STATISTICAL ASYMPTOTICS

This is a commentary on the APTS module 'Statistical Asymptotics'. Please notify the author of errors in these notes (e-mail alastair.young@imperial.ac.uk).

The material of the module is arranged in **three** chapters, of which the first constitutes background material, and the preliminary reading for the module. Some of the key *statistical* ideas of this chapter will be reviewed as necessary during the module, and may have been covered in the APTS module 'Statistical Inference'. However, the probability material should be treated as prerequisite. The material in Sections 1.9 and 1.10 is included to provide a more complete picture, but is non-essential.

The key reference for the module is Young and Smith (2005). A useful background text, which presents basic ideas and techniques of inference, is Casella and Berger (1990). Davison (2003) is another excellent reference: Chapters 4 and 7 represent further very suitable preliminary reading and Chapter 12 is particularly relevant to the course.

Chapters 1 and 2 follow Barndorff-Nielsen and Cox (1994) quite closely. The introductory chapters of Cox and Hinkley (1974) are also drawn on. Some of the material, in particular the large-sample theory in Chapter 2, expands upon components of the APTS module 'Statistical Inference'. The heart of the module is Chapter 3, which is drawn from Young and Smith (2005), and is intended to give a snapshot of important current ideas in asymptotic inference. Many results are stated without proof. Some of the derivations are hard, and beyond the scope of the course.

Another excellent book for the module is Pace and Salvan (1997). The book by Severini (2000) is also strongly recommended, as being a bit more accessible than Barndorff-Nielsen and Cox (1994).

Analytic methods used in the course are detailed by Barndorff-Nielsen and Cox (1989).

The objectives of the module are: (i) to provide an overview of central asymptotic theory of statistical inference, in particular of likelihood-based approaches; (ii) to provide an introduction to analytic methods and tools, in particular approximation techniques that are widely used in the development of statistical theory and methodology; (iii) to provide exposure to key ideas in contemporary statistical theory; and (iv) to provide practice in application of key techniques to particular problems.

References

- Barndorff-Nielsen, O.E. and Cox, D.R. (1989) *Asymptotic Techniques for Use in Statistics*, Chapman and Hall.

- Barndorff-Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics*, Chapman and Hall.
- Casella, G. and Berger, R.L. (1990) *Statistical Inference*, Wadsworth & Brooks/Cole.
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman and Hall.
- Davison, A.C. (2003) *Statistical Models*, Cambridge University Press.
- Pace, L. and Salvan, A. (1997) *Principles of Statistical Inference from a Neo-Fisherian Perspective*, World Scientific.
- Severini, T.A. (2000) *Likelihood Methods in Statistics*, Oxford University Press.
- Young, G.A. and Smith, R.L. (2005) *Essentials of Statistical Inference*, Cambridge University Press.

1 Concepts and Principles

1.1 Introduction

The representation of experimental and observational data as outcomes of random variables provides a structure for the systematic treatment of inference from data, by which inductive conclusions from the particular to the general can be drawn. Such a systematic treatment involves first the formalisation, in mathematical terms, of several basic concepts about data as observed values of random variables. The aim of this chapter is to introduce these concepts and provide a formal basis for the methods of inference discussed in Chapters 2 and 3.

We wish to analyse observations, $y_1, \ldots, y_n$, collected as an $n \times 1$ vector $y = (y_1, \ldots, y_n)^T$. Then:

1. We regard y as the observed value of a random variable $Y = (Y_1, \ldots, Y_n)^T$ having an (unknown) probability distribution conveniently specified by a probability density function $f(y) = f_Y(y)$, with respect to an appropriate measure, usually Lebesgue measure on $\mathbb{R}^n$ or counting measure (so that Y is either a continuous or discrete random variable).

2. We restrict the unknown density to a suitable family $\mathcal{F}$. We are concerned primarily with the case where the density is of known analytical form, but involves a finite number of real unknown parameters $\theta = (\theta^1, \ldots, \theta^d)^T$. We specify the region $\Omega_\theta \subseteq \mathbb{R}^d$ of possible values of θ, called the parameter space. To indicate the dependency of the density on θ we write $f(y; \theta)$ and refer to this as the 'model function'.

3. We assume that the objective of the inference is one or more of:

   (a) assessing some aspects of θ, for example the value of a single component θ^b, say;

   (b) predicting the value of some as yet unobserved random variable whose distribution depends on θ;

   (c) examining the adequacy of the model specified by $\mathcal{F}$ and Ω_θ.

We will be concerned predominantly with (a). There are three main types of inference we might be interested in, point estimation, interval estimation and hypothesis testing. In point estimation, a single value is computed from the data y, and used as an estimate of the parameter of interest. Interval estimation provides a range of values which have some predetermined

high probability of including the true, but unknown, value of the parameter. Hypothesis testing sets up specific hypotheses regarding the parameter of interest and assesses the plausibility of any specified hypothesis by seeing whether the data y supports or refutes that hypothesis. It is assumed that the reader is familiar with basic procedures of inference, which can be evaluated in terms of formal optimality criteria.

Our objective in these notes is to provide a framework for the relatively systematic analysis of a wide range of possible $\mathcal{F}$. We do not do this by aiming to satisfy various formal optimality criteria, but rather by focusing on fundamental elements of the theory of statistical inference, in particular the likelihood function and quantities derived from it: a 'neo-Fisherian' approach to inference.

1.2 Special models

Two general classes of models particularly relevant in theory and practice are exponential families and transformation families.

1.2.1 Exponential families

Suppose that the distribution of Y depends on m unknown parameters, denoted by $\phi = (\phi^1, \ldots, \phi^m)^T$, to be called natural parameters, through a density of the form

$$f_Y(y;\phi) = h(y)\exp\{s^T\phi - K(\phi)\}, \quad y \in \mathcal{Y}, \tag{1.1}$$

where $\mathcal{Y}$ is a set not depending on ϕ. Here $s \equiv s(y) = (s_1(y), \ldots, s_m(y))^T$, are called natural statistics. The value of m may be reduced if the components of ϕ satisfy a linear constraint, or if the components of s are (with probability one) linearly dependent. So assume that the representation (1.1) is minimal, in that m is as small as possible. Provided the natural parameter space Ω_ϕ consists of all ϕ such that

$$\int h(y)\exp\{s^T\phi\}dy < \infty,$$

we refer to the family $\mathcal{F}$ as a full exponential model, or an (m, m) exponential family.

Usually, we can measure ϕ from some suitable origin $\phi_0 \in \Omega_\phi$, by rewriting (1.1) as

$$f_Y(y;\phi) = f_Y(y;\phi_0)\exp[s^T(\phi - \phi_0) - \{K(\phi) - K(\phi_0)\}].$$

We refer to $f_Y(y;\phi)$ as the (m,m) exponential family generated from the baseline $f_Y(y;\phi_0)$, by exponential tilting via s. We generate all the members of the family by tilting a single baseline density. This exponential tilting idea will be used later, in Chapter 3.

We have from (1.1) that the moment generating function of the random variable $S = s(Y)$ corresponding to the natural statistic s is, writing $t = (t_1, \ldots, t_m)$,

$$\begin{aligned} M(S;t,\phi) &= E\{\exp(S^T t)\} \\ &= \int h(y)\exp\{s^T(\phi+t)\}dy \times \exp\{-K(\phi)\} \\ &= \exp\{K(\phi+t) - K(\phi)\}, \end{aligned}$$

from which we obtain

$$E(S_i;\phi) = \left.\frac{\partial M(s;t,\phi)}{\partial t_i}\right|_{t=0} = \frac{\partial K(\phi)}{\partial \phi^i}\,,$$

or

$$E(S;\phi) = \nabla K(\phi),$$

where ∇ is the gradient operator $(\partial/\partial\phi^1, \ldots, \partial/\partial\phi^m)^T$. Also,

$$\mathrm{cov}(S_i, S_j;\phi) = \frac{\partial^2 K(\phi)}{\partial\phi^i\partial\phi^j}\,.$$

To compute $E(S_i)$ etc. it is only necessary to know the function $K(\phi)$.

Let $s(y) = (t(y), u(y))$ be a partition of the vector of natural statistics, where t has k components and u is $m-k$ dimensional. Consider the corresponding partition of the natural parameter $\phi = (\tau, \xi)$. The density of a generic element of the family can be written as

$$f_Y(y;\tau,\xi) = \exp\{\tau^T t(y) + \xi^T u(y) - K(\tau,\xi)\}h(y).$$

Two key results hold, which make exponential families particularly attractive, as they allow inference about selected components of the natural parameter, in the absence of knowledge about the other components.

First, the family of marginal distributions of $U = u(Y)$ is an $m-k$ dimensional exponential family,

$$f_U(u;\tau,\xi) = \exp\{\xi^T u - K_\tau(\xi)\}h_\tau(u),$$

say.

Secondly, the family of conditional distributions of $T = t(Y)$ given $u(Y) = u$ is a k dimensional exponential family, and the conditional densities are free of ξ, so that

$$f_{T|U=u}(t; u, \tau) = \exp\{\tau^T t - K_u(\tau)\}h_u(t),$$

say.

A proof of both of these results is given by Pace and Salvan (1997, p. 190). The key is to observe that the family of distributions of the natural statistics is an m dimensional exponential family, with density

$$f_{T,U}(t, u; \tau, \xi) = \exp\{\tau^T t + \xi^T u - K(\tau, \xi)\}p_0(t, u),$$

where $p_0(t, u)$ denotes the density of the natural statistics when $(\tau, \xi) = (0, 0)$, assuming without loss of generality that $0 \in \Omega_\phi$.

In the situation described above, both the natural statistic and the natural parameter lie in m-dimensional regions. Sometimes, ϕ may be restricted to lie in a d-dimensional subspace, $d < m$. This is most conveniently expressed by writing $\phi = \phi(\theta)$ where θ is a d-dimensional parameter. We then have

$$f_Y(y; \theta) = h(y) \exp[s^T \phi(\theta) - K\{\phi(\theta)\}]$$

where $\theta \in \Omega_\theta \subseteq \mathbb{R}^d$. We call this system an (m, d) exponential family, noting that we required that no element of ϕ is a linear combination of the other elements, so that $(\phi^1, \ldots, \phi^m)$ does not belong to a v-dimensional linear subspace of $\mathbb{R}^m$ with $v < m$: we indicate this by saying that the exponential family is curved. Think of the case $m = 2, d = 1$: $\{\phi^1(\theta), \phi^2(\theta)\}$ defines a curve in the plane, rather than a straight line, as θ varies.

The following simple model, which is actually of some importance in many biological and agricultural problems, is useful as an illustration of many of the ideas of this chapter. It concerns a normal distribution 'of known coefficient of variation'. Full details of analysis are developed on the problem sheet.

Example: normal distribution, known coefficient of variation. The normal distribution, $N(\mu, \sigma^2)$, represents an example of a full exponential family model. However, when the variance σ^2 is constrained to be equal to the square of the mean, μ^2, so that the coefficient of variation, the ratio of the mean to the standard deviation, is known to equal 1, the distribution represents an example of a curved exponential family. Let $Y_1, \ldots, Y_n$ be IID $N(\mu, \mu^2)$. The joint density in curved exponential family form is

$$P_Y(y; \mu) \propto \exp\left\{-\frac{1}{2\mu^2} \sum y_i^2 + \frac{1}{\mu} \sum y_i - n \log \mu\right\}.$$

Interest in curved exponential families stems from two features, related to concepts to be discussed. The maximum likelihood estimator is not a sufficient statistic, so that there is scope for conditioning on an ancillary statistic. Also, it can be shown that any sufficiently smooth parametric family can be approximated, locally to the true parameter value, to some suitable order, by a curved exponential family.

1.2.2 Transformation families

The basic idea behind a transformation family is that of a group of transformations acting on the sample space, generating a family of distributions all of the same form, but with different values of the parameters.

Recall that a group G is a mathematical structure having a binary operation $\circ$ such that

- if $g, g' \in G$, then $g \circ g' \in G$;
- if $g, g', g'' \in G$, then $(g \circ g') \circ g'' = g \circ (g' \circ g'')$;
- G contains an identity element e such that $e \circ g = g \circ e = g$, for each $g \in G$; and
- each $g \in G$ possesses an inverse $g^{-1} \in G$ such that $g \circ g^{-1} = g^{-1} \circ g = e$.

In the present context, we will be concerned with a group G of transformations acting on the sample space $\mathcal{X}$ of a random variable X, and the binary operation will simply be composition of functions: we have $(g_1 \circ g_2)(x) = g_1(g_2(x))$ and the identity element is defined by $e(x) \equiv x$.

The group elements typically correspond to elements of a parameter space Ω_θ, so that a transformation may be written as, say, g_θ. The family of densities of $g_\theta(X)$, for $g_\theta \in G$, is called a **(group) transformation family**.

Setting $x \approx x'$ iff there is a $g \in G$ such that $x = g(x')$ defines an equivalence relation, which partitions $\mathcal{X}$ into equivalence classes called *orbits*. These may be labelled by an index a, say. Two points x and x' on the same orbit have the same index, $a(x) = a(x')$. Each $x \in \mathcal{X}$ belongs to precisely one orbit, and might be represented by a (which identifies the orbit) and its position on the orbit.

1.2.3 Maximal invariant

We say that the statistic t is **invariant** to the action of the group G if its value does not depend on whether x or $g(x)$ was observed, for any $g \in G$: $t(x) = t(g(x))$. An example is the index a above.

The statistic t is **maximal invariant** if every other invariant statistic is a function of it, or equivalently, $t(x) = t(x')$ implies that $x' = g(x)$ for some $g \in G$. A maximal invariant can be thought of (Davison, 2003, Section 5.3) as a reduced version of the data that represents it as closely as possible while remaining invariant to the action of G. In some sense, it is what remains of X once minimal information about the parameter values has been extracted.

1.2.4 Equivariant statistics and a maximal invariant

As described, typically there is a one-to-one correspondence between the elements of G and the parameter space Ω_θ, and then the action of G on $\mathcal{X}$ requires that Ω_θ itself constitutes a group, with binary operation $*$ say: we must have $g_\theta \circ g_\phi = g_{\theta * \phi}$. The group action on $\mathcal{X}$ induces a group action on Ω_θ. If $\bar{G}$ denotes this induced group, then associated with each $g_\theta \in G$ there is a $\bar{g}_\theta \in \bar{G}$, satisfying $\bar{g}_\theta(\phi) = \theta * \phi$.

If t is an invariant statistic, the distribution of $T = t(X)$ is the same as that of $t(g(X))$, for all g. If, as we assume here, the elements of G are identified with parameter values, this means that the distribution of T does not depend on the parameter and is known in principle. T is said to be *distribution constant*.

A statistic $S = s(X)$ defined on $\mathcal{X}$ and taking values in the parameter space Ω_θ is said to be **equivariant** if $s(g_\theta(x)) = \bar{g}_\theta(s(x))$ for all $g_\theta \in G$ and $x \in \mathcal{X}$. Often S is chosen to be an estimator of θ, and it is then called an *equivariant estimator*.

A key operational point is that an equivariant estimator can be used to construct a maximal invariant.

Consider $t(X) = g^{-1}_{s(X)}(X)$. This is invariant, since

$$\begin{aligned} t(g_\theta(x)) &= g^{-1}_{s(g_\theta(x))}(g_\theta(x)) = g^{-1}_{\bar{g}_\theta(s(x))}(g_\theta(x)) = g^{-1}_{\theta * s(x)}(g_\theta(x)) \\ &= g^{-1}_{s(x)}\{g_\theta^{-1}(g_\theta(x))\} = g^{-1}_{s(x)}(x) = t(x). \end{aligned}$$

If $t(x) = t(x')$, then $g^{-1}_{s(x)}(x) = g^{-1}_{s(x')}(x')$, and it follows that $x' = g_{s(x')} \circ g^{-1}_{s(x)}(x)$, which shows that $t(X)$ is maximal invariant.

The statistical importance of a maximal invariant will be illuminated in Chapter 3. In a transformation family, a maximal invariant plays the role of the ancillary statistic in the conditional inference on the parameter of interest indicated by a Fisherian approach. The above direct construction of a maximal invariant from an equivariant estimator facilitates identification of an appropriate ancillary statistic in the transformation family context.

1.2.5 A simple example

Suppose $Y = \tau\epsilon$, where $\tau > 0$ and ϵ is a random variable with known density f. This scale model constitutes a transformation model. Based on an independent, identically distributed sample $Y = \{Y_1, \dots Y_n\}$, $s(Y) = (\sum Y_i^2)^{1/2}$ is equivariant and the corresponding maximal invariant statistic is $(Y_1, \dots, Y_n)/(\sum Y_i^2)^{1/2}$.

Note that a particular case of this model concerns the normal distribution with known coefficient of variation: X distributed as $N(\mu, \mu^2)$ can be represented as μZ, where Z is a random variable with teh known distribution $N(1, 1)$.

1.2.6 An example

An important example is the **location-scale model**. Let $X = \eta + \tau\epsilon$, where ϵ has a known density f, and the parameter $\theta = (\eta, \tau) \in \Omega_\theta = \mathbb{R} \times \mathbb{R}_+$. Define a group action by $g_\theta(x) = g_{(\eta,\tau)}(x) = \eta + \tau x$, so

$$g_{(\eta,\tau)} \circ g_{(\mu,\sigma)}(x) = \eta + \tau\mu + \tau\sigma x = g_{(\eta+\tau\mu,\tau\sigma)}(x).$$

The set of such transformations is closed with identity $g_{(0,1)}$. It is easy to check that $g_{(\eta,\tau)}$ has inverse $g_{(-\eta/\tau,\tau^{-1})}$. Hence, $G = \{g_{(\eta,\tau)} : (\eta, \tau) \in \mathbb{R} \times \mathbb{R}_+\}$ constitutes a group under the composition of functions operation $\circ$ defined above.

The action of $g_{(\eta,\tau)}$ on a random sample $X = (X_1, \dots, X_n)$ is $g_{(\eta,\tau)}(X) = \eta + \tau X$, with $\eta \equiv \eta 1_n$, where 1_n denotes the $n \times 1$ vector of 1's, and X is written as an $n \times 1$ vector.

The induced group action on Ω_θ is given by $\bar{g}_{(\eta,\tau)}((\mu, \sigma)) \equiv (\eta, \tau) * (\mu, \sigma) = (\eta + \tau\mu, \tau\sigma)$.

