Conditional Inference and Cauchy Models

Peter McCullagh

Biometrika, Vol. 79, No. 2. (Jun., 1992), pp. 247-259.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28199206%2979%3A2%3C247%3ACIACM%3E2.0.CO%3B2-4

Biometrika is currently published by Biometrika Trust.

Biometrika (1992), **79**, 2, *pp.* 247–59
Printed in Great Britain

Conditional inference and Cauchy models

BY PETER MCCULLAGH
Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637, U.S.A.

SUMMARY

Many computations associated with the two-parameter Cauchy model are shown to be greatly simplified if the parameter space is represented by the complex plane rather than the real plane. With this convention we show that the family is closed under Möbius transformation of the sample space: there is a similar induced transformation on the parameter space. The chief raison d'être of the paper, however, is that the two-parameter Cauchy model provides an example of a nonunique configuration ancillary in the sense of Fisher (1934), such that the maximum likelihood estimate together with either ancillary is minimal sufficient. Some consequences for Bayesian inference and non-Bayesian conditional inference are explored. In particular, it is shown that conditional coverage probability assessments depend on the choice of ancillary. For moderate deviations, the effect occurs in the $O_p(n^{-1})$ term: for large deviations the relative effect is $O_p(n^{\frac{1}{2}})$.

Some key words: Ancillarity; Bayesian inference; Cauchy model; Complex-valued parameter; Conditional inference; Confidence density; Configuration; Fiducial distribution; Group transformation model; Maximal invariant; Möbius transformation; Non-Bayesian inference; Pivotal statistic; Probable error; Projective group.

1. PARAMETERIZATIONS OF THE CAUCHY FAMILY

Following the usual conventions, we say that the random variable Y has the Cauchy distribution with parameters θ_1, θ_2 if the density function has the form

$$f_Y(y;\theta)=\frac{1}{\pi|\theta_2|\{1+(y-\theta_1)^2/\theta_2^2\}} \tag{1}$$

for $-\infty<y<\infty$. It is conventional to take as parameter space the largest set for which (1) is defined and nondegenerate, namely the upper half plane $R\times R^+$ with $\theta_2>0$. Our purposes can be better served, however, by taking as parameter space the entire complex plane, so that $\theta=\theta_1+i\theta_2$ and $\bar{\theta}=\theta_1-i\theta_2$ represent the same Cauchy density (1). Real-valued parameters with $\theta_2=0$ thus represent degenerate distributions in which all the mass is concentrated at θ_1. We adopt the short-hand notation $Y\sim C(\theta)$ to mean that Y has the Cauchy density (1) with median $\mathbb{R}(\theta)$ and probable error $|\mathbb{I}(\theta)|$, as given by the real and imaginary parts of the parameter.

With this convention, the density (1) can be written in the more compact form

$$f_Y(y;\theta)=\frac{|\theta_2|}{\pi|y-\theta|^2}$$

provided that $\theta_2\neq 0$.

While there is no obvious reason not to use the complex plane to represent a parameter space, this choice is decidedly unusual in statistics. Why introduce complex numbers

when only real-valued random variables are involved? To this question I can give no better answer than to present the curious result that

$$Y^* = \frac{aY+b}{cY+d} \sim C\left(\frac{a\theta+b}{c\theta+d}\right) \tag{2}$$

for all real numbers a, b, c and d. What I find curious about this result is not the fact that the Cauchy family is preserved under linear transformation and reciprocals. Those properties follow immediately from the definition of the Cauchy as the ratio of two correlated normals with zero mean. No, the curiosity lies in the fact that the induced transformation on the parameter space has the same fractional linear form as the transformation on the sample space only if the parameter space is taken to be the complex plane. Without the introduction of complex-valued parameters the essential simplicity of (2) is not so easy to detect. In particular, the median and probable error of Y^* are respectively

$$\mathbb{R}\left(\frac{a\theta+b}{c\theta+d}\right) = \frac{(a\theta_1+b)(c\theta_1+d)+ac\theta_2^2}{(c\theta_1+d)^2+c^2\theta_2^2},$$

$$\left|\mathbb{I}\left(\frac{a\theta+b}{c\theta+d}\right)\right| = \left|\frac{(ad-bc)\theta_2}{(c\theta_1+d)^2+c^2\theta_2^2}\right|.$$

Expression (2) is related to characterization results obtained by Menon (1962, 1966), Pitman & Williams (1967) and Williams (1969). Knight (1976) shows that the Cauchy family is the only univariate location-scale family that is closed under projective transformation, i.e. (2) with a, b, c and d real. Knight & Meyer (1976), and also Dunau & Senateur (1987), prove a similar result for projective transformations in R^n.

