrika* **52**, 55–65.

Fraser, D. A. S. (1966). Structural probability and a generalization. *Biometrika* **53**, 1–9.

Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 65–76.

Fraser, D. A. S. & Reid, N. (1989). Adjustments to profile likelihood. *Biometrika* **76**, 477–88.

Fraser, D. A. S., Reid, N. & Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–64.

Frydenberg, M. & Jensen, J. L. (1989). Is the 'improved likelihood ratio statistic' really improved in the discrete case? *Biometrika* **76**, 655–61.

Frydenberg, M. & Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* **76**, 539–55.

Fujikoshi, Y. & Satoh, K. (1997). Modified aic and C_p in multivariate linear regression. *Biometrika* **84**, 707–16.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58**, 453–67.

Gail, M. H., Lubin, J. H. & Rubinstein, L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* **68**, 703–7.

Gelfand, A. E., Sahu, S. K. & Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika* **82**, 479–88.

George, E. I. & Robert, C. P. (1992). Capture–recapture estimation via Gibbs sampling. *Biometrika* **79**, 677–83.

Gilmour, A. R., Anderson, R. D. & Rae, A. L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**, 593–9.

Glad, I. K. & Sebastiani, G. (1995). A Bayesian approach to synthetic magnetic resonance imaging. *Biometrika* **82**, 237–50.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277–84.

Godambe, V. P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter. *Biometrika* **67**, 155–62.

Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika* **72**, 419–28.

Godambe, V. P. (1991). Orthogonality of estimating functions and nuisance parameters. *Biometrika* **78**, 143–51.

Godambe, V. P. & Thompson, M. E. (1984). Robust estimation through estimating equations. *Biometrika* **71**, 115–25.

Goffinet, B. (1987). Alternative conditions for ignoring the process that causes missing data. *Biometrika* **74**, 437–9.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43–56.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika* **74**, 430–1.

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation. *Biometrika* **76**, 622–3.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika* **78**, 45–51.

Goldstein, M. (1976). Bayesian analysis of regression problems. *Biometrika* **63**, 51–8.

Goldstein, M. (1980). The linear Bayes regression estimator under weak prior assumptions. *Biometrika* **67**, 621–8.

Goldstein, M. & Wooff, D. A. (1998). Adjusting exchangeable beliefs. *Biometrika* **85**, 39–54.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika* **60**, 179–92.

Goutis, C. & Robert, C. P. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika* **85**, 29–37.

Gower, J. C. (1990). Three-dimensional biplots. *Biometrika* **77**, 773–85.

Gower, J. C. (1992). Generalized biplots. *Biometrika* **79**, 475–93.

Gower, J. C. & Harding, S. A. (1988). Nonlinear biplots. *Biometrika* **75**, 445–55.

Green, P. J. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika* **72**, 527–37.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

Gûnel, E. & Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61**, 545–57.

Haldane, J. B. S. & Maynard Smith, S. (1956). The sampling distribution of a maximum-likelihood estimate. *Biometrika* **43**, 96–103.

Hall, P. (2001). *Biometrika* Centenary: Nonparametrics. *Biometrika* **88**, 143–65.

Hanfelt, J. J. & Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika* **82**, 461–77.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika* **76**, 675–84.

Harris, P. (1986). A note on Bartlett adjustments to likelihood ratio tests. *Biometrika* **73**, 735–7.

Hartley, H. O. & Rao, J. N. K. (1967). Maximum-likelihood estimates for the mixed analysis of variance model. *Biometrika* **54**, 93–108.

Harvill, J. L. & Newton, H. J. (1995). Saddlepoint approximations for the difference of order statistics. *Biometrika* **82**, 226–31.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–5.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81**, 701–8.

Heyde, C. C. & Morton, R. (1993). On constrained quasi-likelihood estimation. *Biometrika* **80**, 755–61.

Hill, J. R. (1990). A general framework for model-based statistics. *Biometrika* **77**, 115–26.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika* **56**, 495–504.