The sample mean and standard deviation are equivariant, because with $s(X) = (\bar{X}, V^{1/2})$, where $V = (n-1)^{-1}\sum(X_j - \bar{X})^2$, we have

$$\begin{aligned} s(g_{(\eta,\tau)}(X)) &= \left(\overline{\eta + \tau X}, \left\{(n-1)^{-1}\sum(\eta + \tau X_j - \overline{(\eta + \tau X)})^2\right\}^{1/2}\right) \\ &= \left(\eta + \tau\bar{X}, \left\{(n-1)^{-1}\sum(\eta + \tau X_j - \eta - \tau\bar{X})^2\right\}^{1/2}\right) \\ &= \left(\eta + \tau\bar{X}, \tau V^{1/2}\right) \\ &= \bar{g}_{(\eta,\tau)}(s(X)). \end{aligned}$$

A maximal invariant is $A = g^{-1}_{s(X)}(X)$, and the parameter corresponding to $g^{-1}_{s(X)}$ is $(-\bar{X}/V^{1/2}, V^{-1/2})$. Hence a maximal invariant is the vector of residuals

$$A = (X - \bar{X})/V^{1/2} = \left(\frac{X_1 - \bar{X}}{V^{1/2}}, \ldots, \frac{X_n - \bar{X}}{V^{1/2}}\right)^T,$$

called the *configuration*. It is easily checked directly that the distribution of A does not depend on θ. Any function of A is also invariant. The orbits are determined by different values a of the statistic A, and X has a unique representation as $X = g_{s(X)}(A) = \bar{X} + V^{1/2}A$.

1.3 Likelihood

1.3.1 Definitions

We have a parametric model, involving a model function $f_Y(y;\theta)$ for a random variable Y and parameter $\theta \in \Omega_\theta$. The likelihood function is

$$L_Y(\theta; y) = L(\theta; y) = L(\theta) = f_Y(y;\theta).$$

Usually we work with the log-likelihood

$$l_Y(\theta; y) = l(\theta; y) = l(\theta) = \log f_Y(y;\theta),$$

sometimes studied as a random variable

$$l_Y(\theta; Y) = l(\theta; Y) = \log f_Y(Y;\theta).$$

In likelihood calculations, we can drop factors depending on y only, or additive terms depending only on y may be dropped from log-likelihoods. This idea can be formalised by working with the normed likelihood $\bar{L}(\theta) =$

$L(\theta)/L(\hat{\theta})$, where $\hat{\theta}$ is the value of θ maximising $L(\theta)$. We define the score function by

$$\begin{aligned} u_r(\theta; y) &= \frac{\partial l(\theta; y)}{\partial \theta^r} \\ u_Y(\theta; y) &= u(\theta; y) = \nabla_\theta l(\theta; y), \end{aligned}$$

where $\nabla_\theta = (\partial/\partial\theta^1, \ldots, \partial/\partial\theta^d)^T$.

To study the score function as a random variable (the 'score statistic') we write

$$u_Y(\theta; Y) = u(\theta; Y) = U(\theta) = U.$$

These definitions are expressed in terms of arbitrary random variables Y. Often the components Y_j are mutually independent, in which case both the log-likelihood and the score function are sums of contributions:

$$l(\theta; y) = \sum_{j=1}^{n} l(\theta; y_j),$$

$$u(\theta; y) = \sum_{j=1}^{n} \nabla_\theta l(\theta; y_j) = \sum_{j=1}^{n} u(\theta; y_j),$$

say, and where $l(\theta; y_j)$ is found from the density of Y_j.

Quite generally, even for dependent random variables, if $Y_{(j)} = (Y_1, \ldots, Y_j)$, we may write

$$l(\theta; y) = \sum_{j=1}^{n} l_{Y_j | Y_{(j-1)}}(\theta; y_j \mid y_{(j-1)}),$$

each term being computed from the conditional density given all the previous values in the sequence.

1.3.2 Score function and information

For regular problems for which the order of differentiation with respect to θ and integration over the sample space can be reversed, we have

$$E\{U(\theta); \theta\} = 0. \tag{1.2}$$

To verify this, note that a component of the left-hand side is

$$\begin{aligned} & \int \left\{ \frac{\partial \log f_Y(y; \theta)}{\partial \theta^r} \right\} f_Y(y; \theta) dy \\ = & \int \frac{\partial f_Y(y; \theta)}{\partial \theta^r} \, dy \\ = & \frac{\partial}{\partial \theta^r} \int f_Y(y; \theta) dy = \frac{\partial}{\partial \theta^r} 1 = 0. \end{aligned}$$

Also, when (1.2) holds,

$$\begin{aligned}\operatorname{cov}\{U_r(\theta), U_s(\theta); \theta\} &= E\left\{\frac{\partial l(\theta;Y)}{\partial \theta^r}\frac{\partial l(\theta;Y)}{\partial \theta^s}; \theta\right\} \\ &= E\left\{-\frac{\partial^2 l(\theta;Y)}{\partial \theta^r \partial \theta^s}; \theta\right\}.\end{aligned}$$

More compactly, the covariance matrix of U is

$$\operatorname{cov}\{U(\theta); \theta\} = E\{-\nabla\nabla^T l; \theta\}.$$

This matrix is called the expected information matrix for θ, or sometimes the Fisher information matrix, and will be denoted by $i(\theta)$. The Hessian $-\nabla\nabla^T l$ is called the observed information matrix, and is denoted by $j(\theta)$. Note that $i(\theta) = E\{j(\theta)\}$.

In the (m, m) exponential family (1.1),

$$U(\phi) = \nabla l = S - \nabla K(\phi)$$

and $\nabla\nabla^T l = -\nabla\nabla^T K(\phi)$.

Note that the score $u(\theta; y)$ and the information $i(\theta)$ depend not only on the value of the parameter θ, but also on the parameterisation. If we change from θ to ψ by a smooth one-to-one transformation and calculate the score and information in terms of ψ, then different values will be obtained.

Write $(U^{(\theta)}, i^{(\theta)})$ and $(U^{(\psi)}, i^{(\psi)})$ for quantities in the θ- and ψ-parameterisation respectively. Using the summation convention whereby summation is understood to take place over the range of an index that appears two or more times in an expression, the chain rule for differentiation gives

$$\begin{aligned}U_a^{(\psi)}(\psi; Y) &= \frac{\partial l\{\theta(\psi); Y\}}{\partial \psi^a} \\ &= U_r^{(\theta)}(\theta; Y)\frac{\partial \theta^r}{\partial \psi^a},\end{aligned}$$

or

$$U^{(\psi)}(\psi; Y) = \left[\frac{\partial \theta}{\partial \psi}\right]^T U^{(\theta)}(\theta; Y),$$

where $\partial\theta/\partial\psi$ is the Jacobian of the transformation from θ to ψ, with (r, a) element $\partial\theta^r/\partial\psi^a$.

Similarly,

$$i^{(\psi)}_{ab}(\psi) = \frac{\partial \theta^r}{\partial \psi^a} \frac{\partial \theta^s}{\partial \psi^b} i^{(\theta)}_{rs}(\theta),$$

or

$$i^{(\psi)}(\psi) = \left[\frac{\partial \theta}{\partial \psi}\right]^T i^{(\theta)}(\theta) \left[\frac{\partial \theta}{\partial \psi}\right].$$

The notion of parameterisation invariance is a valuable basis for choosing between different inferential procedures. Invariance requires that the conclusions of a statistical analysis be unchanged by reformulation in terms of ψ, any reasonably smooth one-to-one function of θ.

Consider, for example, the exponential distribution with density $\rho e^{-\rho y}$. It would for many purposes be reasonable to reformulate in terms of the mean $1/\rho$ or, say, $\log \rho$. Parameterisation invariance would require, for example, the same conclusions about ρ to be reached by: (i) direct formulation in terms of ρ, application of a method of analysis, say estimating ρ; (ii) formulation in terms of $1/\rho$, application of a method of analysis, estimating $1/\rho$, then taking the reciprocal of this estimate.

Invariance under reparameterisation can usefully be formulated much more generally. Suppose that $\theta = (\psi, \chi)$, with ψ the parameter of interest and χ a nuisance parameter. It is reasonable to consider one-to-one transformations from θ to $\tilde{\theta} = (\tilde{\psi}, \tilde{\chi})$, where $\tilde{\psi}$ is a one-to-one function of ψ and $\tilde{\chi}$ is a function of both ψ and χ. Such transformations are called interest-respecting reparameterisations.

1.3.3 Pseudo-likelihoods

Typically we consider a model parameterised by a parameter θ which may be written as $\theta = (\psi, \lambda)$, where ψ is the parameter of interest and λ is a nuisance parameter. In order to draw inferences about the parameter of interest, we must deal with the nuisance parameter.

Ideally, we would like to construct a likelihood function for ψ alone. The simplest method for doing so is to construct a likelihood function based on a statistic T such that the distribution of T depends only on ψ. In this case, we may form a genuine likelihood function for ψ based on the density function of T; this is called a **marginal likelihood**, since it is based on the marginal distribution of T.

Another approach is available whenever there exists a statistic S such that the conditional distribution of the data X given $S = s$ depends only on ψ. In this case, we may form a likelihood function for ψ based on the conditional

density function of X given $S = s$; this is called a **conditional likelihood** function. The drawback of this approach is that we discard the part of the likelihood function based on the marginal distribution of S, which may contain information about ψ.

Conditional and marginal likelihoods are particular instances of **pseudo-likelihood functions**. The term pseudo-likelihood is used to indicate any function of the data which depends only on the parameter of interest and which behaves, in some respects, as if it were a genuine likelihood (so that the score has zero null expectation, the maximum likelihood estimator has an asymptotic normal distribution etc.).

Formally, suppose that there exists a statistic T such that the density of the data X may be written as

$$f_X(x; \psi, \lambda) = f_T(t; \psi) f_{X|T}(x|t; \psi, \lambda).$$

Inference can be based on the marginal distribution of T which does not depend on λ. The marginal likelihood function based on t is given by

$$L(\psi; t) = f_T(t; \psi).$$

The drawback of this approach is that we lose the information about ψ contained in the conditional density of X given T. It may, of course, also be difficult to find such a statistic T.

To define formally a conditional log-likelihood, suppose that there exists a statistic S such that

$$f_X(x; \psi, \lambda) = f_{X|S}(x|s; \psi) f_S(s; \psi, \lambda).$$

The statistic S is sufficient (see Section 1.4) in the model with ψ held fixed. A conditional likelihood function for ψ may be based on $f_{X|S}(x|s; \psi)$, which does not depend on λ. The conditional log-likelihood function may be calculated as

$$l(\psi; x \mid s) = l(\theta) - l(\theta; s),$$

where $l(\theta; s)$ denotes the log-likelihood function based on the marginal distribution of S and $l(\theta)$ is the log-likelihood based on the full data X. Note that we make two assumptions here about S. The first is that S is not sufficient in the general model with parameters (ψ, λ), for if it was, the conditional likelihood would not depend on either ψ or λ. The other is that S, the sufficient statistic when ψ is fixed, is the same for all ψ; S does not depend on ψ.

Note that factorisations of the kind that we have assumed in the definitions of conditional and marginal likelihoods arise essentially only in exponential families and transformation families. Outside these cases more general notions of pseudo-likelihood must be found.

1.4 Sufficiency

1.4.1 Definitions

Let the data y correspond to a random variable Y with density $f_Y(y;\theta), \theta \in \Omega_\theta$. Let $s(y)$ be a statistic such that if $S \equiv s(Y)$ denotes the corresponding random variable, then the conditional density of Y given $S = s$ does not depend on θ, for all s, so that

$$f_{Y|S}(y \mid s;\theta) = g(y,s) \tag{1.3}$$

for all $\theta \in \Omega_\theta$. Then S is said to be sufficient for θ.

The definition (1.3) does not define S uniquely. We usually take the minimal S for which (1.3) holds, the minimal sufficient statistic. S is minimal sufficent if it is sufficient and is a function of every other sufficient statistic.

The determination of S from the definition (1.3) is often difficult. Instead we use the factorisation theorem: a necessary and sufficient condition that S is sufficient for θ is that for all y, θ

$$f_Y(y;\theta) = g(s,\theta)h(y),$$

for some functions g and h. Without loss of generality, $g(s,\theta)$ may be taken as the unconditional density of S for given θ.

The following result is easily proved and useful for identifying minimal sufficient statistics. A statistic T is minimal sufficient iff

$$T(x) = T(y) \Leftrightarrow \frac{L(\theta_1;x)}{L(\theta_2;x)} = \frac{L(\theta_1;y)}{L(\theta_2;y)}, \quad \forall \theta_1, \theta_2 \in \Omega_\theta.$$

Example: normal distribution, known coefficient of variation. Let $Y_1, \ldots, Y_n$ be IID $N(\mu, \mu^2)$. It is easily seen from the form of the joint density that a minimal sufficient statistic is $(\sum Y_i, \sum Y_i^2)$.

1.4.2 Further Examples

Exponential models Here the natural statistic S is a (minimal) sufficient statistic. In a curved (m,d) exponential model the dimension m of the sufficient statistic exceeds that of the parameter.

Transformation models Except in special cases, such as the normal distribution, where the model is also an exponential family model, there is no reduction of dimensionality by sufficiency: the minimal sufficient statistic has the same dimension as the data vector $Y = (Y_1, \ldots, Y_n)$.

1.5 Conditioning

In connection with methods of statistical inference, probability is used in two quite distinct ways. The first is to define the stochastic model assumed to have generated the data. The second is to assess uncertainty in conclusions, via significance levels, confidence regions, posterior distributions etc. We enquire how a given method would perform if, hypothetically, it were used repeatedly on data derived from the model under study. The probabilities used for the basis of inference are long-run frequencies under hypothetical repetition. The issue arises of how these long-run frequencies are to be made relevant to the data under study. The answer lies in conditioning the calculations so that the long run matches the particular set of data in important respects.

1.5.1 The Bayesian stance

In a Bayesian approach the issue of conditioning is dealt with automatically. Recall that the key idea of Bayesian inference is that it is supposed that the particular value of θ is the realised value of a random variable Θ, generated by a random mechanism giving a known density $\pi_\Theta(\theta)$ for Θ, the prior density. Then Bayes' Theorem gives the posterior density

$$\pi_{\Theta|Y}(\theta \mid Y = y) \propto \pi_\Theta(\theta) f_{Y|\Theta}(y \mid \Theta = \theta),$$

where now the model function $f_Y(y; \theta)$ is written as a conditional density $f_{Y|\Theta}(y \mid \Theta = \theta)$. The insertion of a random element in the generation of θ allows us to condition on the whole of the data y: relevance to the data is certainly accomplished. This approach is uncontroversial if a meaningful prior can be agreed. In many applications, there may be major obstacles to specification of a meaningful prior and we are forced to adopt a less direct route to conditioning.

1.5.2 The Fisherian stance

Suppose first that the whole parameter vector θ is of interest. Reduce the problem by sufficiency. If, with parameter dimension $d = 1$, there is a one-dimensional sufficient statistic, we have reduced the problem to that of one

observation from a distribution with one unknown parameter and there is little choice but to use probabilities calculated from that distribution. The same notion occurs if there is a d-dimensional θ of interest and a d-dimensional sufficient statistic. If the dimension of the (minimal) sufficient statistic exceeds that of the parameter, there is scope and need for ensuring relevance to the data under analysis by conditioning.

We therefore aim to

1. partition the minimal sufficient statistic s in the form $s = (t, a)$, so that $\dim(t) = \dim(\theta)$ and A has a distribution not involving θ;
2. use for inference the conditional distribution of T given $A = a$.

Conditioning on $A = a$ makes the distribution used for inference involve (hypothetical) repetitions like the data in some respects.

In the next section we extend this discussion to the case where there are nuisance parameters.

1.5.3 An example

Consider again the normal distribution of known coefficient of variation: $Y_1, \ldots, Y_n$ are IID $N(\mu, \mu^2)$. With $T_1 = \bar{Y}$ and $T_2 = \sqrt{n^{-1}\sum_{i=1}^n Y_i^2}$, we know that (T_1, T_2) is minimal sufficient. The transformation structure enables us to see easily that $Z = T_1/T_2$ is ancillary. The minimal sufficient statistic may be equivalently expressed as (V, Z), where $V = \sqrt{n}T_2$. Since Z is ancillary, we should base inference on the conditional distribution of V given Z, which is straightforward, but fiddly, to obtain.

1.5.4 A more complicated example

Suppose that $Y_1, \ldots, Y_n$ are independent and identically uniformly distributed on $(\theta-1, \theta+1)$. The (minimal) sufficient statistic is the pair of order statistics $(Y_{(1)}, Y_{(n)})$, where $Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$ and $Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$. Suppose we make a (one-to-one) transformation to the mid-range $\bar{Y} = \frac{1}{2}(Y_{(1)} + Y_{(n)})$ and the range $R = Y_{(n)} - Y_{(1)}$. The sufficient statistic may equivalently be expressed as $(\bar{Y}, R)$. A direct calculation shows that R has a distribution not depending on θ, so we have the situation where the dimension of the sufficient statistic exceeds the dimension of θ and the statistic R, being distribution constant, plays the role of A. Inference should be based on the conditional distribution of $\bar{Y}$, given $R = r$, which it is easily checked to be uniform over $(\theta - 1 + \frac{1}{2}r, \theta + 1 - \frac{1}{2}r)$.

1.6 Ancillarity and the Conditionality Principle

A component a of the minimal sufficient statistic such that the random variable A is distribution constant is said to be ancillary, or sometimes ancillary in the simple sense.

The Conditionality Principle says that inference about a parameter of interest θ is to be made conditional on $A = a$ i.e. on the basis of the conditional distribution of Y given $A = a$, rather than from the model function $f_Y(y; \theta)$.

An important convention should be flagged here. Later, specifically in Chapter 3, we will use the term ancillary to mean a distribution constant statistic which, together with the maximum likelihood estimator, constitutes a sufficient statistic.

The Conditionality Principle is discussed most frequently in the context of transformation models, where the maximal invariant is ancillary.

1.6.1 Nuisance parameters

In our previous discussion, the argument for conditioning on $A = a$ rests not so much on the distribution of A being known as on its being totally uninformative about the parameter of interest.

Suppose, more generally, that we can write $\theta = (\psi, \chi)$, where ψ is of interest. Suppose that

1. $\Omega_\theta = \Omega_\psi \times \Omega_\chi$, so that ψ and χ are variation independent;
2. the minimal sufficient statistic $s = (t, a)$;
3. the distribution of T given $A = a$ depends only on ψ;
4. one or more of the following conditions holds:
   (a) the distribution of A depends only on χ and not on ψ;
   (b) the distribution of A depends on (ψ, χ) in such a way that from observation of A alone no information is available about ψ;

Then the extension of the Fisherian stance of Section 1.5.2 argues that inference about ψ should be based upon the conditional distribution of T given $A = a$, and we would still speak of A as being ancillary. The most straightforward extension corresponds to (a). In this case A is said to be a cut and to be S-ancillary for ψ and S-sufficient for χ. The arguments for conditioning on $A = a$ when ψ is the parameter of interest are as compelling as in the

case where A has a fixed distribution. Condition (b) is more problematical to qualify. See the discussion in Barndorff-Nielsen and Cox (1994, pp.38–41) for detail and examples. The same authors discuss problems associated with existence and non-uniqueness of ancillary statistics.

1.7 Sample space derivatives

The log-likelihood is, except possibly for a term not depending on the parameter, a function of a sufficient statistic s and parameter θ. If the dimensions of s and θ are equal, the maximum likelihood estimator $\hat{\theta}$ is usually a one-to-one function of s and then $\hat{\theta}$ is minimal sufficient if and only if s is minimal sufficient. We can then take the log-likelihood as $l(\theta;\hat{\theta})$, it being the same as if the data consisted solely of $\hat{\theta}$ or s. If $s = (t, a)$ where t has the dimension of θ and a is ancillary, then we can generally write the log-likelihood as $l(\theta;\hat{\theta}, a)$.

