Statistical Asymptotics
Part I: Background Material

Andrew Wood

School of Mathematical Sciences
University of Nottingham

APTS, April 11-15, 2016

Structure of the Chapter

This chapter covers methods and results needed in subsequent chapters.

Topics (mathematical): relevant limit theorems from probability, multivariable Taylor expansions, delta method.

Topics (statistical): exponential families, likelihood, sufficiency, Bayesian inference.

Motivation

Basic question: why study statistical asymptotics?

- (a) To motivate approximations: to derive useful practical approximations for statistical inference and probability calculations in settings where exact solutions are not available.
- (b) Theoretical insight: to gain theoretical insight into complex problems, e.g. to identify which aspects of a problem are of most importance.
- (c) Theories of optimality: various theories of optimality can be developed in an asymptotic setting.

We shall mainly be focused on (a) in this module, but (b) will also be relevant from time to time. For some aspects of (c), see e.g. the book by van der Vaart (1998).

Random Vectors

Let $Y = (Y^1, \ldots, Y^p)^\top$ denote an p-dimensional random vector, where each component is a real-valued random variable.

Note the use of superscripts to label components.

The mean, $\mu = E(Y)$, of Y, when it exists, is given by

$$\mu \equiv (\mu^1, \ldots, \mu^p)^\top = \{E(Y^1), \ldots, E(Y^p)\}^T \equiv E(Y).$$

The covariance matrix, $\Sigma = \mathrm{Cov}(Y)$, when it exists, is given by

$$\Sigma = E\{(Y-\mu)(Y-\mu)^\top\} = [E\{(Y^i-\mu^i)(Y^j-\mu^j)\}]_{i,j=1}^p = \mathrm{Cov}(Y)$$

Note that Σ is a symmetric, non-negative definite $p \times p$ matrix.

The distribution function $F(y)$ of Y is defined by

$$F(y) \equiv F_Y(y) = P(Y^1 \leq y^1, \ldots, Y^p \leq y^p), \quad y = (y^1, \ldots y^p)^\top \in \mathbb{R}^p.$$

Multivariate Normal Distribution

Definition: a p-dimensional random vector Y is said to have a multivariate normal distribution $N_p(\mu, \Sigma)$ if, for each fixed (i.e. non-random) $a = (a^1, \ldots, a^p)^T \in \mathbb{R}^p$,

$$a^\top Y \equiv \sum_{i=1}^{p} a^i Y^i \sim N(a^\top \mu, a^\top \Sigma a) \quad \text{[i.e. univariate normal]}.$$

If $Y \sim N_p(\mu, \Sigma)$ then $E(Y) = \mu$ and $\mathrm{Cov}(Y) = \Sigma$.

When Σ is positive definite, the probability density function (pdf) of $N_p(\mu, \Sigma)$ is given by

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

where $|\Sigma|$ is the determinant of Σ.

Quadratic forms in normal variables

Suppose that $Y \sim N_p(0_p, \Sigma)$ where 0_p is the p-vector of zeros.

If Σ is positive definite, then

$$Y^\top \Sigma^{-1} Y \sim \chi^2_p, \tag{1}$$

where χ^2_p is the chi-squared distribution with p degrees of freedom.

To see this, consider a general linear transformation of $Z = TY$ where T is non-singular. Then $Z \sim N_p(0_p, T\Sigma T^\top)$ and

$$Z^\top (TVT^\top)^{-1} Z = Y^\top T^\top (T^\top)^{-1} \Sigma^{-1} T^{-1} TY = Y^\top \Sigma^{-1} Y, \tag{2}$$

i.e. the quadratic form in normal variables is invariant with respect to non-singular transformations T.

If we choose T to be the orthogonal matrix whose columns are orthogonal unit eigenvectors of Σ, then (1) follows.

Difference of two normal quadratic forms

Suppose now that $Y = (Y_1^\top, Y_2^\top)^\top \sim N_p(0_p, \Sigma)$, where Y_1 and Y_2 are of dimension p_1 and p_2 respectively, with $p_1 + p_2 = p$, and, in obvious notation,

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then, provided Σ is non-singular,

$$Y^\top \Sigma^{-1} Y - Y_2^\top \Sigma_{22}^{-1} Y_2 \sim \chi^2_{p_1}.$$

To see why, consider a linear transformation T which maps $Y = (Y_1^\top, Y_2^\top)^\top$ to $\tilde{Y} = (\tilde{Y}_1^\top, \tilde{Y}_2^\top)^\top$, where $\tilde{Y}_2 = Y_2$ and $\tilde{Y}_1$ and $\tilde{Y}_2$ are uncorrelated.

