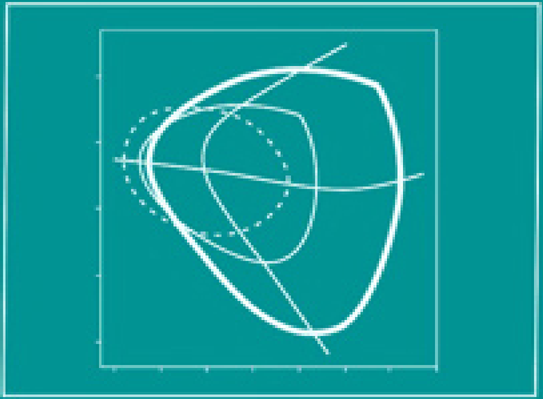
Cambridge Series in Statistical and Probabilistic Mathematics



Applied Asymptotics

Case Studies in Small-Sample Statistics

A. R. Brazzale, A. C. Davison and N. Reid

CAMBRIDGE www.cambridge.org/9780521847032

This page intentionally left blank

Applied Asymptotics: Case Studies in Small-Sample Statistics

In fields such as biology, medical sciences, sociology and economics researchers often face the situation where the number of available observations, or the amount of available information, is sufficiently small that approximations based on the normal distribution may be unreliable. Theoretical work over the last quarter-century has led to new likelihood-based methods that yield very accurate approximations in finite samples, but this work has had limited impact on statistical practice. This book illustrates by means of realistic examples and case studies how to use the new theory, and investigates how and when it makes a difference to the resulting inference. The treatment is oriented towards practice and is accompanied by code in the R language which enables the methods to be applied in a range of situations of interest to practitioners. The analysis includes some comparisons of higher order likelihood inference with bootstrap and Bayesian methods.

ALESSANDRA BRAZZALE is a Professor of Statistics at the Università degli Studi di Modena e Reggio Emilia.

ANTHONY DAVISON is a Professor of Statistics at the Ecole Polytechnique Fédérale de Lausanne.

NANCY REID is a University Professor of Statistics at the University of Toronto.

CAMBRIDGE SERIES IN STATISTICAL AND
PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

Already published

1. *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley
2. *Markov Chains*, by J. Norris
3. *Asymptotic Statistics*, by A. W. van der Vaart
4. *Wavelet Methods for Time Series Analysis*, by Donald B. Percival and Andrew T. Walden
5. *Bayesian Methods*, by Thomas Leonard and John S. J. Hsu
6. *Empirical Processes in M-Estimation*, by Sara van de Geer
7. *Numerical Methods of Statistics*, by John F. Monahan
8. *A User's Guide to Measure Theoretic Probability*, by David Pollard
9. *The Estimation and Tracking of Frequency*, by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics using R*, by John Maindonald and John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by D. Ruppert, M. P. Wand and R. J. Carroll
13. *Exercises in Probability*, by Loic Chaumont and Marc Yor
14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems*, by Anton Bovier
20. *Random Graph Dynamics*, by Rick Durrett
21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler

Applied Asymptotics Case Studies in Small-Sample Statistics

A. R. Brazzale
Università degli Studi di Modena e Reggio Emilia

A. C. Davison
Ecole Polytechnique Fédérale de Lausanne

N. Reid
University of Toronto

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521847032

First published in print format 2007

ISBN-13 978-0-511-28522-6 eBook (Adobe Reader)
ISBN-10 0-511-28282-6 eBook (Adobe Reader)

ISBN-13 978-0-521-84703-2 hardback
ISBN-10 0-521-84703-6 hardback

Cambridge University Press has no responsibility for the persistence or accuracy of urls for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

Preface

The likelihood function plays a central role in both statistical theory and practice. Basic results about likelihood inference, which we call first order asymptotics, were developed in fundamental work by R. A. Fisher during the 1920s, and now form an essential and widely taught part of both elementary and advanced courses in statistics. It is less well known that Fisher later proposed a more refined approach, which has been developed over the past three decades into a theory of higher order asymptotics. While this theory leads to some extremely accurate methods for parametric inference, accounts of the theory can appear forbidding, and the results may be thought to have little importance for statistical practice.

The purpose of this book is dispel this view, showing how higher order asymptotics may be applied in realistic examples with very little more effort than is needed for first order procedures, and to compare the resulting improved inferences with those from other approaches. To do this we have collected a range of examples and case studies, provided details on the implementation of higher order approximations, and compared the resulting inference to that based on other methods; usually first order likelihood theory, but where appropriate also methods based on simulation. Our examples are nearly all derived from regression models for discrete or continuous data, but range quite widely over the types of models and inference problems where likelihood methods are applied.

In order to make higher order methods accessible, we have striven for as simple an exposition as we thought feasible, aiming for heuristic explanation rather than full mathematical rigour. We do not presuppose previous knowledge of higher order asymptotics, key aspects of which are explained early in the book. The reader is assumed to have knowledge of basic statistics including some central classes of models, and some experience of standard likelihood methods in practice. We intend that the book be useful for students of statistics, practising statisticians, and data analysts, as well as researchers interested in a more applied account of the methods than has so far been available. Our effort has been made practicable by software developed by Alessandra Brazzale and Ruggero Bellio over many years, of which the `hoa` package bundle now available in `R` is the culmination. This software is extensively used throughout the book, and the ideas behind the `hoa` packages, described in Chapter 9, formed the basis for our approaches to programming when new software was needed for some of the examples. The hoa package bundle and other materials may be obtained from the book's web page

http://statwww.epfl.ch/AA

This book could not have been written without the help of many colleagues. We thank particularly Sven Bacher, Douglas Bates, Ruggero Bellio, Nathalie Chèvre, David Cox, Don Fraser, Torsten Hothorn, Alessandra Salvan, Ana-Maria Staicu, Charlotte Vandenberghe, and members of the R Development Core Team, for access to data, fruitful collaboration, helpful discussion, valuable comments on the material, and help with computational aspects of our work. David Tranah and Diana Gillooly of Cambridge University Press have been supportive and patient editors. We thank also the following institutions for financial and material support: the EPFL; the University of Padova; the University of Toronto; the Italian National Research Council, and in particular its Institute of Biomedical Engineering, where much of Alessandra Brazzale's work was performed; the Italian Ministry of Education, University and Research; the Canadian Natural Sciences and Engineering Research Council; and the Swiss National Science Foundation. We thank also our friends and families for their enduring encouragement and support.