Similarly, the observed information can, in the scalar parameter case, be written as

$$j(\theta;\hat{\theta}, a) = -\partial^2 l(\theta;\hat{\theta}, a)/\partial\theta^2.$$

In practice, θ being unknown, this would be evaluated at $\theta = \hat{\theta}$, as $j(\hat{\theta};\hat{\theta}, a)$.

For a vector parameter we use $-\nabla_\theta \nabla_\theta^T l(\theta;\hat{\theta}, a)$.

An alternative expression for the observed information uses the notion of 'sample space derivatives', obtained by differentiating $l(\theta;\hat{\theta}, a)$ with respect to $\hat{\theta}$.

The maximum likelihood equation is

$$\frac{\partial l(\theta;\hat{\theta}, a)}{\partial\theta}\,|_{\theta=\hat{\theta}} = 0,$$

so that

$$\frac{\partial l(t; t, a)}{\partial\theta} = 0,$$

identically in t. Differentiating this with respect to t, and evaluating at $t = \hat{\theta}$ we have

$$\left[\frac{\partial^2 l(\theta;\hat{\theta}, a)}{\partial\theta^2} + \frac{\partial^2 l(\theta;\hat{\theta}, a)}{\partial\theta\partial\hat{\theta}}\right]_{\theta=\hat{\theta}} = 0,$$

so that

$$j(\hat{\theta};\hat{\theta}, a) = \left[\frac{\partial^2 l(\theta;\hat{\theta}, a)}{\partial\theta\partial\hat{\theta}}\right]_{\theta=\hat{\theta}}$$

or, for a vector parameter,

$$j(\hat{\theta};\hat{\theta},a) = [\nabla_\theta \nabla_{\hat{\theta}}^T l(\theta;\hat{\theta},a)]_{\theta=\hat{\theta}}.$$

1.8 Parameter Orthogonality

We work now with a multi-dimensional parameter θ. There are a number of advantages, which we will study later, if the matrix $i(\theta) \equiv [i_{rs}(\theta)]$ is diagonal.

1.8.1 Definition

Suppose that θ is partitioned into components $\theta = (\theta^1, \ldots, \theta^{d_1}; \theta^{d_1+1}, \ldots, \theta^d) = (\theta_{(1)}, \theta_{(2)})$ say, such that $i_{rs}(\theta) = 0$ for all $r = 1, \ldots, d_1$; $s = d_1 + 1, \ldots, d$, for all $\theta \in \Omega_\theta$. The matrix $i(\theta)$ is block diagonal and we say that $\theta_{(1)}$ is orthogonal to $\theta_{(2)}$.

1.8.2 An immediate consequence

Orthogonality implies that the corresponding components of the score statistic are uncorrelated.

1.8.3 The case $d_1 = 1$

For this case, write $\theta = (\psi, \lambda^1, \ldots, \lambda^q)$, with $q = d - 1$. If we start with an arbitrary parameterisation $(\psi, \chi^1, \ldots, \chi^q)$ with ψ given, it is always possible to find $\lambda^1, \ldots, \lambda^q$ as functions of $(\psi, \chi^1, \ldots, \chi^q)$ such that ψ is orthogonal to $(\lambda^1, \ldots, \lambda^q)$.

Let l^* and i^* be the log-likelihood and information matrix in terms of $(\psi, \chi^1, \ldots, \chi^q)$ and write $\chi^r = \chi^r(\psi, \lambda^1, \ldots, \lambda^q)$. Then

$$l(\psi, \lambda) \equiv l^*\{\psi, \chi^1(\psi, \lambda), \ldots, \chi^q(\psi, \lambda)\}$$

and use of the chain rule for differentiation gives

$$\begin{aligned}\frac{\partial^2 l}{\partial\psi\partial\lambda^r} &= \frac{\partial^2 l^*}{\partial\psi\partial\chi^s}\frac{\partial\chi^s}{\partial\lambda^r} + \frac{\partial^2 l^*}{\partial\chi^t\partial\chi^s}\frac{\partial\chi^s}{\partial\lambda^r}\frac{\partial\chi^t}{\partial\psi} \\ &\quad + \frac{\partial l^*}{\partial\chi^s}\frac{\partial^2\chi^s}{\partial\psi\partial\lambda^r},\end{aligned}$$

where we have used the summation convention over the range $1, \ldots, q$. Now take expectations.

The final term vanishes and orthogonality of ψ and λ then requires

$$\frac{\partial \chi^s}{\partial \lambda^t}\left(i^*_{\psi s} + i^*_{rs}\frac{\partial \chi^r}{\partial \psi}\right) = 0.$$

Assuming that the Jacobian of the transformation from (ψ, χ) to (ψ, λ) is non-zero, this is equivalent to

$$i^*_{rs}\frac{\partial \chi^r}{\partial \psi} + i^*_{\psi s} = 0. \tag{1.4}$$

These partial differential equations determine the dependence of λ on ψ and χ, and are solvable in general. However, the dependence is not determined uniquely and there remains considerable arbitrariness in the choice of λ.

1.8.4 An example

Let (Y_1, Y_2) be independent, exponentially distributed with means $(\chi, \psi\chi)$. Then equation (1.4) becomes

$$2\chi^{-2}\frac{\partial \chi}{\partial \psi} = -(\psi\chi)^{-1},$$

the solution of which is $\chi\psi^{1/2} = g(\lambda)$, where $g(\lambda)$ is an arbitrary function of λ. A convenient choice is $g(\lambda) \equiv \lambda$, so that in the orthogonal parameterisation the means are $\lambda/\psi^{1/2}$ and $\lambda\psi^{1/2}$.

1.8.5 The case $d_1 > 1$

When $\dim(\psi) > 1$ there is no guarantee that a λ may be found so that ψ and λ are orthogonal.

If, for example, there were two components ψ^1 and ψ^2 for which it was required to satisfy (1.4), there would in general be no guarantee that the values of $\partial\chi^r/\partial\psi^1$ and $\partial\chi^r/\partial\psi^2$ so obtained would satisfy the compatibility condition

$$\frac{\partial^2 \chi^r}{\partial\psi^1\partial\psi^2} = \frac{\partial^2 \chi^r}{\partial\psi^2\partial\psi^1}.$$

1.8.6 Further remarks

Irrespective of the dimension of ψ, orthogonality can be achieved locally at $\theta = \theta_0$ via a linear transformation of parameters with components depending on $i(\theta_0)$. More generally, for a fixed value ψ_0 of ψ it is possible to determine λ so that $i_{\psi\lambda}(\psi_0, \lambda) = 0$ identically in λ.

If λ is orthogonal to ψ, then any one-to-one smooth function of ψ is orthogonal to any one-to-one smooth function of λ.

1.9 General principles

The previous sections have introduced a number of fundamental concepts of statistical inference. In this section we outline the role played by these concepts in various abstract principles of inference. These principles are included here largely for the sake of interest. The formal role that they play in different approaches to statistical inference is sketched in Section 1.10 : further discussion is given by Cox and Hinkley (1974, pp.48–56).

1.9.1 Sufficiency principle

Suppose that we have a model according to which the data y correspond to a random variable Y having p.d.f. $f_Y(y;\theta)$ and that S is minimal sufficient for θ. Then, according to the sufficiency principle, so long as we accept the adequacy of the model, identical conclusions should be drawn from data y_1 and y_2 with the same value of S.

1.9.2 Conditionality principle

Suppose that C is an ancillary statistic, either in the simple sense described at the beginning of Section 1.6, or the extended sense of Section 1.6.1 where nuisance parameters are present. Then the conditionality principle is that the conclusion about the parameter of interest is to be drawn as if C were fixed at its observed value c.

1.9.3 Weak likelihood principle

The weak likelihood principle is that two observations with proportional likelihood functions lead to identical conclusions, so if y_1 and y_2 are such that for all θ

$$f_Y(y_1;\theta) = h(y_1, y_2) f_Y(y_2;\theta),$$

then y_1 and y_2 should lead to identical conclusions, as long as we accept the adequacy of the model.

This is identical with the sufficiency principle.

1.9.4 Strong likelihood principle

Suppose that two different random systems are contemplated, the first giving observations y corresponding to a random variable Y and the second giving observations z on a random variable Z, the corresponding p.d.f.'s being $f_Y(y;\theta)$ and $f_Z(z;\theta)$, with the same parameter θ and the same parameter space Ω_θ. The strong likelihood principle is that if y and z give proportional

likelihood functions, the conclusions drawn from y and z should be identical, assuming adequacy of both models. If, for all $\theta \in \Omega_\theta$,

$$f_Y(y;\theta) = h(y,z) f_Z(z;\theta),$$

identical conclusions about θ should be drawn from y and z.

A simple example concerning Bernoulli trials illustrates this. The log likelihood function corresponding to r successes in n trials is essentially the same whether (i) only the number of successes in a prespecified number of trials is recorded or (ii) only the number of trials necessary to achieve a prespecified number of successes is recorded, or (iii) whether the detailed results of individual trials are recorded, with an arbitrary data-dependent stopping rule. A further example is given in Section 2.7.

The strong likelihood principle may be deduced from the sufficiency principle plus some form of conditionality principle. Bayesian methods of inference satisfy the strong likelihood principle. Nearly all others do not.

1.9.5 Repeated sampling principle

This principle, like that in Section 1.9.6, is concerned with interpretation of conclusions, rather than what aspects of the data and model are relevant. According to the repeated sampling principle, inference procedures should be interpreted and evaluated in terms of their behaviour in hypothetical repetitions under the same conditions. Measures of uncertainty are to be interpreted as hypothetical frequencies in long run repetitions and criteria of optimality are to be formulated in terms of sensitive behaviour in hypothetical repetitions.

1.9.6 Bayesian coherency principle

In the Bayesian approach to inference, all uncertainties are described by probabilities, so that unknown parameters have probabilities both before the data are available and after the data have been obtained. It is justified by the supposition that:

(a) any individual has an attitude to every uncertain event which can be measured by a probability, called a subjective probability;

(b) all such probabilities for any one individual are comparable;

(c) these subjective probabilities can be measured by choice in hypothetical betting games.

The Bayesian coherency principle is that subjective probabilities should be such as to ensure self-consistent betting behaviour. This implies that subjective probabilities for any one individual should be manipulated by the ordinary laws of probability, in particular Bayes' Theorem. The principle implies that conclusions about unknown parameters in models have to be in the form of probability statements. This implies all the principles of 1.9.1–1.9.4, in particular the strong likelihood principle.

1.9.7 Principle of coherent decision making

In problems where an explicit decision is involved, parallel arguments to Section 1.9.6 show that for any individual each decision and true parameter value have an associated 'utility' such that the optimum decision is found by maximising expected utility.

1.10 Approaches to Statistical Inference

We have set out four principles (sufficiency, conditionality, weak likelihood, strong likelihood) which concern the way in which the data should affect the conclusions. They do not concern the exact form and interpretation of the conclusions. Interpretation is governed by the other principles. We are then in a position to describe briefly the main approaches to inference.

There are four broad approaches to statistical inference, via sampling theory, likelihood theory, Bayesian theory and decision theory.

1.10.1 Sampling theory

In this approach primary emphasis is placed on the repeated sampling principle, on ensuring that procedures have an interpretation in terms of frequencies in hypothetical repetitions under the same conditions. An example is construction of a confidence interval for the mean μ of a normal distribution. This approach does not satisfy the strong likelihood principle.

1.10.2 Likelihood theory

In this approach the likelihood function itself is used directly as a summary of information. In particular, ratios of likelihoods or differences in log-likelihoods give the relative plausibilities of two parameter values, say θ_1 and θ_2. This approach clearly satisfies the weak and strong likelihood principles, and the conditionality principle is implicitly satisfied.

1.10.3 Bayesian theory

This approach was sketched in Section 1.5.1. Inference about the parameter of interest θ is derived from the posterior density. If the prior distribution arises from a physical random mechanism with known properties, the posterior distribution can be regarded as a hypothetical frequency distribution, and the principles 1.9.1–1.9.4 are all satisfied. To apply the Bayesian approach more generally, we may invoke the Bayesian coherency principle. Then the prior is taken as measuring the investigator's subjective opinion about the parameter from evidence other than the data under analysis.

1.10.4 Decision theory

This approach emphasises the action to be taken in the light of data. If for each parameter value the consequences of each possible action can be measured by a utility (or loss), then we can evaluate the expected utility of the possible methods of action. We can then rule out certain methods of action on the grounds that they lead to uniformly lower expected utility for all parameter values. A unique optimal action will be defined if a prior distribution is available, in which case the expected utility, averaged with respect to the prior distribution, can be maximised over the set of possible actions. The principle of coherent decision making is explicitly applicable.

1.11 Some Essential Mathematical Material

1.11.1 Background

Consider a random vector Y with a known distribution, and suppose that the distribution of the statistic $f(Y)$ is needed, for some real-valued function f. In most situations, finding the exact distribution of $f(Y)$ is impossible or impractical. The approach then is to use an asymptotic approximation to the distribution of the statistic, which then allows us to approximate distributional quantities of interest, such as quantiles or moments. Much of the module (Chapter 3 in particular) is concerned with methods for obtaining such approximations. An attractive feature of the approximations is that they take just a few basic and general forms, and therefore provide a quite general distribution theory. The current section revises the key notions of probability theory that are essential to an understanding of the nature and properties of these approximations.

1.11.2 Some probability results

A sequence of (scalar) random variables $\{Y_1, Y_2, \ldots\}$ is said to converge in distribution if there exists a distribution function F such that

$$\lim_{n\to\infty} P(Y_n \leq y) = F(y)$$

for all y that are continuity points of the limiting distribution F. If F is the distribution function of the random variable Y, we write $Y_n \xrightarrow{d} Y$.

The extension to random vectors is immediate. Let $\{Y_1, Y_2, \ldots\}$ be a sequence of random vectors, each of dimension d, and let Y denote a random vector of dimension d. For each $n = 1, 2, \ldots$, let F_n denote the distribution function of Y_n, and let F denote the distribution function of Y. Then the sequence Y_n converges in distribution to Y as $n \to \infty$ if

$$\lim_{n\to\infty} F_n(y) = F(y),$$

for all $y \in \mathbb{R}^d$ at which F is continuous.

A sequence of (scalar) random variables $\{Y_1, Y_2, \ldots\}$ is said to converge in probability to a random variable Y if, for any $\epsilon > 0$

$$\lim_{n\to\infty} P(|Y_n - Y| > \epsilon) = 0.$$

We write $Y_n \xrightarrow{p} Y$. [Note that for this to make sense, for each n, Y and Y_n must be defined on the same sample space, a requirement that does not arise in the definition of convergence in distribution.] The extension to $d-$dimensional random vectors is again immediate: the sequence of random vectors Y_n converges in probability to Y if, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(\| Y_n - Y \| > \epsilon) = 0,$$

where $\| \cdot \|$ denotes Euclidean distance on $\mathbb{R}^d$.

An important relationship is that convergence in probability implies convergence in distribution. An important special case is where the sequence converges in probability to a constant, c, $Y_n \xrightarrow{p} Y$, where $P(Y = c) = 1$. Then convergence in probability is equivalent to convergence in distribution.

A stronger yet mode of convergence is almost sure convergence. A sequence of random vectors $\{Y_1, Y_2, \ldots\}$ is said to converge almost surely to Y if

$$P(\lim_{n\to\infty} \| Y_n - Y \| = 0) = 1.$$

We write $Y_n \xrightarrow{a.s} Y$.

Finally, a sequence of random vectors $\{Y_1, Y_2, \ldots\}$ is said to converge to Y in L_p (or p-th moment) if

$$\lim_{n\to\infty} E(\| Y_n - Y \|^p) = 0,$$

where $p > 0$ is a fixed constant. We write $Y_n \xrightarrow{L_p} Y$.

A very useful result is Slutsky's Theorem which states that if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a finite constant, then: (i) $X_n + Y_n \xrightarrow{d} X + c$, (ii) $X_n Y_n \xrightarrow{d} cX$, (iii) $X_n/Y_n \xrightarrow{d} X/c$, if $c \neq 0$.

Let $X_1, \ldots, X_n$ be independent, identically distributed (scalar) random variables with finite mean μ. The strong law of large numbers (SLLN) says that the sequence of random variables $\bar{X}_n = n^{-1}(X_1 + \cdots + X_n)$ converges almost surely to μ if and only if the expectation of $|X_i|$ is finite. The weak law of large numbers (WLLN) says that if the X_i have finite variance, $\bar{X}_n \xrightarrow{p} \mu$. The central limit theorem (CLT) says that, under the condition that the X_i are of finite variance σ^2, then a suitably standardised version of $\bar{X}_n$, $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$, converges in distribution to a random variable Z having the standard normal distribution $N(0, 1)$. We write $Z_n \xrightarrow{d} N(0, 1)$.

Another useful result is the 'delta-method': if Y_n has a limiting normal distribution, then so does $g(Y_n)$, where g is any smooth function. Specifically, if $\sqrt{n}(Y_n - \mu)/\sigma \xrightarrow{d} N(0, 1)$, and g is a differentiable function such that $g'(\mu) \neq 0$, then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{d} N(0, 1).$$

1.11.3 Mann-Wald notation

In asymptotic theory, the so-called Mann-Wald notation is useful, to describe the order of magnitude of specified quantities. For two sequences of positive constants $(a_n), (b_n)$, we write $a_n = o(b_n)$ when $\lim_{n\to\infty}(a_n/b_n) = 0$, and $a_n = O(b_n)$ when $\limsup_{n\to\infty}(a_n/b_n) = K < \infty$. For sequences of random variables $\{Y_n\}$, we write $Y_n = o_p(a_n)$ if $Y_n/a_n \xrightarrow{p} 0$ as $n \to \infty$ and $Y_n = O_p(a_n)$ when Y_n/a_n is bounded in probability as $n \to \infty$, i.e. given $\epsilon > 0$ there exist $k > 0$ and n_0 such that, for all $n > n_0$,

$$\Pr(|Y_n/a_n| < k) > 1 - \epsilon.$$

In particular, $Y_n = c + o_p(1)$ means that $Y_n \xrightarrow{p} c$.

1.11.4 Moments and cumulants

The moment generating function of a scalar random variable X is defined by $M_X(t) = \mathbb{E}\{\exp(tX)\}$, whenever this expectation exists. Note that $M_X(0) = 1$, and that the moment generating function is defined in some interval containing 0. If $M_X(t)$ exists for t in an open interval around 0, then all the moments $\mu'_r = \mathbb{E}X^r$ exist, and we have the Taylor expansion

$$M_X(t) = 1 + \mu'_1 t + \mu'_2 \frac{t^2}{2!} + \cdots + \mu'_r \frac{t^r}{r!} + O(t^{r+1}),$$

as $t \to 0$.

