CHALMERS/GOTHENBURG UNIVERSITY
Principles of statistical inference
Outline notes

D.R. Cox
Department of Statistics and Nuffield College, Oxford
April 2004

1 Preliminaries

1.1 Starting point

We typically start with a subject-matter question. Data are or become available to address this question. After preliminary screening and simple tabulations and graphs, more formal analysis starts with a provisional model. The data are typically split in two parts $(y : z)$, where y is regarded as the observed value of a vector random variable Y and z is treated as fixed. A model, or strictly a family of models, specifies the density of Y to be

$$f_Y(y : z; \theta),$$

where $\theta \subset \Omega_\theta$ is unknown. The distribution may depend also on design features of the study that generated the data. We typically simplify the notation to $f_Y(y; \theta)$.

Further, for each aspect of the research question we partition θ as (ψ, λ), where ψ is called the *parameter of interest* and λ is included to complete the

specification and commonly called a *nuisance parameter*. Usually but not necessarily ψ and λ are *variation independent* in that Ω_θ is the Cartesian product $\Omega_\psi \times \Omega_\lambda$. The choice of ψ is a subject-matter question. In most applications it is best to arrange that ψ is a scalar parameter, i.e. to break the research question of interest into simple components, but this is not necessary for the theoretical discussion.

We concentrate on problems where Ω_θ is a subset of R^d, so-called *fully parametric* problems. Other possibilities are to have semi-parametric problems or fully nonparametric problems. These typically involve fewer assumptions of structure and distributional form but usually contain strong assumptions about independencies.

1.2 Role of formal theory of inference

The formal theory of inference takes the family of models as given and the objective to answer questions about the model in the light of the data. Choice of the family of models is obviously crucial but outside the scope of the present lectures. More than one choice may be needed to answer different questions.

1.3 Some simple examples

General notation is often not best suited to special cases and so we use more conventional notation where appropriate.

Example 1. Linear regression. Here the data are n pairs $(y_1, z_1), \ldots, (y_n, z_n)$ and the model is that $Y_1, \ldots, Y_n$ are independently normally distributed with variance σ^2 and with

$$E(Y_k) = \alpha + \beta z_k.$$

Here typically but not necessarily $\psi = \beta$ and $\lambda = (\alpha, \sigma^2)$.

Example 2. Semiparametric (second-order) linear regression. In Example 1 replace the assumption of normality by an assumption that the Y_k are uncorrelated with constant variance.

Example 3. Linear model. We have a $n \times 1$ vector Y and a $n \times q$ matrix z of fixed constants such that

$$E(Y) = z\beta, \quad \text{cov}(Y) = \sigma^2 I,$$

with in the analogue of Example 1 the components independently normally distributed. A relatively simple but important generalization has $\text{cov}(Y) = \sigma^2 V$. Here z is in initial discussion at least assumed of full rank $q < n$ and V is a given positive definite matrix.

Example 4. Exponential distribution. Here the data are $(y_1, \ldots, y_n)$ and the model takes $Y_1, \ldots, Y_n$ to be independently exponentially distributed with density $\rho e^{-\rho y}$, where $\rho > 0$ is an unknown rate parameter. Note that possible parameters of interest are ρ, $\log \rho$ and $1/\rho$ and the issue will arise of possible invariance or equivariance of the inference under reparameterization, i.e. shifts from, say ρ to $1/\rho$. The observations might be intervals between successive points in a Poisson process of rate ρ.

Example 5. Comparison of binomial probabilities. Suppose that the data are (r_0, n_0) and (r_1, n_1), where r_k denotes the number of successes in n_k binary trials under condition k. Take as model that the trials are mutually independent with probabilities of success π_0 and π_1. Then the random variables R_0 and R_1 have independent binomial distributions. We want to compare the probabilities and for this may take various forms for the parameter of interest, for example

$$\psi = \log\{\pi_1/(1-\pi_1)\} - \log\{\pi_0/(1-\pi_0)\}, \qquad \pi_1 - \pi_0,$$

and so on. For many purposes it is immaterial how we define the complementary parameter λ.

The object is to provide a set of ideas that will deal systematically with the above relatively simple situations and, more importantly, enable us to deal with new models that arise in new situations.

1.4 Formulation of objectives

We can formulate possible objectives in two parts as follows.

Part I takes the family of models as given.

- give intervals or in general sets of values within which ψ is in some sense likely to lie
- assess the consistency of the data with a particular parameter value ψ_0
- predict as yet unobserved random variables from the same random system that generated the data
- use the data to choose one of a given set of decisions $\mathcal{D}$, there being given a utility function $U(.,.)$ on $\Omega_\psi \times \mathcal{D}$ often expressed via the decision loss or regret. Utility is that measure of gain or value taking the expectation of which is the appropriate basis for decision-making.

Part II uses the data to examine the family of models

- does the secondary structure of the model need improvement?
- does the parameter of interest need redefinition in detail without change of the research question
- does the focus of the research need major change?

As examples of the first two, in Example 1 it might be necessary to represent that the variance of Y changes with z or that the linear regression should be of $\log Y$ on $\log z$.

We shall concentrate on the first two of the objectives in Part I.

1.5 Broad approaches

Consider the first objective above. There are two broad approaches, both with variants.

First we may aim to treat θ, and hence ψ, as having a probability distribution. This raises two issues: what does probability mean in such a context and how do we obtain numerical values for the relevant probabilities?

Secondly we may regard θ as an unknown constant and appeal to some notion of the long run frequency properties of our procedures in hypothetical repetitions: what would the properties of this method of analysis be if we were to apply it again and again?

In the first method if we can treat θ as the realized but unobserved value of a random variable, all is in principle straightforward. By Bayes's theorem

$$f_{\Theta|Y}(\theta \mid y) = f_{Y|\Theta}(y \mid \theta) f_{\Theta}(\theta) / \int f_{Y|\Theta}(y \mid \phi) f_{\Theta}(\phi) d\phi.$$

The left-hand side is called the *posterior density* of Θ and the two terms in the numerator are one determined by the model and the other forms the *prior distribution* summarizing information about Θ other than from y. The posterior distribution for Ψ is formed by integration. Any method of inference treating the unknown parameter as having a probability distribution is called *Bayesian*. The intuitive idea is that in such cases all relevant information is then contained in the conditional distribution of the parameter given the data, that this is determined by the elementary formulae of probability theory and that remaining problems are essentially computational.

Example 6. Normal mean. Suppose that $Y_1, \ldots, Y_n$ are independently normally distributed with mean μ and known variance σ_0^2 and that the prior for μ is normal with mean m and variance v. Then the posterior density for μ is proportional to

$$\exp\{-\Sigma(y_k - \mu)^2/(2\sigma^2) - (\mu - m)^2/(2v)\}$$

considered as a function of μ. There results a normal distribution of mean and variance respectively

$$\frac{\bar{y}/(\sigma_0^2/n) + m/v}{1/(\sigma_0^2/n) + 1/v},$$
$$\frac{1}{1/(\sigma_0^2/n) + 1/v}.$$

a weighted mean. Note that if v is much greater than σ_0^2/n then approximately μ given the data is $N(\bar{y}, \sigma_0^2/n)$. The limiting prior with constant density over R is an example of an *improper prior*, but special issues with these are best evaded.

In the second method we may base the argument solely on the operating characteristics of the procedure, i.e. on the properties of the procedure under repeated application, or may appeal to principles such as sufficiency to be discussed later.

