Pivotal Inference and the Conditional View of Robustness (Why have we for so long managed with normality assumptions?)

G. A. Barnard

0. SUMMARY.

A major reason for normality assumptions has been that standard, unconditional theories of inference require such assumptions to avoid excessive computational problems. The pivotal approach, like the Bayesian approach, is conditional, and as such does not need such assumptions, given facilities now available on hand-held computers. When there is doubt about distributional assumptions, therefore, a range of such assumptions can be tested for their effect on the inferences of interest. Some samples will be robust with respect to possible changes in distribution, while other samples will not be. When the latter is the case, the statistician should draw attention to the fact, in accordance with the principle that it is at least as important for the statistician to tell his clients what they do not know as it is to tell them what they do know from their data. In sampling from 'normal-looking' distributions, such as the Cauchy, treating 'normal-looking' samples as if they were normal produces errors unlikely to be of practical importance. The non-robust samples are those presenting non-normal features, such as skewness. In the past such samples have been treated by ad hoc methods. Statisticians should make it their business to acquire empirical knowledge of the types of distribution to be met

ISBN 0-12-121160-6

with the areas of application with which they are concerned. Skewness of distribution is particularly to be watched for.

1. Problems of uncertainty of distributional form arise almost exclusively in connection with continuous observables. For any such set $\underline{x} = (x_1, x_2, \ldots, x_n)$ a probabilistic model will amount to asserting that there is a parameter $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ such that, given $\underline{\theta}$, $\underline{x}$ has probability density $\phi(\underline{x}, \underline{\theta})$, where ϕ is approximately known. If, from the marginal density $\phi_1(x_1; \underline{\theta})$ we derive the probability integral transformation

$$p_1 = p_1(x_1, \underline{\theta}) = \int_{-\infty}^{x_1} \phi_1(t; \underline{\theta})dt$$

as is well known, $p_1(x_1, \underline{\theta})$ will be uniformly distributed between 0 and 1. Then we can form the (marginal) density of x_2, given x_1, $\phi_{12}(x_2; \theta, x_1)$, and thence obtain

$$p_2 = p_2(x_2, x_1, \underline{\theta}) = \int_{-\infty}^{x_2} \phi_{12}(t; \underline{\theta}, x_1)dt$$

again uniformly distributed between 0 and 1. Continuing in this way we can find a set $\underline{p} = (p_1, p_2, \ldots, p_n)$ of functions of the $\underline{x}$ and of $\underline{\theta}$ such that, to say that $\underline{x}$, given $\underline{\theta}$, has the density ϕ is equivalent to saying that the specified functions $\underline{p}$ of $\underline{x}$ and $\underline{\theta}$ are uniformly distributed in the unit cube. Following Fisher we call a function of observations and parameters whose distribution is known, a pivotal quantity, and the vector function $\underline{p}(\underline{x}, \underline{\theta})$ will be (a form of) the basic pivotal of the model we are discussing. This mode of expressing a probability model by means of pivotal functions uniformly distributed in the unit cube was first put forward in 1938 by Irving Segal. Insofar as we think of ϕ as only approximately known, so the pivotal $\underline{p}$ will be only approximately uniformly distributed. More generally, it may be convenient to take the basic pivotal $\underline{p}(\underline{x}, \underline{\theta})$ to have an approximately known distribution other than uniform -- for example, in the cases to which, for simplicity, we restrict ourselves, where

$k = 2$ and $\underline{\theta} = (\lambda, \sigma)$, with λ a location parameter and σ a scale parameter, we can set

$$p_i = (x_i - \lambda)/\sigma \qquad (1)$$

and express our information about the distributional form by saying that $\underline{p}$ has the density $f(\underline{p})$, where f might be standard normal, or standard Cauchy, or some other standard distribution. We shall consider the problem of estimating λ, regarding σ as a nuisance parameter. Our arguments can be extended without essential change to general regression models; models of still greater generality will typically require approximations into the details of which we do not enter here.

The pivotal formulation of a model has a double advantage. First, the meaning of the parameters θ is defined by reference to the way they enter the basic pivotal $\underline{p}$, rather than by reference to a particular feature, such as the mean, the mode or the center of symmetry, of the distribution. Thus the difficulty faced by studies such as that of Andrews et al., in which the location parameter had to be taken as the center of symmetry -- thus limiting consideration to cases where the distribution could be taken to be exactly symmetrical -- can be avoided. Second, as was first stressed by Dempster, specifying the distribution by means of pivotals enables us to reduce, if not eliminate, the sharpness of the distinction between observables and parameters; indeed, we may extend the pivotal model by saying that $\theta = (\underline{\theta}_0, \underline{\theta}_1)$ and that the basic pivotal is $(\underline{p}, \underline{\theta}_1)$, with density $f(\underline{p})\pi(\underline{\theta}_1)$, corresponding to the assertion that the part θ_1 of the parameter θ has the prior density $\pi(\theta_1)$. If the part $\underline{\theta}_0$ were empty we would then be formulating a fully Bayesian model, while if $\underline{\theta}_1$ were empty our model would be fully non-Bayesian. Since our mode of reasoning is the same, we need not commit ourselves in advance to being Bayesian or non-Bayesian.

