STATISTICAL ASYMPTOTICS

0.1 Background

These notes are preliminary reading for the APTS module 'Statistical Asymptotics' which will take place in Nottingham during the week 15-19 April 2013.

Two immediate questions that may come to mind are: (i) What is statistical asymptotics about? and (ii) Why study this topic?

For the purposes of this module, the answer to question (i) is that statistical asymptotics is about the study of statistical models and statistical procedures, such as parameter estimation and hypothesis testing, when the sample size n goes to infinity. Properties such as consistency and asymptotic normality of parameter estimates fall within the domain of statistical asymptotics as they entail statements about what happens as $n \to \infty$.

There are at least 3 distinct answers to question (ii).

1. **Motivating practical approximations**. In most statistical models in a frequentist setting, exact inference is not feasible and so there is a need to develop approximations. Asymptotic (i.e. large n) results often provide convenient approximations which are sufficiently accurate for practical purposes. In Bayesian inference, Laplace's method (an asymptotic procedure) often provides useful approximations to posterior distributions.

2. **Theoretical insight**. Asymptotic analyses often provide valuable theoretical insight into complex models and complex problems. For example, it may help us to identify which aspects of a problem are of most importance.

3. **Theories of optimality**. There is sufficient mathematical tractability within an asymptotic framework to develop various theories of statistical optimality. We shall not have much to say about optimality in this module; see van der Vaart (1998) for discussion of aspects of optimality.

0.2 Module objectives

The objectives of the module are: (i) to provide an overview of the first-order asymptotic theory of statistical statistical inference, with a focus mainly on likelihood-based approaches, but with brief consideration of the more general setting of M-estimators; (ii) to provide an introduction to analytic methods and tools, in particular approximation techniques that are potentially useful in applications; (iii) to provide exposure to key ideas in contemporary statistical theory; and (iv) to provide practice in application of key techniques to particular problems.

0.3 Contents of these notes

These preliminary notes consist of three chapters. The first presents limit theorems from probability theory, multivariable Taylor expansions, plus material on exponential families, likelihood, sufficiency and Bayesian inference. Chapter 2 provides a fairly detailed discussion of the first-order asymptotic theory for parametric likelihoods and maximum likelihood estimators. At the end of this chapter we briefly consider the more general theory of M-estimators. Chapter 3 presents introductory material on higher-order methods including Edgeworth expansions, saddlepoint approximations, Laplace approximations and Bayesian asymptotics. Many results are stated without proof, as some of the derivations are hard, and beyond the scope of the course. However, whenever possible, we shall attempt to provide insight into what is required in the proof.

0.4 How to read these notes

The two main purposes of these notes are:

(i) to summarise the prerequisite material (the locations of the prerequisite material are given below);

(ii) to provide a first pass through the main topics covered in the module (of course, the second pass through the main topics of the module will be during the APTS week itself).

Most of the prerequisite material is in Chapter 1, but you are also strongly encouraged to study §2.3. However, Chapter 1 also contains more advanced material so please do not be surprised or alarmed if there are a number of topics or results you have not seen before. My advice is to read through the notes, trying to understand as much as you can, but identifying those topics, sections or subsections that you have difficulty with. I will be happy to respond to questions on the notes before the APTS week; my email address is given below. Alternatively, please feel free to save your questions for the APTS week itself.

The following subsections of the notes constitute prerequisite material:

- §1.2.1 on modes of convergence in probability theory;
- §1.2.2 on laws of large numbers and the central limit theorem in the $1D$ case (note: this subsection goes beyond prerequisite material with

coverage of multivariate limit theorems, which are straightforward generalisations of their univariate versions, and topics such as the delta method);

- §1.2.4 on $o(.)$, $O(.)$, $o_p(.)$ and $O_p(.)$ notation;
- §1.3.1 on Taylor expansions in the univariate and multivariate cases (note: the index notation may be unfamiliar);
- §1.4.1 on exponential families;
- §1.5.1 on likelihoods;
- §1.5.2 on score functions and information;
- §1.5.4 on sufficiency;
- §1.6 on Bayesian inference;
- §2.3 on asymptotic normality of maximum likelihood estimators.

0.5 Literature

In the last 20 years or so a number of excellent books on statistical inference and statistical asymptotics have been published. A selection of these includes (in alphabetical order): Barndorff-Nielsen and Cox (1989, 1994), Davison (2003), Ferguson (1995), Pace and Salvan (1997), Severini (2000) and Young and Smith (2005). A fine book on statistical asymptotics which is less rooted in parametric inference is van der Vaart (1998), which covers a huge amount of ground in concise fashion. An accessible and extensive account of saddlepoint and Laplace approximations is given in the book by Butler (2007). Depending on the interests of the student, any of these books could provide an excellent basis for further study.

0.6 Acknowledgement

I am grateful to Professor Alastair Young of Imperial College, the previous APTS lecturer for Statistical Asymptotics, for kindly making all his notes and slides available to me. However, with a topic such as this, it is inevitable that any two lecturers will want to present the material in somewhat different ways. This has indeed been the case here and consequently, although much of the overall structure is similar, there have been some quite substantial changes to these notes. Please email feedback or notification of errors to **Andrew.Wood@nottingham.ac.uk**

1 Background Material

1.1 Introduction

Statistical asymptotics draws from a variety of sources including (but not restricted to) probability theory, analysis (e.g. Taylor's theorem), and of course the theory of statistical inference. In this opening chapter we give a brief and selective summary of the material from these sources that will be needed in the module.

In §1.2 we review basic results from probability theory on different modes of convergence relevant to sequences of random variables and random vectors, and then go on to state the classical laws of large numbers and the central limit theorem. The moment generating function and the cumulant generating function are also defined, and the o_p and O_p notation for quantifying the limiting behaviour of a sequence of random variables is introduced. In §1.3 we present some further mathematical material including details on Taylor expansions. For many readers, the index notation introduced in §1.3.1 may be unfamiliar. The remaining sections of the chapter review key elements of statistical inference including exponential families, likelihood, sufficiency and Bayesian inference.

1.2 Basic results from probability theory

1.2.1 Modes of convergence

Various different types of convergence that arise in probability theory are recalled below.

Convergence in distribution: univariate case. A sequence of real-valued random variables $\{Y_1, Y_2, \ldots\}$ is said to converge in distribution if there exists a (cumulative) 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 $Y_n \xrightarrow{d} Y$.

Convergence in distribution: vector case. The extension to random vectors is immediate once we have defined the distribution function of a random vector: if $Y = (Y_1, \ldots, Y_d)^{\mathrm{T}}$ is a random vector and $y = (y_1, \ldots, y_d)^{\mathrm{T}}$ then the distribution function of Y is given by

$$F(y) = P(Y_1 \leq y_1, \ldots, Y_d \leq y_d), \quad y \in \mathbb{R}^d.$$

Now 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.