1.6 An outline history

The following notes merely mention some key contributors to the main styles of statistical inference.

Some early work

Laplace (1749-1827). Used flat priors.

Boole (1815-1864). Criticized flat priors.

K.Pearson (1857-1936). The chi-squared test, 1901.

Bases of current theory

R.A.Fisher (1890-1962). Foundational papers, 1922,1925.

J. Neyman (1894-1981) and E.S. Pearson (1885-1980). Joint papers 1928-initially clarifying Fisher.

A.Wald (1902-1950). *Statistical decision functions*, 1950.

Probability as reasonable degree of belief

H.Jeffreys (1891-1989). Earlier book and then *Theory of probability*, 1939,1961.

Probability as coherent personalistic degree of belief

F.P. Ramsey (1902-1928). Essay on personalistic probability, 1926

B.de Finetti (1906-1985). Independent development in particular emphasizing exchangeability.

L.J.Savage (1917-1971). Independent development more focused on statistical applications.

2 Some more technical ideas

2.1 Likelihood

The *likelihood* for observations y is defined as

$$\text{lik}(\theta; y) = f_Y(y; \theta)$$

considered in the first place as a function of θ for given y. Mostly we work with its logarithm, $l(\theta; y)$ often abbreviated to $l(\theta)$. Sometimes it is treated as a function of the random vector Y rather than of y. The log form is convenient in particular because f will often be a product of component terms. Occasionally we work directly with the likelihood function itself. In some contexts the likelihood is better defined as the equivalence class of all functions obtained by multiplying by a positive function of y.

Any calculation of a posterior density, whatever the prior distribution, uses the data only via the likelihood. Beyond that, there is some intuitive appeal in the idea that $l(\theta)$ measures the relative effectiveness of different parameter values θ in explaining the data. Some have elevated this into a principle sometimes called the law of the likelihood.

2.2 Sufficiency

2.2.1 Definition

The term *statistic* is often used (rather oddly) to mean any function of the observed random variable Y or its observed counterpart. A statistic $S = S(Y)$ is called *sufficient* under the model if the conditional distribution of Y given $S = s$ is independent of θ for all s, θ. Equivalently

$$l(\theta; y) = \log h(s, \theta) + \log m(y),$$

for suitable functions h and m. The equivalence forms the Neyman factorization theorem. For the essence of the proof, see Cox and Hinkley (1974, p.22). We can take $h(s, \theta)$ to be the density of S.

We use the minimal form of S, i.e. extra components could always be added to a given S and we do not do that.

Any Bayesian inference uses the data only via the minimal sufficient statistic.

In other contexts there are two implications of sufficiency. One is that given the model and disregarding considerations of robustness or computational simplicity the data should be used only via s. The second is that the *known* distribution $f_{Y|S}(y \mid s)$ is available to examine the adequacy of the model. Sufficiency may reduce the dimensionality of a problem but we still have to determine what to do with the sufficient statistic once found.

2.2.2 Simple examples

Example 7. Exponential distribution. The likelihood for Example 4 is

$$\rho^n \exp(-\rho \Sigma y_k)$$

so that the log likelihood is

$$n \log \rho - \rho \Sigma y_k,$$

and involves the data only via Σy_k or equivalently via $\bar{y} = \Sigma y_k / n$. By the factorization theorem the sum (or mean) are therefore sufficient. Note that had the sample size been random the sufficient statistic would have been $(n, \Sigma y_k)$.

Example 8. Linear model. A minimal sufficient statistic for the linear model, Example 3 consists of the least squares estimates and the residual sum of squares.

Example 9. Rectangular distribution. Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed in a rectangular (uniform) distribution on (θ_1, θ_2). Let $Y_{(1)}, Y_{(n)}$ be the largest and smallest order statistics. Then the likelihood involves the data only via $(y_{(1)}, y_{(n)})$ which are therefore minimal sufficient. In general care is needed in writing down the likelihood and in evaluating the resulting procedures when the form of the likelihood has discontinuities.

2.2.3 Exponential family

Suppose that θ takes values in a well-behaved region in R^d and that we can find a d dimensional statistic s and a parameterization ϕ, i.e. a (1,1) transformation of θ, such that the model has the form

$$m(y)\exp\{s^T\phi - k(\phi)\}.$$

Then S is sufficient; subject to some important regularity conditions this is called a *regular or full* (d, d) *exponential family* of distributions. The statistic S is called the *canonical statistic* and ϕ the *canonical parameter*. The parameter $\eta = E(S; \phi)$ is called the *mean parameter*. Because $m(y)$ is a normalizing function it follows that

$$\log E(\exp(p^T S)) = k(\phi + p) - k(\phi),$$

i.e. $k(.)$ has an interpretation via a cumulant generating function. Thus $\eta = \nabla k(\phi)$.

Example 10. Binomial distribution. If R denotes the number of successes in n independent binary trials each with probability of success π, its density can be written

$$\{n!/\{r!(n-r)!\}\pi^r(1-\pi)^{n-r} = m(r)\exp\{r\phi - n\log(1+e^{\phi})\},$$

where $\phi = \log\{\pi/(1-\pi)\}$ is the canonical parameter and r the canonical statistic.

Note that the likelihood as a function of ϕ is maximized by solving the equation in $\hat{\phi}$

$$s = \nabla k(\phi)_{\phi=\hat{\phi}} = \hat{\eta},$$

where it can be shown that for the regular exponential family the solution is unique and indeed corresponds to a maximum unless s lies on the boundary of its support. We call $\hat{\phi}$ and $\hat{\eta}$ the *maximum likelihood estimates* of ϕ and η respectively. Essentially we equate the canonical statistic to its expectation to form an estimating equation.

The central importance of this will appear later.

Essentially all well-behaved models with sufficient statistic having the same dimension as the parameter space are of this form. For a few more details see Barndorff-Nielsen and Cox (1994, section 1.3).

A further notion is that of a curved exponential family, Suppose that θ is of dimension d but that the minimal sufficient statistic is of dimension k, where $k > d$. Then the exponential family form will still hold but now with ϕ a $k \times 1$ vector, a function of d originating parameters. That is ϕ is constrained to lie on a curved surface in its k-dimensional space. This is called a (k, d) *curved exponential family*. The statistical properties of curved exponential families are much less simple than those of full exponential families.

Example 11. Fisher's hyperbola. Suppose that the model is that the pairs (Y_{k1}, Y_{k2}) are independently normally distributed with unit variance and with means $(\theta, c/\theta)$, where c is a known constant.

Example 12. Pure birth process. Suppose that a pure birth process with birth rate ρ is observed for a given time. Each interval between birth and reproduction contributes a factor to the likelihood as does each incomplete period at the end. The likelihood is thus

$$\rho^n \exp(-\rho t),$$

where n is the number of births and t the total time at risk. This is a $(2, 1)$ family. Note, however, that if we observed until either one of n and t took on a preassigned value then it would reduce to a full, i.e. $(1, 1)$ family.