The following choice achieves this goal:

$$\tilde{Y}_1 = Y_1 - \Sigma_{12}\Sigma_{22}^{-1} Y_2$$

Difference of two normal quadratic forms (continued)

Straightforward calculation shows that, with this choice of $\tilde{Y}_1$,

$$\mathrm{Cov}(\tilde{Y}_1) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{3}$$

Then, using the invariance property established in (2), some further calculations show that

$$\begin{aligned} Y^\top \Sigma^{-1} Y - Y_2^\top \Sigma_{22}^{-1} Y_2 &= \tilde{Y}_1^\top (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \tilde{Y}_1 \\ &\qquad + Y_2^\top \Sigma_{22}^{-1} Y_2 - Y_2^\top \Sigma_{22}^{-1} Y_2 \\ &= \tilde{Y}_1^\top (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \tilde{Y}_1 \\ &\sim \chi^2_{p_1}, \end{aligned}$$

where in the final step we have used (1) and (3).

Convergence in Distribution

Let $(Y_n)_{n=1}^{\infty}$ denote a sequence of p-dimensional random vectors, where $Y_n = (Y_n^1, \ldots, Y_n^p)^\top$ has distribution function $F_n(y)$, i.e.

$$F_n(y) = P(Y_n^1 \leq y^1, \ldots, Y_n^p \leq y^p), \quad y = (y^1, \ldots, y^p)^\top \in \mathbb{R}^p.$$

Also, let $Y = (Y^1, \ldots, Y^p)^\top$ denote a random vector with distribution function $F(y)$.

Definition: we say that the sequence $(Y_n)_{n=1}^{\infty}$ of random vectors converges in distribution to Y as $n \to \infty$ if

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

for all $y \in \mathbb{R}^p$ at which F is continuous. We write $Y_n \xrightarrow{d} Y$.

Remark: in all cases of convergence in distribution that we shall encounter in this module, the limit distribution will be a familiar one, such as multivariate normal or χ^2.

Convergence in Probability

Definition: a sequence of random vectors $(Y_n)_{n=1}^{\infty}$ converges to a random vector Y in probability if, for each $\epsilon > 0$,

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

where $||.||$ denotes Euclidean distance on $\mathbb{R}^p$,

i.e. $||a|| = (a^\top a)^{1/2}$ for $a \in \mathbb{R}^p$.

We write $Y_n \xrightarrow{p} Y$.

Comments

- Convergence in probability implies convergence in distribution.
- Convergence in distribution implies convergence in probability if the convergence is to a constant, but not otherwise.
- Two other modes of convergence, mentioned in the preliminary notes, are almost sure convergence and L^p convergence, both of which imply convergence in probability.
- In these lectures we shall only require convergence in distribution and convergence in probability.

Multivariate CLT and WLLN

We now state the weak law of large numbers (WLLN) and the central limit theorem (CLT) in the multivariate independent and identically distributed (IID) case.

WLLN. If $(Y_i)_{i=1}^{\infty}$ is an IID sequence of p-dimensional random vectors with finite mean μ, then $n^{-1}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mu$.

Multivariate CLT. Let $(Y_i)_{i=1}^{\infty}$ denote a sequence of IID p-dimensional random vectors with common mean vector $\mu = E(Y_1)$ and variance matrix $\Sigma = \mathrm{Cov}(Y_1)$. Then as $n \to \infty$,

$$n^{-1/2}\sum_{i=1}^{n}(Y_i - \mu) \xrightarrow{d} N_p(0_p, \Sigma),$$

where 0_p is the zero vector in $\mathbb{R}^p$.

Continuous Mapping Theorem

A useful supplement to convergence in distribution is the continuous mapping theorem.

Continuous Mapping Theorem (CMT). Suppose that p-dimensional random vectors Y and $(Y_n)_{n=1}^{\infty}$ are such that $Y_n \xrightarrow{d} Y$. Let $g : \mathbb{R}^p \to \mathbb{R}^q$ denote a continuous function. Then $g(Y_n) \xrightarrow{d} g(Y)$.

Remarks

- The result may not hold if g is not continuous.
- A much more general version of the CMT, set in function space, is given by van der Vaart (1998).

Continuous Mapping Theorem: Example

Example Let $(Y_n)_{n=1}^{\infty}$ is a sequence of p-dimensional random vectors such that $Y_n \xrightarrow{d} N_p(0_p, I_p)$ where I_p is the $p \times p$ identity matrix.

Consider the function $g : \mathbb{R}^p \to \mathbb{R}$ defined by

$$g(y) = y^\top y, \quad y \in \mathbb{R}^p.$$

Note that g is continuous.