Remark. In this module, the limiting distribution will usually be a familiar one, e.g. a univariate or multivariate normal distribution or a chi-squared distribution.

Convergence in probability. A sequence $\{Y_1, Y_2, \ldots\}$ of real random variables 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: for this definition 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 $\mathbb{R}^d$.

Remark. Convergence in probability implies convergence in distribution. An important special case is where the sequence converges in probability to a constant, c; i.e. $Y_n \xrightarrow{p} Y$, where $P(Y = c) = 1$. In this case convergence in probability is equivalent to convergence in distribution.

Almost sure convergence. A stronger mode of convergence is almost sure convergence, in the sense that almost sure convergence implies convergence in probability but the implication does not go the other way. 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 $Y_n \xrightarrow{a.s} Y$.

$\mathbf{L^p}$ convergence. 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 $Y_n \xrightarrow{L_p} Y$. L^p convergence implies convergence in probability (cf. the Markov and Chebychev inequalities; see below).

Exercise. *Chebychev's inequality states that, for any real random variable Y, and any $\epsilon > 0$, $P(|Y| > \epsilon) \leq E(Y^2)/\epsilon^2$. Use this inequality to prove that, if $\{Y_1, Y_2, \ldots\}$ is a sequence of random variables converging to 0 in L_2, then $Y_n \xrightarrow{p} 0$.*

Remark. 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 $c \neq 0$.

Remark. In this module, the most relevant modes of convergence will be convergence in distribution and convergence in probability.

1.2.2 Laws of large numbers and the CLT

These classical results, and more general versions of them, play a fundamental role in statistical asymptotics.

Laws of large numbers. Let $X_1, \ldots, X_n$ be independent, identically distributed (IID) real-valued random variables with finite mean μ. The strong law of large numbers (SLLN) states 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) states that if $E|X_1| < \infty$ and $E(X_1) = \mu$, then $\bar{X}_n \xrightarrow{p} \mu$.

Central limit theorem. The central limit theorem (CLT) states that, under the condition that the X_i are IID and have finite variance σ^2, and $E(X_1) = \mu$, a suitably standardised version of $\bar{X}_n$, $Z_n = n^{1/2}(\bar{X}_n - \mu)/\sigma$, converges in distribution to a random variable Z having the standard normal distribution $N(0, 1)$. We write

$$Z_n = n^{1/2}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0, 1).$$

Delta method. Another useful result is the 'delta method', which is derived via a Taylor expansion (see §1.3.1 below): 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 with derivative g' continuous at μ, then

$$n^{1/2}(g(Y_n) - g(\mu)) \xrightarrow{d} N(0, \{g'(\mu)\}^2\sigma^2).$$

The CLT and the delta method have natural multivariate generalisations.

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

$$n^{-1/2} \sum_{i=1}^{n} (X_i - EX_1) \xrightarrow{d} N_d(0, \Sigma),$$

where $N_d(\xi, \Sigma)$ is the d-dimensional mutivariate normal distribution with mean $\xi \in \mathbb{R}^d$ and $d \times d$ variance matrix Σ; see §1.3.3.

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 continuously differentiable at c, then if Y is $N_d(0, \Sigma)$, we have

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

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

A further important result is the following.

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 $g(X_n) \xrightarrow{d} g(X)$.

Remark. There are many extensions of these classical results which have important applications to statistical asymptotics in the context of general statistical models. For example, Martingale CLTs extend the classical CLT to a broad class of settings in which the random variables concerned are neither independent nor identically distributed. However, in this module we shall mainly be working within the IID setting and the above results will generally be sufficient.

1.2.3 Moments and cumulants

The moment generating function of a scalar random variable X is defined by $M_X(t) = E\{\exp(tX)\}$, whenever this expectation exists. Note that $M_X(0) = 1$, and that the moment generating function is defined in some

interval containing 0 (possibly containing just 0 itself). If $M_X(t)$ exists for t in an open interval $(-c, d)$ with $c, d > 0$, then all the moments $\mu'_r = \mathbb{E}X^r$ exist, and for $t \in (-c, d)$ we have the absolutely convergent expansion

$$M_X(t) = 1 + \sum_{r=1}^{\infty} \mu'_r \frac{t^r}{r!}.$$

The cumulant generating function $K_X(t)$ is defined by $K_X(t) = \log\{M_X(t)\}$, and is finite on the same interval as $M_X(t)$. Provided $M_X(t)$ exists in an open interval $(-c, d)$, for $t \in (-c, d)$ we have the absolutely convergent expansion

$$K_X(t) = \sum_{r=1}^{\infty} \kappa_r \frac{t^r}{r!}.$$

where the coefficient κ_r is defined as the rth cumulant of X.

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, $\kappa_1 = E(X) = \mu'_1$ 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.

Exercise.

(i) *By equating coefficients of powers of t on both sides of the identity $M_X(t) = \exp\{K_X(t)\}$, find expressions for μ'_3 and μ'_4 in terms of the cumulants $\kappa_1, \kappa_2, \ldots$.*

(ii) *By equating coefficients of powers of t on both sides of the identity $K_X(t) = \log\{M_X(t)\}$, find expressions for κ_3 and κ_4 in terms of the moments $\mu'_1, \mu'_2, \ldots$. [Hint: recall the expansion* $\log(1+x) = x - x^2/2 + x^3/3 - x^4/4 + \ldots$*]*

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 vector X involves no conceptual complication: see Pace and Salvan (1997, Chapter 3).

1.2.4 Mann-Wald notation

In asymptotic theory, the so-called Mann-Wald notation is useful for describing 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 a sequence of random variables $\{X_n\}$ and a sequence of positive constants $\{a_n\}$, we write $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$ as $n \to \infty$; and $X_n = O_p(a_n)$ when X_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, for a constant c, $Y_n = c + o_p(1)$ means that $Y_n \xrightarrow{p} c$.

Exercise. *Prove, or provide a counter-example to, each of the following statements:*

(i) $o_p(a_n) = a_n o_p(1)$*;*

(ii) $O_p(a_n) = a_n O_p(1)$*;*

(iii) $O_p(a_n)o_p(b_n) = o_p(a_n b_n)$*;*

(iv) if $a_n \to 0$ *and* $X_n = o_p(a_n)$*, then* $X_n^2 = o_p(a_n)$*;*

(v) if $a_n \to \infty$ *and* $X_n = o_p(a_n)$ *then* $(1 + X_n)^{-1} = o_p(1)$.

