STATISTICAL ASYMPTOTICS

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

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

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

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

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

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

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

References

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

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

1 Concepts and Principles

1.1 Introduction

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

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

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

2. We restrict the unknown density to a suitable family $\mathcal{F}$. We are concerned primarily with the case where the density is of known analytical form, but involves a finite number of real unknown parameters $\theta = (\theta^1, \ldots, \theta^d)^T$. We specify the region $\Omega_\theta \subset \mathbb{R}^d$ of possible values of θ, called the parameter space. To indicate the dependency of the density on θ we write $f(y; \theta)$ and refer to this as the model function.

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

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

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

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

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

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

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

1.2 Special models

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

1.2.1 Exponential families

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

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

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

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

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

The exponential form (1.1) is preserved if we apply any fixed nonsingular linear transformation to s, provided we make the inverse transformation to ϕ, leaving $s^T\phi$ invariant. If $0 \in \Omega_\phi$, we can without loss of generality take

$K(0) = 0$ and then $h(y) = f_Y(y; 0)$. In other cases we can measure ϕ from some suitable origin $\phi_0 \in \Omega_\phi$, by rewriting (1.1) as

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

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

We have from (1.1) that the moment generating function of the random variable S corresponding to s is

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

from which we obtain

$$E(S_i; \phi) = \frac{\partial K(\phi)}{\partial \phi^i},$$

or

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

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

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

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

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

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

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

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

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

say.

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

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

say.

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

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

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

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

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

where $\theta \in \Omega_\theta \subset \mathbb{R}^d$. We call this system an (m, d) exponential family, noting that we required that $(\phi^1, \ldots, \phi^m)$ does not belong to a v-dimensional linear subspace of $\mathbb{R}^m$ with $v < m$: we indicate this by saying that the exponential family is curved. Think of the case $m = 2, d = 1$: $\{\phi^1(\theta), \phi^2(\theta)\}$ defines a curve in the plane, rather than a straight line, as θ varies.

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

1.2.2 Transformation families

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

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

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

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

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

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

1.2.3 Maximal invariant

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

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

1.2.4 Equivariant statistics and a maximal invariant

As described, typically there is a one-to-one correspondence between the elements of G and the parameter space Ω_θ, and then the action of G on $\mathcal{X}$

requires that Ω_θ itself constitutes a group, with binary operation $*$ say: we must have $g_\theta \circ g_\phi = g_{\theta * \phi}$. The group action on $\mathcal{X}$ induces a group action on Ω_θ. If $\bar{G}$ denotes this induced group, then associated with each $g_\theta \in G$ there is a $\bar{g}_\theta \in \bar{G}$, satisfying $\bar{g}_\theta(\phi) = \theta * \phi$.

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

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

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

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

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

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

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

1.2.5 An example

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

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

The set of such transformations is closed with identity $g_{(0,1)}$. It is easy to check that $g_{(\eta,\tau)}$ has inverse $g_{(-\eta/\tau,\tau^{-1})}$. Hence, $G = \{g_{(\eta,\tau)} : (\eta, \tau) \in \mathbb{R} \times \mathbb{R}_+\}$

constitutes a group under the composition of functions operation $\circ$ defined above.

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

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

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

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

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

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

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

1.3 Likelihood

1.3.1 Definitions

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

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

Usually we work with the log-likelihood

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

sometimes studied as a random variable

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

In likelihood calculations, we can drop factors depending on y only, or additive terms depending only on y may be dropped from log-likelihoods. This idea can be formalised by working with the normed likelihood $\bar{L}(\theta) = L(\theta)/L(\hat{\theta})$, where $\hat{\theta}$ is the value of θ maximising $L(\theta)$. We define the score function by

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

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

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

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

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

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

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

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

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

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

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

1.3.2 Score function and information

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

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

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

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

Also, when (1.2) holds,

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

More compactly, the covariance matrix of U is

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

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

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

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

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

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

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

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

or

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

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

Similarly,

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

or

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

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

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

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

1.3.3 Pseudo-likelihoods

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

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

Another approach is available whenever there exists a statistic S such that the conditional distribution of the data X given $S = s$ depends only on ψ. In this case, we may form a likelihood function for ψ based on the conditional density function of X given $S = s$; this is called a **conditional likelihood** function. The drawback of this approach is that we discard the part of the likelihood function based on the marginal distribution of S, which may contain information about ψ.