2. Connection with Brownian motion

A simple interpretation can be given to (2) in terms of Brownian motion in the complex plane. The distribution of the first exit point from the upper half-plane of a Brownian particle starting at θ is the Cauchy density on the real line with parameter θ. Result (2) follows from Lévy's theorem concerning the preservation of Brownian paths under analytic transformation. Interested readers are referred to Rogers & Williams (1986, pp. 63–4). The transformation in our case is fractional linear with real coefficients, under which the real axis is invariant. However, result (2) is true even if the coefficients are complex numbers. In that case, the real Cauchy distribution is transformed to the circular Cauchy whose support is the image of the real axis. For example, the Cauchy distribution on the unit circle is

$$f(z;\theta)|dz| = f(z;1/\bar{\theta})|dz| = \frac{|1-|\theta|^2|}{2\pi|z-\theta|^2}|dz|. \tag{3}$$

For $|\theta|<1$, this expression gives the exit density from the unit circle of a Brownian particle starting at θ. If Y has the real Cauchy density (1), then $Z=(1+iY)/(1-iY)$ has the circular Cauchy density (3) with parameter $(1+i\theta)/(1-i\theta)$.

The example that follows is presented in terms of the Cauchy density on the real axis. Much of the discussion applies equally to the circular Cauchy density. The mathematical treatment of configuration ancillaries is in fact considerably simplified by working on the unit circle rather than the real line.

3. CONDITIONAL INFERENCE

3·1. *Foundations*

The theory of conditional inference is founded on a body of examples built up over the past 50 years by statisticians with deep insight into statistical theory and broad experience of applications. Three particularly important papers are Fisher (1934), Pitman (1938) and Cox (1958). The examples are highly varied and the solutions often ingenious, but they have this feature in common. Unconditional significance probability calculations such as those indicated by Neyman–Pearson theory are manifestly inappropriate and possibly misleading for hypotheses concerning the data at hand. By contrast, a conditional probability calculation often provides an answer to which no right-thinking person could object. The overwhelming evidence, so far as significance tests are concerned, is that conditioning on ancillary statistics is good practice.

My own assessment of the case is that, even though the evidence is largely circumstantial, the argument for conditioning is extremely compelling. The fact that the theory is rooted in examples is simultaneously its strength and its weakness. Certainly it seems safer to have a theory that is rooted in examples, however stylized, than a theory based on purely mathematical constructs whose relevance to the needs of the applied statistician is tenuous at best. In connection with ancillarity, however, there is a parallel body of counter-examples. As I understand it, the thrust of these counter-examples is not that conditioning is a bad idea, but that the usual mathematical formulation is in some respects ad hoc and not completely satisfactory. See, for example, Basu (1959, 1964) or Buehler (1982) and the references therein.

One of the principal pillars in the theory of conditional inference is Fisher's (1934) conditional solution to the location and location-scale problems. The key idea is to make a nonsingular transformation from the original Y to $(\hat{\theta}, A)$, where A is the maximal invariant under the location or location-scale group as appropriate. This invariant is automatically ancillary. By the conditionality principle, we require the conditional distribution of $\hat{\theta}$ given the observed value $A = a$, which is called the sample configuration. Conditional confidence intervals obtained via this route are equivalent to Bayesian intervals based on a certain improper, but not obviously nonsensical, prior. The latter point, however, is entirely incidental and fortuitous, and is rarely advanced in support of the conditional frequency-theory solution.

The need to make significance probability calculations relevant to the particular data at hand has been recognized by many authors, Cox (1958, 1980), Fraser (1968, p. 72), Cox & Hinkley (1974, pp. 38, 39), Efron & Hinkley (1978), Barndorff-Nielsen (1980, 1983, 1984), Hinkley (1980) and McCullagh (1984) to name a few. The example in § 3·5 illustrates one difficulty or ambiguity that arises in Fisher's (1934) conditional solution to the location-scale problem.