Recall that if $Y \sim N_p(0_p, I_p)$, then $Y^\top Y \sim \chi^2_p$.

Therefore we may conclude from the CMT that as $n \to \infty$, $g(Y_n) \xrightarrow{d} \chi^2_p$.

Mann-Wald $o_p(.)$ and $O_p(.)$ notation

Mann-Wald notation provides useful shorthand for making probability statements. We first give the non-random versions.

Let $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ denote two sequences of positive numbers.

We write

1. $a_n = o(b_n)$ when $\lim_{n\to\infty} a_n/b_n = 0$.
2. $a_n = O(b_n)$ when $\limsup_{n\to\infty} a_n/b_n = K < \infty$,

i.e. when a_n/b_n remains bounded as $n \to \infty$.

Mann-Wald $o_p(.)$ and $O_p(.)$ notation (continued)

If $(Y_n)_{n=1}^{\infty}$ is a sequence of random vectors, then

1. $Y_n = o_p(a_n)$ means that $a_n^{-1} Y_n \xrightarrow{p} 0_p$, the zero vector in $\mathbb{R}^p$.

2. $Y_n = O_p(a_n)$ means that $a_n^{-1} Y_n$ is bounded in probability; i.e. for any $\epsilon > 0$, there exist $k < \infty$ and $n_0 < \infty$ such that, for all $n > n_0$,

$$P(||a_n^{-1} Y_n|| > k) < \epsilon.$$

Slutsky's Theorem

An elementary but important result, which we shall use frequently in the next chapter, is Slutsky's theorem.

Slutsky's Theorem

Suppose that as $n \to \infty$, $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, a finite constant. Then:

- $X_n + Y_n \xrightarrow{d} X + c$;
- $X_n Y_n \xrightarrow{d} cX$; and
- if $c \neq 0$ then $X_n / Y_n \xrightarrow{d} X/c$.

Note the requirement that one of the sequences converges to a constant.

Taylor's Theorem: multivariable case

Consider a function $f : \mathbb{R}^p \to \mathbb{R}$, and suppose $x, h \in \mathbb{R}^p$.

Assuming all 3rd order partial derivatives of f are continuous, the 4-term Taylor expansion of $f(x+h)$ about $h = 0_p$ is given by

$$f(x+h) = f(x) + \sum_{i=1}^{p} f_i(x)h^i + \frac{1}{2!}\sum_{i,j=1}^{p} f_{ij}(x)h^i h^j + R(x,h), \quad (4)$$

where

$$f_i(x) = \frac{\partial f}{\partial x^i}(x), \qquad f_{ij}(x) = \frac{\partial^2 f}{\partial x^i \partial x^j}(x), \qquad \sum_{i,j=1}^{p} \equiv \sum_{i=1}^{p}\sum_{j=1}^{p},$$

and the remainder term $R(x,h)$ is given by

$$R(x,h) = \frac{1}{3!}\sum_{i,j,k=1}^{p} f_{ijk}(x^*)h^i h^j h^k, \quad (5)$$

where $x^* = x + \theta h$ for some $\theta \in [0,1]$.

Taylor's Theorem (continued)

An equivalent way of writing the expansion (4) is

$$f(x+h) = f(x) + h^\top \nabla f(x) + \frac{1}{2!} h^\top \nabla \nabla^\top f(x) h + R(x,h),$$

where $R(x,h)$ is given in (5),

$$\nabla f(x) = (f_1(x), \ldots, f_p(x))^\top$$

is the vector of partial derivatives of f evaluated at x, and

$$\nabla \nabla^\top f(x) = \{f_{ij}(x)\}_{i,j=1}^p$$

is the $p \times p$ Hessian of f, evaluated at x.

Index notation and summation convention

The Taylor expansion (4) can be written more simply if we adopt the following summation convention.

Summation convention: when an index appears in the same expression as both a subscript and superscript, then summation over that index is implied.

Applying the summation convention to (4) we obtain

$$f(x+h) = f(x) + f_i(x)h^i + \tfrac{1}{2!}f_{ij}(x)h^i h^j + \tfrac{1}{3!}f_{ijk}(x^*)h^i h^j h^k,$$

where x^* is defined as in (5).

Moments and Cumulants

The Moment Generating Function (MGF) of a random variable X is defined by $M_X(t) = E\{\exp(tX)\}$, $t \in \mathbb{R}$.

When $M(t) < \infty$ for all t in an open neighbourhood containing $t = 0$, the MGF for such t has an absolutely convergent expansion

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

where $\mu'_r = E(X^r)$, $r = 1, 2, \ldots$, are the (uncentred) power moments of X.

