Pivotal inference and the Bayesian controversy

G.A. BARNARD
University of Waterloo

SUMMARY

The theory of pivotal inference applies when parameters are defined by reference to their effect on observations rather than their effect on distributions. It is shown that pivotal inference embraces both Bayesian and frequentist reasoning.

Keywords: INFERENCE; PIVOTAL; ROBUST; BAYES.

1. PIVOTAL INFERENCE

1. A *pivotal model* of an inference situation arises typically when we have a relatively precise idea of the way in which the parameters are related to the observations, and a less precise idea of just how the observations are distributed. Thus for example, we may have observations x_i $(i=1,2,\ldots,n)$ for which μ and σ, respectively, are location and scale parameters, but we may not be sure as to the precise form of their distribution. Then we know that the

$$p_i = (x_i - \mu)/\sigma \tag{1}$$

have a distribution which does not involve the parameters, but we may not know exactly what this distribution is. If we suppose that the x_i are *nearly* distributed independently, each in a double exponential distribution, we might suppose that the joint density ot the p_i could be expressed, sufficiently accurately, in the form

$$\phi_\lambda(p) = (1-\epsilon)(\tfrac{1}{2})^n \exp -\sum_i |p_i| + \epsilon(\sqrt{2\pi})^{-n} \exp -\tfrac{1}{2}\sum_i (p_i - a_i)^2 \tag{2}$$

for some ϵ between 0 and 10^{-6}, and for some vector a with i^{th} component a_i. This would correspond with an idea that, less often than once in a million times, the observations were from a 'rogue' normal distribution; but it will become apparent that the role of this small mixture of normality is to be viewed rather differently, as indicating perhaps only part of the small uncertainty in the form of the distribution.*

We use λ to denote the pair (ϵ,a) which serves to specify exactly which member of the family (2) applies in a specific case. Although λ would ordinarily be called a parameter, we call it, instead, a *label*, because its logical role in the inference is different from that of the pair (μ,σ). And the term 'nuisance parameter', which might be used instead of label, we wish to reserve for a somewhat different concept.

The term 'pivotal' was introduced by Fisher, to denote a quantity such as Student's t:

$$t = (\bar{x} - \mu)\sqrt{n}/s_x \tag{3}$$

which is a function of the observations and of the parameters whose distribution does not involve the parameters. We use the term in the same sense.

2. The elements of a pivotal model of an inference situation are five in number: $\{S,\Omega,p,P,D,\}$. S is the usual sample space, of possible observations and Ω is the usual parameter space, of possible parameter values. p is a mapping from $S \times \Omega$ to P, the pivotal space. p is called the *basic pivotal*. We suppose that measures are given on S and on P, and that for each θ in Ω the inverse mapping $p^{-1}(.,\theta):P \rightarrow S$ is 1-1 and measurable. D is a set of probability distributions on P, specified by density functions ϕ_λ. It is convenient, though not logically necessary, to assume the distributions in D to be absolutely continuous with respect to each other.

3. For any specified label λ the pivotal model defines a likelihood model L_λ, consisting of the usual triplet $\{S,\Omega,\psi_\lambda\}$ of sample space, parameter space, and probability function ψ_λ

$$\psi_\lambda(x,\theta) = \phi_\lambda(p(x,\theta)).\partial p\ (x,\theta)/\partial x. \tag{4}$$

In accordance with our usage, a function $F(x,\theta)$ will be pivotal in L_λ iff its distribution, derived from ψ_λ, does not involve θ.

* if there were such a thing as a 'fuzzy distribution', this would convey the idea better.

Now if $F(x,\theta) = G(p(x,\theta))$, for some function G, then it is evident that F will be pivotal in L_λ *for every* λ. F will then be called a *robust pivotal*-- defined as a function of observations and parameters which is pivotal in L_λ for every λ.

4. We now introduce the concept of a *separating family* of distributions. The family D is said to be *separating* iff the only robust pivotals are functions of the basic pivotal-- i.e. iff $F(x,\theta)$ pivotal in L_λ for every λ implies that there exists a G such that $F(x,\theta) = G(p(\times,\theta))$.

In the pivotal model for which $S = R^n$, $\Omega = R^1 \times R^+$, $P = R^n$, and the i^{th} component of p is p_i in (1) above, we use Lebesgue measure, and D is the family given by ϕ_λ in (2) above, the family D is separating. The steps in proving this are:

(i) If D is complete (in the sense of Lehmann) it is separating. (I owe this remark to Barndorff-Nielsen.)