A. R. Brazzale, A. C. Davison and N. Reid
Reggio Emilia, Lausanne and Toronto

1

Introduction

This book is about the statistical analysis of data, and in particular approximations based on the likelihood function. We emphasize procedures that have been developed using the theory of higher order asymptotic analysis and which provide more precise inferences than are provided by standard theory. Our goal is to illustrate their use in a range of applications that are close to many that arise in practice. We generally restrict attention to parametric models, although extensions of the key ideas to semi-parametric and non-parametric models exist in the literature and are briefly mentioned in contexts where they may be appropriate. Most of our examples consist of a set of independent observations, each of which consists of a univariate response and a number of explanatory variables.

Much application of likelihood inference relies on *first order asymptotics*, by which we mean the application of the central limit theorem to conclude that the statistics of interest are approximately normally distributed, with mean and variance consistently estimable from the data. There has, however, been great progress over the past twenty-five years or so in the theory of likelihood inference, and two main themes have emerged. The first is that very accurate approximations to the distributions of statistics such as the maximum likelihood estimator are relatively easily derived using techniques adapted from the theory of asymptotic expansions. The second is that even in situations where first order asymptotics is to be used, it is often helpful to use procedures suggested by these more accurate approximations, as they provide modifications to naive approaches that result in more precise inferences. We refer throughout to these two developments as *higher order asymptotics*, although strictly speaking we mean 'higher order asymptotic theory applied to likelihood inference' – of course there are many developments of higher order asymptotics in other mathematical contexts.

Asymptotic theory in statistics refers to the limiting distribution of a summary statistic as the amount of information in the data increases without limit. In the simplest situations this means that the sample size increases to infinity, and in more complex models entails a related notion of accumulation of information, often through independent replications of an assumed model. Such asymptotic theory is an essential part of statistical methodology, as exact distributions of quantities of interest are rarely available. It serves in the first instance to check if a proposed inferential method is sensible, that is, it provides what would be regarded as correct answers if there were an unlimited amount of data related to the problem under study. Beyond this minimal requirement asymptotic theory serves to provide approximate answers when exact ones are unavailable. Any achieved amount of information is of course finite, and sometimes quite small, so these approximations need to be checked against exact answers when possible, to be verified by simulation, and to stand the test of practical experience.

The form of the limiting distribution of a statistic is very often obtained from the first term in an asymptotic expansion, higher order terms decreasing to zero as the sample size or amount of information becomes infinite. By considering further terms in the expansion, we may hope to derive approximations that are more accurate for cases of fixed sample size or information. An example of this is the analysis of the distribution of the average of a sample of independent, identically distributed random variables with finite mean and variance. Under some conditions, an asymptotic expansion of the moment generating function of the average, suitably standardized, has as leading term the moment generating function of a standard normal random variable. Incorporating higher order terms in the expansion directly leads to the Edgeworth approximation, and indirectly to the saddlepoint approximation; both of these underlie the theory used in this book. Asymptotic expansions are not convergent series, so including further terms in an expansion does not guarantee a more accurate approximation. In the absence of any uniform error bounds on the omitted terms, detailed examination of examples is needed. Among several asymptotically equivalent approximations, one will sometimes emerge as preferable, perhaps on the basis of detailed numerical work, or perhaps from more general arguments.

One intriguing feature of the theory of higher order likelihood asymptotics is that relatively simple and familiar quantities play a central role. This is most transparent in the tail area approximations for a single scalar parameter, where a combination of the likelihood ratio statistic and the score or Wald statistic leads to remarkably more accurate results than those available from first order theory, but is also the case in more complex models. In a very general way, the likelihood function emerges as an approximately pivotal quantity, that is, a function of the data and the parameter of interest that has a known distribution. In this sense the approximations can be viewed as generalizing Fisher's (1934) result for location models.

In nearly all applications of statistical inference the models used are provisional, in the sense that they are not derived from a precise and widely-accepted scientific theory, although the most useful models will be broadly consistent with theoretical predictions. Both the usual first order likelihood inferences and the type of higher order approximations we shall discuss depend for their validity on the assumed correctness of the model. The point is often made, quite reasonably, that the gains from improved approximations are potentially outweighed by sensitivity to the modelling assumptions. This will have more or less force in particular examples, depending on the model and the application. For example, linear regression has proved to be a useful starting point in innumerable situations, without any notion of a linear model being the 'true' underlying mechanism, so it is worthwhile to ensure the accuracy of the inferential procedure, insofar as this is possible. Furthermore, the availability of easily computed higher order approximations can enable an investigation of the stability of the conclusions to a variety of models. However if it is likely that data may contain one or more extreme outliers, then understanding the origin and influence of these will usually be of more practical importance than highly accurate calculation of confidence limits in a model that ignores outliers.

Most of the higher order approximations we discuss in the examples are for P-values; that is, the probability of observing a result as or more extreme than that observed, under an assumed value for the parameter of interest. The P-value function provides confidence limits at any desired level of confidence. In problems with discrete data the issue of continuity correction arises, and this will be discussed as needed, particularly in Chapter 4.

In Chapter 2 we give a very brief introduction to the main approximations used in the examples. We review first order results, and describe the construction of first and higher order approximations to P-values and posterior probabilities.

The heart of the book is a collection of examples, or case studies, illustrating the use of higher order asymptotics. We have tried to choose examples that are simple enough to be described briefly, but complex enough to be suggestive for applied work more generally. In Chapter 3 we present some elementary one-parameter models, chosen to illustrate the potential accuracy of the procedures in cases where one can perform the calculations easily, and some one- and two-sample problems. Chapter 4 illustrates regression with categorical responses; in particular logistic regression, several versions of 2×2 tables, and Poisson regression. Chapter 5 illustrates linear and non-normal linear regression, nonlinear regression with normal errors, and nonlinear regression with non-constant variance. In Chapters 3, 4 and 5 we try to emphasize the data as much as possible, and let the models and analysis follow, but the data sets are generally chosen to illustrate particular methods. Chapter 6 treats in depth analysis of data that arose in collaboration with colleagues. In these examples the emphasis is on model building and inference; while higher order approximations are used they are not the main focus of the discussion. Chapter 7 takes a different approach: in order to illustrate the wide range of examples that can be treated with higher order asymptotics, we discuss a number of model classes, and use the data simply to illustrate the calculations.