1.3 Some Further Mathematical Material

1.3.1 Taylor's theorem; index notation

Let $f(x)$, $x \in \mathbb{R}$, denote a function with continuous $(n+1)$th derivative. The Taylor expansion of f 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 $c \in [a, x]$.

Some particular expansions therefore are:

$$\begin{aligned} \log(1+x) &= x - x^2/2 + x^3/3 - x^4/4 \ldots (|x| < 1) \\ \exp(x) &= 1 + x + x^2/2! + x^3/3! + x^4/4! \ldots (x \in \mathbb{R}) \\ f(x+h) &= f(x) + f'(x)h + f''(x)h^2/2! + \ldots (x \in \mathbb{R}) \end{aligned}$$

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

$$f_x(a,b) = \left.\frac{\partial f}{\partial x}\right|_{x=a,y=b} \quad \text{and} \quad f_{xy}(a,b) = \left.\frac{\partial^2 f}{\partial x \partial y}\right|_{x=a,y=b},$$

and similarly for the other terms. More generally,

$$f(x+h) = f(x) + \nabla f(x)^{\mathrm{T}} h + h^{\mathrm{T}} \nabla \nabla^{\mathrm{T}} f(x) h/2! + \ldots (x \in \mathbb{R}^p),$$

where $\nabla f(x)$ is the vector of partial derivatives of f and $\nabla \nabla^{\mathrm{T}} f(x)$ is the Hessian matrix of second partial derivatives of f, all evaluated at x.

When higher-order terms are needed, a more convenient way to represent Taylor expansions in the multivariable case is to use index notation combined with a summation convention. Suppose now that $f : \mathbb{R}^d \to \mathbb{R}$, and let $x = (x^r)_{r=1}^d$, $\delta = (\delta^r)_{r=1}^d$, i.e. we use superscripts here rather than subscripts for components of vectors. Our goal is to Taylor expand $f(x+\delta)$ about $f(x)$.

In index notation, the third-order Taylor expansion, assuming that all third-order derivatives of f are continuous, is written

$$f(x+\delta) - f(x) = f_r \delta^r + \frac{1}{2!} f_{rs} \delta^r \delta^s + \frac{1}{3!} f^*_{rst} \delta^r \delta^s \delta^t,$$

where f_r and f_{rs} are evaluated at x; f^*_{rst} is evaluated at a point $x + \lambda\delta$ for some $\lambda \in [0, 1]$; and we use the convention that whenever an index appears

as a subscript and superscript in the same expression, summation is implied. So, in the above,

$$f_r\delta^r = \sum_{r=1}^{d} \frac{\partial f}{\partial x^r}\delta^r \quad \text{and} \quad f_{rs}\delta^r\delta^s = \sum_{r=1}^{d}\sum_{s=1}^{d} \frac{\partial^2 f}{\partial x^r \partial x^s}\delta^r\delta^s.$$

Index notation is particularly useful in higher-order expansions involving several variables; see, for example, Barndorff-Nielsen and Cox (1989).

1.3.2 Inverse of a block matrix

The following result from linear algebra will be needed later. 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,

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

1.3.3 Multivariate normal distribution

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

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

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

1. The covariance form is invariant under non-singular transformation of Y. This is because if $\tilde{Y} = AY$ where A is a non-singular $p \times p$ matrix, then $\tilde{Y} \sim N_p(0,\tilde{\Sigma})$, where $\tilde{\Sigma} = A\Sigma A^{\mathrm{T}}$, and

$$Q_{\tilde{Y}} = \tilde{Y}^{\mathrm{T}}\tilde{\Sigma}^{-1}\tilde{Y} = Y^{\mathrm{T}}A^{\mathrm{T}}(A\Sigma A^{\mathrm{T}})^{-1}AY = Y^{-1}\Sigma^{-1}Y = Q_Y.$$

2. Y can be transformed to independent components of unit variance (set $Z = \Sigma^{-1/2}Y$, where $\Sigma^{-1/2}$ is a matrix square root of Σ^{-1}).

3. The chi-squared distribution then follows directly, $Q_Y \equiv Q_Z = Z^{\mathrm{T}}Z$.

Now suppose that Y is partitioned into two parts $Y^{\mathrm{T}} = (Y_{(1)}^{\mathrm{T}}, Y_{(2)}^{\mathrm{T}})$ 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 $Q_{Y_{(1)}}$. To see this, 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.

1.4 Exponential families

1.4.1 (m, m) exponential families

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

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

where $\mathcal{Y}$ is a set not depending on ϕ. Here $s \equiv s(y) = (s_1(y), \ldots, s_m(y))^{\mathrm{T}}$, 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

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

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

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

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

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 can 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 joint moment generating function of the random variable $S = (S_1,\ldots,S_m)^{\mathrm{T}} = s(Y)$ corresponding to the natural statistic s is, writing $t = (t_1,\ldots,t_m)$,

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

from which we obtain

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

or

$$E_\phi(S) = \nabla_\phi K(\phi),$$

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

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

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

1.4.2 Inference in exponential families

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

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

Two important 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,

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

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

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

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

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

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

1.4.3 Curved exponential families

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

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

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 $\theta \in \mathbb{R}$ varies.

The following simple model, which is actually of some importance in many biological and agricultural problems, is a simple example of a curved exponential family. It concerns a normal distribution 'of known coefficient of variation'.

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

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

1.5 Likelihood and Sufficiency

1.5.1 Definitions

Consider 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; equivalently, 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 $\nabla_\theta = (\partial/\partial\theta^1, \ldots, \partial/\partial\theta^d)^{\mathrm{T}}$.

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 assumed to be IID, 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.

Example: log-likelihood in (m, m) exponential models. Let $X_1, \ldots, X_n$ be an independent sample from a full (m, m) exponential density

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

Ignoring an additive constant, the log-likelihood is

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

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

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

the log-likelihood may be written

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

i.e. it is a function of θ and $\hat{\theta}$ only.

1.5.2 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

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

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,

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

More compactly, the covariance matrix of U is

$$\mathrm{cov}_\theta\{U(\theta)\} = E\{-\nabla_\theta\nabla_\theta^{\mathrm{T}} l\}.$$

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 $-\nabla_\theta\nabla_\theta^T l$ 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 model (1.1),

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

and $\nabla_\phi\nabla_\phi^T l = -\nabla_\phi\nabla_\phi^{\mathrm{T}} K(\phi)$.

1.5.3 Change of parametrisation

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 twice in an expression (see §1.3.1), 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

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

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, say estimating $1/\rho$, then taking the reciprocal of this estimate.