While in Bayesian theory choice of prior is in principle not an issue of achieving mathematical simplicity, nevertheless there are gains in using reasonably simple and flexible forms. In particular if the likelihood has the full exponential family form

$$m(y)\exp\{s^T\phi - k(\phi)\}$$

a prior proportional to

$$\exp\{s_0^T\phi - a_0 k(\phi)\}$$

leads to a posterior proportional to

$$\exp\{(s+s_0)^T)\phi - (1+a_0)k(\phi)\}.$$

Such a prior is called *conjugate to the likelihood*, or sometimes *closed under sampling.*

Example 13. Binomial sampling. If the prior for π is proportional to

$$\pi^{r_0}(1-\pi)^{n_0-r_0},$$

i.e. is a beta distribution, then the posterior is another beta distribution corresponding to $r + r_0$ successes in $n + n_0$ trials. Thus both prior and posterior are beta distributions. It may help to think of r_0, n_0 as fictitious data!

2.2.4 Exponential family and conditioning

Suppose that we are interested in the first component of ϕ or more generally in a collection of components. Note that a linear transformation of ϕ is allowable. Then we can write the density of the observations in the form

$$m(y)\exp\{s_1\phi_1 + s_2^T\phi_2 - k(\phi)\}.$$

It follows that the conditional distribution of S_1 given $S_2 = s_2$ has the form

$$h(s, \phi_1) \exp(s_1 \phi_1),$$

where $h(s_2)$ is a normalizing constant arising from the Jacobian of the transformations involved or their discrete analogue.

That this is essentially the only way to obtain a distribution not depending on ϕ_2 follows from the completeness of the full exponential family, i.e. from the uniqueness of the generalized Laplace transform.

This prepares the way for inference about $\psi = \phi_1$ treating $\lambda = \phi_2$ as a nuisance parameter.

Example 14. Comparison of Poisson distributions or processes. Suppose that Y_0, Y_1 have independent Poisson distributions with means μ_0, μ_1 and that interest focuses on μ_1/μ_0. In particular, the observations might correspond to counting how many points occur in two Poisson processes of rates ρ_0, ρ_1 observed for times t_0, t_1, when $\mu_i = \rho_i t_i$. In exponential family form the joint distribution is

$$(y_1! y_0!)^{-1} \exp(y_1 \log \mu_1 + y_0 \log \mu_0 - \mu_1 - \mu_0).$$

If we write $\psi = \log \mu_1 - \log \mu_0, \quad \lambda = \log \mu_1 + \log \mu_0$, then the operative part of the density is

$$(y_1 - y_0)\psi + (y_0 + y_1)\lambda$$

and thus a distribution for inference about ψ free of λ is obtained from the conditional distribution of $Y_1 - Y_0$ given $Y_0 + Y_1 = t$, say. Equivalently we may use the distribution of Y_1 given t and this is binomial corresponding to t trials with probability of "success" $\mu_1/(\mu_1 + \mu_0)$. This has logistic transform ψ.

Example 15. Linear logistic regression. Let $Y_1, \ldots, Y_n$ be independent binary random variables with

$$\log\{P(Y_k = 1)/P(Y_k = 0)\} = \beta^T z_k,$$

where z_k is a vector of explanatory variables attached to Y_k. Then the likelihood has exponential family form with canonical parameter β, so that in principle exact conditional inference is possible about any component of β.

Example 16. Log Poisson regression. Suppose that $Y_{10}, Y_{11}, z_1, \ldots, Y_{n0}, Y_{n1}, z_n$ are n independent pairs of Poisson variables with a vector of explanatory variables attached to each pair and such that

$$E(Y_{k0}) = \mu_{k0}, \quad E(Y_{k1}) = \mu_{k0} \exp(\beta^T z_k),$$

where the μ_{k0} are not of direct interest. Then conditionally on $T_k = Y_{k1} + Y_{k0} = t_k$ the Y_{k1} are binomial random variables corresponding to t_k trials with probability of "success" having logistic transform $\beta^T z_k$. Thus exact inference about any component of β is possible.

2.3 Pivots

The following definition is helpful in applications.

Suppose for simplicity that ψ is one-dimensional. Suppose that there is a statistic t and a function $p(t, \psi)$ such that for all $\theta \subset \Omega_\theta$ the random variable $p(T, \psi)$ has a fixed and known continuous distribution and that for all t the function $p(t, \psi)$ is strictly increasing in ψ. Then p is called a *pivot* for inference about ψ and the fixed distribution is called the *pivotal distribution.*

We have that for all ψ and all c, $0 < c < 1$ we can find p_c such that

$$P\{p(T, \psi) \leq p_c\} = 1 - c,$$

implying that

$$P\{\psi \leq q(T, c)\} = 1 - c.$$

We call $q(t, c)$ a c *level upper limit* for ψ with obvious changes for lower limits and intervals. The collection of limits for all c encapsulates our information about ψ and the associated uncertainty. In many applications it is convenient to summarize this by giving an upper c limit and a lower c limit forming a $1-2c$ *equi-tailed confidence interval*, usually specifying this for one or two conventionally chosen values of c. The use of equal tails is essentially a presentational simplification. Clearly other than equal tails could be used if there were good reason for doing so.

Example 17. Normal mean. The data are $(y_1, \ldots, y_n)$ and the model is that the Y_k are independent and identically normally distributed with unknown mean μ, that will take the role of ψ and known variance σ_0^2. Let $\bar{Y} = \Sigma Y_k/n$. Then a pivot for μ is

$$\frac{\mu - \bar{Y}}{\sigma_0/\sqrt{n}}$$

and the pivotal distribution is the standard normal. Thus $p_c = \Phi^{-1}(1 - c)$, where $\Phi(.)$ is the standard normal distribution function and we have for all ψ that

$$P(\mu < \bar{Y} + p_c\sigma_0/\sqrt{n}) = 1 - c,$$

so that $\bar{y} + p_c\sigma_0/\sqrt{n}$ is a c level upper limit for μ.

Example 17 can be generalized in many ways. If the standard deviation is unknown and σ_0 is replaced by its standard estimate derived from the residual sum of squares the pivotal distribution becomes the Student t distribution with $n - 1$ degrees of freedom. Further the mean could be replaced by the

median or by any other location estimate with, however, a modification of the pivotal distribution. The argument applies directly to the estimation of any linear parameter in the linear model of Example 2 and in the second-order linear model the pivotal distribution is asymptotically standard normal by the Central Limit Theorem and, if the variance is estimated, also by appeal to the Weak Law of Large Numbers.

Note that the limits of Example 17 correspond to the upper c point of the posterior distribution of μ on taking the limit in Example 6 as nv/σ_0^2 increases. But a warning about interpretation of apparently uninformative priors is essential.

Example 18. Stein's paradox; the noncentral chi-squared distribution. Let $Y_1, \ldots, Y_n$ be independently normally distributed with unit variance and means $\mu_1, \ldots, \mu_n$ referring to independent situations and therefore with independent flat matching priors. Suppose that interest focuses on $\Delta^2 = \Sigma\mu_k^2$. Its posterior distribution is noncentral chi-squared with n degrees of freedom and noncentrality $D^2 = \Sigma Y_k^2$. This implies that, for large n, Δ^2 is with high probability $D^2 + n + O_p(\sqrt{n})$. But this is absurd in that whatever the true value of Δ^2, the statistic D^2 is with high probability $\Delta^2 + n + O_p(\sqrt{n})$. A very flat prior in one dimension gives good results from almost all viewpoints, whereas a very flat prior and independence in many dimensions do not.