(ii) The family of spherical normal densities with arbitrary centre is complete.

(iii) If D' is complete, and ϕ is arbitrary, then for any $\delta > 0$,

$$D = \{\phi_\lambda : (\exists \varepsilon)\phi_\lambda = (1-\varepsilon)\phi + \varepsilon\phi_\lambda',\ 0 \leq \varepsilon > \delta,\ \phi_\lambda' \text{ in } D'\}$$

is complete.

We may also note the obvious

(iv) If D is separating in a given pivotal model, and if $D' \supset D$, then in the pivotal model in which D' replaces D, D' is separating. All this implies that a very small element of uncertainty in the form of the distribution of the basic pivotal is enough to ensure that the family of distributions is separating.

From now on we assume that the family D is separating.

5. The basic inferential steps which justify the term 'pivotal inference' are of two kinds: (i) Making 1-1 transformations which amount to no more than renaming the entities involved; (ii) conditioning steps. These latter make use of what I have called 'Modus ponens probabilitatis' (MPP), by analogy with Modus ponens of classical logic:

Modus ponens	**Modus ponens probabilitatis**
We know 'A implies B'.	We know* $Pr(B \text{ given A}) = q$.
We know 'A' is true.	We know that A is true.
Therefore 'B' is true.	Therefore $Pr(B) = q$.

* or 'agree' - see Sec. 10 bellow.

The general procedure of pivotal inference thus consists in transforming the basic pivotal p, 1-1, to another pivotal q which splits into two parts:

$$q = \begin{bmatrix} q_1 \\ \\ q_2 \end{bmatrix}$$

The second part, q_2, is *ancillary*, that is, it is constant on the parameter space, so that its value is known when the observations are known. Then the original pivotal model can be replaced by one for which the basic pivotal is q_1, endowed with the conditional distribution which it has, given the observed value of q_2.

The role of the concept of 'separating family' can now be seen. It is to guarantee that there is an essentially *unique maximally informative ancillary* (MIA). For any two functions f,g, we say that f is more informative than (mit) g iff there exists h such that $g = h(f)$ i.e. $g = h \circ f$ in Bourbaki notation; that is, if the value of g can be calculated when the value of f is known, but not necessarily conversely. If f mit g and g mit f then f and g convey the same information and are regarded as equivalent. The relation 'mit' is a partial ordering on the set of functions of the basic pivotal; and if $f(p)$ and $g(p)$ are both ancillary, the vector-valued function $\binom{f(p)}{g(p)}$ is also ancillary, and it 'mit' each for f and g. It follows that the maximally informative ancillary is unique up to equivalence.

6. The 'conclusion' of a pivotal inference is then a statement of the conditional distribution of the pivotal q_1, together with a statement of the values of the functions of the observations which enter into q_1. From this statement, if desired, a confidence level of, say, 95% can be chosen, and corresponding confidence sets for the parameters can be found; but such an 'arbitrary' choice of confidence level (and 'arbitrary' choice of e.g. 'shortest', or 'one-sided', for the form of the confidence set) means that information is lost at this stage. Thus, it is suggested that the conclusion should be expressed in the form of the conditional distribution of q_1, with the necessary functions of the observations, allowing each reader of the conclusion to form confidence sets in accordance with his specific interests.

7. To illustrate, we consider the case of the example of section 1, where the parameters are location and scale. Here we transform to

$$q_1 = \begin{pmatrix} \bar{p} \\ \\ s_p \end{pmatrix}$$

(where $\bar{}$ denotes, as usual, mean and $s^2_{(\,)}$ denotes variance)

q_2 with ith component

$$q_{2i} = (p_i - \bar{p})/s_p. \tag{5}$$

Then since, from (1)

$$\bar{p} = (\bar{x} - \mu)/\sigma, \; s_p = s_x/\sigma \tag{6}$$

and

$$(p_i - \bar{p})/s_p = (x_i - \bar{x})/s_x \tag{7}$$

it easily follows that q_2 is the maximal ancillary. The Jacobian of the transformation from p to q is

$$J = (n(n-1)/|q_{2,n} - q_{2,n-1}|)s_p^{n-2} \tag{8}$$

if the last two components of q_2 are regarded as functions of the first n-2 components. Thus, the joint density of the transformed basic pivotals is