Suppose that $\theta = (\psi, \chi)$, with ψ the parameter of interest and χ a nuisance parameter. A nuisance parameter is one which is not of primary interest but is needed in the model. For example, if we wish to compare two nested hypotheses H_0: $\psi = \psi_0$, χ unrestricted, and H_1: ψ, χ both unrestricted, then we would normally think of χ as being a nuisance parameter.

In such cases 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; see, for example, Barndorff-Nielsen and Cox (1994).

1.5.4 Sufficiency

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

$$T(x) = T(y) \Leftrightarrow \frac{L(\theta;x)}{L(\theta;y)} \text{ is independent of } \theta \in \Omega_\theta.$$

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 $(\sum Y_i, \sum Y_i^2)$.
Remark In exponential models, the natural statistic S is a (minimal) sufficient statistic. In a curved (m,d) exponential model with $d < m$, the dimension m of the sufficient statistic exceeds that of the parameter.

1.6 Bayesian Inference

In the Bayesian approach to statistical inference, the parameter θ in a model $f_Y(y|\theta)$ is itself regarded as a random variable. The main idea is that we represent our prior knowledge about θ through a probability distribution with pdf $\pi(\theta)$, and then use Bayes' Theorem to determine the posterior pdf $\pi(\theta|Y=y)$ for θ. More specifically, by Bayes' Theorem,

$$\pi(\theta|y) \propto f_Y(y|\theta)\pi(\theta),$$

where the constant of proportionality is $\{\int f_Y(y|\theta)\pi(\theta)d\theta\}^{-1}$.

Later in the module we shall see that in broad generality the posterior $\pi(\theta|y)$ is asymptotically normal as the sample size n increases, where $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. We shall also see how the Laplace's approximation (an asymptotic procedure) is often useful in posterior calculations.

2 Large Sample Theory

2.1 Motivation

In many situations, statistical inference depends on being able to derive approximations because exact answers are not available. Asymptotic (i.e. large n) results often provide convenient and sufficiently accurate approximations for practical purposes. Among the most important of such results are the asymptotic normality of maximum likelihood estimators (MLEs), and the χ^2 approximation for the null distribution of a log-likelihood ratio for nested models. In this chapter we look in some detail at the derivation of some of these results. We first look at simple situations with no nuisance parameters and then consider the more typical situation where nuisance parameters are present.

In the final section of the chapter we briefly study a broader class of estimators which possess some of the properties of MLEs, the so-called M-estimators. For this class of estimators the proof of asymptotic normality is rather similar to that for MLEs. However, an important difference is that the asymptotic variance matrix for M-estimators is given by the so-called sandwich variance formula rather than the inverse of the Fisher information matrix.

2.2 Some important likelihood statistics

Recall the definitions of the score function and expected and observed information in §1.5.1 and §1.5.2.

Denote by l_r the rth component of $U(\theta)$, l_{rs} the (r,s)th component of $\nabla_\theta \nabla_\theta^{\mathrm{T}} l$, 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'. In the above,

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

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, asymptotically, 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: 1-dimensional case

We now look more closely at the derivation of asymptotical normality of the MLE in the case of a single parameter θ. Consider an IID probability model

$$L(\theta) = L(\theta; x) = p_X(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \quad x = (x_1, \ldots, x_n)^\top,$$

where for simplicity we assume for the moment that $\theta \in \mathbb{R}$. We define the log-likelihood by

$$l(\theta) = l(\theta; x) = \log\{L(\theta)\},$$

and the score statistic

$$U(\theta) = \frac{\partial l}{\partial \theta}(\theta) = \sum_{i=1}^{n} \frac{\partial \log f}{\partial \theta}(\theta; x_i). \tag{2.7}$$

Our goal is to derive the asymptotic (i.e. large sample/large n) distribution of $\hat{\theta}$, the maximum likelihood estimator (MLE) of θ. We write the 'true' value of θ as θ_0.

Assumptions: we shall assume (i) that $l(\theta)$ has a continuous third derivative (in θ), and (ii) that the model is sufficiently regular for the standard results

$$E_\theta[U(\theta)] = 0, \quad \text{and} \quad \mathrm{Var}_\theta\{U(\theta)\} = i(\theta) \tag{2.8}$$

to hold, where $i(\theta)$ is the Fisher information.

In regular models the MLE will satisfy

$$U(\hat{\theta}) = 0.$$

Using a 3-term Taylor expansion in which we expand $U(\hat{\theta})$ about θ_0 we obtain

$$\begin{aligned} 0 = U(\hat{\theta}) &= U(\theta_0) - j(\theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2}\frac{\partial^3 l}{\partial \theta^3}(\theta^*)(\hat{\theta} - \theta_0)^2 \\ &= n^{-1/2}U(\theta_0) - \{n^{-1}j(\theta_0)\}n^{1/2}(\hat{\theta} - \theta_0) + n^{-1/2}R(\theta^*), \end{aligned} \tag{2.9}$$

where by Taylor's theorem, θ^* lies between θ_0 and $\hat{\theta}$, $U(\theta_0)$ is the score statistic, θ_0 is the solution to the first equation in (2.8),

$$j(\theta_0) = -\frac{\partial^2 l}{\partial \theta^2}(\theta_0) = -\sum_{i=1}^{n} \frac{\partial^2 \log f}{\partial \theta^2}(\theta_0; x_i)$$

is the *observed* Fisher information, and

$$R(\theta^*) = \frac{1}{2}\frac{\partial^3 l}{\partial \theta^3}(\theta^*)(\hat{\theta} - \theta_0)^2$$

is the remainder term.

We now study the terms on the RHS of (2.9) one by one.

The score statistic. By (2.7) and (2.8), $U(\theta_0)$ is a sum of IID random variables with mean 0 and variance $i(\theta_0)$. Consequently, by the CLT for IID random variables (see §1.2.2),

$$n^{-1/2}U(\theta_0) \xrightarrow{d} N(0, n^{-1}i(\theta_0)).$$

The observed Fisher information. The quantity $j(\theta_0)$, is a sum of IID random variables with mean $\bar{i}(\theta_0)$, where $\bar{i}(\theta_0) = i(\theta_0)/n$ is the *expected* Fisher information for a single observation. Therefore, by the Strong Law of Large Numbers (SLLN),

$$n^{-1}j(\theta_0) \xrightarrow{a.s} \bar{i}(\theta_0), \quad \text{as} \quad n \to \infty.$$

Recall that the SLLN implies the Weak Law of Large Numbers, so that it is also the case that $n^{-1}j(\theta_0) \xrightarrow{p} \bar{i}(\theta_0)$, which is sufficient for our purposes.