The Cumulant Generating Function (CGF) is defined by

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

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

Moments and Cumulants (continued)

Cumulants can be expressed in terms of moments by equating coefficients in the expansions of $K_X(t)$ and $\log\{M_X(t)\}$.

Note that $\kappa_1 = \mu'_1 = E(X)$ and $\kappa_2 = \mu'_2 - (\mu'_1)^2 = \mathrm{Var}(X)$.

The normal distribution is characterised by the following: $\kappa_2 > 0$ and $\kappa_r = 0$, $r \geq 3$.

Note that, for constants $a, b \in \mathbb{R}$,

$$M_{aX+b}(t) = e^{bt} M_X(at) \text{ and } \quad K_{aX+b} = bt + K_X(at).$$

Statistical framework

Goal: to analyse a sample of observations $\mathcal{S} = \{y_1, \ldots, y_n\}$. For the moment we shall assume that the sample is IID.

- Assume the sample $\mathcal{S}$ is drawn from an unknown probability distribution specified by a probability density function (pdf) or probability mass function (pmf).
- Restrict the unknown density to a suitable family $\mathcal{F}$, of known analytical form, involving a finite number of real unknown parameters $\theta = (\theta^1, \ldots, \theta^d)^T$. The region $\Omega_\theta \subset \mathbb{R}^d$ of possible values of θ is called the parameter space. To indicate dependency of the density on θ write $f(y; \theta)$, the 'model function'.
- Assume that the objective of the analysis is to assess some aspect of θ, for example the value of a single component θ^i.

Statistical framework (continued)

We want to provide a framework for the relatively systematic analysis of a wide range of possible $\mathcal{F}$.

Quite a lot of the module will be focused on a likelihood-based approach.

We do not aim to satisfy formal optimality criteria.

Exponential Families

An important class of models, particularly relevant in theory and practice, is the exponential family of models.

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

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

where $\mathcal{Y}$ is a set not depending on ϕ.

Here $s \equiv s(y) = (s_1(y), \ldots, s_m(y))^T$, are called natural statistics.

The value of m may be reduced if either $s = (s_1, \ldots, s_m)^T$ or $\phi = (\phi^1, \ldots, \phi^m)^T$ satisfies a linear constraint.

Assume that representation is minimal, in that m is as small as possible.

Full Exponential Family

Provided the natural parameter space Ω_ϕ consists of all ϕ such that

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

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

If y is discrete, the integral above is replaced by a sum over $y \in \mathcal{Y}$.

Properties of exponential families

Let $s(y) = (t(y), u(y))$ 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

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

Two key results hold which allow inference about components of the natural parameter, in the absence of knowledge about the other components.

Result 1

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

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

say.

Result 2

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

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

say.

Curved exponential families

In the above, both the natural statistic and the natural parameter lie in m-dimensional regions.

Sometimes, ϕ may be restricted to lie in a d-dimensional subspace, $d < m$.

This is most conveniently expressed by writing $\phi = \phi(\theta)$ where θ is a d-dimensional parameter.

We then have

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

where $\theta \in \Omega_\theta \subset \mathbb{R}^d$.

We call this system an (m, d) exponential family, or curved exponential family, noting we require that $(\phi^1, \ldots, \phi^m)$ does not belong to a v-dimensional linear subspace of $\mathbb{R}^m$ with $v < m$.

Think of the case $m = 2, d = 1$: $\{\phi^1(\theta), \phi^2(\theta)\}$ describes a curve as θ varies.

Likelihood

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

The likelihood function is

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

Log-likelihood

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

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

Score function

We define the score function by

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

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

To study the score function as a random variable we write

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

Score function and information

For regular problems we have

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

Here, $E_\theta\{.\}$ means expectation with respect to $f_Y(y;\theta)$.

Observed and expected information

The covariance matrix of U is

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

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 matrix $-\nabla_\theta \nabla_\theta^\top l$ is called the observed information matrix, and is denoted by $j(\theta)$.

Note that $i(\theta) = E\{j(\theta)\}$.

Logistic regression example

We now introduce a logistic regression example that we shall return to from time to time.

Suppose $Y_1, \ldots, Y_n$ are independent Bernoulli random variables, with

$$P(Y_i = 1) = p_i, \qquad P(Y_i = 0) = 1 - p_i.$$

Assume a logistic model for p_i, i.e.

$$p_i = \frac{e^{\beta^\top x_i}}{1 + e^{\beta^\top x_i}},$$

where $\beta = (\beta_1, \ldots, \beta_d)^\top$ is a vector of parameters and, for each i, $x_i = (x_{i1}, \ldots, x_{id})^\top$ is a covariate vector.