A more detailed discussion of the theoretical aspects of higher order approximation is given in Chapter 8. This topic has a large literature, and some of it is rather formidable. Issues that must be faced almost immediately are the role of conditioning in inference, and definitions of sufficiency and ancillarity in the presence of nuisance parameters. It is also necessary to consider the likelihood function as a function of both the parameter and the data, a viewpoint that is usually unfamiliar to those outside the area. Our goal is to make this literature somewhat more accessible for use in applications of reasonable sophistication, both to assess to what extent this is useful, and to provide some guidance for those seeking to apply these methods in related problems. For the derivations of the results, we direct the reader to a number of books and review papers in the bibliographic notes.

In Chapter 9 we provide some details of our numerical work, which is based on the R package bundle hoa originally developed for S-PLUS and described by Brazzale (2000). There are a number of general points relevant to any implementation, and some slightly more specialized issues in this approach. For some examples, we needed additional code, and we have provided this on the book's web page (see Preface) where we thought it may be useful.

Chapter 10 contains a variety of problems and further results based on the material, and an appendix sketches the asymptotic methods that form the basis for the development of higher order asymptotic inference.

Bibliographic notes

We have been strongly motivated in this effort by the book *Applied Statistics* by Cox and Snell (1981), whose introductory chapters provide an excellent introduction to the role of statistical models and theory in applied work. See also the introductory chapter of Cox and Wermuth (1996), and Davison (2003, Chapter 12). The role of asymptotics in the theory of statistics is surveyed by Reid (2003) and Skovgaard (2001).

We give more thorough notes related to the literature on higher order asymptotics at the end of Chapter 8, but the main book-length treatments of likelihood-based asymptotics are the monographs by Barndorff-Nielsen and Cox (1989, 1994), Pace and Salvan (1997) and Severini (2000a). An overview is given in Brazzale (2000). The hoa package bundle is described in Brazzale (2005) and the computing strategy for higher order approximations is discussed in Brazzale (1999) and Bellio and Brazzale (2003).

2

Uncertainty and approximation

2.1 Introduction

In the examples in later chapters we use parametric models almost exclusively. These models are used to incorporate a key element of statistical thinking: the explicit recognition of uncertainty. In frequentist settings imprecise knowledge about the value of a single parameter is typically expressed through a collection of confidence intervals, or equivalently by computation of the P-values associated with a set of hypotheses. If prior information is available then Bayes' theorem can be employed to perform posterior inference.

In almost every realistic setting, uncertainty is gauged using approximations, the most common of which rely on the application of the central limit theorem to quantities derived from the likelihood function. Not only does likelihood provide a powerful and very general framework for inference, but the resulting statements have many desirable properties.

In this chapter we provide a brief overview of the main approximations for likelihood inference. We present both first order and higher order approximations; first order approximations are derived from limiting distributions, and higher order approximations are derived from further analysis of the limiting process. A minimal amount of theory is given to structure the discussion of the examples in Chapters 3 to 7; more detailed discussion of asymptotic theory is given in Chapter 8.

2.2 Scalar parameter

In the simplest situation, observations $y_1, \ldots, y_n$ are treated as a realization of independent identically distributed random variables $Y_1, \ldots, Y_n$ whose probability density function $f(y; \theta)$ depends on an unknown scalar parameter θ. Let $\ell(\theta) = \sum \log f(y_i; \theta)$ denote the log likelihood based on the observations, $\widehat{\theta}$ the maximum likelihood estimator, and $j(\theta) = -\partial^2 \ell(\theta)/\partial\theta^2$ the observed information function. Below we adopt the convention that additive constants that do not depend on the parameter may be neglected when log likelihoods are defined, and we suppress them without further comment.

Likelihood inference for θ is typically based on the

$$\text{likelihood root,}\quad r(\theta) = \operatorname{sign}(\widehat{\theta} - \theta)\left[2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\}\right]^{1/2}; \tag{2.1}$$

$$\text{score statistic,}\quad s(\theta) = j(\widehat{\theta})^{-1/2}\partial\ell(\theta)/\partial\theta; \qquad \text{or} \tag{2.2}$$

$$\text{Wald statistic,}\quad t(\theta) = j(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta). \tag{2.3}$$

These quantities are functions of the data and the parameter, so strictly speaking should not be called *statistics* unless θ is fixed at a particular value. Note the role of the observed Fisher information, $j(\widehat{\theta})$, in calibrating the distance of $\widehat{\theta}$ from the true value θ, and the distance of the score function $\ell_\theta(\theta) = \partial\ell(\theta)/\partial\theta$ from its expectation of zero.

Under suitable conditions on the parametric model and in the limit as $n \to \infty$, each of these statistics has, under $f(y;\theta)$, an asymptotic standard normal, $N(0, 1)$, distribution. A closely related quantity, the likelihood ratio statistic

$$w(\theta) = r(\theta)^2 = 2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\}, \tag{2.4}$$

has an asymptotic chi-squared distribution with one degree of freedom, χ^2_1. Sometimes the log likelihood is multimodal, so it is useful to graph it or equivalently $w(\theta)$. In most cases, however, $\ell(\theta)$ has a single prominent mode around $\widehat{\theta}$, and then $r(\theta)$, $s(\theta)$ and $t(\theta)$ are decreasing functions of θ in the region of that mode. Various alternative forms of $s(\theta)$ and $t(\theta)$ may be defined by replacing $j(\widehat{\theta})$ by $j(\theta)$, $i(\widehat{\theta})$ or $i(\theta)$, where $i(\theta) = \mathrm{E}\{j(\theta)\}$ is the expected information; in fact the name 'Wald statistic' used in (2.3) is a slight misnomer as the version standardized with $i(\theta)^{1/2}$ is more correctly associated with Abraham Wald. Under suitable conditions similar distributional results apply far beyond independent identically distributed observations, and for vector parameters – in particular, $w(\theta)$ has an asymptotic χ^2_d distribution when θ is of fixed dimension d. For now, however, we continue to suppose that θ is scalar.