$$J\phi_\lambda\left((\bar{p} + q_{21}s_p, \ldots, \bar{p} + q_{2n}s_p)\right) \tag{9}$$

and if the observed values of the ancillaries are $c_1, c_2, \ldots, c_n$,

$$c_i = (x_i - \bar{x})/s_x \tag{10}$$

the conditional density of $q_1 = \begin{pmatrix} \bar{p} \\ s_p \end{pmatrix}$ is

$$K(.)s_p^{n-2}\phi_\lambda\left((\bar{p} + s_p c_1, \ldots, \bar{p} + s_p c_n)\right) \tag{11}$$

where $K(.)$ here, as later, denotes a normalising constant whose value is determined by the condition that the integral of the whole expression, over the whole range of the variables $\bar{p}, s_p$, should come to 1.

If now C is a set in the space of $(\bar{p}, s_p)$, such that the integral of (11) over the set C is 0.95, we have

$$Pr\left((\bar{p}, s_p) \in C \mid q_2 = c\right) = 0.95$$

and so, by the usual argument, if we assert that in our case $(\bar{p}, s_p) \in C$, i.e. that for our observed $\bar{x}, x_x$,

$$((\bar{x} - \mu)/\sigma, s_x/\sigma) \in C$$

we have a joint 95% confidence set for (μ,σ), having the usual coverage frequency property.

7. To express the conclusion of our inference in a convenient and easily understood form, without destroying its full informativeness and uniqueness, I propose we should revert to the practice still common in the physical sciences of expressing our information about a parameter in terms of a 'preferred value' and a 'standard error', for example:

$$\mu = x_0 \pm b \tag{12}$$

which, strictly interpreted, means that our knowledge of μ is equivalent to knowing that $(x - \mu)/b$ is distributed in a standard normal distribution, and that the observed value of x is x_0. A natural extension of this notation to the example we have been considering would be:

$$\begin{aligned} \mu &= \bar{x}_0 -. \sigma\bar{p} \\ & \qquad\qquad \psi(\bar{p},s_p) \\ ln\sigma &= lns_{x0} -. ln\ s_p \end{aligned} \tag{13}$$

to be interpreted as meaning that $\bar{p} = (\bar{x} - \mu)/\sigma$ and $s_p = s_x/\sigma$ have the joint distribution $\psi(\bar{p},s_p)$, and that the observed value of $\bar{x}$ is $\bar{x}_0$ and the observed value of s_x is s_{x0}. The sign '$-.$' is intended to *suggest* subtraction (thought what precedes '$-.$' is a number, and what comes after is a random variable). However, such a mode of expression suffers from the disadvantage that there can be a wide variety of densities ψ, whose properties may be by no means easy to discern from their analytical expression. It seems reasonable, in cases such as the example we are considering, to relocate the distribution so that its mode is at the origin, and then to make a linear transformation of the pivotals if necessary, to secure that in the neighbourhood of the mode the density can be treated as approximately that of two independent standard normal deviates. This means that the second derivatives of the logarithm of the density ψ, taken at the mode, should be unity for the repeated derivatives and zero for the cross derivative. If this is done, the 'preferred values' would be the maximum likelihood estimates of the parameters, and the matrix multiplying the pivotal vector would be the inverse of the information matrix. This would lead to a 'justification' of the method of maximum likelihood in its wider context (i.e. as it is used in situations other than those to which pivotal inference applies), as an approximation, in a certain sense, to an exact

pivotal inference. It is important, however, to realise that maximum likelihood estimates here have a direct justification, as those points in the parameter space which will be contained in *any* shortest confidence sets, quite separate from the justification for the use of maximum likelihood in more general cases.

When, as with the example we have been considering, one of the parameters appears as a factor in the error of estimate of the other, special issues arise into which we do not enter in this summary account. This is where we need the term 'nuisance parameter', reserved in section 1 above.