The remainder term. The heuristic reasoning goes as follows. If $n^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$, as would be the case if $n^{1/2}(\hat{\theta} - \theta_0)$ were asymptotically normal with finite asymptotic variance, then $(\hat{\theta} - \theta_0)^2 = O_p(n^{-1/2})^2 = O_p(n^{-1})$.

Moreover, since, for each θ, $\partial^3 l(\theta)/\partial\theta^3$ is a sum of n IID random variables, it is reasonable to hope that $\partial^3 l(\theta)/\partial\theta^3$ is $O_p(n)$, provided $E_\theta\{\partial^3 l(\theta)/\partial\theta^3\}$ is finite for θ in a neighbourhood of θ_0. In this case $R(\theta^*) = O_p(n^{-1})O_p(n) = O_p(1)$, i.e. $R(\theta^*)$ is bounded in probability, and so $n^{-1/2}R(\theta^*)$ is $O_p(n^{-1/2})$.

We now return to (2.9). Dividing through by $n^{-1}j(\theta_0)$ and rearranging, and assuming that the remainder term $R(\theta^*)$ is bounded in probability, it is seen that

$$\begin{aligned} n^{1/2}(\hat\theta - \theta_0) &= \{n^{-1}j(\theta_0)\}^{-1} n^{-1/2} U(\theta_0) + O_p(n^{-1/2}) \\ &\xrightarrow{d} \bar{i}(\theta_0)^{-1} N\{0, \bar{i}(\theta_0)\} \quad \text{[Slutsky; see §1.2.2]} \\ &\stackrel{d}{=} N\{0, \bar{i}(\theta_0)^{-1}\}, \end{aligned} \tag{2.10}$$

i.e. the limit distribution of $n^{1/2}(\hat\theta - \theta_0)$ is normal with mean 0, variance $\bar{i}(\theta_0)^{-1}$.

Remark. Although this conclusion holds in broad generality, it is important to note that it is a rather non-trivial matter to make the argument that $R(\theta^*) = O_p(1)$ fully rigorous, in the sense that further assumptions and some powerful mathematical machinery is needed. A serious issue concerns the asymptotic existence, uniqueness and consistency of the maximum likelihood estimate. There are no entirely 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. See for example van der Vaart (1998, Chapter 5) for further details. For simplicity we assume from here on that $\hat\theta$ is well defined and consistent.

2.4 Further asymptotic likelihood calculations

We now return to the case where θ is a vector. By the multivariate CLT, we conclude that $U(\theta) = U(\theta; Y) = [l_r(\theta)]_{r=1}^d$ is asymptotically $N_d(0, i(\theta))$, formally

$$n^{-1/2}\bar{i}(\theta)^{-1/2} U(\theta) \xrightarrow{d} N_d\big(0, I_d\big), \tag{2.11}$$

where I_d the $d\times d$ identity matrix, and with $\bar{i}(\theta)^{-1/2}$ interpreted as the (unique symmetric positive definite) matrix square root of $\bar{i}(\theta)^{-1}$. 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

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

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 $j^{rs}/i^{rs} \xrightarrow{p} 1$ for $r, s = 1, \ldots, d$, 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 (2.11) that $[\bar{\delta}^r]$ is asymptotically normal with zero mean and covariance matrix $[\bar{i}^{rs}]$. We have

$$\sqrt{n\bar{i}(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} N_d\big(0, I_d\big). \tag{2.12}$$

Note that the normality relations (2.11) and (2.12) are asymptotically parametrisation invariant. This means, in particular, that to show normality for arbitrary parametrisations it is enough to do so for one parametrisation. The consequence is simplification of theoretical derivations in many circumstances.

It is now explained why the asymptotic χ^2 distribution of $w = w(\theta) = 2\{l(\hat{\theta}) - l(\theta)\}$ follows 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 $w(\theta) \xrightarrow{d} \chi^2_d$. Note that this entails the application of several results from §1.2.1 and §1.2.2: the WLLN, the CLT, Slutsky's theorem and the

continuous mapping theorem. 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.

2.5 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 such that $\psi(\theta) = \psi$.

The log profile likelihood is $l_{\mathrm{p}} = \log L_{\mathrm{p}}$. It may be written as $l_{n\mathrm{p}}$ 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

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

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 $L_{\mathrm{p}}(\psi)$ 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 $2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\}$ equals the log-likelihood ratio statistic for $H_0 : \psi = \psi_0$,

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

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 $(\psi, \hat{\chi}_\psi)$,

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

where j_{p} denotes observed profile information, minus the matrix of second-order derivatives of l_{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 $l_{\mathrm{p}}(\psi) = l(\psi, \hat{\chi}_\psi)$ 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

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

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 §1.3.2, 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,

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

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

2.6 Effects of parameter orthogonality

Assume 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 and estimated is the same as that where λ is known. In this case ψ and λ are said to be **orthogonal**; see Cox and Reid (1987). 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

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

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 §1.3.3.

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 $o_p(1)$),

$$\begin{aligned} w(\psi_0) &= \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix}^{\mathrm{T}} i(\psi_0, \chi) \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} \\ &\quad -(\hat{\chi}_0 - \chi)^{\mathrm{T}} i_{\chi\chi}(\psi_0, \chi)(\hat{\chi}_0 - \chi). \end{aligned} \tag{2.13}$$

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 $\hat{\chi}_0 - \chi = i_{\chi\chi}^{-1} U_\chi$, to first-order. Then, we see from (2.13) that to first-order

$$w(\psi_0) = [U_\psi^{\mathrm{T}} U_\chi^{\mathrm{T}}] \begin{bmatrix} i^{\psi\psi} & i^{\psi\chi} \\ i^{\chi\psi} & i^{\chi\chi} \end{bmatrix} \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix} - U_\chi^{\mathrm{T}} i_{\chi\chi}^{-1} U_\chi. \tag{2.14}$$

From (2.14), in the notation of subsection 1.3.3,

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

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

$$w_p(\psi_0) = [i^{\psi\psi} U_\psi + i^{\psi\chi} U_\chi]^{\mathrm{T}} (i^{\psi\psi})^{-1} [i^{\psi\psi} U_\psi + i^{\psi\chi} U_\chi]. \tag{2.15}$$

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

$$w_U(\psi_0) = (U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi)^{\mathrm{T}} i^{\psi\psi} (U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi). \tag{2.16}$$

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 subsection 1.3.2, that the three quantities (2.14), (2.15) and (2.16) 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 (2.14) gives

$$w(\psi_0) = U_1^{\mathrm{T}} i^{11} U_1 + U_2^{\mathrm{T}} i^{21} U_1 + U_1^{\mathrm{T}} i^{12} U_1 + U_2^{\mathrm{T}} [i^{22} - i_{22}^{-1}] U_2. \tag{2.17}$$