The cumulant generating function $K_X(t)$ is defined by $K_X(t) = \log\{M_X(t)\}$, defined on the same interval as $M_X(t)$. Provided $M_X(t)$ exists in an open interval around 0, the Taylor series expansion

$$K_X(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \cdots + \kappa_r \frac{t^r}{r!} + O(t^{r+1}),$$

as $t \to 0$, defines the rth cumulant κ_r.

The rth cumulant κ_r can be expressed in terms of the rth and lower-order moments by equating coefficients in the expansions of $\exp\{K_X(t)\}$ and $M_X(t)$. We have, in particular, $\kappa_1 = \mathbb{E}(X) = \mu'_1$ and $\kappa_2 = \mathrm{var}(X) = \mu'_2 - \mu'^2_1$. The third and fourth cumulants are called the skewness and kurtosis respectively. For the normal distribution, all cumulants of third and higher order are 0.

Note that, for $a, b \in \mathbb{R}, K_{aX+b}(t) = bt + K_X(at)$, so that if $\tilde{\kappa}_r$ is the rth cumulant of $aX + b$, then $\tilde{\kappa}_1 = a\kappa_1 + b, \tilde{\kappa}_r = a^r \kappa_r, r \geq 2$. Also, if $X_1, \ldots, X_n$ are independent and identically distributed random variables with cumulant generating function $K_X(t)$, and $S_n = X_1 + \ldots + X_n$, then $K_{S_n}(t) = nK_X(t)$.

Extension of these notions to multivariate X involves no conceptual complication: see Pace and Salvan (1997, Chapter 3).

1.11.5 Some reminders

The Taylor expansion for a function $f(x)$ of a single real variable about $x = a$ is given by

$$f(x) = f(a) + f^{(1)}(a)(x-a) + \frac{1}{2!} f^{(2)}(a)(x-a)^2 + \ldots + \frac{1}{n!} f^{(n)}(a)(x-a)^n + R_n,$$

where

$$f^{(l)}(a) = \left. \frac{d^l f(x)}{dx^l} \right|_{x=a},$$

and the remainder R_n is of the form

$$\frac{1}{(n+1)!}f^{(n+1)}(c)(x-a)^{n+1},$$

for some $c \in [a, x]$.

The Taylor expansion is generalised to a function of several variables in a straightforward manner. For example, the expansion of $f(x, y)$ about $x = a$ and $y = b$ is given by

$$\begin{aligned} f(x,y) &= f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) \\ &+ \frac{1}{2!}\{f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(a,b)(y-b)^2\} + \ldots, \end{aligned}$$

where

$$\begin{aligned} f_x(a,b) &= \left.\frac{\partial f}{\partial x}\right|_{x=a,y=b} \\ f_{xy}(a,b) &= \left.\frac{\partial^2 f}{\partial x \partial y}\right|_{x=a,y=b}, \end{aligned}$$

and similarly for the other terms.

Some particular expansions therefore are:

$$\begin{aligned} \log(1+x) &= x - x^2/2 + x^3/3 - x^4/4 \ldots (|x| < 1) \\ \exp(x) &= 1 + x + x^2/2! + x^3/3! + x^4/4! \ldots (x \in \mathbb{R}) \\ f(x+h) &= f(x) + f'(x)h + f''(x)h^2/2! + \ldots (x \in \mathbb{R}) \\ f(x+h) &= f(x) + f'(x)^T h + h^T f''(x)h/2! + \ldots (x \in \mathbb{R}^p). \end{aligned}$$

The sign function sgn is defined by

$$\mathrm{sgn}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0 \end{cases}$$

Suppose we partition a matrix A so that $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, with A^{-1} correspondingly written $A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}$. If A_{11} and A_{22} are non-singular, let

$$A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21},$$

and

$$A_{22.1} = A_{22} - A_{21}A_{11}^{-1}A_{12}.$$

Then,

$$\begin{aligned} A^{11} &= A_{11.2}^{-1}, \quad A^{22} = A_{22.1}^{-1}, \quad A^{12} = -A_{11}^{-1}A_{12}A^{22}, \\ A^{21} &= -A_{22}^{-1}A_{21}A^{11}. \end{aligned}$$

1.11.6 Multivariate normal distribution

Of particular importance is the multivariate normal distribution, which, for nonsingular Σ, has density

$$f(y;\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\tfrac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\}$$

for $y \in \mathbb{R}^p, \mu \in \mathbb{R}^p$. We write this as $N_p(\mu,\Sigma)$. If $Y \sim N_p(\mu,\Sigma)$ then $EY = \mu$, $\operatorname{var} Y = \Sigma$.

If $Y \sim N_p(0,\Sigma)$, call $Q_Y = Y^T\Sigma^{-1}Y$ the covariance form associated with Y. Then a key result is that $Q_Y \sim \chi^2_p$. To see this, note

1. the covariance form is invariant under non-singular transformation of Y;

2. Y can be transformed to independent components of unit variance (set $Z = \Sigma^{-1/2}Y$);

3. the chi-squared distribution then follows directly, $Q_Y \equiv Q_Z = Z^TZ$.

Now suppose that Y is partitioned into two parts $Y^T = (Y_{(1)}^T, Y_{(2)}^T)$ where $Y_{(j)}$ is $p_j \times 1, p_1 + p_2 = p$. It is immediate that $Q_{Y_{(1)}} \sim \chi^2_{p_1}$, but in addition

$$Q_{Y_{(1)}.Y_{(2)}} = Q_Y - Q_{Y_{(1)}} \sim \chi^2_{p_2}$$

independently of $Q_{Y_{(1)}}$. Apply a transformation to Y so that the first p_1 components are $Y_{(1)}$ and the last p_2 components, $Y'_{(2)}$ say, are independent of $Y_{(1)}$. Then, by the invariance of the covariance form under non-singular transformation of Y,

$$Q_Y = Q_{Y_{(1)}} + Q_{Y'_{(2)}},$$

so that $Q_{Y'_{(2)}} \equiv Q_{Y_{(1)}.Y_{(2)}}$. The stated properties of $Q_{Y'_{(2)}}$ clearly hold.

2 Large Sample Theory

2.1 Motivation

In many situations, statistical inference depends on being able to approximate, using asymptotic theory, to densities or distribution functions. Exact answers are rarely available. The approximations used are based on results of probability theory, as revised in Section 1.11.2.

Further, potentially useful results worth highlighting are as follows.

[Continuous mapping theorem] Suppose the sequence $X_1, X_2, \ldots$ of random $d-$dimensional vectors is such that $X_n \xrightarrow{d} X$ and g is a continuous function. Then $g(X_n) \xrightarrow{d} g(X)$.

[Multivariate CLT] Let $X_1, \ldots, X_n$ be independent, identically distributed random $d-$dimensional vectors with $\mathrm{var}\,(X_1) = \Sigma$ a finite matrix. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - EX_1) \xrightarrow{d} N_d(0, \Sigma).$$

[Multivariate delta-method] Let $X_1, X_2, \ldots, Y$ be random $d-$dimensional vectors satisfying $a_n(X_n - c) \xrightarrow{d} Y$, where $c \in \mathbb{R}^d$ and $\{a_n\}$ is a sequence of positive numbers with $a_n \to \infty$ as $n \to \infty$. If g is a function from $\mathbb{R}^d$ to $\mathbb{R}$ which is differentiable at c, then if Y is $N_d(0, \Sigma)$, we have

$$a_n[g(X_n) - g(c)] \xrightarrow{d} N(0, [\nabla g(c)]^T \Sigma [\nabla g(c)]),$$

where $\nabla g(x)$ denotes the $d-$vector of partial derivatives of g at x.

Details of these results and generalizations are described by Barndorff-Nielsen and Cox (1989, Chapter 2).

Theory based on limit results of this kind is combined in 'statistical asymptotics' with asymptotic techniques from analysis and development of asymptotic expansions. Often a first-order approximation as may, say, arise from application of the CLT, can be improved by incorporating higher-order terms in an asymptotic expansion. Chapter 2 will be concerned with first-order theory of statistical quantities based on the likelihood function, while Chapter 3 will examine higher-order approximations that refine first-order results. Note that theory underlying approximation techniques is valid as some quantity, typically the sample size n [or more generally some 'amount of information'], goes to infinity, but the approximations obtained can be very accurate even for extremely small sample sizes.

In this chapter we discuss first-order asymptotic theory in regular parametric models. We will focus on models with independent component random variables. For details of more general cases see Barndorff-Nielsen and Cox (1994, Chapter 3). A key result is that in the independent component case the score function is, by the central limit theorem, asymptotically normal.

2.2 No nuisance parameter case

Recall the definitions of the score function and expected and observed information of Sections 1.3.1 and 1.3.2.

Denote by l_r the rth component of $U(\theta)$, l_{rs} the (r,s)th component of $\nabla_\theta \nabla_\theta^T l$, and denote the (r,s)th component of the inverse of the matrix $[l_{rs}]$ by l^{rs}. The maximum likelihood estimate for given observations y is, for regular problems, defined as the solution, assumed unique, of the 'likelihood equation'

$$u(\hat{\theta}; y) = 0.$$

Consider testing the null hypothesis $H_0 : \theta = \theta_0$, where θ_0 is an arbitrary, specified, point in Ω_θ. We can test H_0 in many ways equivalent to first-order, i.e. using statistics that typically differ by $O_p(n^{-1/2})$. Three such statistics are:

1. the likelihood ratio statistic

$$w(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\}, \tag{2.1}$$

2. the score statistic

$$w_U(\theta_0) = U^T(\theta_0) i^{-1}(\theta_0) U(\theta_0), \tag{2.2}$$

3. the Wald statistic

$$w_p(\theta_0) = (\hat{\theta} - \theta_0)^T i(\theta_0)(\hat{\theta} - \theta_0). \tag{2.3}$$

In (2.3) the suffix p warns that a particular parameterisation is involved.

For a scalar θ, (2.1) may be replaced by

$$r(\theta_0) = \mathrm{sgn}(\hat{\theta} - \theta_0)\sqrt{w(\theta_0)}, \tag{2.4}$$

the directed likelihood or 'signed root likelihood ratio statistic'. Also (2.2) and (2.3) may be replaced by

$$r_U(\theta_0) = U(\theta_0)/\sqrt{i(\theta_0)} \tag{2.5}$$

and

$$r_p(\theta_0) = (\hat{\theta} - \theta_0)\sqrt{i(\theta_0)} \tag{2.6}$$

respectively.

In a first-order asymptotic theory, the statistics (2.1)–(2.3) have, asymptotically, the chi-squared distribution with $d_\theta = \dim(\Omega_\theta)$ degrees of freedom. The signed versions (2.4)–(2.6) have an $N(0,1)$ distribution.

Confidence regions at level $1-\alpha$ are formed approximately as, for example,

$$\{\theta : w(\theta) \leq \chi^2_{d_\theta,\alpha}\},$$

where $\chi^2_{d_\theta,\alpha}$ is the upper α point of the relevant chi-squared distribution.

Note that in (2.5) $\sqrt{i(\theta_0)}$ is the exact standard deviation of $U(\theta_0)$, while in (2.6) $1/\sqrt{i(\theta_0)}$ is the approximate standard deviation of $\hat{\theta}$ when $\theta = \theta_0$.

In asymptotic calculations, because $U(\theta_0)$ and $i(\theta_0)$ refer to the total vector Y of dimension n, then as $n \to \infty$ and subject to some general conditions:

$$\begin{aligned} U(\theta_0) &\equiv \sqrt{n}\bar{U}(\theta_0) = O_p(n^{1/2}), \\ i(\theta_0) &\equiv n\bar{i}(\theta_0) = O(n), \\ \hat{\theta} - \theta_0 &= O_p(n^{-1/2}), \end{aligned}$$

where $\bar{i}(\theta_0)$ is the average information per observation and $\bar{U}(\theta_0)$ is a normalised score function. If the observations are IID, $\bar{i}$ is the information for a single observation.

Note that, as $n \to \infty$, we have in probability that, provided $i(\theta)$ is continuous at $\theta = \theta_0$,

$$\begin{aligned} j(\hat{\theta})/n &\to \bar{i}(\theta_0), \\ j(\theta_0)/n &\to \bar{i}(\theta_0). \end{aligned}$$

Therefore, in the definitions of the various statistics, $i(\theta_0)$ can be replaced by $i(\hat{\theta})$, $j(\hat{\theta})$, $j(\theta_0)$ etc. etc., in the sense that, if $\theta = \theta_0$, the various modified statistics differ typically by $O_p(n^{-1/2})$, so that their limiting distributions are the same under H_0.

2.3 Distribution theory for 2.2

Here we outline the asymptotic distribution theory that justifies the procedures of Section 2.2.

A serious issue concerns the asymptotic existence, uniqueness and consistency of the maximum likelihood estimate. There are no very satisfactory general theorems on such questions. A general result on the existence of a solution of the maximum likelihood equation asymptotically close to the true parameter value is possible, but is less than is required. We assume that $\hat{\theta}$ is well defined and consistent.

A key result is that in considerable generality U is asymptotically normal with zero mean and variance $i(\theta)$. For IID components this is a trivial consequence of the additive form of U and the CLT, together with the assumed finiteness of var (U). Very generally, the asymptotic distribution is a consequence of a martingale property of the score vector. For details see Barndorff-Nielsen and Cox (1994, pp. 85–86).

Suppose that $U(\theta) = U(\theta; Y) = [l_r(\theta)]$ has been shown to be asymptotically $N_d(0, i(\theta))$, formally

$$U(\theta)/\sqrt{n\bar{i}(\theta)} \xrightarrow{d} N_d\big(0, I_d\big), \tag{2.7}$$

with I_d the $d \times d$ identity matrix, and with $\sqrt{\ }$ interpreted as the matrix square root. We review what this implies about $\hat{\theta}$. Now adopt the summation convention and expand the score $l_r(\theta)$ in a Taylor series around θ, writing

$$\begin{aligned} l_r(\theta) &= U_r(\theta) = \sqrt{n}\bar{l}_r(\theta) = \sqrt{n}\bar{U}_r(\theta), \\ l_{rs}(\theta) &= n\bar{l}_{rs}(\theta) = -j_{rs}(\theta) = -n\bar{j}_{rs}(\theta), \\ \bar{\delta}^r &= \sqrt{n}(\hat{\theta}^r - \theta^r), l_{rst}(\theta) = n\bar{l}_{rst}(\theta), \\ i(\theta) &= n\bar{i}(\theta), \text{ etc.} \end{aligned}$$

Then, $l_r(\hat{\theta}) = 0$, so

$$\begin{aligned} \sqrt{n}\bar{l}_r(\theta) &+ n\bar{l}_{rs}(\theta)\bar{\delta}^s/\sqrt{n} \\ &+ \tfrac{1}{2}n\bar{l}_{rst}(\theta)\bar{\delta}^s\bar{\delta}^t/n + \cdots = 0, \end{aligned}$$

so that to a first-order approximation, ignoring the third term, we have

$$\begin{aligned} \bar{\delta}^r &= -\bar{l}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}) \\ &= \bar{j}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}). \end{aligned}$$

Now $j^{rs}/i^{rs} \xrightarrow{p} 1$, so

$$\bar{\delta}^r = \bar{i}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}),$$

a linear function of asymptotically normal variables of zero mean. It follows from (2.7) that $[\bar{\delta}^r]$ is asymptotically normal with zero mean and covariance matrix $[\bar{i}^{rs}]$. We have

$$\sqrt{n\bar{i}(\theta)}(\hat{\theta}-\theta) \xrightarrow{d} N_d\big(0, I_d\big). \tag{2.8}$$

Note that the normality relations (2.7) and (2.8) are asymptotically parameterisation invariant. This means, in particular, that to show normality for arbitrary parameterisations it is enough to do so for one parameterisation. The consequence is simplification of theoretical derivations in many circumstances.

The asymptotic χ^2 distribution of $w = w(\theta) = 2\{l(\hat{\theta}) - l(\theta)\}$ follows directly from the above. By direct expansion in θ around $\hat{\theta}$ we have, writing $\hat{j} \equiv j(\hat{\theta}) = [\hat{j}_{rs}]$,

$$w(\theta) = \hat{j}_{rs}(\hat{\theta}-\theta)^r(\hat{\theta}-\theta)^s + o_p(1)$$

or equivalently

$$w(\theta) = i^{rs}l_r l_s + o_p(1),$$

so $w(\theta) \xrightarrow{d} \chi^2_d$. The asymptotic χ^2 distribution of the Wald and score statistics follows similarly.

When the dimension of θ is $d = 1$, we have that the signed root likelihood ratio statistic

$$r = \mathrm{sgn}(\hat{\theta}-\theta)\sqrt{w(\theta)}$$

satisfies

$$r = \hat{j}^{-1/2}U + o_p(1)$$

so that $r \xrightarrow{d} N(0,1)$. Also, $i(\hat{\theta})^{1/2}(\hat{\theta}-\theta)$ is asymptotically $N(0,1)$, so that an approximate $100(1-\alpha)\%$ confidence interval for θ is

$$\hat{\theta} \mp i(\hat{\theta})^{-1/2}\Phi^{-1}(1-\alpha/2),$$

in terms of the $N(0,1)$ distribution function Φ.

2.4 Multiparameter problems: profile likelihood

Consider again the multiparameter problem in which $\theta = (\theta^1, \ldots, \theta^d) \in \Omega_\theta$, an open subset of $\mathbb{R}^d$.

Typically, interest lies in inference for a subparameter or parameter function $\psi = \psi(\theta)$. The *profile likelihood* $L_{\mathrm{p}}(\psi)$ for ψ is defined by

$$L_{\mathrm{p}}(\psi) = \sup_{\{\theta:\psi(\theta)=\psi\}} L(\theta),$$

the supremum of $L(\theta)$ over all θ that are consistent with the given value of ψ.

The log profile likelihood is $l_{\rm p} = \log L_{\rm p}$. It may be written as $l_{n{\rm p}}$ if it is to be stressed that it is based on a sample of size n.

Often ψ is a component of a given partition $\theta = (\psi, \chi)$ of θ into sub-vectors ψ and χ of dimension $d_\psi = d - d_\chi$ and d_χ respectively, and we may then write

$$L_{\rm p}(\psi) = L(\psi, \hat{\chi}_\psi),$$

where $\hat{\chi}_\psi$ denotes the maximum likelihood estimate of χ for a given value of ψ. We assume this is the case from now on.

The profile likelihood $L_{\rm p}(\psi)$ can, to a considerable extent, be thought of and used as if it were a genuine likelihood. In particular, the maximum profile likelihood estimate of ψ equals $\hat{\psi}$, the first d_ψ components of $\hat{\theta}$. Further, the profile log-likelihood ratio statistic $2\{l_{\rm p}(\hat{\psi}) - l_{\rm p}(\psi_0)\}$ equals the log-likelihood ratio statistic for $H_0 : \psi = \psi_0$,

$$2\{l_{\rm p}(\hat{\psi}) - l_{\rm p}(\psi_0)\} \equiv 2\{l(\hat{\psi}, \hat{\chi}) - l(\psi_0, \hat{\chi}_0)\} \equiv w(\psi_0),$$

where $l \equiv l_n$ is the log-likelihood and we have written $\hat{\chi}_0$ for $\hat{\chi}_{\psi_0}$. The asymptotic null distribution of the profile log-likelihood ratio statistic is $\chi^2_{d_\psi}$: this follows from general distribution theory considered later.