Example 19. Exchange paradox. There are two envelopes one with ϕ Skr and the other with 2ϕ Skr. One is chosen at random and given to you and when opened it is found to contain 10^3 Skr. You now have a choice. You may keep the 10^3 Skr or you may open the other envelope in which case you keep its contents. You argue that the other envelope is equally likely to contain 500 Skr or 2×10^3 Skr and provided utility is linear in money the expected utility of the new envelope is 1.25×10^3 Skr and so you take the

new envelope. This argument does not depend on the particular value 10^3 so that there was no need to open the first envelope.

The conclusion is clearly wrong. The error stems from attaching a probability in a non-Bayesian context to the content of the new envelope given the observations; the only probability is that of the observation given the parameter, i.e. given the contents. Both possible values for the content of the new envelope assign the same likelihood to the data, but this is not the same as the data assigning equal probabilities to the possible values. A Bayesian discussion is possible.

3 Some interpretational issues

3.1 General

We can now consider some issues involved in formulating and comparing the different approaches.

In some respects the Bayesian formulation is the simplest and in other respects the most difficult. Once a likelihood and a prior are specified to a reasonable approximation all problems are in principle at least straightforward. The resulting posterior distribution can be manipulated in accordance with the ordinary laws of probability. The difficulties centre on the meaning of the prior and then on its numerical specification to sufficient accuracy.

Sometimes, as in certain genetical problems, it is reasonable to think of θ as generated by a stochastic mechanism. Then there is no dispute that the Bayesian approach is a right one. In other cases to use the formulation in a literal way we have to regard probability as measuring uncertainty in a sense not necessarily directly linked to frequencies. We return to this briefly later. There is another justification of some Bayesian methods is that they

provide an algorithm for procedures whose fundamental justification comes from frequentist considerations.

The emphasis in these notes is on the close relation between answers possible from different approaches. This does not imply that the differences between different views never conflict or that the differences are never conceptually important.

3.2 Frequentist interpretation of upper limits

Now consider the interpretation of upper limits obtained for example from a suitable pivot. We take the simplest example, namely the normal mean, but the considerations are quite general. The upper limit

$$\bar{y} + p_c\sigma_0/\sqrt{n}$$

derived here from the probability statement

$$P(\mu < \bar{Y} + p_c\sigma_0/\sqrt{n}) = 1 - c$$

is a particular instance of a *hypothetical* long run of statements a proportion $1 - c$ of which will be true, always of course assuming our model sound. We can, at least in principle, make such a statement for each c and thereby generate a collection of statements with the formal properties of a distribution, sometimes called a *confidence distribution*. There is no restriction to a single c, so long as some compatibility requirements hold.

This has the formal properties of a distribution for μ called by Fisher the *fiducial distribution* and sometimes the fiducial probability distribution. A crucial question is whether this distribution can be interpreted and manipulated like an ordinary probability distribution. The position is:

- a single set of limits for μ from some data can in some respects be considered just like a probability distribution for μ

- such probability distributions cannot in general be combined or manipulated by the laws of probability
- a proper probability distribution for μ would allow us to evaluate the chance that μ exceeds some *given* constant, for example zero. This is not allowed here.

As a single statement a $1-c$ upper limit has the evidential force of a statement of a unique event within a probability system. But the rules for manipulating probabilities in general do not apply. The limits are of course directly based on probability calculations.

Note finally that the identical limits are obtained and have a direct Bayesian probabilistic interpretation in the limit for a very extended (flat) prior. Such a limiting prior is said to be *exactly matching*.

At an applied level it is, I think, a mistake to put too much emphasis on these relatively subtle issues; they will be discussed again in connection with the interpretation of probability.

3.3 Fisherian reduction

The essence of one approach to simple problems is as follows.

- find the likelihood function
- reduce to a sufficient statistic S of the same dimension as θ
- find a function of S that has a distribution depending only on ψ
- place it in pivotal form
- invert to obtain limits for ψ at an arbitrary set of probability levels.

Effective application is largely confined to simple regular exponential family problems.

There is sometimes an extension of the method that works when the reduction is only to (k, d) curved exponential family form. In this

- rewrite the k dimensional statistic, if $k > d$, in the form S, A such that S is of dimension d and A has a distribution not depending on θ
- consider the distribution of S given $A = a$ and proceed as before. The statistic A is called *ancillary.*

Example 20. Linear regression. Suppose that in the linear regression model of Example 1 the z_k have a known distribution. Then the sufficient statistic is augmented by $\Sigma z_k, \Sigma z_k^2$, i.e. we have a (3, 5) exponential family. But $\Sigma z_k, \Sigma z_k^2$ have a fixed distribution and play the role of A and are conditioned on, as is the standard practice in developing the theory of linear regression. Now the condition that the distribution of the z_k is *known* is not really used in this argument. It would be enough that the distribution of the z_k depended on a parameter ξ, say, variation-independent of θ.

There are obvious limitations to these methods especially over the second and third components and one is driven to approximate, i.e. asymptotic, arguments for problems of reasonable complexity and sometimes even for simple problems.

3.4 Neyman-Pearson operational criteria

Here we wish to find, for example, upper limits for ψ with the appropriate frequency interpretation, i.e. derived from a property such as

$$P(\mu < \bar{Y} + p_c \sigma_0/\sqrt{n}) = 1 - c.$$

Initially we may require that exactly (or to a sufficient approximation) for all θ

$$P\{\psi < T(Y;c)\} = 1 - c,$$

where $T(Y;c)$ is a function of the observed random variable Y. This ensures that the right coverage probability is achieved. There are likely to be many ways of achieving this, some in a sense more efficient than others. It is convenient to define optimality by requiring

$$P\{\psi' < T(Y;c)\}$$

to be minimal subject to correct coverage for all $\psi' > \psi$.

Essentially this strong optimality requirement is satisfied only when a simple Fisherian reduction is possible.

The approach via direct study of hypothetical long-run frequency properties has the very great advantage is that it provides a way of comparing procedures that may have compensating properties of robustness, simplicity and directness and of considering the behaviour of procedures when the underlying assumptions are not satisfied. There remains an important issue; however. Is it clear in each case what is the relevant long run of repetitions?

Example 21. Two measuring instruments. With probability 1/2 a measurement of an unknown mean μ is obtained having low precision, i.e. normally distributed with mean μ and large variance σ_0^2 and with probability 1/2 the observation has high precision, i.e. small variance σ_1^2. The observed random variables are (Y, I), where Y represents the measurement and I is an indicator of which instrument is used, i.e. we know which variance applies to our observation.

The minimal sufficient statistics is (Y, I) with an obvious generalization if repeat observations are made. We have a (2, 1) exponential family and by

the Fisherian reduction we condition on the observed variance, i.e. we use the normal distribution we know actually obtained and take no account of the fact that in repetitions we might have obtained a quite different precision.

Unfortunately the Neyman-Pearson approach does not yield this result automatically; it conflicts superficially at least with the sensitivity requirement for optimality. We need a supplementary principle to define the appropriate long run of repetitions that determines the statistical procedure. Note that this is necessary even if the repetitive process were realized physically, for example if the above measuring procedure took place every day over a long period.

Example 21 can be regarded as a pared-down version of Example 20. The use of A to define a conditional distribution is at first sight conceptually different from that in Example 14 to obtain inference free of a nuisance parameter.

4 Nature of probability

The previous section is in effect about the nature of probability, an issue that for mathematicians was largely if not wholly sidelined by Kolmogorov's axiomization of probability theory. Note, however, both that his axioms were motivated by a frequency-based theory and also that towards the end of his life he developed a quite different notion of probability based on complexity.