Multiplying out (2.15) gives

$$w_p(\psi_0) = U_1^{\mathrm{T}} i^{11} U_1 + U_1^{\mathrm{T}} i^{12} U_2 + U_2^{\mathrm{T}} i^{21} U_1 + U_2^{\mathrm{T}} i^{21} (i^{11})^{-1} i^{12} U_2, \tag{2.18}$$

since $(i_{11} - i_{12} i_{22}^{-1} i_{21})^{-1} = i^{11}$. Equivalence of (2.17) and (2.18) 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

$$f(z; \theta) = \theta^n \exp\{-\theta \sum_{j=1}^{n} z_j\}. \tag{2.19}$$

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

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

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

$$g(y; \theta) = \theta^r \exp\{-\theta \sum_{j=1}^{n} y_j\}, \tag{2.20}$$

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

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

Under (2.20) 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 (2.20) an approximate $100(1-\alpha)\%$ confidence interval for θ based on $\bar{i}(\hat{\theta})$ is

$$\frac{r}{n\bar{y}} \mp \frac{1}{n^{1/2}(n\bar{y}/r)[1 - \exp\{-cr/(n\bar{y})\}]^{1/2}} \Phi^{-1}(1 - \alpha/2). \tag{2.21}$$

Under (2.19) 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

$$\frac{1}{\bar{z}} \mp \frac{1}{n^{1/2}\bar{z}} \Phi^{-1}(1 - \alpha/2). \tag{2.22}$$

When no censoring occurs (2.21) reduces to

$$\frac{1}{\bar{z}} \mp \frac{1}{n^{1/2}\bar{z}\{1 - \exp(-c/\bar{z})\}^{1/2}} \Phi^{-1}(1 - \alpha/2), \tag{2.23}$$

which is wider than (2.22).

The difference between (2.22) and (2.23) 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 (2.22) when censoring does not actually occur.

2.9 Asymptotic behaviour of M-estimators

So far in this chapter we have focused on maximum likelihood estimators (MLEs). We now consider a broader class of estimators called M-estimators which contains, and is much larger than, the class of MLEs.

The asymptotic theory of M-estimators is relevant in a number of contexts.

(i) *Misspecified models.* Here, we would like to know what happens when the 'wrong' likelihood is maximised.

(ii) *Estimating equations.* If for some reason (e.g. because the likelihood is too complicated) we prefer not to work with the likelihood for the model but rather to set up an alternative system of equations to estimate θ. An example of this is partial likelihood in survival analysis.

(iii) *Robust estimation.* In this setting, we may wish to set up a system of equations for estimating θ which produces an estimator of θ which is insensitive to outliers.

Let $X_1, \ldots, X_n$ denote an IID sample from a population with distribution function F. Suppose that we wish to construct an estimator of a parameter vector $\theta \in \Omega_\theta \subseteq \mathbb{R}^d$ based on the sample $X_1, \ldots, X_n$ and using an estimating function given by

$$G(\theta) \equiv \sum_{i=1}^{n} G_i(\theta) \equiv \sum_{i=1}^{n} G(X_i, \theta), \tag{2.24}$$

where $G(X_i, \theta) \in \mathbb{R}^d$, i.e. G has the same dimension as θ.

Assume θ_0 is such that

$$E_F\{G(X_1, \theta_0)\} \equiv \int G(x, \theta_0) dF(x) = 0, \tag{2.25}$$

and consider the sequence of estimating equations for θ given by

$$G(\theta) \equiv \sum_{i=1}^{n} G_i(\theta) = 0, \quad n = d, d+1, \ldots \tag{2.26}$$

Theorem. *Under mild conditions, (2.26) admits a sequence of solutions $(\hat{\theta}_n)_{n=d}^{\infty}$ with the following properties: as $n \to \infty$,*

(i) $\hat{\theta}_n \xrightarrow{p} \theta_0$, i.e. $\hat{\theta}_n$ is a consistent estimator of θ_0;

(ii) $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_d(0, H(\theta_0)V(\theta_0)H(\theta_0)^\top)$ *where*

$$V(\theta) = \mathrm{Var}\{G(X_1, \theta)\} \quad \textit{and} \quad H(\theta) = [E_F\{\nabla_\theta^T G(X, \theta)\}]^{-1}.$$

Remark. From the results stated in §1.2 we can deduce that (ii) implies (i), but nevertheless (i) is worth stating explicitly.

This theorem is an important result in first-order statistical asymptotics. Some comments are given below.

Comments.

1. The slightly vague wording '...(2.26) admits a sequence of solutions...' is necessary because, in general, for each n the solution to (2.26) may not be unique.

2. The even vaguer wording 'Under mild conditions...' is to avoid having to state conditions which are cumbersome and/or may be difficult to check. For further discussion of technical details and proofs, see for example van der Vaart (1998, Chapter 5).

3. In the case in which $\hat{\theta}_n$ is an MLE, G is the score statistic S so that, from standard likelihood theory, $V = \bar{i}(\theta_0)$, the expected Fisher information for a single observation, and $H(\theta_0) = \bar{i}(\theta_0)^{-1}$, so that the standard result $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_d(0, \bar{i}(\theta_0)^{-1})$ is recovered.

4. The expression $HVH^\top$ is known as the 'sandwich variance formula'. It depends on quantities that will usually be unknown but, provided the X_i are independent, in broad generality it can be estimated consistently by

$$\hat{H}\hat{V}\hat{H}^T = [n^{-1}\nabla_\theta^T G(\hat{\theta}_n)]^{-1}\{n^{-1}\sum_{i=1}^{n} G_i(\hat{\theta}_n)G_i(\hat{\theta}_n)^T\}[n^{-1}\nabla_\theta G(\hat{\theta}_n)^T]^{-1}, \tag{2.27}$$

which depends only on sample (i.e. observed) quantities.

5. Approximate confidence intervals for individual parameters and confidence regions for subsets of the parameters can be derived using the asymptotic normality given in the theorem.

6. The theorem generalises to many situations where the observations $X_1, \ldots, X_n$ are neither independent nor identically distributed.

3 Higher-order Theory

In this chapter we investigate various topics in higher-order statistical asymptotics including Edgeworth expansions, saddlepoint approximations, Laplace approximations, Bartlet correction and Bayesian asymptotics. In different ways each of the above goes beyond the most basic first-order asymptotic analysis. Moreover, these approaches lead to numerically accurate approximation and inference in a wide variety of applications.

3.1 Asymptotic expansions

Various technical tools are of importance in the development of higher-order statistical theory and approximation. 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, by 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 later state some of the key formulae in distribution function approximation.