An important variant of the likelihood root is the *modified likelihood root*

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)}\log\left\{\frac{q(\theta)}{r(\theta)}\right\}, \tag{2.5}$$

which is based on so-called *higher order asymptotics* that are described in Chapter 8. The modified likelihood root combines the likelihood root with the score statistic, if $q(\theta) = s(\theta)$, with the Wald statistic, if $q(\theta) = t(\theta)$, or with variants of these which depend on the context. We shall see below that normal approximation to the distribution of $r^*(\theta)$ can provide almost exact inferences for θ, when these are available.

Any of (2.1)–(2.5) or their variants may be used to set confidence intervals or compute a P-value for θ. Suppose, for instance, that it is desired to use the Wald statistic (2.3) to test the null hypothesis that θ equals some specified θ_0 against the alternative $\theta > \theta_0$. As a large positive value of $t(\theta_0)$ relative to the $N(0, 1)$ distribution will give evidence against the null hypothesis, the corresponding P-value is $1 - \Phi\{t(\theta_0)\}$, where Φ denotes the standard normal distribution function. Likewise a confidence interval may be constructed using those values of θ most consistent with this distribution; for example, the limits of an equi-tailed $(1-2\alpha)$ interval $(\theta_\alpha, \theta^\alpha)$ for θ_0 are given by

$$\Phi\{t(\theta_\alpha)\} = 1-\alpha, \quad \Phi\{t(\theta^\alpha)\} = \alpha,$$

or equivalently

$$t(\theta_\alpha) = z_{1-\alpha}, \quad t(\theta^\alpha) = z_\alpha,$$

where z_α is the α quantile of the standard normal distribution. The resulting interval,

$$(\theta_\alpha, \theta^\alpha) = (\widehat{\theta} - z_{1-\alpha} j(\widehat{\theta})^{-1/2}, \widehat{\theta} - z_\alpha j(\widehat{\theta})^{-1/2}),$$

is highly convenient and widely used because it can be computed for any desired α using only $\widehat{\theta}$ and $j(\widehat{\theta})^{-1/2}$. Often in practice a 95% confidence interval is sought; then $\alpha = 0.025$, $z_\alpha = -1.96$, and the interval has limits of the familiar form $\widehat{\theta} \pm 1.96 j(\widehat{\theta})^{-1/2}$.

Similar computations yield intervals based on $r(\theta)$, $s(\theta)$ and $r^*(\theta)$, though typically numerical solution is required. Suppose, for example, that a $(1-2\alpha)$ confidence interval is to be based on $r^*(\theta)$. Then one possibility is to compute the values of $r^*(\theta)$ for a grid of values of θ, to fit a spline or other interpolating curve to the resulting pairs $(r^*(\theta), \theta)$, and to read off those values of θ corresponding to $r^*(\theta) = z_\alpha, z_{1-\alpha}$. We often plot such curves or the corresponding probabilities $\Phi\{r^*(\theta)\}$, and call the plots *profiles*. An alternative to $\Phi\{r^*(\theta)\}$ which has the same asymptotic properties is the *Lugannani–Rice formula*

$$\Phi\{r(\theta)\} + \left\{\frac{1}{r(\theta)} - \frac{1}{q(\theta)}\right\} \phi\{r(\theta)\}, \tag{2.6}$$

where ϕ denotes the standard normal density function.

When the likelihood is unimodal, confidence intervals based on $r(\theta)$ and on $w(\theta)$ are the same, and equal to

$$\left\{\theta : w(\theta) \le c_{1,1-2\alpha}\right\}. \tag{2.7}$$

Here and below we use $c_{\nu,1-2\alpha}$ to denote the $(1-2\alpha)$ quantile of the χ^2_ν distribution. If $\ell(\theta)$ is multimodal then (2.7) may be a union of disjoint intervals. If the random variable Y is continuous then the chi-squared approximation to the distribution of $w(\theta)$ is improved if a *Bartlett adjustment* is used: $w(\theta)$ is replaced in (2.7) by $w(\theta)/(1+b/n)$, where the *Bartlett correction* b is computed from $\mathrm{E}\{w(\theta)\}$.

A quantity which depends both on the data and on the parameter, and whose distribution is known, is called a *pivot*, or *pivotal quantity*. The previous paragraphs describe the use of the quantities $r(\theta)$, $s(\theta)$, $t(\theta)$, $r^*(\theta)$ and $w(\theta)$, regarded as functions of the data and of θ, as *approximate* pivots; their asymptotic distributions are used to set confidence intervals or compute P-values for inference on θ. For brevity we will refer to $r(\theta)$, $s(\theta)$ and so forth as pivots even when their exact distributions are unknown. We call the functions of θ obtained by computing the P-value for a range of values of θ *significance functions*; an example is $\Phi\{r(\theta)\}$. As we now illustrate, the usefulness of these significance functions will depend on the accuracy of the underlying distributional approximations.

Illustration: Exponential data

Suppose that a sample $y_1, \ldots, y_n$ is available from the exponential density

$$f(y;\theta)=\theta\exp(-\theta y),\quad y>0,\theta>0,$$

and that a 95% confidence interval is required for θ. The log likelihood is

$$\ell(\theta)=n(\log\theta-\theta\overline{y}),\quad \theta>0,$$

where $\overline{y}=(y_1+\cdots+y_n)/n$ is the sample average. Here $\ell(\theta)$ is unimodal with maximum at $\widehat{\theta}=1/\overline{y}$ and observed information function $j(\theta)=n/\theta^2$, and

$$\begin{aligned} r(\theta) &= \operatorname{sign}(1-\theta\overline{y})\left[2n\left\{\theta\overline{y}-\log(\theta\overline{y})-1\right\}\right]^{1/2}, \\ s(\theta) &= n^{1/2}\{1/(\theta\overline{y})-1\}, \\ t(\theta) &= n^{1/2}(1-\theta\overline{y}). \end{aligned}$$

In this exponential family model it is appropriate to take $q(\theta)=t(\theta)$ in the construction of $r^*(\theta)$. The Bartlett correction is readily obtained on noting that $\mathrm{E}\{w(\theta)\}=2n\{\log n-\Psi(n)\}$, where $\Psi(n)=\mathrm{d}\log\Gamma(n)/\mathrm{d}n$ is the digamma function.

The quality of the approximations outlined above may be assessed using the exact pivot $\theta\sum Y_i$, whose distribution is gamma with unit scale and shape parameter n.