The inverse of the observed profile information equals the ψ component of the full observed inverse information evaluated at $(\psi, \hat{\chi}_\psi)$,

$$j_{\rm p}^{-1}(\psi) = j^{\psi\psi}(\psi, \hat{\chi}_\psi),$$

where $j_{\rm p}$ denotes observed profile information, minus the matrix of second-order derivatives of $l_{\rm p}$, and $j^{\psi\psi}$ is the $\psi\psi$-block of the inverse of the full observed information j.

For scalar ψ, this result follows on differentiating $l_{\rm p}(\psi) = l(\psi, \hat{\chi}_\psi)$ twice with respect to ψ. Let l_ψ and l_χ denote the partial derivatives of $l(\psi, \chi)$ with respect to ψ, χ respectively. The profile score is $l_\psi(\psi, \hat{\chi}_\psi)$, on using the chain rule to differentiate $l_p(\psi)$ with respect to ψ, noting that $l_\chi(\psi, \hat{\chi}_\psi) = 0$. The second derivative is, following the notation, $l_{\psi\psi}(\psi, \hat{\chi}_\psi) + l_{\psi\chi}(\psi, \hat{\chi}_\psi)\frac{\partial}{\partial\psi}\hat{\chi}_\psi$. Now use the result that

$$\partial\hat{\chi}_\psi/\partial\psi = -j_{\psi\chi}(\psi, \hat{\chi}_\psi) j^{-1}_{\chi\chi}(\psi, \hat{\chi}_\psi).$$

This latter formula follows by differentiating the likelihood equation $l_\chi(\psi, \hat{\chi}_\psi) = 0$ with respect to ψ. This gives

$$l_{\chi\psi}(\psi, \hat{\chi}_\psi) + l_{\chi\chi}(\psi, \hat{\chi}_\psi)\frac{\partial}{\partial\psi}\hat{\chi}_\psi = 0,$$

from which

$$\frac{\partial}{\partial\psi}\hat{\chi}_\psi = -(l_{\chi\chi}(\psi,\hat{\chi}_\psi))^{-1}l_{\chi\psi}(\psi,\hat{\chi}_\psi).$$

It follows that

$$j_p(\psi) = -(l_{\psi\psi} - l_{\psi\chi}(l_{\chi\chi})^{-1}l_{\chi\psi}),$$

where all the derivatives are evaluated at $(\psi,\hat{\chi}_\psi)$. Then, using the formulae for the inverse of a partitioned matrix, as given in Section 1.11.5, the result is proved. The vector case follows similarly.

When ψ is scalar, this implies that the curvature of the profile log-likelihood is directly related to the precision of $\hat{\psi}$. We have seen that a key property of the log-likelihood $l(\theta)$ when there are no nuisance parameters is that the observed information $j(\hat{\theta})$ can be as an estimate of the inverse asymptotic covariance matrix of $\hat{\theta}$ (which is actually $i(\theta)$). The above result shows that the corresponding function computed from the profile log-likelihood,

$$j_{\rm p}(\hat{\psi}) = -[\nabla_\psi\nabla_\psi^T l_{\rm p}(\psi)]_{\psi=\hat{\psi}}$$

determines an estimate of the inverse asymptotic covariance matrix for $\hat{\psi}$.

2.5 Multiparameter problems: further statistics

For testing $H_0 : \psi = \psi_0$, in the presence of a nuisance parameter χ, the forms of the score statistic and the Wald statistic corresponding to the profile log-likelihood ratio statistic $w(\psi_0)$ are obtained by partitioning the maximum likelihood estimate, the score vector, the information matrix and its inverse:

$$\begin{aligned} U(\theta) &= \begin{pmatrix} U_\psi(\psi,\chi) \\ U_\chi(\psi,\chi) \end{pmatrix}, \\ i(\theta) &= \begin{bmatrix} i_{\psi\psi}(\psi,\chi)\ i_{\psi\chi}(\psi,\chi) \\ i_{\chi\psi}(\psi,\chi)\ i_{\chi\chi}(\psi,\chi) \end{bmatrix}, \\ i^{-1}(\theta) &= \begin{bmatrix} i^{\psi\psi}(\psi,\chi)\ i^{\psi\chi}(\psi,\chi) \\ i^{\chi\psi}(\psi,\chi)\ i^{\chi\chi}(\psi,\chi) \end{bmatrix}. \end{aligned}$$

We know that $\hat{\psi}$ is asymptotically normally distributed with mean ψ_0 and covariance matrix $i^{\psi\psi}(\psi_0,\chi_0)$, which can be replaced by $i^{\psi\psi}(\psi_0,\hat{\chi}_0)$, yielding a version of the Wald test statistic for this nuisance parameter case:

$$w_p(\psi_0) = (\hat{\psi}-\psi_0)^T[i^{\psi\psi}(\psi_0,\hat{\chi}_0)]^{-1}(\hat{\psi}-\psi_0).$$

Cox and Hinkley (1974, pp. 323–324) give a detailed derivation of a version of the score statistic for testing $H_0 : \psi = \psi_0$:

$$w_u(\psi_0) = U_\psi(\psi_0, \hat{\chi}_0)^T i^{\psi\psi}(\psi_0, \hat{\chi}_0) U_\psi(\psi_0, \hat{\chi}_0).$$

This test has the advantage that the maximum likelihood estimator only has to be obtained under H_0, and is derived from the asymptotic normality of U.

Both $w_p(\psi_0)$ and $w_u(\psi_0)$ have asymptotically a chi-squared distribution with d_ψ degrees of freedom, as will be shown in Section 2.7, by showing their first-order equivalence to $w(\psi_0)$.

2.6 Effects of parameter orthogonality

Assume that it is possible to make the parameter of interest ψ and the nuisance parameter, now denoted by λ, orthogonal. This is always possible if ψ is one-dimensional. Any transformation from, say, (ψ, χ) to (ψ, λ) necessary to achieve this leaves the profile log-likelihood invariant.

Now the matrices $i(\psi, \lambda)$ and $i^{-1}(\psi, \lambda)$ are block diagonal. Therefore, $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent and the asymptotic variance of $\hat{\psi}$ where λ is unknown is the same as that where λ is known. A related property is that $\hat{\lambda}_\psi$, the MLE of λ for specified ψ, varies only slowly in ψ in the neighbourhood of $\hat{\psi}$, and that there is a corresponding slow variation of $\hat{\psi}_\lambda$ with λ. More precisely, if $\psi - \hat{\psi} = O_p(n^{-1/2})$, then $\hat{\lambda}_\psi - \hat{\lambda} = O_p(n^{-1})$. For a nonorthogonal nuisance parameter χ, we would have $\hat{\chi}_\psi - \hat{\chi} = O_p(n^{-1/2})$.

We sketch a proof of this result for the case where both the parameter of interest and the nuisance parameter are scalar. If $\psi - \hat{\psi} = O_p(n^{-1/2})$, $\chi - \hat{\chi} = O_p(n^{-1/2})$, we have

$$\begin{aligned} &l(\psi, \chi) = l(\hat{\psi}, \hat{\chi}) \\ &-\tfrac{1}{2}\{\hat{j}_{\psi\psi}(\psi - \hat{\psi})^2 + 2\hat{j}_{\psi\chi}(\psi - \hat{\psi})(\chi - \hat{\chi}) + \hat{j}_{\chi\chi}(\chi - \hat{\chi})^2\} + O_p(n^{-1/2}). \end{aligned}$$

It then follows that

$$\begin{aligned} \hat{\chi}_\psi - \hat{\chi} &= \frac{-\hat{j}_{\psi\chi}}{\hat{j}_{\chi\chi}}(\psi - \hat{\psi}) + O_p(n^{-1}) \\ &= \frac{-i_{\psi\chi}}{i_{\chi\chi}}(\psi - \hat{\psi}) + O_p(n^{-1}). \end{aligned}$$

Then, because $\psi - \hat{\psi} = O_p(n^{-1/2})$, $\hat{\chi}_\psi - \hat{\chi} = O_p(n^{-1/2})$ unless $i_{\psi\chi} = 0$, the orthogonal case, when the difference is $O_p(n^{-1})$.

Note also that, so far as asymptotic theory is concerned, we can have $\hat{\chi}_\psi = \hat{\chi}$ independently of ψ only if χ and ψ are orthogonal. In this special case we can write $l_{\rm p}(\psi) = l(\psi, \hat{\chi})$. In the general orthogonal case, $l_{\rm p}(\psi) = l(\psi, \hat{\chi}) + o_p(1)$, so that a first-order theory could use $l^*_{\rm p}(\psi) = l(\psi, \hat{\chi})$ instead of $l_{\rm p}(\psi) = l(\psi, \hat{\chi}_\psi)$.

2.7 Distribution theory in nuisance parameter case

First-order asymptotic distribution theory when nuisance parameters are present follows from basic properties of the multivariate normal distribution given in Section 1.11.6.

The log-likelihood ratio statistic $w(\psi_0)$ can be written as

$$w(\psi_0) = 2\big\{l(\hat{\psi}, \hat{\chi}) - l(\psi_0, \chi)\big\} - 2\big\{l(\psi_0, \hat{\chi}_0) - l(\psi_0, \chi)\big\},$$

as the difference of two statistics for testing hypotheses without nuisance parameters.

Taylor expansion about (ψ_0, χ), where χ is the true value of the nuisance parameter, gives, to first-order (i.e. ignoring terms of order $o_p(1)$),

$$\begin{aligned} w(\psi_0) &= \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix}^T i(\psi_0, \chi) \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} \\ &\qquad -(\hat{\chi}_0 - \chi)^T i_{\chi\chi}(\psi_0, \chi)(\hat{\chi}_0 - \chi). \end{aligned} \tag{2.9}$$

Note that the linearised form of the maximum likelihood estimating equations is

$$\begin{bmatrix} i_{\psi\psi} & i_{\psi\chi} \\ i_{\chi\psi} & i_{\chi\chi} \end{bmatrix} \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} = \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix},$$

so

$$\begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} = \begin{bmatrix} i^{\psi\psi} & i^{\psi\chi} \\ i^{\chi\psi} & i^{\chi\chi} \end{bmatrix} \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix}.$$

Also $\hat{\chi}_0 - \chi = i^{-1}_{\chi\chi} U_\chi$, to first-order. Then, we see from (2.9) that to first-order

$$w(\psi_0) = [U_\psi^T U_\chi^T] \begin{bmatrix} i^{\psi\psi} & i^{\psi\chi} \\ i^{\chi\psi} & i^{\chi\chi} \end{bmatrix} \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix} - U_\chi^T i^{-1}_{\chi\chi} U_\chi. \tag{2.10}$$

From (2.10), in the notation of Section 1.11.6,

$$w(\psi_0) \sim Q_U - Q_{U_\chi} = Q_{U_\psi . U_\chi},$$

and is thus asymptotically $\chi^2_{d_\psi}$.

The Wald statistic $w_p(\psi_0)$ is based directly on the covariance form of $\hat{\psi} - \psi_0$, and so can be seen immediately to be asymptotically $\chi^2_{d_\psi}$. Note that to first-order we have

$$w_p(\psi_0) = [i^{\psi\psi}U_\psi + i^{\psi\chi}U_\chi]^T (i^{\psi\psi})^{-1} [i^{\psi\psi}U_\psi + i^{\psi\chi}U_\chi]. \tag{2.11}$$

Correspondingly, we can express the statistic $w_U(\psi_0)$ in terms of the score vector U. To first-order we have

$$w_U(\psi_0) = (U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi)^T i^{\psi\psi} (U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi). \tag{2.12}$$

This follows since, to first-order,

$$\begin{aligned} U_\psi(\psi_0, \hat{\chi}_0) &= U_\psi + \frac{\partial U_\psi}{\partial \chi} (\hat{\chi}_0 - \chi) \\ &= U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi. \end{aligned}$$

The equivalence of the three statistics, and therefore the asymptotic distribution of $w_U(\psi_0)$, follows on showing, using results for partitioned matrices given in Section 1.11.5, that the three quantities (2.10), (2.11) and (2.12) are identical.

As an illustration, write

$$\begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \quad \begin{bmatrix} i_{\psi\psi} & i_{\psi\chi} \\ i_{\chi\psi} & i_{\chi\chi} \end{bmatrix} = \begin{bmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{bmatrix}$$

for ease of notation.

Multiplying out (2.10) gives

$$w(\psi_0) = U_1^T i^{11} U_1 + U_2^T i^{21} U_1 + U_1^T i^{12} U_1 + U_2^T [i^{22} - i_{22}^{-1}] U_2. \tag{2.13}$$

Multiplying out (2.11) gives

$$w_p(\psi_0) = U_1^T i^{11} U_1 + U_1^T i^{12} U_2 + U_2^T i^{21} U_1 + U_2^T i^{21} (i^{11})^{-1} i^{12} U_2, \tag{2.14}$$

since $(i_{11} - i_{12} i_{22}^{-1} i_{21})^{-1} = i^{11}$. Equivalence of (2.13) and (2.14) follows on noting that

$$i^{21}(i^{11})^{-1} i^{12} = i_{22}^{-1} i_{21} i_{11}^{-1} i_{12} i^{22} = i_{22}^{-1} [i_{22} - (i^{22})^{-1}] i^{22} = i^{22} - i_{22}^{-1}.$$

2.8 An example: possible censoring

Suppose that we observe a realization z of $Z = (Z_1, ..., Z_n)$, where the Z_i are independent, identically distributed exponential random variables, with parameter θ, so that the likelihood is

$$f(z;\theta) = \theta^n \exp\{-\theta \sum_{j=1}^{n} z_j\}. \tag{2.15}$$

Now suppose that the observations are censored at $c > 0$, so that instead of z we actually observe y, where

$$y_j = z_j I(z_j \leq c) + cI(z_j > c), \quad j = 1, ..., n.$$

The y_j are realizations of independently distributed random variables Y_j which have density $\theta \exp(-\theta x)$ if $x < c$, and equal c with probability $P(Z_j > c) = e^{-\theta c}$. Thus in this censored case, the likelihood is

$$g(y;\theta) = \theta^r \exp\{-\theta \sum_{j=1}^{n} y_j\}, \tag{2.16}$$

where $r = \sum_{j=1}^{n} I(z_j \leq c)$ is the random number of uncensored observations.

If we draw a sample in which none of the observations is actually actually greater than c, no censoring occurs and we have $z_j = y_j, r = n$ and

$$g(y;\theta) = f(z;\theta).$$

The (strong) likelihood principle asserts that we should make the same inferences about θ in the two cases. That is, if censoring is possible but does not occur, inference should be the same as when censoring is impossible.

Under (2.16) the Fisher information in a single observation is

$$\bar{i}(\theta) \equiv i(\theta)/n = E\{r/(n\theta^2)\} = \frac{1 - e^{-\theta c}}{\theta^2}.$$

The likelihood is maximized at $\hat{\theta} = r/(n\bar{y})$. The observed information is $\bar{j}(\hat{\theta}) \equiv j(\hat{\theta})/n = n\bar{y}^2/r$. Therefore, under (2.16) an approximate $100(1-\alpha)\%$ confidence interval for θ based on $\bar{i}(\hat{\theta})$ is

$$\frac{r}{n\bar{y}} \mp \frac{1}{n^{1/2}(n\bar{y}/r)[1 - \exp\{-cr/(n\bar{y})\}]^{1/2}} \Phi^{-1}(1 - \alpha/2). \tag{2.17}$$

Under (2.15) the likelihood is maximized by $\hat{\theta} = 1/\bar{z}$. The expected and observed Fisher information are equal and $\bar{i}(\hat{\theta}) \equiv \bar{j}(\hat{\theta}) = 1/\hat{\theta}^2 = \bar{z}^2$. An approximate $100(1-\alpha)\%$ confidence interval for θ is

$$\frac{1}{\bar{z}} \mp \frac{1}{n^{1/2}\bar{z}}\Phi^{-1}(1-\alpha/2). \tag{2.18}$$

When no censoring occurs (2.17) reduces to

$$\frac{1}{\bar{z}} \mp \frac{1}{n^{1/2}\bar{z}\{1-\exp(-c/\bar{z})\}^{1/2}}\Phi^{-1}(1-\alpha/2), \tag{2.19}$$

which is wider than (2.18), so that the use of (2.19) conflicts with the likelihood principle.

The difference between (2.18) and (2.19) is that the asymptotic variances based on the expected Fisher information reflect the dependence on the sampling scheme. If we use the observed information $\bar{j}(\hat{\theta}) = r/(n\hat{\theta}^2)$ in the censored case, we find that an approximate $100(1-\alpha)\%$ confidence interval for θ is

$$\frac{r}{n\bar{y}} \mp \frac{r^{1/2}}{n\bar{y}}\Phi^{-1}(1-\alpha/2),$$

which reduces to (2.18) when censoring does not actually occur. Use of observed information in a confidence interval obeys the likelihood principle because the maximum likelihood estimate and the observed information are identical for any two models with proportional likelihoods.

3 Higher-order Theory

The refinements to the asymptotic theory of Chapter 2 to be discussed here have two main origins. One motivation is to improve on the first-order limit results of Chapter 2, so as to obtain approximations whose asymptotic accuracy is higher by one or two orders. The other is the Fisherian proposition that inferences on the parameter of interest should be obtained by conditioning on an ancillary statistic, rather than from the original model.

3.1 Asymptotic expansions

Various technical tools are of importance in the development of statistical theory. Key methods, which we describe in subsequent sections, used to obtain higher-order approximations to densities and distribution functions are Edgeworth expansion, saddlepoint approximation and Laplace's method. Here we consider first two important general ideas, those of asymptotic expansion, and stochastic asymptotic expansion.

Asymptotic expansions typically arise in the following way. We are interested in a sequence of functions $\{f_n(x)\}$, indexed by n, and write

$$f_n(x) = \gamma_0(x)b_{0,n} + \gamma_1(x)b_{1,n} + \gamma_2(x)b_{2,n} + \ldots + \gamma_k(x)b_{k,n} + o(b_{k,n}),$$

as $n \to \infty$, where $\{b_{r,n}\}_{r=0}^{k}$ is a sequence, such as $\{1, n^{-1/2}, n^{-1}, \ldots, n^{-k/2}\}$ or $\{1, n^{-1}, n^{-2}, \ldots, n^{-k}\}$. An essential condition is that $b_{r+1,n} = o(b_{r,n})$ as $n \to \infty$, for each $r = 0, 1, \ldots, k-1$.

Often the function of interest $f_n(x)$ will be the exact density or distribution function of a statistic based on a sample of size n at a fixed x, and $\gamma_0(x)$ will be some simple first-order approximation, such as the normal density or distribution function. One important feature of asymptotic expansions is that they are not in general convergent series for $f_n(x)$ for any fixed x: taking successively more terms, letting $k \to \infty$ for fixed n, will not necessarily improve the approximation to $f_n(x)$.

We will concentrate here on asymptotic expansions for densities, but describe some of the key formulae in distribution function estimation.