Hinkley, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67**, 287–92.

Hosking, J. R. M. (1984). Testing whether the shape parameter is zero in the generalized extreme-value distribution. *Biometrika* **71**, 367–74.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–77.

Hsu, P. L. (1940). On generalized analysis of variance (I). *Biometrika* **31**, 221–37.

Hurvich, C. M., Shumway, R. & Tsai, C.-L. (1990). Improved estimators for Kullback–Leibler information for autoregressive model selection in small samples. *Biometrika* **77**, 709–19.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Hurvich, C. M. & Tsai, C.-L. (1991). Bias of the corrected aic criterion for underfitted regression and time series models. *Biometrika* **78**, 499–509.

Hurvich, C. M. & Tsai, C.-L. (1995). Relative rates of convergence for efficient model selection criteria in linear regression. *Biometrika* **82**, 418–25.

Hurvich, C. M. & Tsai, C.-L. (1998). A crossvalidatory aic for hard wavelet thresholding in spatially adaptive function estimation. *Biometrika* **85**, 701–10.

Jensen, J. L. (1986). Similar tests and the standardized log likelihood ratio statistic. *Biometrika* **73**, 567–72.

Jensen, J. L. (1992). The modified signed likelihood statistics and saddlepoint approximations. *Biometrika* **79**, 693–703.

Jensen, S. T., Johansen, S. & Lauritzen, S. L. (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika* **78**, 867–77.

Jing, B.-Y., Feuerverger, A. & Robinson, J. (1994). On the bootstrap saddlepoint approximations. *Biometrika* **81**, 211–5.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika* **57**, 239–51.

Jørgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70**, 19–28.

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika* **62**, 251–68.

Kalbfleisch, J. D. & Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**, 267–78.

Kass, R. E. (1990). Data-translated likelihood and Jeffreys's rules. *Biometrika* **77**, 107–14.

Kass, R. E., Tierney, L. & Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76**, 663–74.

Kelly, F. P. & Ripley, B. D. (1976). A note on Strauss's model for clustering. *Biometrika* **63**, 357–60.

Kendall, M. G. (1949). On the reconciliation of theories of probability. *Biometrika* **36**, 101–16.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69**, 19–27.

Kent, J. T. (1986). The underlying structure of nonnested hypothesis tests. *Biometrika* **73**, 333–43.

Kohn, R. (1977). Note concerning the Akaike and Hannan estimation procedures for an autoregressive-moving average process. *Biometrika* **64**, 622–5.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83**, 299–313.

Kowalski, J., Tu, X. M., Day, R. S. & Mendoza-Blanco, J. R. (1997). On the rate of convergence of the ECME algorithm for multiple regression models with t-distributed errors. *Biometrika* **84**, 269–81.

Krzanowski, W. J. (1979). Some linear transformations for mixtures of binary and continuous variables, with particular reference to linear discriminant analysis. *Biometrika* **66**, 33–40.

Künsch, H. R. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika* **74**, 517–24.

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–35.

Laird, N. M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**, 581–90.

Larimore, W. E. (1983). Predictive inference, sufficiency, entropy and an asymptotic likelihood principle. *Biometrika* **70**, 175–81.

Lawley, D. N. (1938). A generalization of Fisher's z test. *Biometrika* **30**, 180–7.

Lawley, D. N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika* **43**, 295–303.

Levenbach, H. (1972). Estimation of autoregressive parameters from a marginal likelihood function. *Biometrika* **59**, 61–71.

Li, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika* **80**, 741–53.

Liang, K.-Y. (1983). On information and ancillarity in the presence of a nuisance parameter. *Biometrika* **70**, 607–12.

Liang, K.-Y. (1984). The asymptotic efficiency of conditional likelihood methods. *Biometrika* **71**, 305–13.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika* **64**, 207–14.

Lindsay, B. (1982). Conditional score functions: Some optimality results. *Biometrika* **69**, 503–12.

Lipsitz, S. R., Laird, N. M. & Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**, 153–60.

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295–304.