Consider an exponential sample with $n=1$ and $\overline{y}=1$; then $j(\widehat{\theta})=1$. The log likelihood $\ell(\theta)$, shown in the left-hand panel of Figure 2.1, is unimodal but strikingly asymmetric, suggesting that confidence intervals based on an approximating normal distribution for $\widehat{\theta}$ will be poor. The right-hand panel is a chi-squared probability plot in which the ordered values of simulated $w(\theta)$ are graphed against quantiles of the χ_1^2 distribution – if the simulations lay along the diagonal line $x=y$, then this distribution would be a perfect fit. The simulations do follow a straight line rather closely, but with slope $(1+b/n)$, where $b=0.1544$. This indicates that the distribution of the Bartlett-adjusted likelihood ratio statistic $w(\theta)/(1+b/n)$ would be essentially χ_1^2. The 95% confidence intervals for θ based on the unadjusted and adjusted likelihood ratio statistics are $(0.058, 4.403)$ and $(0.042, 4.782)$ respectively.

The left-hand panel of Figure 2.2 shows the pivots $r(\theta)$, $s(\theta)$, $t(\theta)$, and $r^*(\theta)$. The limits of an equi-tailed 95% confidence interval for θ based on $r(\theta)$ are given by the intersections of the horizontal lines at ±1.96 with the curve $r(\theta)$, yielding $(0.058, 4.403)$; this equals the interval given by $w(\theta)$, as of course it should. The 95% interval based on the exact pivot $\theta\sum y_i$ is $(0.025, 3.689)$, while the interval based on $t(\theta)$ is

$$\widehat{\theta}\pm 1.96 j(\widehat{\theta})^{-1/2}=1\pm 1.96=(-0.96, 2.96),$$

which is a very unsatisfactory statement of uncertainty for a positive quantity. The 95% confidence interval based on $r^*(\theta)$ is $(0.024, 3.705)$, very close to the exact one. This illustrates one important advantage of $r(\theta)$ and $r^*(\theta)$; intervals based on these

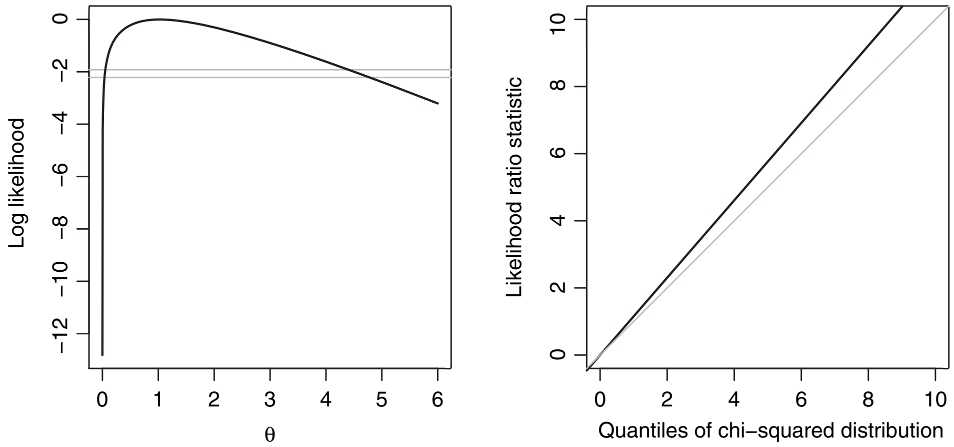


Figure 2.1 Likelihood inference for exponential sample of size $n = 1$. Left: log likelihood $\ell(\theta)$. Intersection of the function with the two horizontal lines gives two 95% confidence intervals for θ: the upper line is based on the χ_1^2 approximation to the distribution of $w(\theta)$, and the lower line is based on the Bartlett-adjusted statistic. Right: comparison of simulated values of likelihood ratio statistic $w(\theta)$ with χ_1^2 quantiles. The χ_1^2 approximation is shown by the line of unit slope, while the $(1 + b/n)\chi_1^2$ approximation is shown by the upper straight line.

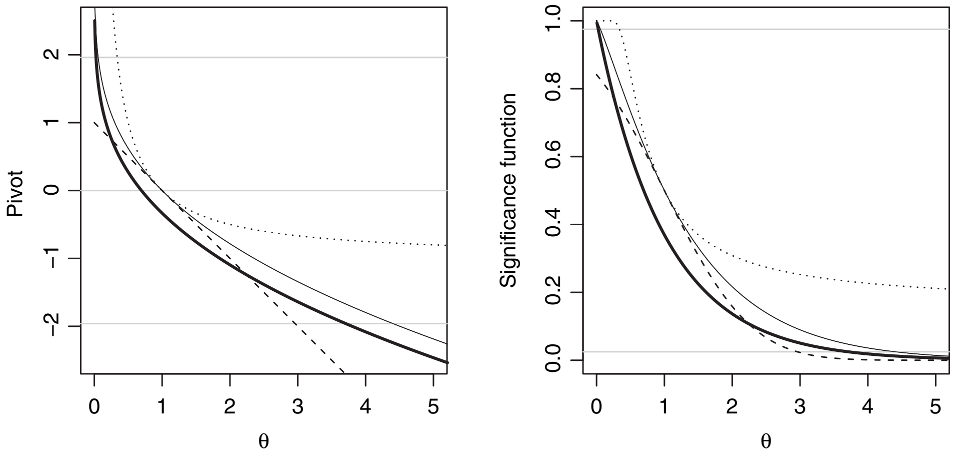


Figure 2.2 Approximate pivots and P-values based on an exponential sample of size $n = 1$. Left: likelihood root $r(\theta)$ (solid), score pivot $s(\theta)$ (dots), Wald pivot $t(\theta)$ (dashes), modified likelihood root $r^*(\theta)$ (heavy), and exact pivot $\theta \sum y_i$ indistinguishable from the modified likelihood root. The horizontal lines are at 0, ±1.96. Right: corresponding significance functions, with horizontal lines at 0.025 and 0.975.

must lie within the parameter space. A further advantage is the invariance of these intervals to parameter transformation: for example, the likelihood root for θ^{-1} yields the 95% confidence interval $(1/4.403, 1/0.057)$, which is obtained simply by applying the reciprocal transformation to the likelihood root interval for θ. The interval based on $t(\theta)$

is not similarly transformation-invariant, while that based on the score statistic $s(\theta)$ has upper limit at infinity, and is also unsatisfactory.

