Statistical Asymptotics

G. Alastair Young

Department of Mathematics
Imperial College London

APTS, April 2008

Statistical objective

Statistical objective

To analyse observations $y = (y_1, \ldots, y_n)$.

- Regard y as observed value of random variable $Y = (Y_1, \ldots, Y_n)$ having an (unknown) probability distribution specified by a probability density function, or probability mass function, $f(y)$.

Statistical objective

To analyse observations $y = (y_1, \ldots, y_n)$.

- Regard y as observed value of random variable $Y = (Y_1, \ldots, Y_n)$ having an (unknown) probability distribution specified by a probability density function, or probability mass function, $f(y)$.
- 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'.

Statistical objective

To analyse observations $y = (y_1, \ldots, y_n)$.

- Regard y as observed value of random variable $Y = (Y_1, \ldots, Y_n)$ having an (unknown) probability distribution specified by a probability density function, or probability mass function, $f(y)$.
- 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 assessing some aspect of θ, for example the value of a single component θ^i.

Neo-Fisherian Statistics

Provide a framework for the relatively systematic analysis of a wide range of possible $\mathcal{F}$.

Neo-Fisherian Statistics

Provide a framework for the relatively systematic analysis of a wide range of possible $\mathcal{F}$.

We do not aim to satisfy formal optimality criteria.

Neo-Fisherian Statistics

Provide a framework for the relatively systematic analysis of a wide range of possible $\mathcal{F}$.

We do not aim to satisfy formal optimality criteria.

Focus on the likelihood function and quantities derived from it: a 'neo-Fisherian' approach to inference.

Special Models

Two general classes of models particularly relevant in theory and practice are:

Special Models

Two general classes of models particularly relevant in theory and practice are:

- exponential families

Special Models

Two general classes of models particularly relevant in theory and practice are:

- exponential families
- transformation families

Exponential Families

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 (with probability one). 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.

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 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 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.

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, $d < m$.

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 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 that we required that $(\phi^1, \ldots, \phi^m)$ does not belong to a v-dimensional linear subspace of $\mathbb{R}^m$ with $v < m$.

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

Transformation families

A transformation family is defined by a group of transformations acting on the sample space which generates a family of distributions all of the same form, but with different values of the parameters.

Present context

Concerned with group G of transformations acting on sample space $\mathcal{Y}$ of random variable Y, binary operation $\circ$ is composition of functions. 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, transformation may be written as g_θ. The family of densities of $g_\theta(Y)$, for $g_\theta \in G$ is called a (group) transformation family.

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 y or $g(y)$ was observed, for any $g \in G : t(y) = t(g(y))$.

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 y or $g(y)$ was observed, for any $g \in G : t(y) = t(g(y))$.

The statistic t is maximal invariant if every other invariant statistic is a function of it, or equivalently, $t(y) = t(y')$ implies that $y' = g(y)$ for some $g \in G$.

Group action on Ω_θ

Typically, there is a one-to-one correspondence between the elements of G and the parameter space Ω_θ.

Group action on Ω_θ

Typically, there is a one-to-one correspondence between the elements of G and the parameter space Ω_θ.

Assume this.

Then the action of G on $\mathcal{Y}$ requires that Ω_θ itself constitutes a group, with binary operation $*$ say: we must have $g_\theta \circ g_\phi = g_{\theta * \phi}$.

Then the action of G on $\mathcal{Y}$ requires that Ω_θ itself constitutes a group, with binary operation $*$ say: we must have $g_\theta \circ g_\phi = g_{\theta * \phi}$.

Group action on $\mathcal{Y}$ induces group action on Ω_θ. If $\bar{G}$ denotes induced group, associated with each $g_\theta \in G$ is a $\bar{g}_\theta \in \bar{G}$, satisfying $\bar{g}_\theta(\phi) = \theta * \phi$.

Distribution constant statistic

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

Distribution constant statistic

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

T is said to be distribution constant.

Equivariant statistic

A statistic $S = s(Y)$ defined on $\mathcal{Y}$ and taking values in the parameter space Ω_θ is said to be equivariant if $s(g_\theta(y)) = \bar{g}_\theta(s(y))$ for all $g_\theta \in G$ and $y \in \mathcal{Y}$.

Often S is chosen to be an estimator of θ, and it is then called an equivariant estimator. An equivariant estimator can be used to construct a maximal invariant.

Often S is chosen to be an estimator of θ, and it is then called an equivariant estimator. An equivariant estimator can be used to construct a maximal invariant.

Consider $t(Y) = g^{-1}_{s(Y)}(Y)$.

Often S is chosen to be an estimator of θ, and it is then called an equivariant estimator. An equivariant estimator can be used to construct a maximal invariant.

Consider $t(Y) = g^{-1}_{s(Y)}(Y)$.

Then $t(Y)$ is maximal invariant.

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

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

Observed and expected information

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

The Hessian matrix $-\nabla\nabla^T l$ is called the observed information matrix, and is denoted by $j(\theta)$.

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

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.

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.

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

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

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

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

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, for all s, 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 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.

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.

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: sufficient statistic has same dimension as Y.

Conditioning

In methods of statistical inference, probability is used in two quite distinct ways.

Conditioning

In methods of statistical inference, probability is used in two quite distinct ways.

- To define the stochastic model assumed to have generated the data.

Conditioning

In methods of statistical inference, probability is used in two quite distinct ways.

- To define the stochastic model assumed to have generated the data.
- To assess uncertainty in conclusions. The probabilities used for the basis of inference are long-run frequencies under hypothetical repetition from the assumed model.

The issue arises of how these long-run frequencies are to be made relevant to the data under study.

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.

The Bayesian stance

In a Bayesian approach conditioning is dealt with automatically.

The Bayesian stance

In a Bayesian approach conditioning is dealt with automatically.

The particular value of θ is itself 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 $f_{Y|\Theta}(y \mid \Theta = \theta)$.

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.

The Fisherian stance

Suppose first that the whole parameter vector θ is of interest.

The Fisherian stance

Suppose first that the whole parameter vector θ is of interest.

Reduce the problem by sufficiency.

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.

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.

Ancillarity, 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.

Ancillarity, 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 parameter of interest, θ, is to be made conditional on $A = a$ i.e. on the basis of the conditional distribution of Y given $A = a$, its observed value, rather than from the model function $f_Y(y; \theta)$.

Nuisance parameter case

Suppose, more generally, that we can write $\theta = (\psi, \chi)$, where ψ is of interest and χ is nuisance. 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. either:
   - (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;

A Conditionality Principle

Inference about ψ should be based upon the conditional distribution of T given $A = a$. Still refer to A as ancillary.

A Conditionality Principle

Inference about ψ should be based upon the conditional distribution of T given $A = a$. Still refer to A as ancillary.

The most straightforward case corresponds to (a). 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.

A Conditionality Principle

Inference about ψ should be based upon the conditional distribution of T given $A = a$. Still refer to A as ancillary.

The most straightforward case corresponds to (a). 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.

An important modern convention

Often we use the term ancillary to mean a distribution constant statistic which, together with the MLE, constitutes a (minimal) sufficient statistic.

An important modern convention

Often we use the term ancillary to mean a distribution constant statistic which, together with the MLE, constitutes a (minimal) sufficient statistic.

Then we can write the log-likelihood as $l(\theta; \hat{\theta}, a)$.

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.

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