Little, R. J. A. & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497–512.

Liu, C. & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633–48.

Liu, C., Rubin, D. B. & Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**, 755–70.

Liu, J. S., Wong, W. H. & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 817–27.

Lubin, J. H. (1981). An empirical evaluation of the use of conditional and unconditional likelihoods for case-control data. *Biometrika* **68**, 567–71.

Lund, J. & Rudemo, M. (2000). Models for point processes observed with noise. *Biometrika* **87**, 235–49.

Madigan, D. & Mosurski, K. (1990). An extension of the results of Asmussen and Edwards on collapsibility in contingency tables. *Biometrika* **77**, 315–9. Correction (1999) **86**, 973.

Malec, D. & Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika* **79**, 593–601.

Mallick, B. K. & Walker, S. G. (1997). Combining information from several experiments with nonparametric priors. *Biometrika* **84**, 697–706.

Mardia, K. V. & Kent, J. T. (1991). Rao score tests for goodness of fit and independence. *Biometrika* **78**, 355–63.

Mardia, K. V. & Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–46.

Maritz, J. S. (1966). Smooth empirical Bayes estimation for one-parameter discrete distributions. *Biometrika* **53**, 417–29.

Maritz, J. S. (1967). Smooth empirical Bayes estimation for continuous distributions. *Biometrika* **54**, 435–50.

Maritz, J. S. (1968). On the smooth empirical Bayes approach to testing of hypotheses and the compound decision problem. *Biometrika* **55**, 83–100.

Matthews, J. N. S. (1989). Estimating dispersion parameters in the analysis of data from crossover trials. *Biometrika* **76**, 239–44.

McAleer, M. (1983). Exact tests of a model against nonnested alternatives. *Biometrika* **70**, 285–8.

McCullagh, P. (1984a). Local sufficiency. *Biometrika* **71**, 233–44.

McCullagh, P. (1984b). Tensor notation and cumulants of polynomials. *Biometrika* **71**, 461–76.

McCullagh, P. (1992). Conditional inference and Cauchy models. *Biometrika* **79**, 247–59.

McCulloch, R. E. & Rossi, P. E. (1992). Bayes factors for nonlinear hypotheses and likelihood distributions. *Biometrika* **79**, 663–76.

McKean, J. W. & Hettmansperger, T. P. (1978). A robust analysis of the general linear model based on one step R-estimates. *Biometrika* **65**, 571–79.

McLeish, D. L. & Small, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika* **79**, 93–102.

Meng, X.-L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267–78.

Michael, J. R. (1983). The stabilized probability plot. *Biometrika* **70**, 11–7.

Molenberghs, G., Kenward, M. G. & Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* **84**, 33–44.

Monti, A. C. & Ronchetti, E. (1993). On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate M-estimators. *Biometrika* **80**, 329–38.

Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika* **73**, 583–8.

Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika* **79**, 747–54.

Morton, R. (1981). Efficiency of estimating equations and the use of pivots. *Biometrika* **68**, 227–33.

Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation. *Biometrika* **74**, 247–57.

Morton, R. (1991). Analysis of extra-multinomial data derived from extra-Poisson variables conditional on their total. *Biometrika* **78**, 1–6.

Mukerjee, R. & Chandra, T. K. (1991). Bartlett-type adjustments for the conditional likelihood ratio statistic of Cox and Reid. *Biometrika* **78**, 365–72.

Mukerjee, R. & Dey, D. K. (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: Higher order asymptotics. *Biometrika* **80**, 499–505.

Müller, P. & Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84**, 523–37.

Murray, G. D. (1977). A note on the estimation of probability density functions. *Biometrika* **64**, 150–2.

Nelder, J. A. & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221–32.

Neyman, J. (1941). Fiducial argument and the theory of fiducial confidence intervals. *Biometrika* **32**, 128–50.

Neyman, J. & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika* **20A**, 175–240.

Neyman, J. & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika* **20A**, 263–94.

Neyman, J. & Pearson, E. S. (1931). Further notes on the χ^2 distribution. *Biometrika* **22**, 298–305.