The right-hand panel of Figure 2.2 shows the significance functions $\Phi\{r(\theta)\}$ and $\Phi\{t(\theta)\}$ corresponding to the likelihood root and Wald pivot, with the exact significance function based on the gamma distribution of $\theta \sum y_i$. Also shown is $\Phi\{r^*(\theta)\}$, which coincides with the exact significance function to plotting accuracy.

The remarkable behaviour of $r^*(\theta)$ in this example underlines a major advantage of higher order approximations over more conventional ones. Approximations to distribution functions based on the central limit theorem typically have an error of order $n^{-1/2}$, so in some cases the approximation is very poor – as shown by $t(\theta)$ in Figure 2.2. By contrast the approximation involving $r^*(\theta)$ has relative error of order $n^{-3/2}$ in the centre of the distribution and of order n^{-1} in the tails, although in the absence of detailed information on the constants in these terms it is difficult to completely explain its remarkable accuracy.

This exponential example is one in which exact inference is possible. In most realistic cases approximation based on asymptotic arguments is required, however, and then the quality of the inference will depend on the accuracy of the approximation. As we have seen above, the same ingredients can be combined in different ways to give markedly different results – in particular, inference based on $r^*(\theta)$ appeared to recover the exact results almost perfectly. The applications in later chapters of this book are intended to illustrate when and how such higher order approximations are useful. In the rest of this chapter we sketch a few key notions needed to apply them. A more detailed treatment is given in Chapter 8.

2.3 Several parameters

Almost all realistic models have several unknown parameters. We will assume that the parameter is a $d \times 1$ vector which may be expressed as $\theta = (\psi, \lambda)$, where the focus of scientific enquiry is the *interest parameter* ψ. The vector of *nuisance parameters* λ, though essential for realistic modelling, is of secondary importance. In many cases ψ is scalar or individual components of a vector ψ may be treated one at a time. Occasionally it is more convenient to consider ψ as a constraint on the full vector θ rather than a specific component; this extension is discussed in Section 8.5.3.

A quite general likelihood-based approach to eliminating the nuisance parameter λ is to replace it by the *constrained maximum likelihood estimate* $\widehat{\lambda}_\psi$ obtained by maximizing $\ell(\psi, \lambda)$ with respect to λ for fixed ψ, and then summarizing knowledge about ψ through the *profile log likelihood*

$$\ell_{\mathrm{p}}(\psi) = \max_{\lambda} \ell(\psi, \lambda) = \ell(\psi, \widehat{\lambda}_\psi).$$

The observed information function for the profile log likelihood

$$j_{\mathrm{p}}(\psi) = -\frac{\partial^2 \ell_{\mathrm{p}}(\psi)}{\partial\psi \partial\psi^{\mathrm{T}}}$$

can be expressed in terms of the observed likelihood function for the full log likelihood by means of the identity

$$j_{\mathrm{p}}(\psi) = \{j^{\psi\psi}(\psi, \widehat{\lambda}_\psi)\}^{-1},$$

where $j^{\psi\psi}(\psi, \lambda)$ is the (ψ, ψ) block of the inverse of the observed information matrix $j(\psi, \lambda)$. If ψ is a scalar then

$$j_{\mathrm{p}}(\psi) = \frac{|j(\psi, \widehat{\lambda}_\psi)|}{|j_{\lambda\lambda}(\psi, \widehat{\lambda}_\psi)|},$$

where $|\cdot|$ indicates determinant. To a first order of approximation we can treat $\ell_{\mathrm{p}}(\psi)$ as an ordinary log likelihood, for instance basing confidence intervals for ψ on a chi-squared approximation to the likelihood ratio statistic

$$w_{\mathrm{p}}(\psi) = 2\left\{\ell_{\mathrm{p}}(\widehat{\psi}) - \ell_{\mathrm{p}}(\psi)\right\},$$

or on normal approximation to appropriate versions of approximate pivots r, s and t. However, such approximations can give poor results, particularly if the dimension of λ is high. The difficulty is that $\ell_{\mathrm{p}}(\psi)$ is not in general the logarithm of a density function.

It is possible to define a modified likelihood root r^*, analogous to the modified likelihood root of the previous section, that incorporates both an improved approximation and an adjustment for the elimination of the nuisance parameters. This modified likelihood root is central to the higher order approximations used in the following chapters. It will be seen to be well approximated in distribution by the standard normal distribution, and to combine the likelihood root $r(\psi)$, now defined in terms of the profile log likelihood as $\mathrm{sign}(\widehat{\psi} - \psi) w_{\mathrm{p}}^{1/2}(\psi)$, with an approximate pivot $q(\psi)$. In general the form of $q(\psi)$ is somewhat complex, and detailed consideration is postponed until Chapter 8. However, its ingredients are particularly simple when the log likelihood has exponential family form

$$\ell(\psi, \lambda) = \psi u + \lambda^{\mathrm{T}} v - c(\psi, \lambda), \tag{2.8}$$

where the *canonical parameter* is (ψ, λ) and the *natural observation* is (u, v). In such models the distribution of

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

is well approximated by the standard normal law, with likelihood root and Wald pivot

$$\begin{aligned} r(\psi) &= \mathrm{sign}(\widehat{\psi} - \psi)\left[2\left\{\ell_{\mathrm{p}}(\widehat{\psi}) - \ell_{\mathrm{p}}(\psi)\right\}\right]^{1/2}, \\ t(\psi) &= j_{\mathrm{p}}^{1/2}(\widehat{\psi})(\widehat{\psi} - \psi) \end{aligned}$$

and

$$q(\psi) = t(\psi)\ \rho(\psi, \widehat{\psi}). \tag{2.9}$$

This is a modified form of Wald pivot, where

$$\rho(\psi, \widehat{\psi}) = \left\{ \frac{|j_{\lambda\lambda}(\widehat{\theta})|}{|j_{\lambda\lambda}(\widehat{\theta}_\psi)|} \right\}^{1/2}$$

and we have introduced the shorthand notation $\widehat{\theta} = (\widehat{\psi}, \widehat{\lambda})$ and $\widehat{\theta}_\psi = (\psi, \widehat{\lambda}_\psi)$.