8. In the example we have been considering, we can find a pair of functions of the basic pivotal one of which contains the location parameter and not the scale parameter, while the other pivotal contains the scale parameter and not the location parameter: If

$$t = \bar{p}\sqrt{n}/s_p = (\bar{x} - \mu)\sqrt{n}/s_x, \qquad s_p = s_x/\sigma \tag{14}$$

the Jacobian of the transformation is

$$\partial(\bar{p},s_p)/\partial(t,s_p) = s_p/\sqrt{n}$$

and the joint density (conditional on c) of t,s_p is

$$\psi(t,s_p) = K(.)s_p{}^{n-1}\phi_\lambda(s_p((t/\sqrt{n}) + c_1),\ldots,s_p((t/\sqrt{n}) + c_n)). \tag{15}$$

We can now take the marginal density for t by integrating out s_p (after substituting $u = s_p t$, $s_p = u.t$, $ds_p = du/t$)

$$\zeta(t|c) = \frac{K(.)}{t^n}\int_0^\infty u^{n-1}\phi_\lambda\{u((1/\sqrt{n}) + (c_1/t)),\ldots,u((1/\sqrt{n}) + (c_n/t))\}du \tag{16}$$

(showing that under wide regularity conditions on ϕ_λ the tails of the t density behave like $K/t/^n$).

The step of integrating out s_p is an *information-losing* step. Even if we are really interested only in μ, the use only of the marginal distribution of t means that any external information we may have concerning the value of σ and which could give information about the error in μ, becomes unusable. In fact, if we knew, for example, that σ was distributed with density $\Pi(\sigma)$, we should take the integral of (15) after weighting by $\Pi(\sigma)$. While if we knew that, say, $\sigma = 2$, to a sufficient approximation, we should take the distribution of t *conditional* on $s_p = s_x/2$.

9. The possibility that we have, or may acquire, information which enables us to assign a density to σ will be taken into account in the general theory by noting that if σ is assumed to have a known (prior) density $\Pi(\sigma)$ then σ satisfies the definition of a pivotal and should be included in the basic pivotal, which thus becomes (p,σ), with density

$$\phi_\lambda(p\,|\,\sigma)\Pi(\sigma). \tag{17}$$

The maximal ancillary is now larger than before. We can transform from (p,σ) to (d,σ,q_2,s_x), with q_2 defined as in (5) above, and

$$d = \bar{p}\sigma = \bar{x} - \mu,\ \sigma = \sigma, \text{ and } s_x = s_p\sigma. \tag{18}$$

The new maximal ancillary is (q_2,s_x). Making the 1-1 transformation, and conditioning on the observed values c for q_2 and s_x for s_x we obtain, for the joint conditional density of d and σ:

$$\psi(d,\sigma\,|\,c,s_x) = K(.)(1/\sigma)^{n-1}\phi_\lambda\ ((d+s_xc_1)/\sigma),\ldots,((d+s_xc_n)/\sigma))\ldots \tag{19}$$

With this additional information about σ we can improve our confidence statements about μ by basing them upon the marginal distribution of d derived from (19) by integrating out σ. Alternatively, if it is σ we are interested in, we can integrate out d, and obtain a 'quasi-posterior' density for σ which can serve to derive confidence limits for σ if required. This 'quasi-posterior' will be identical with the 'posterior' for σ which would be obtained from the 'improper' uniform prior for μ, independent of σ.

Finally, of course, we may assume a known prior density for both μ and σ, so that the basic pivotal becomes (p,μ,σ). The maximal ancillary will then be the whole set of sample values, or equivalently $\bar{x}$, s_x and q_2, and our conditional distribution will be for (μ,σ), given the sample. It will clearly be identical with the posterior distribution derived in accordance with the usual Bayesian rules.

10. The fact that pivotal inference, as formulated here, *includes*, without *requiring* the use of the standard form of Bayes' theorem is important from the point of view of the Bayesian controversy. The present writer goes a very long way with de Finetti's arguments concerning the way we should react to uncertainty as individuals; as a follower of Wittgenstein I lay less stress on the mental material dichotomy than de Finetti seems to do, but my disagreements here come at a philosophical level remote from applications in statistical or decision making practice. What does differentiate me from many of those

who call themselves Bayesian is a respect in which I agree with de Finetti when he stresses the distinction between what he calls the Bayesian standpoint, on the one hand, and Bayesian techniques, on the other. By the latter, which he condemns along with other 'ad hockery', he means the formal applications of Bayes theorem to a prior distribution chosen, not because it corresponds to any individual's actual prior beliefs, but because it has some convenient mathematical property, such as 'smoothness' or 'conjugacy'. An essential part of the true Bayesian standpoint is the careful investigation of the prior beliefs of the individual concerned, in the expectation that these prior beliefs will turn out to be peculiar to the individual in question.