Ng, V. M. (1980). On the estimation of parametric density functions. *Biometrika* **67**, 505–6.

Paik, M. C. (1996). Quasi-likelihood regression models with missing covariates. *Biometrika* **83**, 825–34.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–54.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13–27.

Pearl, J. (1995). Causal diagrams for empirical research (with Discussion). *Biometrika* **82**, 669–88.

Pearson, E. S. (1928). The distribution of frequency constants in small samples from symmetrical populations (Assisted by N. K. Adyanthaya). *Biometrika* **20A**, 356–60.

Pearson, E. S. (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations (Assisted by N. K. Adyanthaya and others). *Biometrika* **21**, 259–86.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika* **23**, 114–33.

Pearson, E. S. (1947). The choice of statistical test illustrated on the interpretation of data classed in a 2×2 table. *Biometrika* **34**, 139–67.

Pearson, E. S. & Wilks, S. S. (1933). Methods of statistical analysis appropriate for k samples of two variables. *Biometrika* **25**, 353–78.

Peers, H. W. (1971). Likelihood ratio and associated test criteria. *Biometrika* **58**, 577–87.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607–12.

Pettitt, A. N. (1983). Approximate methods using ranks for regression with censored data. *Biometrika* **70**, 121–32.

Phillips, P. C. B. (1978). Edgeworth and saddlepoint approximations in the first-order noncircular autoregression. *Biometrika* **65**, 91–8.

Pillai, K. C. S. (1956). On the distribution of the largest or the smallest root of a matrix in multivariate analysis. *Biometrika* **43**, 122–7.

Pitman, E. J. G. (1938). The estimation of location and scale parameters of a continuous population of any given form. *Biometrika* **30**, 391–421.

Pitman, E. J. G. (1939). Tests of hypotheses concerning location and scale parameters. *Biometrika* **31**, 200–15.

Plackett, R. L. (1947). An exact test for the equality of variances. *Biometrika* **34**, 311–9.

Plackett, R. L. (1977). The marginal totals of a 2×2 table. *Biometrika* **64**, 37–42.

Porteous, B. T. (1985). Improved likelihood ratio statistics for covariance selection models. *Biometrika* **72**, 97–101.

Preisser, J. S. & Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika* **83**, 551–62.

Raftery, A. E. (1988). Inference for the binomial N parameter: A hierarchical Bayes approach. *Biometrika* **75**, 223–8.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 251–66.

Raftery, A. E. & Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73**, 85–9.

Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika* **35**, 58–79.

Rider, P. R. (1929). On the distribution of the ratio of mean to standard deviation in small samples from non-normal universes. *Biometrika* **21**, 124–43.

Roberts, G. O. & Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83**, 95–110.

Robins, J. M. & Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–24.

Robinson, G. K. (1975). Some counterexamples to the theory of confidence intervals. *Biometrika* **62**, 155–62.

Roeder, K., Escobar, M., Kadane, J. B. & Balazas, I. (1998). Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika* **85**, 269–87.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rotnitzky, A. & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485–97.

Roverato, A. & Whittaker, J. (1998). The Isserlis matrix and its application to non-decomposable graphical Gaussian models. *Biometrika* **85**, 711–25.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–90.

Rubin, D. B. & Szatrowski, T. H. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika* **69**, 657–60.

Rue, H. & Hurn, M. A. (1999). Bayesian object identification. *Biometrika* **86**, 649–60.

Ryall, T. A. (1981). Extensions of the concept of local ancillarity. *Biometrika* **68**, 677–83.

Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika* **74**, 385–91.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–27.

Severini, T. A. (1993). Local ancillarity in the presence of a nuisance parameter. *Biometrika* **80**, 305–20.

Severini, T. A. (1998a). An approximation to the modified profile likelihood function. *Biometrika* **85**, 403–11.

Severini, T. A. (1998b). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85**, 507–22.

Severini, T. A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika* **86**, 235–47.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45–54.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43–9.