It can sometimes be useful to decompose the modified likelihood root as

$$r^*(\psi) = r(\psi) + r_{\mathrm{INF}}(\psi) + r_{\mathrm{NP}}(\psi), \tag{2.10}$$

where r_{INF} makes an adjustment allowing for non-normality of r, and r_{NP} compensates r for the presence of the nuisance parameters; expressions for these terms are given in Section 8.6. Graphs of r_{INF} and r_{NP} can be useful in diagnosing the causes of any strong divergences between inferences based on r and on r^*.

Depending on the form of the model $f(y; \psi, \lambda)$, a more direct approach to elimination of nuisance parameters may be available. The most common instances of this are a *conditional likelihood* or a *marginal likelihood*, taken to be the first terms of the decompositions

$$f(y; \psi, \lambda) = \begin{cases} f_{\mathrm{c}}(y \mid u; \psi) f_{\mathrm{m}}(u; \psi, \lambda), \\ f_{\mathrm{m}}(u; \psi) f_{\mathrm{c}}(y \mid u; \psi, \lambda), \end{cases}$$

respectively, when such a factorization is possible. Typically conditional likelihood arises in exponential family models and marginal likelihood arises in transformation models, and numerous examples of both will be given in later chapters. Likelihood roots and related quantities can be defined using such a function, if it is available, and may be used for exact inference on ψ simply by applying the discussion in Section 2.2 to the conditional or marginal likelihood function.

It turns out that the corresponding approximations for conditional and marginal log likelihoods, and the higher order approximations based on r^*, are both closely related to the use of an *adjusted profile log likelihood*

$$\ell_{\mathrm{a}}(\psi) = \ell(\psi, \widehat{\lambda}_\psi) - \tfrac{1}{2} \log \left| j_{\lambda\lambda}(\psi, \widehat{\lambda}_\psi) \right|,$$

which may be derived from a log posterior marginal density for ψ. Unfortunately ℓ_{a} is not invariant to reparametrization, and a further, in general rather complicated, term must be added to achieve the desired invariant *modified profile log likelihood* $\ell_{\mathrm{m}}(\psi)$. In most of the cases in this book this additional term simplifies considerably; see Sections 8.5.3 and 8.6. An alternative is to use ℓ_{a} but in an *orthogonal parametrization*, chosen so that the contribution made by this extra term is reduced. This implies setting the model up so that the corner of the expected information matrix corresponding to ψ and λ is identically zero. As a first step in investigating the effect of the estimation of nuisance parameters on inference, it can be useful to compare plots of $\ell_{\mathrm{p}}(\psi)$ and $\ell_{\mathrm{a}}(\psi)$.

Illustration: Gamma data

Suppose that a sample $y_1, \ldots, y_n$ is available from the gamma density

$$f(y; \psi, \lambda) = \frac{\lambda^{\psi} y^{\psi-1}}{\Gamma(\psi)} \exp(-\lambda y), \quad y > 0, \lambda, \psi > 0,$$

and that interest is focused on the shape parameter ψ. The log likelihood may be written as

$$\ell(\psi, \lambda) = \psi u + \lambda v + n\{\psi \log \lambda - \log \Gamma(\psi)\}, \quad \lambda, \psi > 0,$$

where $u = \sum \log y_i$ and $v = -n\bar{y}$, and so $j_{\lambda\lambda}(\psi, \lambda) = n\psi/\lambda^2$. Now $\widehat{\lambda}_{\psi} = \psi/\bar{y}$, giving $\rho(\psi, \widehat{\psi}) = (\psi/\widehat{\psi})^{1/2}$,

$$\ell_{\mathrm{p}}(\psi) = \psi(u - n) + n\{\psi \log(\psi/\bar{y}) - \log \Gamma(\psi)\}, \quad \psi > 0,$$

and $j_{\mathrm{p}}(\psi) = n\{\Psi'(\psi) - 1/\psi\}$, where $\Psi(\psi)$ denotes the digamma function.

Suppose that the five observations 0.2, 0.45, 0.78, 1.28 and 2.28 are available. The left-hand panel of Figure 2.3 shows both the profile log likelihood for ψ and the conditional log likelihood for the distribution of u given v, which gives exact inferences on ψ because it does not involve λ. The right-hand panel shows the likelihood root $r(\psi)$, the Wald pivot $t(\psi) = j_{\mathrm{p}}(\widehat{\psi})^{1/2}(\widehat{\psi} - \psi)$, and the modified likelihood root $r^*(\psi)$. As with the adjusted profile log likelihood, r^* produces intervals which are shifted towards the origin relative to those based on ℓ_{p} and on the likelihood root r. The value of $r^*(\psi)$ is extremely close to the curve corresponding to an exact pivot available in this case; the difference is not

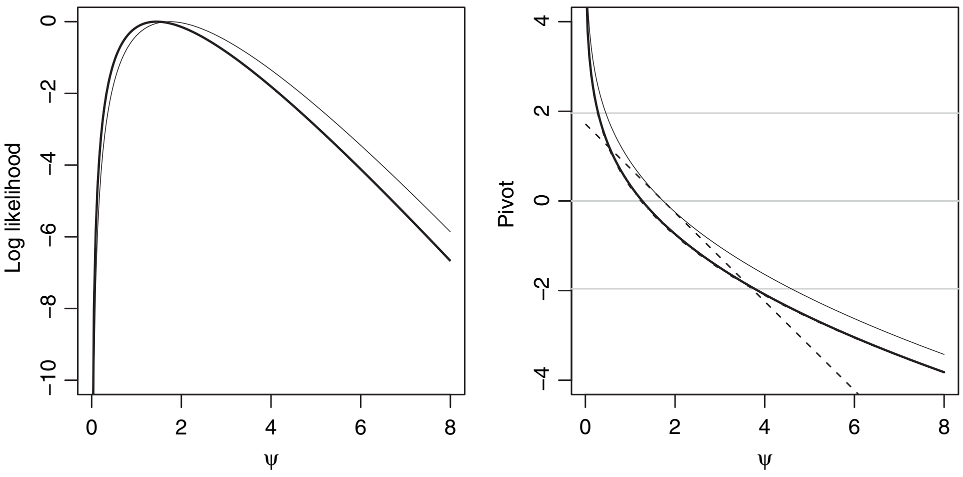


Figure 2.3 Inference for shape parameter ψ of gamma sample of size $n = 5$. Left: profile log likelihood ℓ_{p} (solid) and the log likelihood from the conditional density of u given v (heavy). Right: likelihood root $r(\psi)$ (solid), Wald pivot $t(\psi)$ (dashes), modified likelihood root $r^*(\psi)$ (heavy), and exact pivot overlying $r^*(\psi)$. The horizontal lines are at 0, ±1.96.

visible in Figure 2.3. Intervals based on the Wald pivot are again unsuitable owing to a poor normal approximation to the distribution of $\widehat{\psi}$. The adjusted profile log likelihood function $\ell_{\mathrm{a}}(\psi)$, computed using the orthogonal parametrizaton $(\psi, \mu = \psi/\lambda)$, is in this model identically equal to the likelihood from the exact conditional density of u given v, although this will rarely be the case, even in exponential family models.