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 which are $O_p(1)$ and typically have distributions which are only weakly dependent 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 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

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

in terms of the score $U(\theta)$ and Fisher information $i(\theta)$. The quantity $i(\theta)^{-1/2}U(\theta)$ plays the role of X_0. By the CLT we can write

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

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

3.2 Edgeworth expansion

In this section and in §3.3 we assume, for simplicity, that the random variables concerned are real-valued. Extensions to the multivariate case are straightforward and are summarised, for example, by Severini (2000, Chapter 2).

3.2.1 Edgeworth density approximation

Let $X_1, X_2, \ldots, X_n$ be IID random variables with cumulants $\kappa_1, \kappa_2, \ldots$. Let $S_n = \sum_1^n X_i$, $S_n^* = (S_n - n\mu)/(n^{1/2}\sigma)$ where $\mu \equiv \kappa_1 = E(X_1)$, $\sigma^2 \equiv \kappa_2 = \mathrm{var}\,(X_1)$. 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:

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

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 the 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, but are typically accurate in the centre of the distribution.

Exercise. *For $\alpha, \beta > 0$ let Gamma(α, β) denote the gamma distribution with pdf*

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x} \tag{3.2}$$

on $x > 0$ and zero elsewhere. Let $X_1, \ldots, X_n$ denote a random sample from Gamma(1,β), the Exp(β) distribution.

(i) Show that the moment generating function of the Gamma(α, β) distribution is given by $M(t) = \{\beta/(\beta - t)\}^\alpha$ for $t < \beta$.

(ii) Using (i), deduce that $S_n = \sum_{i=1}^n X_i \sim$ Gamma(n, β).

(iii) Calculate ρ_3, the third standardised cumulant for X_1.

(iv) Putting $\mu = \beta^{-1}$, and noting that $E(X_1) = \mu$ and $Var(X_1) = \mu^2$, define the standardised sum by $S_n^* = (S_n - n\mu)/(n^{1/2}\mu)$. Hence determine the one-term Edgeworth expansion for $f_{S_n^*}(x)$, i.e.

$$\hat{f}_{S_n^*}(x) \equiv \phi(x) \left\{ 1 + \frac{\rho_3}{6\sqrt{n}} H_3(x) \right\}.$$

(v) For several choices of n, e.g. $n = 5, 20, 50, 100$, plot $f_{S_n^*}(x)$ against x and $\hat{f}_{S_n^*}(x)$ against x on the same graph. What are your findings and conclusions? [Warning: part (ii) gives the pdf of S_n, not S_n^*, so you will need to apply a simple linear transformation to the Gamma(n,β) density to obtain the correct expression for the pdf of S_n^*.]

Later we shall consider the same example but using the saddlepoint density approximation.

3.2.2 Edgeworth distribution function approximation

Integrating the Edgeworth expansion (3.1) term by term (a procedure whose justification is non-trivial), 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

$$f_{\bar{X}}(\bar{x}) = \frac{n}{2\pi i}\int_{\tau - i\infty}^{\tau + i\infty} \exp\big[n\{K(\phi) - \phi\bar{x}\}\big]\,d\phi, \tag{3.3}$$

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 approximations

3.3.1 Saddlepoint density approximations

Saddlepoint density approximations, first discussed in a statistical context by Daniels (1954), have some excellent theoretical properties and typically have excellent numerical accuracy. For a full account see the book by Butler (2007).

There are two principal approaches to deriving saddlepoint approximations:

(I) a complex variable approach which involves studying the inversion integral the moment generating function (MGF) via contour integration in the complex plane;

(II) a real-variable approach making use of a procedure known as *exponential tilting*.

For deeper study of saddlepoint approximations it is essential to follow approach (I). However, when first meeting this topic, approach (II) is preferable for a statistical audience because it has a more statistical flavour and is easier to understand than (I).

We focus here on the univariate case. Let X be a continuous random variable with probability density function (pdf) $f_0(x)$ and cumulant generating function (CGF) $K_0(t) = \log[E\{\exp(tX)\}]$. It is assumed that $K_0(t)$ is finite for all $t \in (-a, b)$ for some $a, b > 0$. The reason for the zero subscript in f_0 and K_0 will become clear below when we consider exponential tilting.

The situation of interest here is when $K_0(t)$ is available in closed form, f_0 is not known explicitly, and we would like an approximation to f_0 at the point x. This type of problem arises frequently in statistics and applied probability.

First we define the exponentially tilted pdf $f_t(x)$ by

$$f_t(x) = \exp\{xt - K(t)\} f_0(x). \tag{3.4}$$

Observe that the definition of f_t is 'correct' at $t = 0$ because $K(0) = 0$. Also, f_t is a pdf since (i) it is non-negative, and (ii) it integrates to 1, because

$$\int_{x\in\mathbb{R}} f_t(x)dx = \exp\{-K(t)\}\int_{x\in\mathbb{R}} \exp(tx)f_0(x)dx = \exp\{-K(t)+K(t)\} = 1.$$

Main idea. *For given x, at which we want to approximate $f_0(x)$, choose $\hat{t}$ so that the mean of the distribution with pdf $f_{\hat{t}}$ is x; and then approximate $f_{\hat{t}}(x)$ by a normal density with the same mean and variance as the distribution with pdf $f_{\hat{t}}$. This makes especially good sense when X is itself a sum of continuous IID random variables.*

Remark. The normal approximation is likely to be better at the mean, which is at or near the centre of the distribution, than in the tails, especially when the random variable X is itself a sum of IID random variables.

We shall now show how to put this idea into practice. But first...

Exercise. *Show that, for given x, the CGF of the distribution with pdf $f_{\hat{t}}$ is given by $K_{\hat{t}}(\phi) = K_0(\hat{t}+\phi) - K_0(\hat{t})$. Hence show that the mean and variance of this distribution are given by $K_0^{(1)}(\hat{t})$ and $K_0^{(2)}(\hat{t})$, where the bracketed superscripts indicate the number of derivatives.*

From the main idea, we choose $\hat{t}$ so that the mean of the distribution with pdf $f_{\hat{t}}$ is equal to x, i.e. we choose $\hat{t}$ to solve

$$K_0^{(1)}(t) = x. \tag{3.5}$$

This is known as the saddlepoint equation.