Skovgaard, I. (1986). Successive improvement of the order of ancillarity. *Biometrika* **73**, 516–9.

Smith, A. F. M. (1973). Bayes estimates in one-way and two-way models. *Biometrika* **60**, 319–29.

Smith, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika* **62**, 407–16.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67–90.

Smith, R. L. (1994). Nonregular regression. *Biometrika* **81**, 173–83.

Solomon, P. J. & Taylor, J. M. G. (1999). Orthogonality and transformations in variance components models. *Biometrika* **86**, 289–300.

Spady, R. H. (1991). Saddlepoint approximations for regression models. *Biometrika* **78**, 879–89.

Sprott, D. A. (1975). Marginal and conditional sufficiency. *Biometrika* **62**, 599–606.

Stafford, J. E. & Andrews, D. F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika* **80**, 715–30.

Stefanski, L. A., Carroll, R. J. & Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* **73**, 413–24.

Stephens, M. A. (1977). Goodness of fit for the extreme value distribution. *Biometrika* **64**, 583–8.

Stephens, M. A. (1979). Tests of fit for the logistic distribution based on the empirical distribution function. *Biometrika* **66**, 591–6.

Stigler, S. M. (1970). Estimating the age of a Galton–Watson branching process. *Biometrika* **57**, 505–12.

Stigler, S. M. (1971). The estimation of the probability of extinction and other parameters associated with branching processes. *Biometrika* **58**, 499–508.

Stone, M. & Dawid, A. P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika* **59**, 369–75.

Strauss, D. J. (1975). A model for clustering. *Biometrika* **62**, 467–76.

"Student" (1908). The probable error of a mean. *Biometrika* **6**, 1–25.

Sugden, R. A. & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495–506.

Sun, D. & Berger, J. O. (1998). Reference priors with partial information. *Biometrika* **85**, 55–71.

Sweeting, T. J. (1978). On efficient tests for branching processes. *Biometrika* **65**, 123–8.

Sweeting, T. J. (1992). Parameter-based asymptotics. *Biometrika* **79**, 219–30.

Sweeting, T. J. (1995a). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82**, 1–23.

Sweeting, T. J. (1995b). A Bayesian approach to approximate conditional inference. *Biometrika* **82**, 25–36.

Thisted, R. & Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74**, 445–55.

Thomas, W. & Cook, R. D. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika* **76**, 741–9.

Tiao, G. C. & Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. I: Posterior distribution of variance-components. *Biometrika* **52**, 37–53.

Tiao, G. C. & Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. II: Effect of autocorrelated errors. *Biometrika* **53**, 477–95.

Tiao, G. C. & Zellner, A. (1964). Bayes's theorem and the use of prior knowledge in regression analysis. *Biometrika* **51**, 219–30.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–8.

Tierney, L., Kass, R. E. & Kadane, J. B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika* **76**, 425–33.

Titterington, D. M. (1977). Analysis of incomplete multivariate binary data by the kernel method. *Biometrika* **64**, 455–60.

Titterington, D. M. & Jiang, J.-M. (1983). Recursive estimation procedures for missing data problems. *Biometrika* **70**, 613–24.

Vach, W. & Schumacher, M. (1993). Logistic regression with incompletely observed categorical covariates: A comparison of three approaches. *Biometrika* **80**, 353–62.

Vidoni, P. (1995). A simple predictive density based on the p^*-formula. *Biometrika* **82**, 855–63.

Vlachonikolis, I. G. (1990). Predictive discrimination and classification with mixed binary and continuous variables. *Biometrika* **77**, 657–62.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika* **83**, 447–52.

Wang, N. & Robins, J. M. (1998). Large-sample theory for parametric multiple-imputation procedures. *Biometrika* **85**, 935–48.

Wang, S. (1992). Tail probability approximations in the first-order noncircular autogression. *Biometrika* **79**, 431–4.

Wang, S. (1993). Saddlepoint expansion in finite population problems. *Biometrika* **80**, 583–90.