3·2. *Conditional distribution given the configuration*

Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables with density

$$f(y; \theta) = \theta_2^{-1} f\{(y - \theta_1)/\theta_2\}$$

depending on the unknown parameter θ. For the moment $f(y)$ is an arbitrary known probability density. The usual regularity conditions are not required in what follows, so

$f(.)$ might represent the double triangular density $|y|$ on $(-1, 1)$. In what follows, it is more critical than usual to distinguish between random variables and their observed values. In particular, $\hat{\theta}$ denotes the observed value of the random variable T whose components are T_1 and T_2: we take $T_2 > 0$. Assuming that the maximum-likelihood estimate is unique (Copas, 1975), the usual configuration ancillary is the vector with components $a_i = (y_i - \hat{\theta}_1)/|\hat{\theta}_2|$. The corresponding random variable is denoted by A. Fisher's definition of the configuration was $(Y_{(i)} - Y_{(1)})/(Y_{(2)} - Y_{(1)})$, where $Y_{(i)}$ is the ith order statistic. This is equivalent to the present definition in all essential respects, the only difference having to do with the order of the observations.

Fisher (1934) argues that inference for θ should be based on the conditional distribution of T given the observed value $A = a$ of the configuration ancillary. On transforming from Y to (T, A), the required conditional density is obtained in the form

$$\text{pr}\,(T \in dt | A = a;\, \theta) \propto t_2^{n-2} \prod f(t_1 + a_i t_2;\, \theta)\, dt_1\, dt_2, \tag{4}$$

where dt means $(t_1, t_1 + dt_1) \times (t_2, t_2 + dt_2)$. Note in particular that $\Pi f(\hat{\theta}_1 + a_i \hat{\theta}_2;\, \theta)$ is the likelihood function given the data $(T, A) = (\hat{\theta}, a)$ observed.

As I understand it, Fisher's intention is that the conditional density (4) should be used, *inter alia*, for the evaluation of significance probabilities. Thus, if H_0 is the simple hypothesis specifying $\theta = \theta_0$ and $X = X(Y, \theta_0)$ is a test statistic with observed value $x = X(y, \theta_0)$, the required conditional p-value for testing H_0 is given by

$$\text{pr}\,(X \geqslant x | A = a;\, \theta_0) = \int_{\{t: X \geqslant x\}} \text{pr}\,(T \in dt | A = a;\, \theta_0). \tag{5}$$

See, for example, Pitman (1939). Fisher's argument has some considerable appeal and, in one sense at least, it appears to provide a completely satisfactory answer to the location and location-scale problems. The argument can be extended to other problems where the model is generated by the action of a group on the sample space.

3·3. *Pivotal statistics*

Suppose now, as would often be the case, that confidence intervals or sets are required for the components θ_1 and θ_2 separately. In the conditional density (4) we make a change of variables as follows:

$$Z_1 = (T_1 - \theta_1)/T_2, \quad Z_2 = T_2/\theta_2.$$

The joint density of Z_1, Z_2 given $A = a$ is then

$$f_Z(z_1, z_2 | A = a;\, \theta) \propto z_2^{n-1} \prod f\{z_2(z_1 + a_i)\}, \tag{6}$$

which does not depend on θ. In other words Z_1, Z_2 are jointly pivotal.

The marginal density of either pivot may be obtained by numerical integration of (6) if necessary. In the case of Cauchy models, either marginal distribution can be obtained analytically by contour integration. For example, if $n = 2m + 1$ is odd, the marginal density of Z_1 given A is found to be proportional to

$$\sum_{r=1}^{n} \alpha_r^n \prod_{j \neq r} \frac{\alpha_j^2}{\alpha_j^2 - \alpha_r^2},$$

where $\alpha_r = 1/|z_1 + a_r|$ are taken to be distinct. Obviously, this expression is not in a form suitable for computation, particularly when, as must occur for some values of z_1, two of the α_i are equal.