If it is accepted that the personalistic Bayesian standpoint is concerned with the coherent development of attitudes in a *single* individual, the question arises as to what function the *statistician* has in relation to his *client* or *clients* where at least *two* individuals are involved. It seems to me that it could be argued, by one who accepts the personalistic view, that the statistician has two functions: (i) he has experience of types of random behaviour-- such as, for example, the likely shapes of measurement error distributions to be found in given circumstances-- which enable him to advise his clients about distributional shapes, and thereby effectively communicate additional *empirical* data, (ii) he then should base his reasoning on those probabilities which can be taken as *agreed* by all parties likely to be involved. Such agreement about probabilities may, in a given case, extend to the 'full Bayesian' case, in which (to refer to our example) the basic pivotal is taken as (p,μ,σ); but in another case there may well be room for individuals to differ concerning their assessment of the prior distribution for μ, in which case the agreed probabilities would extend only as far as the joint distribution of (p,σ). And in yet another case agreement may extend only to the approximate specification of the density of p. In each case the pivotal inference procedure of conditioning on known quantities having known (agreed) distributions can be carried through and the result stated in the form suggested in section 7 above, leaving it to individuals, if necessary, to assess, to within sufficient accuracy (which often will not need to be great) their personal priors with which the statement of the statistical inference should be combined.

To sum up this section, we can say that pivotal inference by-passes the Bayesian controversy by making the inference depend on what is *agreed* between individuals as its basis; how far this goes in the direction of a fully Bayesian inference will depend, in a given case, on how much agreement there is among those concerned. There remains, of course, disagreement with those 'ultra-Bayesians' for whom statistics is a branch of psychiatry, concerned only with purely personal coherence, and who consequently insist that there is no need to ask whether or not there is agreement about assigned probabilities;

and there is also disagreement with the 'ultra empiricists', for whom there is no such thing as statistical 'inference', only 'inductive behaviour'. The rule of Modus Ponens Probabilitatis has as much right as its older, narrower correlative to be regarded as a 'principle' of 'inference'.

APPENDIX

1. We give here the details of the proof outlined in Section 4.

(i) **Theorem:** If $D = \{\phi_\lambda\}$ is complete, D is separating.

Proof: Suppose $F(x,\theta)$ is a robust pivotal, then the mean value of F

$$= \int_P F(p^{-1}(u,\theta),\theta)\,\phi_\lambda(u)\,du$$

does not depend on θ. Hence for any fixed $\theta_0 \epsilon \Omega$,

$$\int_D \{F(p^{-1}(u,\theta) - F(p^{-1}(u,\theta_0),\theta_0)\}\,\phi_\lambda(u)\,du$$

vanishes for all λ. Hence, by completeness,

$F(p^{-1}(u,\theta),\theta) - F(p^{-1}(u,\theta_0),\theta_0)$

vanishes for all u. Thus identically

$$F(x,\theta) = F(p^{-1}(u,\theta),\theta) = F(p^{-1}(u,\theta_0),\theta_0) = G(u).$$

(ii) If $\int g(u)(\sqrt{2\pi})^{-n} \exp -\tfrac{1}{2}(u-\alpha)'(u-\alpha).du = 0$, all u,

and $g^x(t)$ is the Fourier transform of $g(u)$, then

$g^x(t).e^{it'\alpha - t't/2} = 0$, all t

so $g(t) = 0$ all t

so $g(u) = 0$ all u.

(iii) If $D' = \{\phi_\alpha\}$ and is complete, if $\lambda = [{}^{\epsilon}_{\alpha}]$ and

$\phi_\lambda = (1-\epsilon)\phi_0 + \epsilon\phi_\alpha$ for $0 \leq \epsilon \leq \delta$ and if $\int g(u)\phi_\lambda(u)\,du = 0$

for all λ then for all ϵ in $(0,\delta)$ and all α,

$$(1-\epsilon)\int_P g(u)\phi_0(u)\,du + \epsilon \int_P g(u)\phi_\alpha(u)\,du = 0$$

so that

$$\int_P g(u)\phi_\alpha(u)\,du = 0, \text{ for all } \alpha$$

which, by completeness of $\{\phi_\alpha\}$ implies $g(u) = 0$.