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

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

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

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

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

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

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

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

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

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

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

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

1.4 Sufficiency

1.4.1 Definitions

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

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

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

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

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

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

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

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

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

1.4.2 Examples

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

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

1.5 Conditioning

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

1.5.1 The Bayesian stance

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

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

where now the model function $f_Y(y; \theta)$ is written as a conditional density $f_{Y|\Theta}(y \mid \Theta = \theta)$. The insertion of a random element in the generation of

θ allows us to condition on the whole of the data y: relevance to the data is certainly accomplished. This approach is uncontroversial if a meaningful prior can be agreed. In many applications, there may be major obstacles to specification of a meaningful prior and we are forced to adopt a less direct route to conditioning.

1.5.2 The Fisherian stance

Suppose first that the whole parameter vector θ is of interest. Reduce the problem by sufficiency. If, with parameter dimension $d = 1$, there is a one-dimensional sufficient statistic, we have reduced the problem to that of one observation from a distribution with one unknown parameter and there is little choice but to use probabilities calculated from that distribution. The same notion occurs if there is a d-dimensional θ of interest and a d-dimensional sufficient statistic. If the dimension of the (minimal) sufficient statistic exceeds that of the parameter, there is scope and need for ensuring relevance to the data under analysis by conditioning.

We therefore aim to

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

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

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

1.5.3 An example

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

1.6 Ancillarity and the Conditionality Principle

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

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

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

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

1.6.1 Nuisance parameters

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

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

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

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

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

1.7 Sample space derivatives

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

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

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

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

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

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

The maximum likelihood equation is

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

so that

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

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

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

so that

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

or, for a vector parameter,

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

1.8 Parameter Orthogonality

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

1.8.1 Definition

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

1.8.2 An immediate consequence

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

1.8.3 The case $d_1 = 1$

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

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

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

and use of the chain rule for differentiation gives

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

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

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

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

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

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

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

1.8.4 An example

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

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

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

1.8.5 The case $d_1 > 1$

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

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

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

1.8.6 Further remarks

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

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

1.9 General principles

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

1.9.1 Sufficiency principle

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

1.9.2 Conditionality principle

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

1.9.3 Weak likelihood principle

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

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

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

This is identical with the sufficiency principle.

1.9.4 Strong likelihood principle

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

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

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

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

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

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

1.9.5 Repeated sampling principle

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

1.9.6 Bayesian coherency principle

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

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

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

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

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

1.9.7 Principle of coherent decision making

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

1.10 Approaches to Statistical Inference

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

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

1.10.1 Sampling theory

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

1.10.2 Likelihood theory

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

1.10.3 Bayesian theory

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

1.10.4 Decision theory

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

1.11 Some Essential Mathematical Material

1.11.1 Background

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

1.11.2 Some probability results

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.11.3 Mann-Wald notation

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

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

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

1.11.4 Moments and cumulants

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

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

as $t \to 0$.

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

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

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

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

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

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

1.11.5 Some reminders

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

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

where

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

and the remainder R_n is of the form

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

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

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

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

where

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

and similarly for the other terms.

Some particular expansions therefore are:

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

The sign function sgn is defined by

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

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

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

and

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

Then,

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

1.11.6 Multivariate normal distribution

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

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

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

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

1. the covariance form is invariant under non-singular transformation of Y;
2. Y can be transformed to independent components of unit variance (set $Z = \Sigma^{-1/2}Y$);
3. the chi-squared distribution then follows directly, $Q_Y \equiv Q_Z = Z^TZ$.

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

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

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

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

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