A typical equi-tailed confidence interval for θ_1 derived via this pivot has the form

$$(\hat{\theta}_1 - \hat{\theta}_2 k_{1-\alpha/2},\ \hat{\theta}_1 - \hat{\theta}_2 k_{\alpha/2}), \tag{7}$$

where $\text{pr}\,(Z_1 < k_\alpha) = \alpha$. Operationally, so far as coverage probability for each component is concerned, we may interpret confidence intervals thus obtained as if

$$\theta_1 = \hat{\theta}_1 - \hat{\theta}_2 Z_1, \quad \theta_2 = \hat{\theta}_2 / Z_2,$$

with $\hat{\theta}_1$, $\hat{\theta}_2$ treated as fixed constants, and θ_2, θ_2 as random variables. With this interpretation, the probability that the random variable θ_1 lies in the nonrandom interval (7) is then exactly $1-\alpha$. Extending this argument in a purely formal way to the joint density, the joint 'fiducial density' for θ is proportional to

$$\theta_2^{-n-1} \prod f\{(y_i - \theta_1)/\theta_2\} \equiv \theta_2^{-1} \prod f(y_i;\ \theta), \tag{8}$$

with $\{y_i\}$ regarded as fixed numbers. For a numerically equivalent treatment from a slightly different point of view, see Pitman (1938, p. 413). The fiducial density and the resulting intervals are numerically identical to those obtained from the Bayesian posterior density based on the improper but commonly used prior

$$\pi(d\theta) \propto d\theta_1\, d\theta_2 / \theta_2.$$

A single-parameter version of this result is given by Lindley (1958). Although it is not clear that he would approve of the eponym, I will refer to this as the Pitman prior. Jeffreys's invariant prior $d\theta_1\, d\theta_2/\theta_2^2$ yields a different posterior that is not equivalent to the conditional frequency-theory solution.

My impression is that, forced to adopt a Bayesian viewpoint, most statisticians would prefer to use the Pitman prior over the Jeffreys prior in the location-scale context (Zellner, 1984). In normal-theory linear problems the Pitman prior gives answers numerically identical to those obtained via the usual t-statistic. Jeffreys's prior gets the degrees of freedom slightly wrong. Jeffreys (1966, p. 182) notes this difficulty and argues in favour of the Pitman prior, particularly for inferences regarding θ_1 or θ_2 separately.

3·4. *Barndorff-Nielsen's formula*

Barndorff-Nielsen (1980) noted that the conditional density (4) can be written directly in terms of the likelihood function

$$L(\theta;\ y) = \theta_2^{-n} \prod f\left(\frac{y_i - \theta_1}{\theta_2}\right).$$

For an arbitrary sample point $t_1 + at_2$ having the same configuration as that observed, the likelihood ratio

$$\frac{L(\theta;\ t_1 + at_2)}{L(t;\ t_1 + at_2)} = t_2^n \theta_2^{-n} \prod \frac{f\{(t_1 + a_i t_2 - \theta_1)/\theta_2\}}{f(a_i)}$$

differs from (4) by t_2^{-2} times a factor depending only on a. Now, the observed information determinant for θ is $\hat{j} \equiv j(\hat{\theta},\ a) = \hat{\theta}_2^{-4} j(a)$. Thus Barndorff-Nielsen's re-expression of Fisher's formula is

$$\text{pr}\,(T \in dt \mid A = a;\ \theta) \propto \{j(t,\ a)\}^{\frac{1}{2}} \frac{L(\theta;\ t_1 + at_2)}{L(t;\ t_1 + at_2)}\, dt_1\, dt_2. \tag{9}$$

This is sometimes written rather cryptically in the form

$$p(\hat{\theta} \mid a;\, \theta) \propto \hat{j}^{\frac{1}{2}} \exp\{l(\theta) - l(\hat{\theta})\},$$

where $l(\theta) \equiv l(\theta;\, y)$ is the log likelihood function in which all possible sample points y having the same configuration must be written as a function of $(t,\, a)$.

The advantage of expression (9) is that it is exact not just for the location-scale problem, but also for all such transformation models generated by the action of a group. Furthermore (9) has an interpretation even in the absence of group structure, though the existence and/or choice of ancillary may pose a problem. For full exponential-family problems the ancillary is immaterial by Basu's theorem, and (9) is in fact the saddlepoint approximation to the distribution of T. The relative error is then $O(n^{-1})$: with re-normalization this may be reduced to $O(n^{-3/2})$. The formula is also approximately correct more generally for large n, at least in regular cases.

Note that in order to evaluate significance probabilities based on (9) it is necessary in principle to evaluate the likelihood function not just for the observed data point $y = \hat{\theta}_1 1 + \hat{\theta}_2 a$ but also for all y in the two-dimensional subspace $t_1 1 + t_2 a$ spanned by the vectors 1 and a. Thus, the choice of ancillary appears to be critical in order that this space be unambiguously defined.