In interpreting data there are two quite different roles for probability. One is phenomenological, as an idealized model to represent *variability* in the real world and is to be regarded as directly or indirectly frequency-based.

The second use is epistemological, as an idealized representation of *uncertainty* of our knowledge. This is really a different although of course strongly related role.

The key issue is: do we need two or more different theories of probability or can one be made to do both tasks?

An extreme Bayesian view of de Finetti and Savage, but not Jeffreys, is that a view of probability as self-consistent personalistic assessments of uncertainty embraces the whole subject. Jeffreys distinguished chances, which are essentially physical frequency-based probabilities, from impersonal assessments of how a reasonable person would measure uncertainty. A key issue in such discussions is to justify the application of the standard results of probability theory, which are directly motivated by frequentist arguments.

Neyman's view in principle at least avoids the whole issue by concentrating on what he called inductive rules of behaviour with specified long-run properties but having no special relevance to any particular case. Fisher's view of the objective, and to some extent Neyman's practice, was the same as that of Jeffreys. Fisher achieved this by a special view of probability which addressed what is sometimes called the *problem of the unique case*. That is, probabilities involved in assessing uncertainty attached to a particular set of data should have a frequency interpretation and should not be members of a sub-ensemble with a different frequency. This implies that the distributions used should be adequately conditional in order to achieve relevance for what Fisher sometimes referred to as the unique set of data under analysis.

The essence is then that appropriately conditional long-run frequency statements are, in some limited respects, like probability statements for unique events.

Example 22. Rainy days in Gothenburg. Consider daily rainfall measured at some defined point in Gothenburg in April. To be specific let W be the event that on a given day the rainfall exceeds 5mm. Ignore climate change, etc and suppose we have a large amount of historical data recording the

occurrence and non-occurrence of W. For simplicity ignore possible dependence between nearby days. Then proportions of occurrence of W when we aggregate will tend to stabilize and we idealize this to a limiting value π_W, the probability of a wet day. This is frequency-based, a physical measure of the weather-system and what Jeffreys called a chance. Now consider the question: will it be wet in Gothenburg tomorrow? Suppose we are interested in this one special day, not in a sequence of predictions. If probability is a degree of belief then there are arguments that in the absence of other evidence the value π_W should apply or if there are other considerations, then they have to be combined with π_W.

But supposing that we stick to a frequentist approach. There are then two lines of argument. One, essentially that of Neyman, is that probability is inapplicable (at least until tomorrow midnight by when the probability will be either zero or one). We may follow a rule of inductive behaviour and say "It will rain tomorrow". If we follow such a rule repeatedly we will be right a proportion π_W of times but no further statement about tomorrow is possible.

Another approach, essentially that of Fisher, is to say that the probability for tomorrow is π_W but only if two conditions are satisfied. The first is that tomorrow is a member of an ensemble of repetitions in a proportion π_W of which the event occurs. The second is that one cannot establish tomorrow as a member of a sub-ensemble with a different proportion, that tomorrow is not a member of a *recognizable subset*. That is, the statement must be adequately conditional. There are substantial difficulties in implementing this notion mathematically and these correspond to serious difficulties in statistical analysis. If we condition too far, every event is unique.

Example 23. The normal mean. We now return to the study of limits

for the mean μ of a normal distribution with known variance; this case is taken purely for simplicity and the argument is really very general. In the Bayesian discussion we derive a distribution for μ which is when the prior is "flat" normal with mean $\bar{y}$ and variance σ_0^2/n. In the frequentist approach we start from the statement

$$P(\mu < \bar{Y} + p_c\sigma_0/\sqrt{n}) = 1 - c.$$

Then we take our data with mean $\bar{y}$ and substitute into the previous equation to obtain a limit for μ, namely $\bar{y} + p_c\sigma_0/\sqrt{n}$. Following the discussion of Example 20 we have two interpretations. The first is that probability does not apply, only the properties of the rule of inductive behaviour. The second is that probability does apply provided there is no further conditioning set available that would lead to a (substantially) different probability.

This second argument led Fisher to claim that μ has a fiducial probability distribution and that it can be manipulated like a random variable, essentially that a Bayesian-like conclusion can be achieved without a prior!

But there is an important difference between the hypothetical repetitions in Examples 22 and 23, namely that in Example 21 but not in Example 23 fixed sets in the sample space are involved. There are a number of arguments why in this context μ cannot for all purposes be treated like a random variable.

A summary would be that for evidential interpretation of a single statement a limit for μ can be regarded just like a single probability statement about a unique event, such as that it will be wet in Gothenburg tomorrow, but that Fisher was mistaken in proposing that the resulting distributions could be manipulated by the ordinary rules of probability theory.

5 Significance tests

5.1 General

Suppose now there is specified a particular value ψ_0 of the parameter of interest and we wish to assess the relation of the data to that value. Often the hypothesis that $\psi = \psi_0$ is called the *null hypothesis* and conventionally denoted by H_0. It may, for example, assert that some effect is zero or takes on a value given by a theory or by previous studies.

There are at least four different situations in which this may arise, namely the following:

- there may be some special reason for thinking that the null hypothesis may be exactly or approximately true
- there may be no special reason for thinking that the null hypothesis is true but it is important because it divides the parameter space into two (or more) regions with very different interpretations. We are then interested in whether the data establish reasonably clearly which region is correct, for example the value of $\text{sgn}(\psi)$
- only the model when $\psi = \psi_0$ is a possible model for interpreting the data and it has been embedded in a richer family only to provide a qualitative basis for assessing departure from the model
- only the model when $\psi = \psi_0$ is defined but there is a qualitative idea of the kinds of departure that are of potential interest.

The last two formulations are especially appropriate for examining model adequacy.

Example 24. Poisson distribution. Let $Y_1, \ldots, Y_n$ be independent Poisson distributions with unknown mean μ. The null hypothesis H_0 is that this

model applies for some μ, i.e. a test of model adequacy is involved. Initially no alternative is explicitly formulated. For some restricted circumstances, a negative binomial distribution would be a clear candidate for the alternative, but other considerations may suggest that particular features of the distribution, for example the relation between variance and mean or the relation between the probability of zero and mean are the real focus of interest.

5.2 Simple significance test

First find a distribution for observed random variables that is, under H_0 free of nuisance parameters, i.e. is completely known. Then find or determine a test statistic T, large (or extreme) values of which indicate departure from the null hypothesis. Then if t_{obs} is the observed value of T we define

$$p_{\text{obs}} = P(T \geq t_{\text{obs}}),$$

the probability being evaluated under H_0, to be the (observed) *p-value* of the test.

It is conventional in many fields to report only very approximate values of p_{obs}, for example that the departure from H_0 is significant just past the 1 per cent level, etc.

Its hypothetical frequency interpretation is as follows. If we were to accept the available data as just decisive evidence against H_0, then we would reject the hypothesis when true a proportion p_{obs} of times.

Put more qualitatively we examine consistency with H_0 by finding the consequences of H_0, in this case a random variable with a known distribution, and seeing whether the prediction is fulfilled.

One way of finding appropriate distributions is by appeal to the second property of sufficient statistics, namely that after conditioning on their observed value the remaining data have a fixed distribution.