For a sequence of random variables $\{Y_n\}$, a **stochastic asymptotic expansion** is expressed as

$$Y_n = X_0 b_{0,n} + X_1 b_{1,n} + \ldots + X_k b_{k,n} + o_p(b_{k,n}),$$

where $\{b_{k,n}\}$ is a given set of sequences and $\{X_0, X_1, \ldots\}$ are random variables having distributions not depending on n.

There are several examples of the use of stochastic asymptotic expansions in the literature, but they are not as well defined as asymptotic expansions, as there is usually considerable arbitrariness in the choice of the coefficient random variables $\{X_0, X_1, \ldots\}$, and it is often convenient to use instead of $X_0, X_1, \ldots$ random variables for which only the asymptotic distribution is free of n. A simple application of stochastic asymptotic expansion is the proof of asymptotic normality of the maximum likelihood estimator, as sketched in Chapter 2: we have

$$\sqrt{i(\theta)}(\hat{\theta} - \theta) = \left\{ \frac{U(\theta)}{\sqrt{i(\theta)}} \right\} + O_p(n^{-1/2}),$$

in terms of the score $U(\theta)$ and Fisher information $i(\theta)$. The quantity $U(\theta)/\sqrt{i(\theta)}$ plays the role of X_0. By the CLT we can write

$$\frac{U(\theta)}{\sqrt{i(\theta)}} = X_0 + O_p(n^{-1/2}),$$

where X_0 is $N(0, 1)$.

3.2 Edgeworth expansion

In this Section and in Section 3.3 we assume, for simplicity, the case of univariate random variables. Extensions to the multivariate case are straightforward and are summarised, for example, by Severini (2000, Chapter 2).

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with cumulant generating function $K_X(t)$ and cumulants κ_r. Let $S_n = \sum_1^n X_i$, $S_n^* = (S_n - n\mu)/\sqrt{n}\sigma$ where $\mu \equiv \kappa_1 = \mathbb{E}X_1, \sigma^2 \equiv \kappa_2 = \operatorname{var} X_1$. Define the rth standardised cumulant by $\rho_r = \kappa_r/\kappa_2^{r/2}$.

The **Edgeworth expansion** for the density of the standardised sample mean S_n^* can be expressed as:

$$\begin{aligned} f_{S_n^*}(x) &= \phi(x) \left\{ 1 + \frac{\rho_3}{6\sqrt{n}} H_3(x) \right. \\ &\quad \left. + \frac{1}{n} \left[\frac{\rho_4 H_4(x)}{24} + \frac{\rho_3^2 H_6(x)}{72} \right] \right\} + O(n^{-3/2}). \end{aligned} \tag{3.1}$$

Here $\phi(x)$ is the standard normal density and $H_r(x)$ is the rth degree Hermite polynomial defined by

$$\begin{aligned} H_r(x) &= (-1)^r \frac{d^r \phi(x)}{dx^r} \Big/ \phi(x) \\ &= (-1)^r \phi^{(r)}(x)/\phi(x), \quad \text{say.} \end{aligned}$$

We have $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$ and $H_6(x) = x^6 - 15x^4 + 45x^2 - 15$. The asymptotic expansion (3.1) holds uniformly for $x \in \mathbb{R}$.

The leading term in the Edgeworth expansion is the standard normal density, as is appropriate from CLT. The remaining terms may be considered as higher order correction terms. The $n^{-1/2}$ term is an adjustment for the main effect of the skewness of the true density, via the standardised skewness ρ_3, and the n^{-1} term is a simultaneous adjustment for skewness and kurtosis. If the density of X_1 is symmetric, $\rho_3 = 0$ and a normal approximation to the density of S_n^* is accurate to order n^{-1}, rather than the usual $n^{-1/2}$ for $\rho_3 \neq 0$. The accuracy of the Edgeworth approximation, say

$$f_{S_n^*}(x) \doteq \phi(x) \left\{ 1 + \frac{\rho_3}{6\sqrt{n}} H_3(x) + \frac{1}{n} \left[\frac{\rho_4 H_4(x)}{24} + \frac{\rho_3^2 H_6(x)}{72} \right] \right\},$$

will depend on the value of x. In particular, Edgeworth approximations tend to be poor, and may even be negative, in the tails of the distribution, as $|x|$ increases.

Integrating the Edgeworth expansion (3.1) term by term (which requires a non-trivial justification), using the properties of the Hermite polynomials, we obtain an expansion for the distribution function of S_n^*:

$$\begin{aligned} F_{S_n^*}(x) &= \Phi(x) - \phi(x) \left\{ \frac{\rho_3}{6\sqrt{n}} H_2(x) \right. \\ &\qquad \left. + \frac{\rho_4}{24n} H_3(x) + \frac{\rho_3^2}{72n} H_5(x) \right\} + O(n^{-3/2}). \end{aligned}$$

Also, if T_n is a sufficiently smooth function of S_n^*, then a formal Edgeworth expansion can be obtained for the density of T_n. Further details and references are given by Severini (2000, Chapter 2).

When studying the coverage probability of confidence intervals, for example, it is often convenient to be able to determine x as x_α say, so that $F_{S_n^*}(x_\alpha) = \alpha$, to the order considered in the Edgeworth approximation to the distribution

function of S_n^*. The solution is known as the Cornish-Fisher expansion and the formula is

$$\begin{aligned} x_\alpha &= z_\alpha + \frac{1}{6\sqrt{n}}(z_\alpha^2 - 1)\rho_3 + \frac{1}{24n}(z_\alpha^3 - 3z_\alpha)\rho_4 \\ &- \frac{1}{36n}(2z_\alpha^3 - 5z_\alpha)\rho_3^2 + O(n^{-3/2}), \end{aligned}$$

where $\Phi(z_\alpha) = \alpha$.

The derivation of the Edgeworth expansion stems from the result that the density of a random variable can be obtained by inversion of its characteristic function. A form of this inversion result useful for our discussion here is that the density for $\bar{X}$, the mean of a set of independent, identically distributed random variables $X_1, \ldots, X_n$, can be obtained as

$$f_{\bar{X}}(\bar{x}) = \frac{n}{2\pi i} \int_{\tau - i\infty}^{\tau + i\infty} \exp\big[n\{K(\phi) - \phi\bar{x}\}\big] d\phi, \tag{3.2}$$

where K is the cumulant generating function of X, and τ is any point in the open interval around 0 in which the moment generating function M exists. For details, see Feller (1971, Chapter 16). In essence, the Edgeworth expansion (3.1) is obtained by expanding the cumulant generating function in a Taylor series around 0, exponentiating and inverting term by term. Details are given in Barndorff-Nielsen and Cox (1989, Chapter 4).

3.3 Saddlepoint expansion

The **saddlepoint expansion** for the density of S_n is

$$f_{S_n}(s) = \frac{1}{\sqrt{2\pi}} \frac{1}{\{nK_X''(\hat{\phi})\}^{1/2}} \times \exp\{nK_X(\hat{\phi}) - \hat{\phi}s\}\{1 + O(n^{-1})\} \tag{3.3}$$

where $\hat{\phi} \equiv \hat{\phi}(s)$ satisfies $nK_X'(\hat{\phi}) = s$.

A detailed analysis shows that the $O(n^{-1})$ term is actually $(3\hat{\rho}_4 - 5\hat{\rho}_3^2)/(24n)$, where $\hat{\rho}_j \equiv \hat{\rho}_j(\hat{\phi}) = K_X^{(j)}(\hat{\phi})/\{K_X''(\hat{\phi})\}^{j/2}$ is the jth standardised derivative of the cumulant generating function for X_1 evaluated at $\hat{\phi}$. A simple change of variable in (3.3) gives a saddlepoint expansion for the density of $\bar{X}_n = S_n/n$:

$$f_{\bar{X}_n}(x) = (2\pi)^{-1/2}\{n/K_X''(\hat{\phi})\}^{1/2} \times \exp\big\{n[K_X(\hat{\phi}) - \hat{\phi}x]\big\}\big(1 + O(n^{-1})\big), \tag{3.4}$$

where $K_X'(\hat{\phi}) = x$.

The saddlepoint expansion is quite different in form from the Edgeworth expansion. In order to use the former to approximate $f_{\bar{X}_n}(x)$ either with the leading term, or the leading term plus n^{-1} correction, it is necessary to know the whole cumulant generating function, not just the first four cumulants. It is also necessary to solve the equation $K'_X(\hat{\phi}) = x$ for each value of x. The leading term in (3.4) is not the normal (or any other) density; in fact it will not usually integrate to 1, although it can be renormalised to do so. The saddlepoint expansion is an asymptotic expansion in powers of n^{-1}, rather than $n^{-1/2}$ as in the Edgeworth expansion. This suggests that the main correction for skewness has been absorbed by the leading term, which is in fact the case.

Observe that, crucially, the saddlepoint expansion is stated with a *relative* error, while the Edgeworth expansion is stated with an *absolute* error.

The approximation obtained from the leading term of (3.4), ignoring the $O(n^{-1})$ correction term, is generally very accurate. In particular, the saddlepoint approximation tends to be much more accurate than an Edgeworth approximation in the tails of the distribution. In distributions that differ from the normal density in terms of asymmetry, such as the gamma distribution, the saddlepoint approximation is extremely accurate throughout the range of x. It is customary to use as an approximation to $f_{\bar{X}_n}(x)$ a renormalised version of (3.4):

$$f_{\bar{X}_n}(x) \doteq c\{n/K''_X(\hat{\phi})\}^{1/2} \exp\big[n\{K_X(\hat{\phi}) - \hat{\phi}x\}\big] \tag{3.5}$$

where c is determined, usually numerically, so that the right-hand side of (3.5) integrates to 1. If the $O(n^{-1})$ correction term is constant in x, (3.5) will be exact. For scalar random variables this happens only in the case of the normal, gamma and inverse Gaussian distributions. In general, the n^{-1} correction term $\{3\hat{\rho}_4(\hat{\phi}) - 5\hat{\rho}_3^2(\hat{\phi})\}/24$ varies only slowly with x and the relative error in the renormalised approximation (3.5) is $O(n^{-3/2})$.

The saddlepoint approximation is usually derived by one of two methods. The first (Daniels, 1954) uses the inversion formula (3.2) and contour integration, choosing the contour of integration to pass through the saddlepoint of the integrand on the line of steepest descent. We sketch instead a more statistical derivation, as described by Barndorff-Nielsen and Cox (1979) .

We associate with the density $f(x)$ for X_1 an exponential family density $f(x;\phi)$ defined by

$$f(x;\phi) = \exp\{x\phi - K_X(\phi)\}f(x)$$

where K_X is the cumulant generating function of X_1, under $f(x)$. Then it is straightforward to check that the sum $S_n = X_1 + \cdots + X_n$ has associated density

$$f_{S_n}(s;\phi) = \exp\{s\phi - nK_X(\phi)\} f_{S_n}(s)$$

from which

$$f_{S_n}(s) = \exp\{nK_X(\phi) - s\phi\} f_{S_n}(s;\phi). \tag{3.6}$$

Now use the Edgeworth expansion to obtain an approximation to the density $f_{S_n}(s;\phi)$, remembering that cumulants all must refer to cumulants computed under the tilted density $f(x;\phi)$. Since ϕ is arbitrary, it is chosen so that the Edgeworth expansion for the tilted density is evaluated at its mean, where the $n^{-1/2}$ term in the expansion is zero. This value is defined by $nK'_X(\hat{\phi}) = s$ and (3.6) becomes

$$f_{S_n}(s) \doteq \exp\{nK_X(\hat{\phi}) - \hat{\phi}s\}\{2\pi nK''_X(\hat{\phi})\}^{-1/2}, \tag{3.7}$$

which is the approximation deriving from (3.3). The factor $\{2\pi nK''_X(\hat{\phi})\}^{-1/2}$ comes from the normal density evaluated at its mean.

A case of special interest is when $f(x)$ is itself in the exponential family, $f(x;\theta) = \exp\{x\theta - c(\theta) - h(x)\}$. Then since $K_X(t) = c(\theta + t) - c(\theta)$, it follows that $\hat{\phi} = \hat{\theta} - \theta$, where $\hat{\theta}$ is the MLE based on $s = x_1 + \cdots + x_n$. Then (3.7) is

$$f_{S_n}(s;\theta) \doteq \exp\big[n\{c(\hat{\theta}) - c(\theta)\} - (\hat{\theta} - \theta)s\big]\{2\pi nc''(\hat{\theta})\}^{-1/2},$$

which can be expressed as

$$c\exp\{l(\theta) - l(\hat{\theta})\}|j(\hat{\theta})|^{-1/2} \tag{3.8}$$

where $l(\theta)$ is the log-likelihood function based on $(x_1, \ldots, x_n)$, or s, and $j(\hat{\theta})$ is the observed Fisher information. Since $\hat{\theta} = \hat{\theta}(s)$ is a one-to-one function of s, with Jacobian $|j(\hat{\theta})|$, (3.8) can be used to obtain an approximation to the density of $\hat{\theta}$

$$f_{\hat{\theta}}(\hat{\theta};\theta) \doteq c\exp\{l(\theta) - l(\hat{\theta})\}|j(\hat{\theta})|^{1/2}. \tag{3.9}$$

This latter approximation is a particular example of the p^*-formula, considered in Section 3.5.

It is not easy to integrate the right-hand side of the saddlepoint approximation (3.3) to obtain an approximation to the distribution function of S_n: see Lugannani and Rice (1980). The **Lugannani-Rice approximation** is

$$F_{S_n}(s) = \Phi(r_s) + \phi(r_s)\Big(\frac{1}{r_s} - \frac{1}{v_s}\Big) + O(n^{-1}),$$

where

$$\begin{aligned} r_s &= \mathrm{sgn}(\hat{\phi})\sqrt{2n\{\hat{\phi}K'_X(\hat{\phi}) - K_X(\hat{\phi})\}} \\ v_s &= \hat{\phi}\sqrt{nK''_X(\hat{\phi})}, \end{aligned}$$

and $\hat{\phi} \equiv \hat{\phi}(s)$ is the saddlepoint, satisfying $nK'_X(\hat{\phi}) = s$. The expansion can be expressed in the asymptotically equivalent form

$$F_{S_n}(s) = \Phi(r_s^*)\{1 + O(n^{-1})\},$$

with

$$r_s^* = r_s - \frac{1}{r_s}\log\frac{r_s}{v_s}.$$

3.4 Laplace approximation of integrals

Suppose $g : \mathbb{R} \to \mathbb{R}$ is a smooth function, and that we wish to evaluate the integral

$$g_n = \int_a^b e^{-ng(y)}dy.$$

The main contribution to the integral, for large n, will come from values of y near the minimum of $g(y)$, which may occur at a or b, or in the interior of the interval (a, b). Assume that $g(y)$ is minimised at $\tilde{y} \in (a, b)$ and that $g'(\tilde{y}) = 0$, $g''(\tilde{y}) > 0$. The other cases may be treated in a similar manner. For a useful summary of Laplace approximation see Barndorff-Nielsen and Cox (1989, Chapter 3).

Then, using Taylor expansion, we can write

$$\begin{aligned} g_n &= \int_a^b e^{-n\{g(\tilde{y})+\frac{1}{2}(\tilde{y}-y)^2 g''(\tilde{y})+\cdots\}}dy \\ &\doteq e^{-ng(\tilde{y})}\int_a^b e^{-\frac{n}{2}(\tilde{y}-y)^2 g''(\tilde{y})}dy \\ &\doteq e^{-ng(\tilde{y})}\sqrt{\frac{2\pi}{ng''(\tilde{y})}}\int_{-\infty}^{\infty}\phi\left(y-\tilde{y};\frac{1}{ng''(\tilde{y})}\right)dy \end{aligned}$$

where $\phi(y - \mu; \sigma^2)$ is the density of $N(\mu, \sigma^2)$. Since ϕ integrates to one,

$$g_n \doteq e^{-ng(\tilde{y})}\sqrt{\frac{2\pi}{ng''(\tilde{y})}}. \tag{3.10}$$

A more detailed analysis gives

$$g_n = e^{-ng(\tilde{y})}\sqrt{\frac{2\pi}{ng''(\tilde{y})}}\left\{1+\frac{5\tilde{\rho}_3^2-3\tilde{\rho}_4}{24n}+O(n^{-2})\right\}, \tag{3.11}$$

where

$$\begin{aligned}\tilde{\rho}_3 &= g^{(3)}(\tilde{y})/\{g''(\tilde{y})\}^{3/2},\\ \tilde{\rho}_4 &= g^{(4)}(\tilde{y})/\{g''(\tilde{y})\}^{2}.\end{aligned}$$

A similar analysis gives

$$\int_a^b h(y)e^{-ng(y)}dy = h(\tilde{y})e^{-ng(\tilde{y})}\sqrt{\frac{2\pi}{ng''(\tilde{y})}}\{1+O(n^{-1})\}. \tag{3.12}$$

A further refinement of the method, which allows $g(y)$ to depend weakly on n, gives

$$\begin{aligned}&\int_a^b e^{-n\{g(y)-\frac{1}{n}\log h(y)\}}dy\\ &= \int_a^b e^{-nq_n(y)}dy, \quad \text{say,}\\ &= e^{-ng(y^*)}h(y^*)\sqrt{\frac{2\pi}{nq_n''(y^*)}}\{1+(5\rho_3^{*2}-3\rho_4^*)/(24n)+O(n^{-2})\},\end{aligned} \tag{3.13}$$

where

$$q_n'(y^*)=0, \rho_j^* = q_n^{(j)}(y^*)/\{q_n''(y^*)\}^{j/2}.$$

The multi-dimensional version of (3.12) is

$$g_n = \int_D h(y)e^{-ng(y)}dy = h(\tilde{y})e^{-ng(\tilde{y})}\frac{(2\pi)^{m/2}}{\sqrt{n|g''(\tilde{y})|}}\{1+O(n^{-1})\},$$

where it is assumed that $g(y)$ takes its minimum in the interior of the region $D\subset\mathbb{R}^m$, where the gradient is zero and the Hessian $g''(\tilde{y})$ is positive definite.

The Laplace approximations are particularly useful in Bayesian inference: see Section 3.9.

3.5 The p^* formula

3.5.1 Introduction

Recall the convention introduced in Chapter 1 that the minimal sufficient statistic based in data x can be re-expressed, by a one-to-one smooth transformation, as $(\hat{\theta}, a)$ where a is ancillary, so that we can write the log-likelihood $l(\theta; x)$ as $l(\theta; \hat{\theta}, a)$. Similarly, we can write the observed information $j(\theta) \equiv j(\theta; x) = j(\theta; \hat{\theta}, a)$.

Under a transformation model, the maximal invariant statistic serves as the ancillary. In a full (m, m) exponential model the MLE is minimal sufficient and no ancillary is called for.