3·5. *Non-uniqueness of ancillary*

Although (4) and (9) are exact, the entire argument is open to criticism on at least two fronts. First, it is not entirely clear what meaning should be attached to (4) if $\hat{\theta}$ is not unique, as in the Cauchy problem with $n = 2$. Secondly, although the maximal invariant under the location-scale group is indeed unique and equal or equivalent to A, nothing in the formulation of the problem forces us to restrict our attention to the location-scale group. Apart from Fraser's (1968) treatment of transformation models, group-theoretic considerations are not usually considered to be an intrinsic part of the model specification. Instead, they emerge in the mathematical process of finding a satisfactory frequency-theory solution. In the usual set-up, therefore, the choice of group may be to some extent arbitrary. The maximal invariant under the group is always ancillary, but there may exist noninvariant ancillaries, or ancillaries that are invariant under a different group than the one initially considered. We now focus our attention on the consequences of nonuniqueness of the configuration ancillary for Cauchy models.

Let $R_i = 1/Y_i$, where $Y_i \sim C(\theta)$. It follows from (2) that $R_i \sim C(\psi)$, with $\psi = 1/\theta$ as complex numbers. For $n \geqslant 3$ the maximum likelihood estimate $\hat{\psi} = 1/\hat{\theta}$ is unique, and the configuration ancillary on the new scale is the vector with components $b_i = (r_i - \hat{\psi}_1)/|\hat{\psi}_2|$. In the algebra that follows, $\hat{\psi}_1$, $\hat{\psi}_2$, r are the observed values of the random variables P_1, P_2, R. By Fisher's argument, the conditional density of P given $B = b$ is

$$\text{pr}\,(P \in dp \mid B = b;\, \theta) \propto p_2^{n-2} \prod f(p_1 + b_i p_2;\, \psi)\, dp_1\, dp_2. \tag{10}$$

Note that, when computed at the observed values $P = \hat{\psi}$ and $T = \hat{\theta}$, the conditional densities (4) and (10) give rise to the same likelihood function because, for the Cauchy density,

$$f(r_i;\, \psi) = \frac{f(1/r_i;\, 1/\psi)}{r_i^2} = f(y_i;\, \theta) y_i^2.$$

Back-transformation from P to $T = 1/P$ yields the conditional density of T given $B = b$ in the form

$$\text{pr}\,(T \in dt \mid B = b;\, \theta) \propto t_2^{n-2} \prod \left\{ f\left(\frac{|t|^2}{t_1 + b_i t_2};\, \theta \right) \frac{|t|^2}{(t_1 + b_i t_2)^2} \right\} dt_1\, dt_2. \tag{11}$$

Evidently, this is not the same as (4).

The constant of integration in (4) depends on a but not on θ: likewise in (11) the constant depends on b but not on θ. Apart from these constants, the ratio of (4) to (11) is the rational function

$$Q(t;\, \theta) = \prod \frac{|t|^4 - 2\theta_1 |t|^2 (t_1 + b_i t_2) + |\theta|^2 (t_1 + b_i t_2)^2}{|t|^2 \{ |\theta|^2 - 2\theta_1 (t_1 + a_i t_2) + (t_1 + a_i t_2)^2 \}}.$$

In order for the conditional densities (4) and (11) to be approximately equal, it is necessary and sufficient that $Q(t;\, \theta)$ be approximately constant in t for all t-values in regions where the probability densities are not negligibly small. The following observations suggest that $Q(t;\, \theta)$ is indeed approximately constant in the region of interest. First, at the true value $t = \theta$ we have

$$Q(\theta;\, \theta) = \prod \frac{1 + b_i^2}{1 + a_i^2}$$

independently of θ. Secondly, at the observed value $t = \hat{\theta}$, and in fact on the entire line $t \propto \hat{\theta}$, we find

$$Q(\hat{\theta};\, \theta) = \prod (|\hat{\theta}|^2 / y_i^2) = \prod \frac{1 + b_i^2}{1 + a_i^2} = Q(\theta;\, \theta).$$