Example 25. Poisson distribution. The sufficient statistic in Example 19 is ΣY_k so we examine the conditional distribution of the data given $\Sigma Y_k = s$. This distribution is zero if $\Sigma y_k \neq s$ and is otherwise

$$\frac{s!}{n^s \Pi y_k!},$$

i.e. is a multinomial distribution with s trial each giving a response equally likely to fall in one of n cells. There remains, except when $n = 2$, the need to choose a test statistic. This is usually taken to be either the dispersion index $\Sigma(Y_k - \bar{Y})^2/\bar{Y}$ or the number of zeros. The former is in this specification equivalent to ΣY_k^2. Note that if, for example, the dispersion test is used no explicit family of alternative models has been specified, only a indication of the kind of discrepancy that is especially important to detect. A more formal procedure might have considered the negative binomial distribution as representing such departures and then used the apparatus of the Neyman-Pearson theory to develop a test sensitive to such departures.

5.3 Choice of test statistic

When only a null hypothesis is specified the test statistic has to be chosen on an informal basis as being sensitive to the types of departure from H_0 thought of interest. In fact a high proportion of the tests used in applications were developed by that route. When a full family of distributions is specified it is natural to base the test on the optimal pivot for inference about ψ within that family. This has sensitivity properties in making the random variable P corresponding to p_{obs} stochastically small under alternative hypotheses.

We have for simplicity concentrated on continuously distributed test statistics. In the discrete case the argument is unchanged, although only a discrete set of p values are achievable in any particular case. Because pre-assigned

values such as 0.05 play no special role, the only difficulty in interpretation is the theoretical one of comparing alternative procedures with different sets of achievable values.

5.4 Interpretation of significance tests

There is a large and ever-increasing literature on the use and misuse of significance tests. This centres on such points as

- often the null hypothesis is almost certainly false and so why is it worth testing it?
- estimation of ψ is usually more enlightening than testing hypotheses about ψ
- failure to "reject" H_0 does not mean that we consider H_0 to be true
- with large amounts of data small departures from H_0 of no subject-matter importance may be highly significant
- what is being examined is consistency with H_0 not acceptance and rejection as such
- p_{obs} is not the probability that H_0 is true

5.5 One- and two-sided tests

In many situations departures of the test statistic into either tail of its distribution represent interpretable, although different, departures from H_0. The simplest procedure is then often to contemplate two tests, one for each tail, in effect taking the more significant, i.e. the smaller tail as the basis for possible interpretation. Operational interpretation of the result is achieved

by doubling the corresponding p, with a slightly more complicated argument in the discrete case.

5.6 Relation with acceptance and rejection

There is a conceptual difference, but essentially no mathematical difference, between the discussion here and the treatment of testing as a two-decision problem, with control over the formal error probabilities. In this we fix in principle the probability of rejecting H_0 when it is true, usually denoted by α, aiming to maximize the probability of rejecting H_0 when false. This approach demands the explicit formulation of alternative possibilities. Essentially it amounts to setting in advance a threshold for p_{obs}. It is of course appropriate when clear decisions are to be made, as for example in some classification problems. The present discussion seems to match more closely scientific practice in these matters, at least for those situations where analysis and interpretation rather than decision making are the focus.

That is, there is a distinction between the Neyman-Pearson formulation of testing as usually given regarded as clarifying the meaning of statistical significance via hypothetical repetitions and that theory regarded as in effect an instruction on how to implement the ideas.

5.7 Relation with interval estimation

While conceptually it is probably best to regard estimation with uncertainty as a simpler and more important mode of analysis than significance testing there are some practical advantages, especially in dealing with relatively complicated problems, in arguing in the other direction. Essentially confidence intervals, or more generally confidence sets, can be produced by testing consistency with every possible value in Ω_ψ and taking all those values not

"rejected" at level c, say, to produce a $1-c$ level interval or region. This procedure has the property that in repeated applications any true value of ψ will be included in the region except in a proportion c of cases. This can be done at various levels c, using the same form of test throughout.

Example 26. Ratio of normal means. Given two independent sets of random variables from normal distributions of unknown means μ_0, μ_1 and with known variance σ_0^2 we first reduce by sufficiency to the sample means $\bar{y}_0, \bar{y}_1$. Suppose that the parameter of interest is $\psi = \mu_1/\mu_0$; note that this is a ratio of canonical parameters, not a difference (or more generally a linear function of the canonical parameters for which the discussion is simpler). Consider the null hypothesis $\psi = \psi_0$. Then we look for a statistic with a distribution under the null hypothesis that does not depend on the nuisance parameter. Such a statistic is

$$\frac{\bar{Y}_1 - \psi_0 \bar{Y}_0}{\sigma_0 \sqrt{(1/n_1 + \psi_0^2/n_0)}};$$

this has a standard normal distribution under the null hypothesis. This with $\psi_0 replaced by \psi$ could be treated as a pivot provided that we can treat $\bar{Y}_0$ as positive.

Note that provided the two distributions have the same variance a similar result with the Student t distribution replacing the standard normal would apply if the variance is unknown and has to be estimated.

We now form a $1-c$ level confidence region by taking all those values of ψ_0 that would not be "rejected" at level c in this test. i.e. we take the set

$$\{\psi : \frac{(\bar{Y}_1 - \psi \bar{Y}_0)^2}{\sigma_0^2(1/n_1 + \psi^2/n_0)} \leq k_{1;c}\},$$

where $k_{1;c}$ is the upper c point of the chi-squared distribution with one degree of freedom.

Thus we find the limits for ψ as the roots of a quadratic equation. If there are no real roots, *all* values of ψ are consistent with the data at the level in question. If the numerator and especially the denominator are poorly determined a confidence interval consisting of the whole line may be the only rational conclusion to be drawn and is entirely reasonable from a testing point of view, even though regarded from a confidence interval perspective it may, wrongly, seem like a vacuous statement.

5.8 Bayesian testing

A Bayesian discussion of significance testing is available only when a full family of models is available. In both cases we work from the posterior distribution of ψ. When the null hypothesis is quite possibly exactly or nearly correct we specify a prior probability π_0 that H_0 is true and need also to specify the conditional prior distribution of ψ when H_0 is false. Some care is needed here because the issue of testing is not likely to arise when massive easily detected differences are present. Thus when, say, ψ can be estimated with a standard error of $\sigma_0/\sqrt{n}$ the conditional prior should have standard error $a\sigma_0/\sqrt{n}$, for some modest value of a.

When the role of H_0 is to divide the parameter space into qualitative different parts a "flat" prior may be appropriate and the relevant posterior probability is that of, say, $\psi > \psi_0$.

In both cases there is often at least qualitative agreement with the frequentist discussion.

6 Some additional issues

6.1 General points

The above discussion has concentrated on the estimation with assessment of uncertainty of key parameters which, hopefully, address the research questions of concern. The following are brief notes on a number of matters that fall outside the earlier discussion. All could be developed in depth.

6.2 Prediction

In prediction problems the target of study is not a parameter but the value of an unobserved random variable. This includes, however, in so-called hierarchical models estimating the value of a random parameter attached to a particular portion of the data. In Bayesian theory the distinction between prediction and estimation largely disappears in that all unknowns have probability distributions. In frequentist theory the simplest approach is to use Bayes's theorem to find the distribution of the aspect of interest and to replace unknown parameters by good estimates. In special cases more refined treatment is possible.

In the special case when the value Y^*, say, to be predicted is conditionally independent of the data given the parameters the Bayesian solution is particularly simple. A predictive distribution is found by averaging the density $f_{Y^*}(y^*;\theta)$ over the posterior distribution of the parameter.