Example 3.1 We consider first the *location model*, which is the simplest example of a transformation model, the general theory of which was described in Chapter 1. We have $X_1, \ldots, X_n$ independent random variables with

$$X_j = \theta + \epsilon_j, \ , j = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent random variables each having the known density function $\exp\{g(\cdot)\}$. The log-likelihood is given by

$$l(\theta) = \sum g(x_j - \theta).$$

Let $a = (a_1, \ldots, a_n)$, where $a_j = x_j - \hat{\theta}$: it is readily shown that a is ancillary. We may write $x_j = a_j + \hat{\theta}$, so that the log- likelihood may be written

$$l(\theta; \hat{\theta}, a) = \sum g(a_j + \hat{\theta} - \theta).$$

Example 3.2 As a further example, let $X_1, \ldots, X_n$ be an independent sample from a full (m, m) exponential density

$$\exp\{x^T\theta - k(\theta) + D(x)\}.$$

The log-likelihood is, ignoring an additive constant,

$$l(\theta) = \sum x_j^T\theta - nk(\theta).$$

Since $\hat{\theta}$ satisfies the likelihood equation

$$\sum x_j - nk'(\theta) = 0,$$

the log-likelihood may be written

$$l(\theta; \hat{\theta}) = nk'(\hat{\theta})^T\theta - nk(\theta).$$

3.5.2 Approximate ancillaries

Outside full exponential family and transformation models it is often difficult to construct an appropriate ancillary a such that $(\hat{\theta}, a)$ is minimal sufficient, and it is usually necessary to work with notions of *approximate ancillarity*. A statistic a is, broadly speaking, approximately ancillary if its asymptotic distribution does not depend on the parameter. Useful approximate ancillaries can often be constructed from signed log-likelihood ratios or from score statistics.

Severini (2000, Section 6.6) gives a summary of techniques for construction of approximate ancillaries. One particularly important approximate ancillary is the **Efron–Hinkley ancillary** (Efron and Hinkley, 1978). Consider the case of a scalar parameter θ and let, as before, i and j be the expected and observed information and let $l_\theta = \frac{\partial l}{\partial \theta}$, $l_{\theta\theta} = \frac{\partial^2 l}{\partial \theta^2}$ etc. Use the notation $\nu_{2,1} = \mathbb{E}(l_{\theta\theta} l_\theta; \theta)$, $\nu_{2,2} = \mathbb{E}(l_{\theta\theta} l_{\theta\theta}; \theta)$, $\nu_2 = \mathbb{E}(l_{\theta\theta})$. Define

$$\gamma = i^{-1}(\nu_{2,2} - \nu_2^2 - i^{-1}\nu_{2,1}^2)^{1/2},$$

and use circumflex to denote evaluation at $\hat{\theta}$. Then the Efron–Hinkley ancillary is defined by

$$a = (\hat{i}\hat{\gamma})^{-1}(\hat{j} - \hat{i}).$$

A particularly powerful result is the following. For a location model with θ as the location parameter, if $\hat{i}$ and $\hat{j}$ denote respectively the Fisher and observed information evaluated at $\hat{\theta}$,

$$\frac{\mathrm{var}\,(\hat{\theta} \mid a) - \hat{j}^{-1}}{\mathrm{var}\,(\hat{\theta} \mid a) - \hat{i}^{-1}} = O_p(n^{-1/2}),$$

where a denotes the Efron–Hinkley ancillary: $\hat{j}^{-1}$ provides a more accurate estimate of the conditional variance of $\hat{\theta}$ given a.

A simple example of construction of this ancillary is provided by the *exponential hyperbola*. Under this model, $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote independent pairs of independent exponential random variables, such that each X_j has mean $1/\theta$ and each Y_j has mean θ. The minimal sufficent statistic for the model may be written as $(\hat{\theta}, a)$, where $\hat{\theta} = (\bar{y}/\bar{x})^{1/2}$ is the MLE and $a = (\bar{x}\bar{y})^{1/2}$ is an (exact) ancillary. Simple calculations show that the Efron–Hinkley ancillary is

$$\sqrt{(2n)}(\bar{y}/\hat{\theta} - 1) = \sqrt{(2n)}\{(\bar{x}\bar{y})^{1/2} - 1\},$$

which is in fact also exactly ancillary.

3.5.3 The formula

A striking result due to Barndorff-Nielsen (1983) is that the conditional density function $f(\hat{\theta};\theta \mid a)$ for the MLE $\hat{\theta}$ given an ancillary statistic a is, in wide generality, exactly or approximately equal to

$$p^*(\hat{\theta};\theta \mid a) = c(\theta,a)|j(\hat{\theta})|^{1/2}\exp\{l(\theta)-l(\hat{\theta})\}, \tag{3.14}$$

i.e.

$$f(\hat{\theta};\theta \mid a) \doteq p^*(\hat{\theta};\theta \mid a).$$

In (3.14), $c(\theta,a)$ is a normalising constant, determined, usually numerically, so that the integral of p^* with respect to $\hat{\theta}$, for fixed a, equals 1.

Equation (3.14) gives the exact conditional distribution of the MLE for a considerable range of models. In particular, this is the case for virtually all transformation models, for which $c(\theta,a)$ is independent of θ. The location-scale model provides a prototypical example, with the configuration statistic as the ancillary. Among models for which (3.14) is exact, but which is not a transformation model, is the inverse Gaussian distribution. Under many of these models the norming constant c equals $(2\pi)^{-d/2}$ exactly, $d = \dim(\theta)$. In general, $c = c(\theta,a) = (2\pi)^{-d/2}\bar{c}$, where $\bar{c} = 1 + O(n^{-1})$. Outside the realm of exactness cases, (3.14) is quite generally accurate to relative error of order $O(n^{-1})$:

$$f(\hat{\theta};\theta \mid a) = p^*(\hat{\theta};\theta \mid a)\big(1+O(n^{-1})\big),$$

for *any fixed* $\hat{\theta}$. For $\hat{\theta}$ of the form $\hat{\theta} = \theta + O_p(n^{-1/2})$, which is the situation we are primarily interested in in practice, the approximation achieves higher accuracy, the relative error in fact being of order $O(n^{-3/2})$. Severini (2000, Section 6.5) provides an account of definitions of approximate ancillarity which are strong enough for the relative error to be of order $O(n^{-1})$ for values of the argument $\hat{\theta}$ of this latter form without a being exactly ancillary.

Comparing (3.9) with (3.14), we see that the p^* formula is equivalent to the saddlepoint approximation in exponential families, with θ the natural parameter.

Integration of the p^* formula in the case of scalar θ to obtain an approximation to the distribution function of the MLE is intricate: a very clear description is given by Barndorff-Nielsen (1990). Write

$$r_t \equiv r_t(\theta) = \mathrm{sgn}(t-\theta)\sqrt{2(l(t;t,a)-l(\theta;t,a))},$$

and let

$$v_t \equiv v_t(\theta) = j(t;t,a)^{-1/2}\{l_{;\hat{\theta}}(t;t,a) - l_{;\hat{\theta}}(\theta;t,a)\},$$

in terms of the sample space derivative $l_{;\hat{\theta}}$ defined by

$$l_{;\hat{\theta}}(\theta;\hat{\theta},a) = \frac{\partial}{\partial\hat{\theta}} l(\theta;\hat{\theta},a),$$

and with j the observed information. Then

$$\Pr_\theta(\hat{\theta} \le t \mid a) = \Phi\{r_t^*(\theta)\}\{1 + O(n^{-3/2})\},$$

where $r_t^*(\theta) = r_t + r_t^{-1}\log\{v_t/r_t\}$, for $t = \theta + O(n^{-1/2})$.

The random quantity $r^*(\theta)$ corresponding to $r_t^*(\theta)$ is an approximate pivot, conditional on the ancillary, in the sense that its distribution is close to normal. We may view $r^*(\theta)$ as a modified form of the signed root likelihood ratio statistic

$$r(\theta) = \mathrm{sgn}(\hat{\theta} - \theta)[2\{l(\hat{\theta};\hat{\theta},a) - l(\theta;\hat{\theta},a)\}]^{1/2}$$

which improves the accuracy of the normal approximation.

To define $r^*(\theta)$ formally,

$$r^*(\theta) = r(\theta) + r(\theta)^{-1}\log\{v(\theta)/r(\theta)\},$$

where

$$v(\theta) = \hat{j}^{-1/2}\{l_{;\hat{\theta}}(\hat{\theta};\hat{\theta},a) - l_{;\hat{\theta}}(\theta;\hat{\theta},a)\},$$

with $\hat{j}$ denoting evaluation of the observed information at $\hat{\theta}$.

We have that $r^*(\theta)$ is distributed as $N(0,1)$ to (relative) error of order $O(n^{-3/2})$:

$$\Pr_\theta\{r^*(\theta) \le t \mid a\} = \Phi(t)\{1 + O(n^{-3/2})\},$$

for $t = O(1)$.

The limits of an approximate $(1-2\alpha)$ confidence interval for θ may be found as those θ such that $\Phi\{r^*(\theta)\} = \alpha, 1-\alpha$.

The above is expressed in terms of a one-parameter model. Versions of the approximation appropriate to inference about a scalar parameter of interest in the presence of a nuisance parameter are more complicated. To present just the key formula, suppose that the model depends on a multi-dimensional parameter $\theta = (\psi, \lambda)$, with ψ a scalar parameter of interest, with λ nuisance. Then the $N(0,1)$ approximation to the distribution of the signed root likelihood ratio statistic $r_{\mathrm{p}} = \mathrm{sgn}(\hat{\psi} - \psi)[2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi)\}]^{1/2}$ is improved by analytically adjusted versions of the form

$$r_{\mathrm{a}}(\psi) = r_{\mathrm{p}}(\psi) + r_{\mathrm{p}}(\psi)^{-1}\log(v_{\mathrm{p}}(\psi)/r_{\mathrm{p}}(\psi)),$$

that are distributed as $N(0,1)$, conditionally on a (and hence unconditionally), to error of order $O(n^{-3/2})$.

Now the statistic $v_{\rm p}$ is defined (Barndorff-Nielsen, 1986) by

$$v_{\rm p}(\psi) = \left| \begin{array}{c} l_{;\hat\theta}(\hat\theta) - l_{;\hat\theta}(\psi,\hat\lambda_\psi) \\ l_{\psi;\hat\theta}(\psi,\hat\lambda_\psi) \end{array} \right| / \{|j_{\psi\psi}(\psi,\hat\lambda_\psi)|^{1/2}|j(\hat\theta)|^{1/2}\}.$$

Here, as previously, the log-likelihood function has been written as $l(\theta;\hat\theta,a)$, with $(\hat\theta,a)$ minimal sufficient and a ancillary, $\hat\lambda_\psi$ denotes the MLE of λ for given ψ, and

$$l_{;\hat\theta}(\theta) \equiv l_{;\hat\theta}(\theta;\hat\theta,a) = \frac{\partial}{\partial\hat\theta} l(\theta;\hat\theta,a), \quad l_{\psi;\hat\theta}(\theta) \equiv l_{\psi;\hat\theta}(\theta;\hat\theta,a) = \frac{\partial^2}{\partial\psi\partial\hat\theta} l(\theta;\hat\theta,a).$$

Again, j denotes the observed information matrix and $j_{\psi\psi}$ denotes the (ψ,ψ) component of the observed information matrix.

A key drawback to use of $r_{\rm a}(\psi)$ (the same comment is true of $r^*(\theta)$) is the need to calculate sample space derivatives, which necessitates explicit specification of the ancillary a. We have commented that this is difficult in general, outside full exponential family and transformation models. Several methods of approximation to $r_{\rm a}(\psi)$ which avoid this by approximating to the sample space derivatives have been developed. A computationally attractive approximation based on orthogonal parameters is described by DiCiccio and Martin (1993): recall that in the case we are assuming here of a scalar parameter of interest it is always possible to find a parameterisation in which the interest parameter ψ and the nuisance parameters λ are orthogonal. The DiCiccio and Martin (1993) approximation replaces $v_{\rm p}(\psi)$ by

$$\tilde v_{\rm p}(\psi) = l_\psi(\psi,\hat\lambda_\psi) \frac{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{1/2} i_{\psi\psi}(\hat\theta)^{1/2}}{|j(\hat\theta)|^{1/2} i_{\psi\psi}(\psi,\hat\lambda_\psi)^{1/2}},$$

with the usual partitioning of the observed information j and the Fisher information i, and with l_ψ denoting, as before, the derivative of the log-likelihood l with respect to the parameter of interest. The corresponding adjusted version of the signed root likelihood ratio statistic,

$$\tilde r_{\rm a}(\psi) = r_{\rm p}(\psi) + r_{\rm p}(\psi)^{-1} \log(\tilde v_{\rm p}(\psi)/r_{\rm p}(\psi)),$$

is distributed as $N(0,1)$ to error of order $O(n^{-1})$, rather than order $O(n^{-3/2})$ for $r_{\rm a}(\theta)$. A further point should be noted, that $r_{\rm a}$ is parameterisation invariant, with respect to interest-respecting reparameterisation, while $\tilde r_{\rm a}$ depends on the orthogonal parameterisation adopted. Other approximations to $r_{\rm a}$, due to various authors and with the same property of being distributed as $N(0,1)$ to error of order $O(n^{-1})$, are detailed by Severini (2000, Chapter 7).

3.5.4 An example: Normal distribution with known coefficient of variation

Let $X_1, \ldots, X_n$ denote independent normally distributed random variables each with mean θ and standard deviation $r\theta$ where $\theta > 0$ and the coefficient of variation r is known; for simplicity take $r = 1$. This distribution is widely assumed in many biological and agricultural problems. The minimal sufficient statistic for the model may be written $(\hat{\theta}, a)$ where

$$a = \sqrt{n}\frac{(\sum x_j^2)^{1/2}}{\sum x_j}$$

is easily seen to be an exactly ancillary statistic and

$$\hat{\theta} = \frac{(\sum x_j^2)^{1/2}}{\sqrt{n}} \; \frac{2|a|}{(1+4a^2)^{1/2} + \mathrm{sgn}(a)}$$

is the maximum likelihood estimator of θ. Assume that $a > 0$, which occurs with probability rapidly approaching 1 as $n \to \infty$.

The log-likelihood function may be written

$$l(\theta; \hat{\theta}, a) = -\frac{n}{2\theta^2}\big[q^2\hat{\theta}^2 - \frac{2q\theta\hat{\theta}}{a}\big] - n \log\, \theta$$

where

$$q = \frac{(1+4a^2)^{1/2} + 1}{2a}.$$

It follows that

$$\begin{aligned} p^*(\hat{\theta}; \theta \mid a) &= \frac{\sqrt{n}\bar{c}}{\sqrt{(2\pi)}} \left(\frac{\hat{\theta}}{\theta}\right)^{n-1} \frac{1}{\theta}(1+q^2)^{1/2} \\ &\quad \times \exp\big\{ -\frac{n}{2}\big[\frac{q^2}{\theta^2}(\hat{\theta}^2 - \theta^2) - \frac{2q}{a\theta}(\hat{\theta} - \theta)\big]\big\}. \end{aligned}$$

This expression may be rewritten as

$$\begin{aligned} p^*(\hat{\theta}; \theta \mid a) &= \frac{\sqrt{n}\bar{c}}{\sqrt{(2\pi)}} \exp\big\{\frac{n}{2}(q - 1/a)^2\big\}(1+q^2)^{1/2} \left(\frac{\hat{\theta}}{\theta}\right)^{n-1} \frac{1}{\theta} \\ &\quad \times \exp\big\{ -\frac{n}{2}q^2(\hat{\theta}/\theta - 1/(aq))^2\big\}. \end{aligned}$$

It may be shown that the exact conditional density of $\hat{\theta}$ given a is of the form

$$p(\hat{\theta}; \theta \mid a) = b(a) \left(\frac{\hat{\theta}}{\theta}\right)^{n-1} \frac{1}{\theta} \exp\big\{ -\frac{n}{2}q^2(\hat{\theta}/\theta - 1/(aq))^2\big\},$$

where $b(a)$ is a normalizing constant depending on a. Hence, the conditional density approximation is exact for this model. A $N(0,1)$ approximation to the conditional distribution of $r^*(\theta)$ is not exact, but highly accurate.

3.5.5 The score function

We now consider the application of the p^*-formula to the score vector. Given an ancillary a, the MLE $\hat{\theta}$ and the score vector $U = \nabla l$, with components l_r, will in general be in one-to-one correspondence for a region of values of $\hat{\theta}$ around the true parameter value θ, and this region will carry all the probability mass, except for an asymptotically negligible amount. The Jacobian of the transformation from $\hat{\theta}$ to the vector of derivatives $l_r = l_r(\theta; \hat{\theta}, a)$ is the matrix $l_;$ of mixed second-order log model derivatives

$$l_{r;s} = l_{r;s}(\theta; \hat{\theta}, a) = \frac{\partial}{\partial \theta^r} \frac{\partial}{\partial \hat{\theta}^s} l(\theta; \hat{\theta}, a).$$

As an example of calculation of these derivatives, consider the *location model*. We saw above that

$$l(\theta; \hat{\theta}, a) = \sum g(a_j + \hat{\theta} - \theta).$$

Then

$$l_; \equiv l_{\theta;\hat{\theta}} = -\sum g''(a_j + \hat{\theta} - \theta).$$

From (3.14) an approximation of high accuracy to the conditional density of the score vector given a is provided by

$$p(u; \theta \mid a) \doteq p^*(u; \theta \mid a),$$

where

$$p^*(u; \theta \mid a) = c(\theta, a) |\hat{j}|^{1/2} |l; |^{-1} e^{l - \hat{l}}.$$

Note than an Edgeworth or saddlepoint apporximation to the marginal distribution of U is easy to obtain in the case when U is a sum of IID variates.

3.6 Conditional inference in exponential families

A particularly important inference problem to which ideas of this Chapter apply concerns inference about the natural parameter of an exponential family model.

Suppose that $X_1, \ldots, X_n$ are independent, identically distributed from the exponential family density

$$f(x; \psi, \lambda) = \exp\{\psi \tau_1(x) + \lambda \tau_2(x) - d(\psi, \lambda) - Q(x)\},$$

where we will suppose for simplicity that the parameter of interest ψ and the nuisance parameter λ are both scalar.

The natural statistics are $T = n^{-1}\sum \tau_1(x_i)$ and $S = n^{-1}\sum \tau_2(x_i)$. We know from the general properties of exponential families (Chapter 1) that the conditional distribution of $X = (X_1, \ldots, X_n)$ given $S = s$ depends only on ψ, so that inference about ψ may be derived from a conditional likelihood, given s.

The log-likelihood based on the full data $x_1, \ldots, x_n$ is

$$n\psi t + n\lambda s - nd(\psi, \lambda),$$

ignoring terms not involving ψ and λ, and the conditional log-likelihood function is the full log-likelihood minus the log-likelihood function based on the marginal distribution of S. We consider an approximation to the marginal distribution of S, based on a saddlepoint approximation to the density of S, evaluated at its observed value s.