At the observed point, the ratio is independent of θ because both (4) and (11) give rise to the same likelihood function as the original data. The likelihood function cancels in the ratio $Q(\hat{\theta};\, \theta)$. Finally, Taylor expansion about $t = \theta$ gives

$$Q(t;\, \theta) = Q(\theta;\, \theta) + O(|t - \theta|^2).$$

The coefficients in the quadratic term are functions of the two configurations a and b. For $n > 3$, numerical evidence shows that, apart from special configurations, these quadratic terms are nonzero, and $Q(t;\, \theta)$ has a saddlepoint at $t = \theta$. In general, it appears that there are several saddlepoints in a neighbourhood of $t = \theta$, but $t = \hat{\theta}$ is not normally a stationary point of the ratio. Figure 1 shows a typical contour plot of $\log Q$, with $y = (0{\cdot}91,\ 1{\cdot}34,\ 2{\cdot}84,\ 3{\cdot}55)$ and $\theta = (1, 1)$, and drawn so that the zero contour passes through θ and $\hat{\theta}$. The contours are at $0, \pm(0{\cdot}1, 0{\cdot}2, 0{\cdot}3, 0{\cdot}5, 0{\cdot}7, 0{\cdot}9)$. The dashed lines are contours of the log likelihood function corresponding to nominal 20%, 50%, 80% and 95% levels.

In the conditional density given $B = b$, the quantities

$$Z_1^* = (P_1 - \psi_1)/P_2, \quad Z_2^* = P_2/\psi_2 \tag{12}$$

are jointly pivotal, whereas (Z_1, Z_2) as defined in § 3·3 are not. By the argument given in § 3·3, the joint fiducial density for ψ based on the pivot (Z_1^*, Z_2^*) is proportional to

$$\psi_2^{-n-1} \prod f\{(r_i - \psi_1)/\psi_2\} = \psi_2^{-1} \prod f(r_i;\, \psi).$$

By transformation of variables, the fiducial density for θ based on the conditional distribution given B is proportional to

$$|\theta|^{-2} \theta_2^{-1} \prod f(y_i;\, \theta),$$

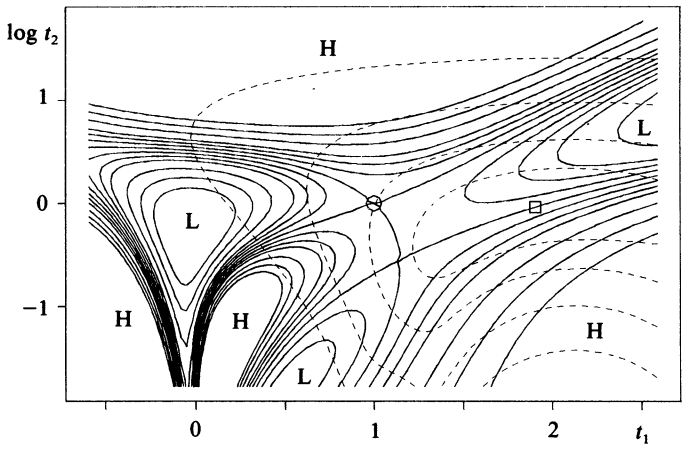

Fig. 1. Contour plot of log density ratio $\log Q(t; \theta)$ for $n = 4$; H and L, high and low values. Circle is at θ; square at $\hat{\theta}$. Dashed lines are contours of the log likelihood.

which is clearly different from (8). This nonuniqueness of bivariate fiducial pivots, previously remarked on by Tukey (1957, 1990), is a major deficiency of the fiducial method.

In general, therefore, A and B are not equivalent ancillaries, nor are they jointly ancillary (McCullagh, 1987, p. 248). The conditional sample spaces are two-dimensional with a one-dimensional intersection corresponding to multiples of the observed y. The two conditional distributions are not compatible because

$$\text{pr}(T \in dt \mid B; \theta) \neq \text{pr}(T \in dt \mid A; \theta).$$

In other words the p-values obtained from (5) do depend on the choice of ancillary. Furthermore, because of the symmetry of the problem, Cox's (1971) criterion cannot distinguish between the two ancillaries unless prior information on θ is available for use.

3·6. *Numerical example*

To take a simple numerical example, suppose $n = 3$ and $y = (0{\cdot}5, 1{\cdot}0, 2{\cdot}0)$ so that $\hat{\theta} = 0{\cdot}92857 \pm 0{\cdot}37115i$ and $a = (-1{\cdot}155, 0{\cdot}192, 2{\cdot}887)$. The conditional density of T given A is

$$\text{pr}(T \in dt \mid A = a; \theta) \propto t_2 \prod f(t_1 + a_i t_2; \theta)\, dt_1\, dt_2.$$