Warnes, J. J. & Ripley, B. D. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika* **74**, 640–2.

Waterman, R. P. & Lindsay, B. G. (1996). Projected score methods for approximating conditional scores. *Biometrika* **83**, 1–13.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61**, 439–47.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27–32.

Weiss, R. E. & Cook, R. D. (1992). A graphical case statistic for assessing posterior influence. *Biometrika* **79**, 51–5.

Welch, B. L. (1933). Some problems in the analysis of regression among k samples of two variables. *Biometrika* **27**, 145–60.

Welch, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29**, 21–52.

Welch, B. L. (1939). Note on discriminant functions. *Biometrika* **31**, 218–20.

Wermuth, N. & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537–52.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434–49.

Wilk, M. B. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55**, 1–17.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika* **24**, 471–94.

Wilks, S. S. (1941). Review of *Theory of Probability* by Harold Jeffreys. *Biometrika* **32**, 192–4.

Williams, E. R. (1986). A neighbour model for field experiments. *Biometrika* **73**, 279–87.

Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. *Biometrika* **20A**, 32–52.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* **80**, 791–5.

Wolpert, R. L. & Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, 251–67.

Wong, W. H. & Li, B. (1992). Laplace expansion for posterior densities of nonlinear functions of parameters. *Biometrika* **79**, 393–8.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621–9.

Zimmerman, D. L. & Harville, D. A. (1989). On the unbiasedness of the Papadakis estimator and other nonlinear estimators of treatment contrasts in field-plot experiments. *Biometrika* **76**, 253–9.

Other references

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Ed. B. N. Petrov and F. Czáki, pp. 267–81. Budapest: Akademiai Kiadó. Reprinted (1992) in *Breakthroughs in Statistics*, Volume 1: *Foundations and Basic Theory*, Ed. S. Kotz and N. L. Johnson, pp. 610–24. New York: Springer-Verlag.

Barndorff-Nielsen, O. E. & Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli* **2**, 319–40.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond.* A **160**, 268–82.

Birnbaum, A. (1962). On the foundations of statistical inference (with Discussion). *J. Am. Statist. Assoc.* **57**, 269–306.

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 359–72.

Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematics, Probability and Statistics*, Vol. 1, Ed. J. Neyman, pp. 105–23. Berkeley: University of California Press.

Cox, D. R. (1968). Notes on some aspects of regression analysis (with Discussion). *J. R. Statist. Soc.* A **131**, 256–79.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and conditional inference (with Discussion). *J. R. Statist. Soc.* B **49**, 1–39.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25**, 631–50.

Daniels, H. E. (1958). Discussion of 'The regression analysis of binary sequences', by D. R. Cox. *J. R. Statist. Soc.* B **20**, 236–8.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. R. Soc. Lond.* A **144**, 285–307.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–12.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Hinkley, D. V. (1979). Predictive likelihood. *Ann. Statist.* **7**, 718–28.

Jeffreys, H. (1939). *Theory of Probability*, 1st ed. Oxford: Oxford University Press.

Kalbfleisch, J. D. & Sprott, D. A. (1970). Applications of likelihood methods to models involving large numbers of parameters (with Discussion). *J. R. Statist. Soc.* B **32**, 175–208.

Liang, K., Zeger, S. L. & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with Discussion). *J. R. Statist. Soc.* B **54**, 3–40.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. R. Statist. Soc.* B **20**, 102–7.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–91.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A **135**, 370–84.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc.* A **231**, 289–337.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil. Magaz.* **2**, 559–72.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any population. *J. R. Statist. Soc. Suppl.* **4**, 119–30.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematics, Probability and Statistics*, Vol. 1, Ed. J. Neyman, pp. 157–63. Berkeley: University of California Press.

Thompson Jr., W. A. (1962). The problem of negative estimates of variance components. *Ann. Math. Statist.* **33**, 273–89.

Welch, B. L. (1958). On confidence limits and sufficiency with particular reference to parameters of location. *Ann. Math. Statist.* **10**, 58–69.