2. The general inference procedure is to make 1-1 transformations on $(\underline{p}, \underline{\theta}_1)$ to bring it, so far as possible, to the form $(\underline{T}, \underline{N}, \underline{A})$. Here $\underline{T}$ is of the form $\underline{T}(\underline{x}, \underline{\theta}_i)$, where θ_i denotes the parameter(s) of interest, and for

each fixed value $\underline{x}_0$ of the observations, $\underline{T}(\underline{x}_0, \underline{\theta}_i)$ defines a 1-1 mapping between the range of T and the range of $\underline{\theta}_i$. $\underline{N}$ is similarly a function of the nuisance parameters, $N(\underline{x}, \underline{\theta}_n)$, while $A(x)$ does not involve the parameters. Since (T,N,A) is a 1-1 function of (p, θ_1), its density is approximately known; then, given the observations $\underline{x} = \underline{x}_0$, the value of $\underline{A}$ will be known, and the relevant density of $(\underline{T}, \underline{N})$ will be the conditional density, given $\underline{A}(x) = \underline{A}(x_0)$. Integrating $\underline{N}$ out from this density will given a density for T which can be used to derive a confidence distribution for θ_i.

This inference procedure will often not be capable of being carried through exactly, and then we must resort to approximations. But in the case of location and scale, and more generally regression problems, we can carry it through exactly, as follows: Taking $\underline{p}$ as in (1), we require

$$\partial T/\partial\sigma = 0, \quad \partial N/\partial\lambda = 0, \quad \text{and} \quad \partial A/\partial\lambda = \partial A/\partial\sigma = 0 . \tag{2}$$

Now

$$\partial A/\partial\lambda = \sum_i (\partial A/\partial p_i)(\partial p_i/\partial\lambda) = (-1/\sigma) \sum_i \partial A/\partial p_i$$

while

$$\partial A/\partial\sigma = \sum_i (\partial A/\partial p_i)\partial p_i/\partial\sigma) = (-1/\sigma) \sum_i p_i \partial A/\partial p_i .$$

so that A must satisfy the PDE's

$$\sum_i \partial A/\partial p_i = 0, \qquad \sum_i p_i \partial A/\partial p_i = 0 \tag{3}$$

the general solution to which is an arbitrary function of the $n - 2$ functionally independent quantitites $c_1, c_2, \ldots, c_{n-2}$, where

$$c_i = (p_i - \bar{p})/s_p = (x_i - \bar{x})/s_x, \quad i = 1,2,\ldots,n \tag{4}$$

and $\bar{x}$ and s_x are used in the customary way to denote the mean and S.D. of a finite set of quantities. Since we clearly want the maximal possible conditioning, we take $\underline{A} = \underline{c}$. We could then use standard methods of the theory of PDE's to arrive at expressions for $\underline{T}$ and $\underline{N}$, but it is obvious that the 1-1 transformation

$$p_i = N(T + c_i) \tag{5}$$

will serve our purpose; for since $\sum c_i = 0$, summing over i gives

$$\bar{p} = NT, \quad p_i - \bar{p} = Nc_i, \quad \text{so} \quad N = s_p = s_x/\sigma \tag{6}$$

and finally $T = (\bar{x} - \lambda)/s_x$. The functions N and T are clearly of the form required.

The transformation (5) can be shown to have Jacobian

$$\sqrt{n(n-1)(n-2)}N^{n-1}/|c_n - c_{n-1}| \tag{7}$$

so that the joint density of $(T,N,\underline{A})$ is

$$\sqrt{n(n-1)(n-2)}N^{n-1}/|c_n - c_{n-1}| \cdot f(N(t\cdot\underline{1} + \underline{c})) \tag{8}$$

Given the observations, the value of $\underline{c} = \underline{c}_0$ is known, so that joint conditional density of (T,N) is

$$KN^{n-1}f(N(T\cdot\underline{1} + \underline{c}_0)$$

and the marginal density of T is

$$\xi(T;\underline{c}) = K\int_0^\infty z^{n-1}f(z(T\cdot\underline{1} + \underline{c}_0))dz \ . \tag{9}$$

where K is determined by the condition that ξ integrates to 1.

Provided $(\bar{x} - \lambda)\sqrt{n}/s_x = t$ has a non-singular distribution, as will very often be the case, we can change the variable in (9) from T to t, and put $u = zt$, $dz = du/t$ in the integral to obtain

$$\xi^x(t;\underline{c}) = (K^*/t^n)\int_0^\infty u^{n-1}f(u(\underline{1} + (1/t)\underline{c}_0))du \tag{10}$$

which is $0(1/t^n)$ as $t \to \infty$. Thus the tail behaviour of the conditional density of t is in this broad class of cases, the same as in the case of normality. This is one reason why interpretations of Student's t as if the original density was normal have often been adequate in practice.