It is customary to treat the x_i as non-random quantities. If they were generated randomly, then we would usually condition on their values.

Logistic regression example (continued)

For a sample $y_1, \ldots, y_n$, the likelihood is given by

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},$$

and a short calculation shows that the log-likelihood is

$$l(\beta) = \sum_{i=1}^{n} \{y_i \beta^\top x_i - \log(1 + e^{\beta^\top x_i})\}.$$

The score statistic and observed information are found to be

$$U(\beta) = \nabla_\beta l(\beta) = \sum_{i=1}^{n} y_i x_i - p_i x_i = \sum_{i=1}^{n} (y_i - p_i) x_i \qquad (6)$$

and

$$j(\beta) = -\nabla_\beta \nabla_\beta^\top l(\beta) = \sum_{i=1}^{n} p_i(1-p_i) x_i x_i^\top . \qquad (7)$$

Logistic regression example (continued)

Comments

1. Note that $U(\beta)$ is a vector of dimension d and that $j(\beta)$ is a $d \times d$ matrix.

2. Because $E(Y_i) = p_i$, it follows that $U(\beta)$, with Y_i replacing y_i satisfies

$$E_\beta[U(\beta)] = 0_d,$$

where 0_d is the d-vector of zeros.

3. Note that in this somewhat special model $j(\beta)$ in (7) does not depend on any random quantities, so the Fisher information equals the observed information, i.e. $i(\beta) = j(\beta)$.

4. It is clear from the RHS of (7) that $j(\beta)$ is non-negative definite and will be positive definite if $x_1, \ldots, x_n$ span $\mathbb{R}^d$.

Pseudo-likelihoods

Consider a model parameterised by a parameter θ which may be written as $\theta = (\psi, \lambda)$, where ψ is the parameter of interest and λ is a nuisance parameter.

A nuisance parameter is a parameter not of primary interest.

For example, when testing

$$H_0 : \psi = \psi_0, \ \lambda \text{ unrestricted} \quad \text{versus} \quad H_A : \psi, \lambda \ \text{ both unrestricted},$$

we would usually regard λ as a nuisance parameter.

To draw inferences about the parameter of interest, we must deal with the nuisance parameter. Ideally, we would like to construct a likelihood function for ψ alone.

Marginal likelihood

Suppose that there exists a statistic T such that the density of the data Y may be written as

$$f_Y(y;\psi,\lambda) = f_T(t;\psi) f_{Y|T}(y|t;\psi,\lambda).$$

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

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

Conditional likelihood

Suppose that there exists a statistic S such that

$$f_Y(y;\psi,\lambda) = f_{Y|S}(y|s;\psi)f_S(s;\psi,\lambda).$$

A likelihood function for ψ may be based on $f_{Y|S}(y|s;\psi)$, which does not depend on λ.

The conditional log-likelihood function may be calculated as

$$l(\psi; y|s) = l(\theta) - l(\theta; s),$$

where $l(\theta; s)$ denotes the log-likelihood function based on the marginal distribution of S and $l(\theta)$ is the log-likelihood based on the full data Y.

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, so that

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

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

Minimal sufficient statistic

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

Factorisation

Determination of S from the definition above 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.

A useful result

To identify minimal sufficient statistics.

A statistic T is minimal sufficient iff

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

Examples

Exponential families

Here the natural statistic S is sufficient.

In a curved (m, d) exponential family the dimension m of the sufficient statistic exceeds that of the parameter.

The Bayesian approach

In the Bayesian approach to statistical inference, the parameter θ in a model $f_Y(y|\theta)$ is itself regarded as a random variable.

Our prior knowledge about θ is represented through a pdf $\pi(\theta)$ known as the prior distribution.

Then Bayes' Theorem gives the posterior density

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

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

Later, we will see that, in a large-sample framework, the posterior is typically asymptotically normal, and that Laplace's method often provides a useful way to approximate the posterior.

Parameter Orthogonality

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

Suppose that θ is partitioned into components

$$\theta = (\theta^1, \ldots, \theta^{d_1}; \theta^{d_1+1}, \ldots, \theta^d)^T = (\theta_{(1)}^T, \theta_{(2)}^T)^T.$$

Suppose that $i_{rs}(\theta) = 0$ for all $r = 1, \ldots, d_1$; $s = d_1 + 1, \ldots, d$, for all $\theta \in \Omega_\theta$, so that $i(\theta)$ is block diagonal.

We say that $\theta_{(1)}$ is orthogonal to $\theta_{(2)}$.

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

The case $d_1 = 1$

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

The case $d_1 > 1$

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