2.4 Further remarks

The most difficult aspect of higher order approximation is determining the approximate pivot $q(\psi)$ to use in the construction of r^*. In linear exponential families, where the parameter of interest is a component of the canonical parameter, the expression (2.9) is readily computable from the fitting of any of the standard generalized linear models, and can be shown in general to lead to excellent approximation of the exact conditional distribution. This is the basis for the `cond` package in the `hoa` bundle for R, and is illustrated in Chapters 3 and 4.

There is a similar direct expression for $q(\psi)$ for inference about β_j or σ in regression-scale models, described in more detail in Chapter 8, and illustrated in Chapters 5 and 6; q is a combination of the score pivot s and a correction for nuisance parameters. The r^* pivot in this case gives very accurate approximation to the appropriate marginal distribution. The `marg` package in the `hoa` bundle implements this.

The ideas above have been extended to more general models in numerous ways. One uses a local exponential family approximation in which the canonical parameter is $\varphi(\theta)$, yielding the expression

$$q(\psi) = \frac{|\varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi) \quad \varphi_\lambda(\widehat{\theta}_\psi)|}{|\varphi_\theta(\widehat{\theta})|} \frac{|j(\widehat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\widehat{\theta}_\psi)|^{1/2}}, \tag{2.11}$$

where, for example, φ_θ denotes the matrix $\partial\varphi/\partial\theta^{\mathrm{T}}$ of partial derivatives. Expression (2.11) is invariant to affine transformation of φ (Problem 4, see Chapter 10), and this can greatly simplify analytical calculations. The numerator of the first term of $q(\psi)$ is the determinant of a $d \times d$ matrix whose first column is $\varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi)$ and whose remaining columns are $\varphi_\lambda(\widehat{\theta}_\psi)$. For a sample of independent observations φ is defined as

$$\varphi(\theta)^{\mathrm{T}} = \sum_{i=1}^{n} \left. \frac{\partial \ell(\theta; y)}{\partial y_i} \right|_{y=y^0} V_i, \tag{2.12}$$

where y^0 denotes the observed data, and $V_1, \ldots, V_n$, whose general construction is described in Section 8.4.3 is a set of $1 \times d$ vectors that depend on the observed data alone. An important special case is that of a log likelihood with independent contributions of curved exponential family form,

$$\ell(\theta) = \sum_{i=1}^{n} \{\alpha_i(\theta) y_i - c_i(\theta)\}, \tag{2.13}$$

for which

$$\varphi(\theta)^{\mathrm{T}} = \sum_{i=1}^{n} \alpha_i(\theta) V_i.$$

Affine invariance implies that if $n = d$, that is, we have a reduction by sufficiency to a set of d variables, we can take $\varphi(\theta)^{\mathrm{T}} = (\alpha_1(\theta), \ldots, \alpha_d(\theta))$. Likewise, if $\alpha_i(\theta) \equiv \alpha(\theta)$, then $\varphi(\theta) = \alpha(\theta)$.

The vectors V_i implement conditioning on an approximately ancillary statistic. When y_i has a continuous distribution, V_i is computed by differentiating y_i with respect to θ, for a fixed pivotal z_i; see Section 8.4.3. The expression for V_i computed this way is, as at (8.19),

$$V_i = \left.\frac{\mathrm{d}y_i}{\mathrm{d}\theta^{\mathrm{T}}}\right|_{\theta=\widehat{\theta}} = -\left(\frac{\partial z_i}{\partial y_i}\right)^{-1}\left.\left(\frac{\partial z_i}{\partial \theta^{\mathrm{T}}}\right)\right|_{\theta=\widehat{\theta}}. \tag{2.14}$$

Inference using (2.11) is relatively easily performed. If functions are available to compute the log likelihood $\ell(\theta)$ and the constructed parameter $\varphi(\theta)$, then the maximisations needed to obtain $\widehat{\theta}$ and $\widehat{\theta}_\psi$ and the derivatives needed to compute (2.11) may be obtained numerically. Explicit formulae for the quantities involved are available for a variety of common classes of models, such as linear regression with non-normal errors and nonlinear regression with normal errors, and these are discussed in later chapters. In many cases standard fitting routines can be exploited and so a negligible additional effort is needed for higher order inference.

Other readily implemented variants of $q(\psi)$ have been proposed. In particular we also use one described in Section 8.5.3 which we call Skovgaard's approximation.

Above, we have supposed that the observations are continuous. If they are discrete, then the ideas remain applicable, but the theoretical accuracy of the approximations is reduced, and the interpretation of significance functions such as $\Phi\{r^*(\theta)\}$ changes slightly; see Sections 3.3, 3.4 and 8.5.4. The construction (2.14) is replaced with

$$V_i = \left.\frac{\mathrm{dE}(y_i;\,\theta)}{\mathrm{d}\theta^{\mathrm{T}}}\right|_{\theta=\widehat{\theta}}; \tag{2.15}$$

see Section 8.5.4 and Problem 55.

There is a close link with analytical approximations useful for Bayesian inference. Suppose that posterior inference is required for ψ and that the chosen prior density is $\pi(\psi, \lambda)$. Then it turns out that using

$$q(\psi) = -\ell_{\mathrm{p}}'(\psi) j_{\mathrm{p}}(\widehat{\psi})^{-1/2} \rho^{-1}(\psi, \widehat{\psi}) \frac{\pi(\widehat{\theta})}{\pi(\widehat{\theta}_\psi)}, \tag{2.16}$$

where ℓ_{p}' is the derivative of $\ell_{\mathrm{p}}(\psi)$ with respect to ψ, in the definition of the modified likelihood root, leads to a Laplace-type approximation to the marginal posteror distribution for ψ; see Section 8.7.