3. It is instructive to study the case when the observations are independent and from a Cauchy distribution. The algebra becomes heavy for n greater than 5 but, for example for $n = 3$ we have

$$f(\underline{p}) = 1/\pi^3(1 + p_1^2)(1 + p_2^2)(1 + p_3^2)$$

and putting $b_i = t + c_i$ in (10) we get

$$\xi^x(t;\underline{c}) = -K^{**} \int_0^\infty z^2 dz / \prod_i (b_i z - i) \prod_i (-b_i z - i)$$

$$= -K^{**} \int_0^\infty z^2 dz / h_3(z) h_3(-z)$$

and this can be evaluated by formulae given in Gradshtyn and Ruzhik (p. 218). We find

$$\xi^x(t;\underline{c}) = K^{**}/(S_1 S_2 - S_3)$$

where S_k stand for the sum of the products, k at a time, of the quantities $b_i = t + c_i$. Plots of $\xi^x(t;\underline{c})$ for various configurations $\underline{c}$ are easily produced using an HP41C with plotter. Comparison with the plot of the density of the normal Student's t with 2 degrees of freedom shows that for the symmetric configuration $\underline{c} = (-\sqrt{3}, 0, +\sqrt{3})$ there will be little error in using the normal tables beyond the values $t = \pm 2.2$ -- and, of course it is this range which will be the one most frequently used. Things are otherwise with the skew configurations, especially with the extremely skew $\underline{c} = (-1.4908, -0.4092, +1.9000)$; but even here, provided one is concerned with two-sided probabilities, and with sufficiently large values of t, the errors involved are not particularly serious.

4. Another density worthy of study is the Barndorff-Nielsen density with

$$\ln f(\underline{p}) = \sum_i (K - \sqrt{(1 + p_i^2)} - \beta p_i)$$

where β is a skewness parameter. When $\beta = 0$ this density approximates the standard normal in the center of

its range, but in the tails it goes down as e^{-p_i} instead of as $e^{-p_i^2}$. It appears that use of normal tables here does little harm provided whatever skewness there is in the sample fairly reflects the skewness in the true distribution; the worst case arises when there is skewness in the sample in a sense opposite to that in the true distribution.

Whenever the density is such that

$$\ln f(\underline{p}) = K - H(\underline{p})$$

where $H(\underline{p})$ is homogeneous of degree α, so that $H(z\underline{p}) = z^{\alpha}H(\underline{p})$ it is possible to evaluate the integral (10) in closed form. Because then

$$\xi^*(t;\underline{c}) = K^* \int_0^{\infty} z^{n-1}\exp -H(z(T\cdot\underline{1} + \underline{c}_0)dz$$

$$= K^* \int_0^{\infty} z^{n-1}\exp -z^{\alpha}L\cdot dz$$

where

$$L = H(T\cdot\underline{1} + \underline{c}_0) \ .$$

The integral becomes a gamma integral by the substitution $v = z^{\alpha}L$ and we find

$$\xi^*(t;\underline{c}) = K^{**}/(H(T\cdot\underline{1} + \underline{c}_0))^{n/\alpha} \ .$$

The independent or dependent normal distribution is a special case of this. In particular, for the independent normal, $H(\underline{p}) = \frac{1}{2}\,\underline{p}'\underline{p}$, $\alpha = 2$, and

$$H(T\cdot\underline{1} + \underline{c}_0) = (T\cdot\underline{1} + \underline{c}_0)'(T\cdot\underline{1} + \underline{c}_0)$$

$$= (nT^2 + \underline{c}_0'\underline{c}\ )$$

$$= (t^2 + (n - 1))$$

giving the usual result for Student's t-density. In the case of the double exponential, $H(p) = \sum_i |p_i|$ and $\alpha = 1$. It will be noted that for the normal distribution, the

resultant t density does not involve $\underline{c}$, so in this case -- and essentially only in this case together with that of the uniform density -- the conditional approach and the marginal approach give the same result.

5. There exists a wide area of extremely useful research to be done in finding which non-normal densities apply to which types of empirical data. Barndorff-Nielsen has shown that his family of hyperbolic distributions fit very well to distributions arising in connection with turbulent flows of various sorts. Karl Pearson and Weldon and their collaborators --including Student himself -- did great work early in the century in connection with biometrical distributions. But for the past fifty years work of this kind has been neglected -- presumably because little was known of how to use such information, and the computing facilities needed, now available on quite small computers, was simply not available.

For those whose limited access to sets of empirical data prevent them from engaging in the useful research indicated, there is the purely theoretical problem still, so far as I know, without a solution -- how far we can determine the form of a density, given an arbitrarily large number of samples of a fixed finite size, with varying and unknown location and scale parameters.

Faculty of Mathematics
Department of Statistics
University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1