2. ON THE BAYESIAN - ANTIBAYESIAN CONTROVERSY

1. It would be foolish to imagine that in the course of what must necessarily be a short paper one could hope to review any more than a few aspects of the issues in a debate which has already gone on for upwards of a century and a half. But of late the controversy seems to have become sharper, with extremists on one side seeming to say that the Bayesian model is the only one which can be used to represent experimental logic, and on the other seeming to say that it should never be used. One is concerned lest such sharp divisions should cause us to lose the respect of the community of experimental scientists which we have only relatively recently gained. It seemed worthwhile to take the opportunity presented by this conference to test whether we are ready to move towards the middle ground.

2. The central aim of the theory of statistical inference I take to be the modelling of the logical structure of experiments with a view to assisting in their interpretation and combination for the advancement of knowledge. In pursuing this aim it has set up many types of logical model, some of which are:

(i) The Significance Test Model (ST model).

Here the elements of the model are the sample space $S = [x]$ of possible experimental results, a 'null hypothesis' H_0 specifying $f_0(x)$, the probability of x if H_0 is true, and a discrepancy function $D(x)$ such that large values of D are thought of as explicable if some alternative to H_0 is true. We calculate the P value, $P = \text{Prob}[D(x) \geq D(x_0):H_0]$ and if this is small we are disposed to give serious consideration to the alternatives to H_0. (Here x_0 is the observed result).

The canonical case for this mode of reasoning is provided by Daniel Bernoulli. Asked to consider why the points on the unit sphere representing the poles of the planetary orbits should lie so close together, and why they do not exactly coincide, he began by testing the significance of the departure from a random (uniform) distribution on the sphere. Here $D(x)$ was a measure of clustering, such as the reciprocal of the radius of the smallest circle containing all the points.

It is of the essence of the situation that Bernoulli did this *before* seriously considering alternatives. And to apply Bayes' Theorem he would have had to have given serious consideration to these alternatives.

(ii) The Bayes Model (B model).

Here the elements are S as before, $\Omega = [\theta]$, the parameter space, $f(x,\theta)$ specifying the probability of x if θ is the true value of the parameter, and $Pr(\theta)$, the prior distribution of θ. We calculate the conditional distribution of θ, given x_0:

$$Pr(\theta:x_0) = f(x_0,\theta)Pr(\theta)/Pr(x_0)$$

and the posterior distribution represents our conclusion.

The inferential step here consists in conditioning on knowing the observed value x_0, the probability of which is completely specified by the model. It should be noted* that if, in addition to the given four elements we also have a discrepancy measure D, we can calculate a P value as in the ST model and if this is small we may be led to modify our B model. The calculation of the posterior belongs to what George Box has called 'model analysis' and the calculation of a P value belongs to what he has called 'model criticism'.

(iii) The Likelihood Model (L model).

Here the elements are as in the B model, except that $Pr(\theta)$ is missing. If special interest attaches to a particular θ_0, and we have a discrepancy measure D_0 associated with this value, then we can again calculate a P value. If, on the other hand, all values of θ are to be considered on an equal footing, and there are no other logical relevant features in the situation, the inference is given in terms of likelihood, the likelihood function being $f(x_0,\theta)$. For any pair of values θ, θ', the ratio $f(x_0,\theta)/f(x_0,\theta')$ measures the relative plausibility of θ as against θ', on the given data.

A principal disadvantage of the L model is that we cannot, in general, derive the plausibility of a disjunction of hypotheses represented by a range of values of θ. This is because, in general, a disjunction of hypotheses does not specify the probability function of x, nor does there in general exist a function $y = y(x)$ whose distribution is specified by the disjunction. Sometimes such reductions are possible. Thus in the case of a sample from a normal distribution with unknown mean μ and unknown standard deviation σ, the disjunction of hypotheses given by $\mu = \delta\sigma$, $0 < \sigma \leq \infty$, for each δ specifies the distribution of $t = \bar{x}\sqrt{n}/s_x$.

Finally, an L model *may* serve to generate a confidence distribution.

* as pointed out by Box

(iv) The Pivotal Model (P model)

Here the elements are S and Ω as before, together with a space $P = [u]$ of values of a *basic pivotal function* $p\ (x,\theta) = u$. The fifth element is a family $F = [\alpha]$ of densities on P representing the range of uncertainty we often are in concerning the form of the distribution of the observations x. The parameter α, indexing the members of the family F, is a *model adjustment parameter.* (MA parameter). It is required that for each θ in Ω the mapping $p\ (.,\theta): S \to P$ is invertible, with inverse $p_\theta^{-1}(u) = x$. For each α the distribution of u specified by α yields a probability function of x, depending on θ, given by

$$f_\alpha(x,\theta) = \alpha(p\ (x,\theta)).\ J(x,\theta)$$

where J is the Jacobian of the invertible transformation $u = p\ (x,\theta)$, for each θ.