The cumulant generating function of $\tau_2(X_i)$ is given by

$$K(z) = d(\psi, \lambda + z) - d(\psi, \lambda).$$

Write $d_\lambda(\psi, \lambda) = \frac{\partial}{\partial\lambda} d(\psi, \lambda)$ and $d_{\lambda\lambda}(\psi, \lambda) = \frac{\partial^2}{\partial\lambda^2} d(\psi, \lambda)$. The saddlepoint equation is then given by

$$d_\lambda(\psi, \lambda + \hat{z}) = s.$$

With s the observed value of S, the likelihood equation for the model with ψ held fixed is

$$ns - nd_\lambda(\psi, \hat{\lambda}_\psi) = 0,$$

so that $\lambda + \hat{z} = \hat{\lambda}_\psi$, where $\hat{\lambda}_\psi$ denotes the maximum likelihood estimator of λ for fixed ψ. Applying the saddlepoint approximation, ignoring constants, we therefore approximate the marginal likelihood function based on S as

$$|d_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2} \exp\{n[d(\psi, \hat{\lambda}_\psi) - d(\psi, \lambda) - (\hat{\lambda}_\psi - \lambda)s]\};$$

the resulting approximation to the conditional log-likelihood function is given by

$$\begin{aligned} &n\psi t + n\hat{\lambda}_\psi^T s - nd(\psi, \hat{\lambda}_\psi) + \frac{1}{2}\log|d_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| \\ &\equiv l(\psi, \hat{\lambda}_\psi) + \frac{1}{2}\log|d_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|. \end{aligned}$$

The form of this conditional log-likelihood indicates that instead of just using the profile log-likelihood of ψ, an adjustment term should be added. This notion is developed in detail in Section 3.8 below.

3.7 Bartlett correction

The first-order χ^2 approximation to the distribution of the likelihood ratio statistic $w(\psi)$ can be expressed as

$$\Pr_\theta\{w(\psi) \leq \omega^\circ\} = \Pr\{\chi^2_q \leq \omega^\circ\}\{1 + O(n^{-1})\},$$

where q is the dimension of ψ and the full parameter vector is $\theta = (\psi, \lambda)$, with λ nuisance. The χ^2 approximation has relative error of order $O(n^{-1})$.

In the case of independent, identically distributed sampling, it can be shown that

$$\mathbb{E}_\theta w(\psi) = q\{1 + b(\theta)/n + O(n^{-2})\},$$

and so $\mathbb{E}_\theta w'(\psi) = q\{1 + O(n^{-2})\}$, where $w' = w/\{1 + b(\theta)/n\}$.

This adjustment procedure, of replacing w by w', is known as **Bartlett correction**. In spite of its simplicity, this device yields remarkably good results under continuous models, the reason being that division by $\{1 + b(\theta)/n\}$ adjusts, in fact, not only the mean but simultaneously all the cumulants—and hence the whole distribution—of w towards those of χ^2_q. It can be shown that

$$\Pr_\theta\{w'(\psi) \leq \omega^\circ\} = \Pr\{\chi^2_q \leq \omega^\circ\}\{1 + O(n^{-2})\}.$$

In practice, because of the (possible) presence of an unknown nuisance parameter λ, $b(\theta)$ may be unknown. If $b(\theta)$ is replaced by $b(\psi, \hat{\lambda}_\psi)$, the above result still holds, even to $O(n^{-2})$. An explicit expression for $b(\theta)$ is given by Barndorff-Nielsen and Cox (1994, Chapter 6).

Note that the effect of the Bartlett correction is due to the special character of the likelihood ratio statistic, and the same device applied to, for instance, the score test does not have a similar effect. Also, under discrete models this type of adjustment does not generally lead to an improved χ^2 approximation.

3.8 Modified profile likelihood

The profile likelihood $L_{\mathrm{p}}(\psi)$ for a parameter of interest ψ can largely be thought of as if it were a genuine likelihood. However, this amounts to behaving as if the nuisance parameter χ over which the maximisation has been carried out were known. Inference on ψ based on treating $L_{\mathrm{p}}(\psi)$ as a proper likelihood may therefore be grossly misleading if the data contain insufficient information about χ, as is likely to happen, for instance, if the dimension of χ is large. Modified profile likelihood is intended as a remedy for this type of problem.

The modified profile likelihood $\tilde{L}_{\rm p}(\psi)$ for a parameter of interest ψ, with nuisance parameter χ, due to Barndorff-Nielsen (1983), is defined by

$$\tilde{L}_{\rm p}(\psi) = M(\psi) L_{\rm p}(\psi), \tag{3.15}$$

where M is a modifying factor

$$M(\psi) = \left| \frac{\partial \hat{\chi}}{\partial \hat{\chi}_\psi} \right| |\hat{j}_\psi|^{-1/2}.$$

Here $|\cdot|$ denotes the absolute value of a matrix determinant, and $\partial\hat{\chi}/\partial\hat{\chi}_\psi$ is the matrix of partial derivatives of $\hat{\chi}$ with respect to $\hat{\chi}_\psi$, where $\hat{\chi}$ is considered as a function of $(\hat{\psi}, \hat{\chi}_\psi, a)$. Also, $\hat{j}_\psi = j_{\chi\chi}(\psi, \hat{\chi}_\psi)$, the observed information on χ assuming ψ is known. An instructive example to look at to grasp the notation is the case of $X_1, \ldots, X_n$ independent, identically distributed $N(\mu, \sigma^2)$. Here we see that $\hat{\sigma}^2_\mu = \frac{1}{n}\sum(X_j - \mu)^2 = \hat{\sigma}^2 + (\hat{\mu} - \mu)^2$.

The modified profile likelihood $\tilde{L}_{\rm p}$ is, like $L_{\rm p}$, parametrisation invariant. An alternative expression for the modifying factor M is

$$M(\psi) = |l_{\chi;\hat{\chi}}(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a)|^{-1} \times |j_{\chi\chi}(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a)|^{1/2}. \tag{3.16}$$

Identity (3.16) follows from the likelihood equation for $\hat{\chi}_\psi$:

$$l_\chi(\psi, \hat{\chi}_\psi(\hat{\psi}, \hat{\chi}, a); \hat{\psi}, \hat{\chi}, a) = 0.$$

Differentiation with respect to $\hat{\chi}$ yields

$$l_{\chi\chi}(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a) \frac{\partial \hat{\chi}_\psi}{\partial \hat{\chi}} + l_{\chi;\hat{\chi}}(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a) = 0,$$

from which (3.16) follows.

Asymptotically, $\tilde{L}_{\rm p}$ and $L_{\rm p}$ are equivalent to first-order. A justification for using $\tilde{L}_{\rm p}$ rather than $L_{\rm p}$ is that (3.15) arises as a higher-order approximation to a marginal likelihood for ψ when such a marginal likelihood function is available, and to a conditional likelihood for ψ when this is available.

Specifically, suppose that the density $f(\hat{\psi}, \hat{\chi}; \psi, \chi \mid a)$ factorises, either as

$$f(\hat{\psi}, \hat{\chi}; \psi, \chi \mid a) = f(\hat{\psi}; \psi \mid a) f(\hat{\chi}; \psi, \chi \mid \hat{\psi}, a) \tag{3.17}$$

or as

$$f(\hat{\psi}, \hat{\chi}; \psi, \chi \mid a) = f(\hat{\chi}; \psi, \chi \mid a) f(\hat{\psi}; \psi \mid \hat{\chi}, a). \tag{3.18}$$

In the case (3.17), (3.15) can be obtained as an approximation (using the p^*-formula) to the marginal likelihood for ψ based on $\hat{\psi}$ and conditional on a, i.e. to the likelihood for ψ determined by $f(\hat{\psi}; \psi \mid a)$. Similarly, under (3.18) the same expression (3.15) is obtained as an approximation to the conditional likelihood for ψ given $\hat{\chi}$ and a i.e. to the likelihood for ψ obtained from $f(\hat{\psi}; \psi \mid \hat{\chi}, a)$. Proofs of both results are given by Barndorff-Nielsen and Cox (1994, Chapter 8).

Sometimes the joint conditional distribution of $\hat{\psi}$ and $\hat{\chi}_\psi$ may be factorised as

$$f(\hat{\psi}, \hat{\chi}_\psi; \psi, \chi \mid a) = f(\hat{\chi}_\psi; \psi, \chi \mid a) f(\hat{\psi}; \psi \mid \hat{\chi}_\psi, a),$$

while (3.18) does not hold. In this case, (3.15) may be obtained as an approximation to $f(\hat{\psi}; \psi \mid \hat{\chi}_\psi, a)$, considered as a pseudo-likelihood for ψ.

Note that if $\hat{\chi}_\psi$ does not depend on ψ,

$$\hat{\chi}_\psi \equiv \hat{\chi}, \tag{3.19}$$

then

$$\tilde{L}_{\mathrm{p}}(\psi) = |\hat{j}_\psi|^{-1/2} L_{\mathrm{p}}(\psi). \tag{3.20}$$

In the case that ψ and χ are orthogonal, which is a weaker assumption than (3.19), we have that (3.19) holds to order $O(n^{-1})$, as does (3.20).

The version of modified profile likelihood defined by (3.20) was first presented by Cox and Reid (1987). It is easy to construct and seems to give reasonable results in applications. It is easier to compute than (3.15), but is not invariant with respect to one-to-one transformations of χ which leave the parameter of interest fixed. A simple Bayesian motivation for (3.20) may be given. Let ψ and the nuisance parameter χ be orthogonal, and let the prior density of ψ and χ be $\pi(\psi, \chi)$. Then the posterior density of ψ is proportional to

$$\int \exp\{l(\psi, \chi)\} \pi(\psi, \chi) d\chi. \tag{3.21}$$

We consider this at a fixed value of ψ. As a function of χ, $l(\psi, \chi)$ achieves its maximum at $\chi = \hat{\chi}_\psi$. Expanding about this point using Laplace's method, as given by (3.10), shows that (3.21) is approximately

$$(2\pi)^{d_\chi/2} \pi(\psi, \hat{\chi}_\psi) \exp\{l(\psi, \hat{\chi}_\psi)\} / |\hat{j}_\psi|^{1/2},$$

with d_χ denoting the dimension of χ. Now argue as follows. As ψ varies in the range of interest, within $O(n^{-1/2})$ of $\hat{\psi}$, $\hat{\chi}_\psi$ varies by $O_p(n^{-1})$, by orthogonality, and therefore so too does the term involving the prior density.

Because of its dependence on ψ, the factor involving the determinant varies by $O(n^{-1/2})$, while the part depending on the likelihood is $O(1)$. Therefore, ignoring an error of order $O(n^{-1})$, inference about ψ can be based on an effective log-likelihood of

$$l(\psi, \hat{\chi}_\psi) - \frac{1}{2} \log |\hat{j}_\psi|,$$

as given by (3.20).

3.9 Bayesian asymptotics

In this section we review briefly the asymptotic theory of Bayesian inference. The results provide demonstration of the application of asymptotic approximations discussed earlier, in particular Laplace approximations. Key references in such use of Laplace approximation in Bayesian asymptotics include Tierney and Kadane (1986) and Tierney, Kass and Kadane (1989).

The key result is that the posterior distribution given data x is asymptotically normal. Write

$$\pi_n(\theta \mid x) = f(x;\theta)\pi(\theta) / \int f(x;\theta)\pi(\theta) d\theta$$

for the posterior density. Denote by $\hat{\theta}$ the MLE.

For θ in a neighbourhood of $\hat{\theta}$ we have, by Taylor expansion,

$$\log \left\{ \frac{f(x;\theta)}{f(x;\hat{\theta})} \right\} \doteq -\tfrac{1}{2}(\theta - \hat{\theta})^T j(\hat{\theta})(\theta - \hat{\theta}).$$

Provided the likelihood dominates the prior, we can approximate $\pi(\theta)$ in a neighbourhood of $\hat{\theta}$ by $\pi(\hat{\theta})$. Then we have

$$f(x;\theta)\pi(\theta) \doteq f(x;\hat{\theta})\pi(\hat{\theta}) \exp\{-\tfrac{1}{2}(\theta - \hat{\theta})^T j(\hat{\theta})(\theta - \hat{\theta})\},$$

so that, to first order,

$$\pi_n(\theta \mid x) \sim N\big(\hat{\theta}, j^{-1}(\hat{\theta})\big).$$

A more natural approximation to the posterior distribution when the likelihood does not dominate the prior is obtained if we expand about the posterior mode $\hat{\theta}_\pi$, which maximises $f(x;\theta)\pi(\theta)$. An analysis similar to the above then gives

$$\pi_n(\theta \mid x) \sim N\big(\hat{\theta}_\pi, j_\pi^{-1}(\hat{\theta}_\pi)\big),$$

where j_π is minus the matrix of second derivatives of $f(x;\theta)\pi(\theta)$.

A more accurate approximation to the posterior is provided by the following. We have

$$\begin{aligned}\pi_n(\theta \mid x) &= f(x;\theta)\pi(\theta) / \int f(x;\theta)\pi(\theta)d\theta \\ &\doteq \frac{c\exp\{l(\theta;x)\}\pi(\theta)}{\exp\{l(\hat\theta;x)\}|j(\hat\theta)|^{-1/2}\pi(\hat\theta)},\end{aligned}$$

by Laplace approximation of the denominator. We can rewrite as

$$\pi_n(\theta \mid x) \doteq c|j(\hat\theta)|^{1/2}\exp\{l(\theta) - l(\hat\theta)\} \times \{\pi(\theta)/\pi(\hat\theta)\};$$

note the similarity to the density approximation (3.14) for $\hat\theta$.

We can consider also use of the Laplace approximation to approximate to the posterior expectation of a function $g(\theta)$ of interest,

$$\mathbb{E}\{g(\theta) \mid x\} = \frac{\int g(\theta)e^{n\bar l_n(\theta)}\pi(\theta)d\theta}{\int e^{n\bar l_n(\theta)}\pi(\theta)d\theta},$$

where $\bar l_n = n^{-1}\sum_{i=1}^n \log f(x_i;\theta)$ is the average log-likelihood function. Recall that such expectations arise as the solutions to Bayes decision problems. It turns out to be more effective to rewrite the integrals as

$$\mathbb{E}\{g(\theta) \mid x\} = \frac{\int e^{n\{\bar l_n(\theta)+q(\theta)/n\}}d\theta}{\int e^{n\{\bar l_n(\theta)+p(\theta)/n\}}d\theta}$$

and to use the version (3.13) of the Laplace approximation. Applying this to the numerator and denominator gives

$$\begin{aligned}E\{g(\theta) \mid x\} &\doteq \frac{e^{n\bar l_n(\theta^*)+q(\theta^*)}}{e^{n\bar l_n(\tilde\theta)+p(\tilde\theta)}} \\ &\quad \times \frac{\{-n\bar l_n''(\tilde\theta) - p''(\tilde\theta)\}^{1/2}}{\{-n\bar l_n''(\theta^*) - q''(\theta^*)\}^{1/2}} \frac{\{1+O(n^{-1})\}}{\{1+O(n^{-1})\}}\end{aligned}$$

where θ^* maximises $n\bar l_n(\theta) + \log g(\theta) + \log\pi(\theta)$ and $\tilde\theta$ maximises $n\bar l_n(\theta) + \log\pi(\theta)$. Further detailed analysis shows that the relative error is, in fact, $O(n^{-2})$. If the integrals are approximated in their unmodified form the result is not as accurate.

Finally, consider the situation where the model depends on a multi-dimensional parameter $\theta = (\psi, \lambda)$, with ψ a scalar interest parameter and λ a nuisance parameter. For values ψ_0 such that $\hat\psi - \psi_0$ is of order $O(n^{-1/2})$, we have

$$Pr(\psi \geq \psi_0 \mid x) = \Phi\{r_{\rm p}(\psi_0)\} + \varphi\{r_{\rm p}(\psi_0)\}\{r_{\rm p}^{-1}(\psi_0) - u_B^{-1}(\psi_0)\} + O(n^{-3/2}),$$

where Φ and φ are the standard normal distribution and density functions respectively, r_p is, as given in Section 3.5.3, the signed root likelihood ratio statistic, and u_B is given by

$$u_B(\psi) = \tilde{\ell}_\psi \frac{\left| -\tilde{\ell}_{\lambda\lambda} \right|^{\frac{1}{2}}}{\left| -\hat{\ell}_{\theta\theta} \right|^{\frac{1}{2}}} \frac{\hat{\pi}}{\tilde{\pi}},$$

where $\pi = \pi(\theta)$ is the prior. Here, letting $\hat{\theta} = \left(\hat{\psi}, \hat{\lambda}\right)$ be the global maximum likelihood estimator of θ and $\tilde{\theta} = \tilde{\theta}(\psi) = (\psi, \hat{\lambda}_\psi)$ be the constrained maximum likelihood estimator of θ for a given value of ψ, evaluation of functions of θ at $\hat{\theta}$ and $\tilde{\theta}$ are denoted by $\hat{}$ and $\tilde{}$, respectively. The value ψ_0 satisfying $\Phi\{r_{\mathrm{p}}(\psi_0)\} + \varphi\{r_{\mathrm{p}}(\psi_0)\}\{r_{\mathrm{p}}^{-1}(\psi_0) - u_B^{-1}(\psi_0)\} = \alpha$ agrees with the posterior $1-\alpha$ quantile of ψ to error of order $O(n^{-2})$.

References

Barndorff-Nielsen, O. E. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–365.

Barndorff-Nielsen, O. E. (1986) Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–322.

Barndorff-Nielsen, O. E. (1990) Approximate interval probabilities. *J.R. Statist. Soc. B* **52**, 485–496.

Barndorff-Nielsen, O. E. and Cox, D. R. (1979) Edgeworth and saddlepoint approximations with statistical applications (with discussion). *J.R. Statist. Soc. B* **41**, 279–312.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989) *Asymptotic Techniques for Use in Statistics*. London: Chapman & Hall.

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

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

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

Daniels, H. E. (1954) Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25**, 631–650.

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

DiCiccio, T. J. and Martin, M. A. (1993) Simple modifications for signed roots of likelihood ratio statistics. *J. Roy. Statist. Soc. B* **55**, 305–316.

Efron, B. and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika* **65**, 457–487.

Feller, W. (1971) *An Introduction to Probability Theory, Volume 2*. New York: Wiley (Second Edition).

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A* **222**, 309–368. [This and the next reference are often debated as the most important single paper in statistical theory. This and the next three references are included here for their importance to the perspective taken in the module.]

Fisher, R. A. (1925) Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700-725.

Fisher, R. A. (1934) Two new properties of mathematical likelihood. *Proc. R. Soc. Lond. A* **144**, 285–307. [The origin of the conditional inference procedure for location-scale families, and therefore of much of the Fisherian viewpoint.]

Fisher, R. A. (1990) *Statistical Methods, Experimental Design and Scientific Inference.* Oxford: Clarendon Press. [A relatively recent reprint of three of Fisher's best known works on statistics.]

Lugannani, R. and Rice, S. (1980) Saddlepoint approximations for the distribution of the sum of independent random variables. *Adv. Appl. Probab.* **12**, 475–490.

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

Severini, T. A. (2000) *Likelihood Methods in Statistics.* Oxford: Clarendon Press.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments amd marginal densities. *J. Amer. Statist. Assoc.* **81**, 82–86.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989) Approximate marginal densities of nonlinear functions. *Biometrika* **76**, 425–433.