In special cases a formally exact predictive distribution is obtained by the following device. Suppose that the value to be predicted has the distribution $f_{Y^*(y^*;\theta^*)}$ whereas the data have the density $f_Y(y;\theta)$. Construct a sensible test of the null hypothesis $\theta = \theta^*$ and take all those values of y^* consistent with the null hypothesis at level c as the prediction interval or region.

Example 27. A new observation from a normal distribution. Suppose that the data correspond to independent and identically distributed in a normal distribution of unknown mean μ and known variance σ_0^2 and that it is required to predict a new observation from the same distribution. Suppose then that the new observation y^* has mean μ^*. Then the null hypothesis is tested by the statistic

$$\frac{y^* - \bar{y}}{\sigma_0\sqrt{(1+1/n)}}$$

so that, for example, a level $1-c$ upper prediction limit for the new observation is

$$\bar{y} + p_c\sigma_0\sqrt{(1+1/n)}.$$

The difference from the so-called plug-in predictor in which errors of estimating μ are ignored are here slight but especially if a relatively complicated model is used as the base for prediction the plug-in estimate may seriously underestimate uncertainty.

6.3 Decision analysis

In many contexts data are analyzed with a view to reaching a decision, for example in a laboratory science context about what experiment to do next. It is, of course, always important to keep the objective of an investigation in mind, but in most cases statistical analysis is used to guide discussion and interpretation rather than as an immediate automatic decision-taking device.

There are at least three approaches to a more fully decision-oriented discussion. Fully Bayesian decision theory supposes available a decision space $\mathcal{D}$, a utility function $U(d,\theta)$, a prior for θ, data and a model for the data, y. The objective is to choose a decision rule maximizing for each y the expected utility averaged over the posterior distribution of y. Such a rule is

called a *Bayes rule*. The arguments for including a prior distribution are strong in that a decision rule should take account of all reasonably relevant information and not be confined to the question: what do these data tell us?

Wald's treatment of decision theory supposed that a utility function but not a prior are available; in some respects this is often an odd assumption in that the choice of the utility may be at least as contentious as that of the prior. Wald showed that the only admissible decision rules, in a natural definition of admissibility, are Bayes rules and limits of Bayes rules. This leaves, of course, an enormous range of possibilities open; a minimax regret strategy is one possibility.

A third approach is to fall back on, for example, a more literal interpretation of the Neyman-Pearson theory of testing hypotheses and to control error rates at prechosen levels and to act differently according as y falls in different regions of the sample space. While there is substantial arbitrariness in such an approach it does achieve some consistency as between different similar applications.

6.4 Point estimation

The most obvious omission from the previous discussion is point estimation, that is estimation of a parameter of interest without an explicit statement of uncertainty. This involves the choice of one particular value when a range of possibilities are entirely consistent with the data. This is best regarded as a decision problem. With the exception outlined below imposition of constraints like *unbiasedness*, i.e. that the estimate T of ψ satisfy

$$E(T;\theta) = \psi,$$

for all θ are somewhat arbitrary.

An important generalization is that of an *unbiased estimating equation.* That is, the estimate is defined implicitly as the solution of the equation

$$g(Y, \tilde{\psi}) = 0,$$

where for all θ we have

$$E\{g(Y, \psi); \theta\} = 0$$

.

The exception when these conditions are immediately appealing is when data are available in sections and analysis proceeds by condensing each section to a summary statistic before entering these summary statistics as data into a final analysis. Then, especially if that final analysis is linear in the summary values, unbiasedness does make sense.

7 Asymptotic theory

7.1 General

The analysis above gives several ways of representing inference about ψ. While the discussion covers many simple problems of importance, the application of the main ideas is severely limited. While in principle once a likelihood and a prior are given the only difficulties with the Bayesian approach are ones of numerical application, the conceptual difficulties with the approach are major.

Therefore we seek to apply the previous ideas in some approximate way. For this we suppose that errors of estimation are relatively small allowing locally linear approximations to key functions and appeal to the Central Limit Theorem to sums of independent or almost independent sums of random variables.

The mathematical formulation of this involved the introduction of a quantity n which can be thought of as indicating the amount of information in the data. It is often but not necessarily interpreted as a sample size. We then consider a sequence of problems like the problem under study in which, however, $n \to \infty$. This allows application of the limit laws of probability theory to obtain limiting estimates and their properties. It is crucial, however, that this is a technical mathematical device for generating approximations and the resulting statistical procedures have in principle to be judged by how well the resulting properties match those obtaining in the data under analysis. There is no physical sense in which n tends to infinity.

7.2 Fisherian reduction

To study the log likelihood function, as a function of θ, we denote the true value of θ by θ_0. By Jensen's inequality

$$E\{l(\theta; Y); \theta_0\} < E\{l(\theta_0 : Y); \theta_0\}$$

for $\theta \neq \theta_0$. Further if the model consists of n independent and identically distributed components and indeed much more generally the difference is $O(n)$ as n increases for any fixed $\theta \neq \theta_0$.

Subject to some mild regularity conditions $E\{l(\theta; Y); \theta_0\}$ will be locally quadratic around θ_0 within an $O(1/\sqrt{n})$ neighbourhood of θ_0. Now the observed log likelihood, considered as a random variable, will be within $O_p(\sqrt{n})$ of its expectation and thus, subject to mild regularity conditions will be quadratic with a maximum within $O_p(1/\sqrt{n})$ of θ_0. We write the likelihood as

$$m_n(y) \exp\{-(\theta - \hat{\theta})^T \hat{j} (\theta - \hat{\theta})/2\}\{1 + o_p(1)\},$$

where $\hat{\theta}$ is the *maximum likelihood estimate* of θ and $\hat{j}$ is the *observed information*. Here

$$\hat{j} = [-\nabla\nabla^T l(\theta; y)]_{\theta=\hat{\theta}}$$

is minus the Hessian matrix of the log likelihood at the maximum.

This is the Fisherian reduction to the first order of asymptotic theory and the first-order asymptotically sufficient statistics are $(\hat{\theta}, \hat{j})$.

Note that any statistic that differed from $\hat{\theta}$ by $O_p(1/n)$ could be used instead of $\hat{\theta}$ and that there are equivalently different forms instead of $\hat{j}$. Further discussion will show that we may treat $\hat{j}$ as essentially constant, or preferably, may condition on its observed value.

To complete the frequentist argument we need to form a pivot for inference about the parameter of interest ψ.

First, however, we note a Bayesian version.

7.3 Bayesian asymptotic theory

To the above order we multiply the asymptotic version of the likelihood by the prior density $f_\Theta(\theta)$ and note that provided this is continuous, differentiable and nonzero at $\hat{\theta}$ then with high probability the posterior for Θ will be asympotically normal with mean $\hat{\theta}$ and covariance matrix $\hat{j}^{-1}$. Then ψ is asymptotically normal with mean $\hat{\psi}$ and covariance matrix given by the (ψ, ψ) section of $\hat{j}^{-1}$, conveniently denoted by $\hat{j}^{\psi\psi}$.