A pivotal model is appropriate typically when we take observations to be normally distributed, when we usually mean that we *think* they are *approximately* normally distributed. Because of the uncertainty in the form of the distribution we can give a precise definition of the parameter θ only by reference to the way in which it affects the observations x rather than by the way it enters the distribution of x.

From a given pivotal model P, for each α we can derive an L model $L(P,\alpha)$, with elements S, Ω, $f\alpha$. In this L model we can define a *pivotal*, following Fisher, as a function of x and θ whose distribution does not depend on θ. In the P model we require, for a pivotal, that it should be pivotal in the L model sense *for every* α. To emphasize this we sometimes call such a function a robust pivotal. It can be shown that, under very weak conditions on the family F, for a function $q\ (x,\theta)$ to be a (robust) pivotal it is necessary and sufficient that it should be a function of the basic pivotal $p\ (x,\theta)$:

$$q\ (x,\theta) = r\ (p(x,\theta)).$$

If q is constant on Ω it is called an ancillary, and if it is constant on S it is called a Bayesian pivotal. If a function $\phi(\theta)$ exists such that $q\ (x,\theta) = q\ (x,\phi)$, and such that for each x the mapping $q\ (x,.)$ from $\phi(S)$ to $r(P)$ is invertible, then q is said to be a confidence pivotal for ϕ. It can be used to generate a confidence distribution for ϕ.

The inference procedure consists in transforming $p\ (x,\theta)$ 1-1 to $q\ (x,\theta)$ where

$$q(x,\theta) = \begin{pmatrix} q_1(\phi_1(\theta)) \\ q_2(x,\theta) \\ q_3(x) \end{pmatrix} \qquad \begin{matrix} \text{is Bayesian} \\ \\ \text{is ancillary.} \end{matrix}$$

Then when the observations are known, $q_3(x)$ is known, and as in Bayes' argument we can condition on this known value to obtain the joint distribution of q_1 and q_2. The former will give a marginal distribution which yields the posterior distribution of ϕ_1, while the latter will often be a confidence pivotal for a function ϕ_2 of θ, and the mapping from θ to (ϕ_1,ϕ_2) will be invertible.

A noteworthy feature of the pivotal model for inference is, that is always unique, in that the maximal ancillary on which we should condition is unique.

An example of pivotal inference is sketched in the Appendix.

The scheme of pivotal inference can be extended to cover cases where the observations consists of classifications of items into categories; but this involves considerable complication and loss of some of the desirable properties of the model, which is best suited to quantitative observations, discrete or continuous. Over this field it can be seen to cover both the model B and the L model. If the basic pivotal contains a Bayesian component for all the parameters involved, then the maximal ancillary will consist of all the observations x, and the inference will be the usual Bayesian posterior; if the basic pivotal contains no Bayesian component, and if F contains only one element, then we obtain a likelihood model. In general we obtain a mixed model.

It is far from my intention to suggest that the four models listed above exhaust the possibilities. For example the 'predictive sample re-use' models of Seymour Geisser have not been mentioned. Our selection has been made with a view to raising some questions which I hope those present will see fit to answer.

3. The questions are these:

(i) Was Daniel Bernoulli right or wrong to argue as he did? Am I wrong in thinking he could not have used Bayes' Theorem? If so, how would he have used the theorem?

(ii) If it be admitted that the personal theory of probability would always provide complete Bayesian pivotals in the P model, are there not instances where a bevy of Bayesians (in Dawid's useful phrase) might agree on parts of the basic pivotal only, so that the inference could not, with agreement, be carried through to a complete Bayesian conclusion? If so, could not the partial analysis be useful in that it might show that remaining differences of opinion are likely to be unimportant?

(iii) Carrying this situation envisaged in (ii) further, could it not happen that the bevy could agree only on the consituents of an L model? If so, how should they proceed?

It should be clear how I would hope these questions will be answered. If they are so, I think it would be worth emphasis that out differences amount to much less than might be thought.