The idea now is to approximate $f_{\hat{t}}(x)$ by a normal density with mean $K_0^{(1)}(\hat{t})$ and variance $K_0^{(2)}(\hat{t})$, evaluated at its mean $K_0^{(1)}(\hat{t})$. From our knowledge of the normal density, the latter is given by $\{2\pi K_0^{(2)}(\hat{t})\}^{-1/2}$. Substituting this into the LHS of (3.4) gives

$$\frac{1}{\{2\pi K_0^{(2)}(\hat{t})\}^{1/2}} \approx \exp\{\hat{t}x - K_0(\hat{t})\}f_0(x);$$

equivalently, by multiplying both sides by $\exp\{K_0(\hat{t}) - \hat{t}x\}$,

$$f_0(x) \approx \hat{f}_0(x) \equiv \frac{1}{\{2\pi K_0^{(2)}(\hat{t})\}^{1/2}} \exp\{K_0(\hat{t}) - \hat{t}x\}, \tag{3.6}$$

where $\hat{f}_0(x)$ defined above is the first-order saddlepoint density approximation.

So far we have not introduced the sample size n. Suppose we replace X in the above by the sum $S_n = (X_1 + \ldots + X_n)$ where the X_i are IID each with cumulant generating function (CGF) $K_0(t)$. Then the CGF of S_n is $nK_0(t)$. Writing $\hat{f}_{S_n}(s)$ for the saddlepoint approximation to $f_{S_n}(s)$, the pdf of S_n, and substituting into (3.6), we obtain

$$f_{S_n}(s) \approx \hat{f}_{S_n}(s) \equiv \frac{1}{\{2\pi n K_0^{(2)}(\hat{t})\}^{1/2}} \exp\{nK_0(\hat{t}) - \hat{t}s\}, \tag{3.7}$$

where now the saddlepoint equation takes the form $K_0(\hat{t}) = s/n$.

Exercise. *As in the exercise in the previous section, let $X_1, \ldots, X_n$ be an IID sample from the Exp(β) distribution or, equivalently, from the Gamma(1,β) distribution in the notation of (3.2).*

(i) Show that the cumulant generating function (CGF) of the sum $S_n = \sum_{i=1}^n X_i$ is $K(t) = n\{\log(\beta) - \log(\beta - t)\}$.

(ii) Derive the saddlepoint equation for this CGF (cf. (3.5)) and then solve it.

(iii) Hence determine the saddlepoint approximation $\hat{f}_{S_n}(s)$ to $f_{S_n}(s)$, the pdf of S_n.

(iv) For various values of n, e.g. $n = 1, 5, 10$, plot $\hat{f}_{S_n}(s)$ against x and plot $f_{S_n}(s)$ against s on the same graph. What are your findings and conclusions? How do the results compare with those for the Edgeworth approximation considered in the previous section?

Theoretically, saddlepoint approximations have good absolute accuracy in the centre of the distribution being approximated, and excellent relative accuracy in the tails. Further details will be given in the lectures. Numerically, saddlepoint approximations are often remarkably accurate, even in many cases which are not asymptotic.

In the lectures we will consider various extensions and further topics including: the multivariate case, where X is a random vector; higher-order saddlepoint approximations obtained via Edgeworth expansion; and Barndorff-Nielsen's p^* formula.

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 $K(t) = c(\theta + t) - c(\theta)$ is the CGF of the corresponding random variable X, it follows that $\hat{\phi} = \hat{\theta} - \theta$,

where $\hat{\theta}$ is the MLE based on $s = x_1 + \cdots + x_n$. Then, following (3.6) and (3.7), in this case the saddlepoint approximation is given by

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

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 $\hat{\theta}$

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

This latter approximation is a particular example of Barndorff-Nielsen's p^*-formula which will be discussed in the lectures.

3.3.2 Lugananni-Rice CDF approximation

It is not easy to integrate the saddlepoint density approximation exactly to obtain an approximation to the distribution function of S_n. An alternative to numerical integration is to use the Lugannani and Rice (1980) approximation

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

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 $g(y)$ has a unique minimum over $[a, b]$ 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 a Taylor expansion about $y = \tilde{y}$ in the exponent, we can write

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

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

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

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

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

Remark. When evaluation the Laplace approximation for an integral such as $\int_a^b h(x)e^{-ng(x)}dx$, the following convention is adopted: the term outside the exponent, in this case $h(x)$, plays no role in the maximisation; it is only the exponent which is maximised, which is equivalent to minimisation of $g(x)$.

A further refinement of the method, available in the common situation where $h(y)$ is strictly positive, and which allows $g(y)$ to depend weakly on n, gives

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

where

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

Exercise: Stirling's approximation. *Stirling's approximation, $\hat{\Gamma}(n)$, to the gamma function, $\Gamma(n)=\int_0^\infty x^{n-1}e^{-x}dx$, is given by $\hat{\Gamma}(n)\equiv\sqrt{(}2\pi)n^{n-\frac{1}{2}}e^{-n}$, as $n\to\infty$. Derive Stirling's approximation using Laplace's method. [Hint: you will need to take care in the specification of the functions $g(x)$ and $h(x)$ in (3.12), remembering that it is only the exponent that is maximised.]*

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 $g(y)$ has a unique 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 approximation is particularly useful in Bayesian inference: see §3.7.

3.5 Conditional inference in exponential families

An 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. Note: given a conditional distribution, the conditional likelihood is simply the likelihood based on this conditional distribution.

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.

3.6 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 IID sampling, and in other settings, 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.

3.7 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 et al. (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,

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

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

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

so that, to first order,

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

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

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

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

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

by Laplace approximation of the denominator.

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

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

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

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

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

$$E\{g(\theta) \mid x\} \approx \frac{e^{n\bar{l}_n(\theta^*)+q(\theta^*)}}{e^{n\bar{l}_n(\tilde{\theta})+p(\tilde{\theta})}} \times \frac{\{-n\bar{l}_n''(\tilde{\theta}) - p''(\tilde{\theta})\}^{1/2}}{\{-n\bar{l}_n''(\theta^*) - q''(\theta^*)\}^{1/2}} \frac{\{1+O(n^{-1})\}}{\{1+O(n^{-1})\}}$$

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

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

where Φ and φ are the standard normal distribution and density functions respectively, r_p is the signed root (profile) likelihood ratio statistic (cf. (2.4)) given by

$$r_p(\psi) = \mathrm{sgn}(\hat{\psi} - \psi)[2\{l_p(\hat{\psi}) - l_p(\psi)\}]^{1/2},$$

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 $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$ 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

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

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

[3] Butler, R. W. (2007). *Saddlepoint Approximations with Applications.* Cambridge: Cambridge University Press.

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

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

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

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

[8] Ferguson, T. S. (1996). *A Course in Large Sample Theory.* London: Chapman & Hall.

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

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

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

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

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

[14] van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

[15] Young, G. A. and Smith, R. L. (2005) *Essentials of Statistical Inference.* Cambridge: Cambridge University Press.