7.4 Frequentist discussion

We now consider the distribution of $\hat{\theta}$ when θ_0 is the true value. We assume on the basis of the previous argument that $\hat{\theta}$ is within $O(1/\sqrt{n})$ of θ_0. Then expansion of $l(\theta)$ around θ_0. Then to the first order and in a slightly

condensed notation

$$0 = \nabla l(\theta_0) + \{\nabla\nabla^T l(\theta_0)\}(\hat{\theta} - \theta_0).$$

The vector $\nabla l(\theta_0)$ is called the *score vector*.

Now subject to differentiation under the expectation sign being legitimate, it follows from

$$\int f(y, \theta) d\mu(y) = 1,$$

where $\mu(.)$ is a dominating measure, that

$$E\{\nabla l(\theta_0, Y)\} = 0,$$
$$\text{cov}\{\nabla l(\theta_0, Y)\} = E\{-\nabla l(\theta_0, Y)\nabla^T l(\theta_0, Y)\}.$$

This covariance matrix, assumed positive definite, is called the *Fisher information matrix* or *expected information matrix*, and is denoted by $i(\theta)$. It is to be distinguished from $\hat{j}$, the observed information matrix. The first order equation shows that the equation defining $\hat{\theta}$, the maximum likelihood estimating equation, has the property of being an *unbiased estimating equation.*

The next steps are guided by the behaviour when $Y_1, \ldots, Y_n$ are independent and identically distributed but the argument is valid much more generally. Then we have that $i(\theta) = n\bar{i}(\theta)$, where $\bar{i}(\theta)$ is the Fisher information per unit n. Further the observed Hessian matrix per unit n evaluated at θ_0 converges in probability to $\bar{i}(\theta_0)$ by the Weak Law of Large Numbers. Provided the information matrix is continuous at θ_0 it follows also the $\hat{j}/i(\theta_0)$ converges in probability to one as n increases.

Next the score vector is the sum of n independent and identically distributed random variables of zero mean and finite covariance matrix and so is asymptotically normally distributed with zero mean and covariance matrix

$i(\theta_0)$. That is, $\nabla l(\theta_0)/\sqrt{n}$ converges in distribution to the zero mean normal distribution with covariance matrix $\bar{i}(\theta_0)$.

It follows that $(\hat{\theta} - \theta_0)\ \sqrt{n}$ is an asymptotically normal random vector of zero mean "weighted" by a matrix converging in probability. Therefore $(\hat{\theta} - \theta)$ is asymptotically normal with covariance matrix $i^{-1}(\theta_0)$ and that in test statistics and pivots derived from this $i(\theta_0)$ can be replaced by $\hat{j}$.

Example 28. Full exponential family. If the full likelihood has the form

$$m(y) \exp\{s^T\theta - k(\theta\}$$

the maximum likelihood estimating equation is $s = \nabla k(\hat{\theta})$ and the observed and expected information are the Hessian matrix of $k(.)$ evaluated respectively at $\hat{\theta}$ and at θ_0.

Thus in particular $\hat{\psi} - \psi_0$ is asymptotically normal with covariance matrix $i^{\psi\psi}(\theta_0)$. If ψ is a scalar it follows that

$$\frac{\psi - \hat{\psi}}{\sqrt{\hat{j}^{\psi\psi}}}$$

is an asymptotic pivot for inference about ψ.

Comparison with the Bayesian result outlined above shows that a very wide class of priors are matching to the first order of asymptotic theory. That raises the issue of the circumstances under which a stronger matching result might be possible.

7.5 Alternative forms

The above discussion applies to any estimate differing by $O_p(1/n)$ from the maximum likelihood estimate and to any covariance matrix asymptotically equivalent to $\hat{j}^{-1}$, that is such that the ratio to $\hat{j}^{-1}$ converges in probability to one.

Even within procedures strongly connected to maximum likelihood estimates there are numerous possibilities. In particular to test the null hypothesis $\theta = \theta_0$, where now θ_0 is known, the statistics

$$(\hat{\theta} - \theta_0)^T \hat{j}(\hat{\theta} - \theta_0),$$
$$\nabla^T l(\theta_0) \hat{j}^{-1} \nabla l(\theta_0),$$
$$2\{l(\hat{\theta}) - l(\theta_0)\}$$

differ by $o_p(1)$ and have asymptotic chi-squared distributions with d_θ degrees of freedom when $\theta = \theta_0$. The equivalence follows by repeating the above expansion arguments and the limiting chi-squared distribution follows from the property of a random vector X having a multivariate normal distribution of zero mean and covariance matrix V that $X^T V^{-1} X$ is reducible to a sum of squares and has a chi-squared distribution.

A disadvantage of the first of the three forms is its non-invariance under nonlinear transformation of θ.

There are corresponding results for the parameter of interest ψ based essentially on properties of

$$l(\psi, \hat{\lambda}_\psi),$$

where $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ for fixed ψ. This function, called the *profile* or *concentrated likelihood* of ψ has some, but by no means all, of the properties of a likelihood function.

This formulation leads to confidence regions based on the profile likelihood, namely

$$\{\psi : 2[l(\hat{\psi}, \hat{\lambda}) - l(\psi, \hat{\lambda}_\psi)] \leq k_{d_\psi, c}\},$$

where $k_{d_\psi, c}$ is the upper c point of the chi-squared distribution with d_ψ degrees of freedom. Thus in the common case where ψ is a scalar parameter $k_{d_\psi, c}$ is the square of the appropriate point in the standard normal distribution.

7.6 Developments

The following developments of the above discussion are now required:

- higher order approximations, both to improve distributional accuracy and to choose between alternative procedures
- discussion of cases where the theory fails; this can happen for a variety of reasons
  - large number of nuisance parameters
  - irregular likelihood
  - singular information matrix
  - incomplete or inaccurate specification
- more detailed study of the relation with Bayesian approaches, in particular of higher order matching criteria
- modifications of the likelihood
- extensions to semi-parametric problems

8 A postscript

Essentially four approaches to statistical inference have been outlined with some emphasis on the strong links between them. They all occur in many variants but in essence are as follows:

- Fisherian frequentist, with application to the unique case under analysis assured by conditioning
- rational degree of belief Bayesian, with concentration on the data achieved by flat priors

These have essentially identical objectives. The difficulty with the latter is with the choice of prior.

Then there are

- Neyman-Pearson frequentist, with a decision-like emphasis, assessment by operating characteristics and a denial of relevance to a single case in formulation if not in application
- personalistic Bayesian, usually presented as having a strong link to decision-taking.

Much of the mathematical formulation of Neyman-Pearson is closely related to that of Fisher and the answers are usually very similar if not identical. The Neyman-Pearson mathematical formulation is broad and powerful. Personalistic Bayesian analysis is much more ambitious and aims to synthesize all information. Unless the prior is evidence-based the conclusions are not the basis for public discussion and this questions the relevance of the whole approach for many purposes, at least at a quantitative level.

Formal Bayesian calculations can be used in connection with any of the approaches.

Finally note that the naming of approaches after people is a convenient short-hand. There are, for example, many variants of what here has broadly been termed the Neyman-Pearson viewpoint.

A FEW REFERENCES

Azzalini, A. (1996). *Statistical inference based on the likelihood.* London: Chapman and Hall.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and asymptotics.* London: Chapman and Hall; see especially Chapters 1-4.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical statistics.* London: Chapman and Hall.

Lehmann, E.L. (1983). *Theory of point estimation.* New York: Wiley.

Lehmann, E.L. (1986). *Testing statistical hypotheses.* New York: Wiley.

Pawitan, Y. (2002). *In all likelihood.* Oxford University Press.