APPENDIX

Pivotal Inference

Example: $S = R^n$, $\Omega = R^1 \times R^+$, $p = R^n$, $p(x,\theta)$ has i^{th} component $p_i = (x_i-\theta_1)/\theta_2$, $F = \{\phi_a:\phi_a(u) = \prod_i K \exp- |u_i|^a + \epsilon, 1 \leq a \leq \infty\}$. Here K, as later, is a normalising constant (not all K's are equal!), ϵ is a small 'error' term expressing uncertainty in ϕ_a sufficient to ensure the 'separating' property -- i.e. that any robust pivotal must be a function of $p(x,\theta)$.

Here the maximal ancillary may be taken as c, with i^{th} component defined by

$$p_i = s_p((t_p/\sqrt{n}) + c_i, \sum_i c_i = 0, \sum_i c_i^2 = n\text{-}1. \ (i = 1,2,\ldots,n)$$

The Jacobian is of the form $J(c)s_p^{n-1}$ and ignoring the error term the joint density is

$$KJ(c)s_p^{n-1} \exp -s_p^a \sum_i |((t_p/\sqrt{n}) + c_i|^a$$

and in terms of the observations and parameters the transformed pivotals are

$$t_p = (\bar{x}\text{-}\theta_1)\sqrt{n}/s_x, \ s_p = s_x/\theta_2, \ c_i = (x_i\text{-}\bar{x})s_x$$

exhibiting the fact that the c_i are ancillary.

For the complete inference we condition on the observed $c = c_0$, obtaining the joint density

$$Ks_p^{n-1} \exp -s_p^a \sum_i |(t/\sqrt{n}) + c_{i0}|^a$$

from which joint confidence sets can be obtained. But if we are interested only in θ_1, and ignore the possibility of further information about θ_2, then we can integrate out s_p and obtain the marginal density of t_p as

$$K/\{\sum_i |(t_p/\sqrt{n}) + c_{i0}|^a\}^{n/a} \tag{4}$$

and we may note that in the case of normality, with $a = 2$, the side conditions on the c_i make this density independent of c_{i0}, and in fact equal to Student's t density on n-1 degrees of freedom. The fact that the condition 1 density in this

case does not involve the c_{i0} corresponds to the fact that when the observations are normally distributed $\bar{x}$ and s_x are jointly sufficient for θ_1 and θ_2.

If we find a set T such that the density (4) integrated over T is equal to 0.95, then if $\bar{x}_0$ and s_{x0} are the observed sample mean and standard deviation, the set $\{\theta_1: t_p \in T\}$ is a 95% confidence set for θ_1. The smallest such set will be obtained if T consist of all points t_p for which the density (4) exceeds some suitably chosen constant.

Box and Tiao have discussed this model from the Bayesian point of view, using a 'non-informative prior' for θ_1 and θ_2. For given a, the posterior distribution they arrive at is the same as the confidence distribution derived from (4). That this is not accidental can be seen if we change our pivotal model so that P becomes $R^{n+1} \times R^+$, and define the first n components of p as before, but add $p_{n+1} = \theta_1$, $p_{n+2} = \theta_2$, and regard the $\phi_a(u)$ as giving the density of $p_1,\ldots,p_n$, given p_{n+1} and p_{n+2}, and giving to these last two components the distribution corresponding to the prior used by Box and Tiao. In so far as strict Bayesians sometimes object to these improper priors, it might be said that the Pivotal analysis given above is more Bayesian than the Bayesian treatment!.

Box and Tiao also assign a prior distribution to a, on the basis of external information to the effect that the observations are nearly normally distributed, though they are careful to examine whether, over the plausible range of the MA parameter* a, the value of makes any drastic difference. This is, of course, a perfectly reasonable way of dealing with an MA parameter, provided the inferences are suitably qualified. As a matter of fact, for the Darwin data examined by Box and Tiao, it appears more probable, from a reading of Darwin's own detailed account of how he obtained his data, that two of his observations have been given the wrong sign, and that the corrected observations are quite closely normal. If, of course, information was available providing an observational basis for a prior for either or both of θ_1 and θ_2 the pivotal analysis could be carried through on this basis.

AKNOWLEDGEMENT

Research supported by the NSERC of Canada.

* MA stands for "Model Adjustment", MA parameter seems a better term, I think, than "label", or "discrepancy parameter".