Fundamental Theory of Statistical Inference

G. Alastair Young

Department of Mathematics
Imperial College London

LTCC, 2017

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

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

Usually we work with the log-likelihood

$$l(\theta; y) = \log f_Y(y;\theta).$$

General case

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

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

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

Likelihood Principle

Two different random systems, the first giving observations y corresponding to a random variable Y and the second giving observations z on a random variable Z, the corresponding densities 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.

Example: Bernoulli trials

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.

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 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.

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.

Completeness

A sufficient statistic $T(Y)$ is complete if for any real function g,

$$\mathbb{E}_\theta\{g(T)\} = 0, \text{ for all } \theta$$

implies

$$\Pr_\theta\{g(T) = 0\} = 1 \text{ for all } \theta.$$

Key consequence

If there exists an unbiased estimator of a scalar parameter θ which is a function of a complete sufficient statistic T, then it is the unique such estimator (except possibly on a set of measure 0).

If $g_1(T)$ and $g_2(T)$ are two such estimators, then $\mathbb{E}_\theta\{g_1(T) - g_2(T)\} = \theta - \theta = 0$, so $g_1(T) = g_2(T)$ with probability 1.

Key example

If $S \equiv s(Y) = (s_1(Y), ..., s_m(Y))$ is the natural statistic for a full exponential family in its natural parametrisation and if Ω_ϕ contains an open rectangle in $\mathbb{R}^m$, then S is complete.

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 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 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 $f_{Y|\Theta}(y \mid \Theta = \theta)$.

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.

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.

An Example

Let $Y_1, \ldots, Y_n$ be IID $U(\theta - 1, \theta + 1)$.

The (minimal) sufficient statistic is the pair of order statistics $(Y_{(1)}, Y_{(n)})$, where $Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$ and $Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$.

Transform 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 be (re-)expressed as $(\bar{Y}, R)$.

R has a distribution not depending on θ.

Inference should be based on the conditional distribution of $\bar{Y}$, given $R = r$, which is $U(\theta - 1 + \frac{1}{2}r, \theta + 1 - \frac{1}{2}r)$.

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.

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.