The marginal density of the pivot $Z_1 = (t_1 - \theta_1)/T_2$ is found by contour integration to be

$$k(a)\{(|z_1 + a_1| + |z_1 + a_2|)(|z_1 + a_1| + |z_1 + a_3|)(|z_1 + a_2| + |z_1 + a_3|)\}^{-1},$$

which is inverse cubic in the tails and inverse linear in the centre. Numerical integration gives $k(a) \doteq 5{\cdot}7024$. Figure 2 shows that, although the density is continuous, cusps occur at the points $-a_1, -a_2, -a_3$. The probability content of the four regions bounded by the cusps is approximately 0·0786, 0·4214, 0·3630 and 0·1370 respectively, so that $-a_2$ is at the median. In general, at least for odd n, it appears that the conditional median of Z_1 is close to, but not exactly equal to, the median of the set $-A$.

The lower and upper 5% points of the density of Z_1 are −3·405 and 2·135 respectively. Thus the conditional equi-tailed 90% confidence interval for θ_1 is

$$(0{\cdot}9286 - 0{\cdot}3712 \times 2{\cdot}135, 0{\cdot}9286 + 0{\cdot}3712 \times 3{\cdot}405) = (0{\cdot}1363, 2{\cdot}1923).$$

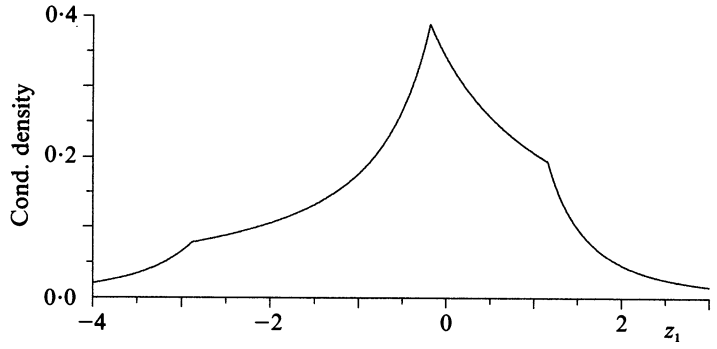

Fig. 2. Conditional density of pivot Z_1 given $n = 3$ and $A = \{-1{\cdot}155, 0{\cdot}192, 2{\cdot}887\}$. Cusps occur at points $-A$. median and mode are equal.

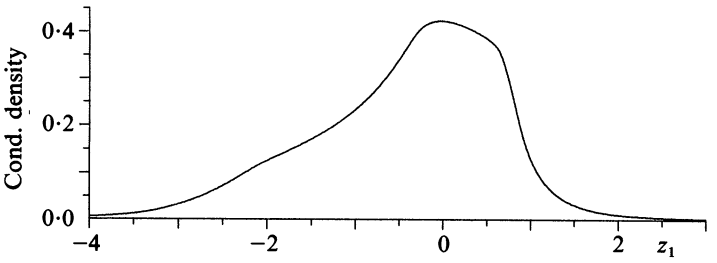

Fig. 3. Conditional density of pivot Z_1 given $n = 5$ and $A = \{-0{\cdot}825, -0{\cdot}639, 0{\cdot}292, 2{\cdot}152, 3{\cdot}082\}$. Cusps occur in the derivative at points $-A$.

The centre of this interval is 1·164, which is almost equal to the mean of the observations. However, the normal-theory 90% interval for the mean is (−1·063, 3·397), which is more than twice as wide as the conditional Cauchy confidence interval for the median.

The analogous pivot for normal translation-scale models has the t-distribution on two degrees of freedom. In both cases, therefore, the first moment of Z_1 exists, but the variance and higher-order moments are infinite. In this example $E(Z_1 | A = a) \doteq -0{\cdot}377$. More generally, it can be shown that, for odd values of $n \geqslant 3$, the moments of Z_1 up to order $n-2$ are finite, while those of order $n-1$ or more are infinite. This result may well be true also for even values of n, but that remains unverified.

Figure 3 shows the distribution of Z_1 for a sample of size 5 in which values 0·4 and 2·5 have been added to the previous sample. The tails are inverse quintic and the density is markedly lumpy, though unimodal. The first three moments of this density are finite.

In the context of robust confidence interval estimation, S. Morgenthaler, in an unpublished paper, has presented diagrams very similar to Figs 2 and 3.

As it happens, $r = 1/y$ is a permutation of y, so we have $|\hat{\theta}| = 1$, $\hat{\psi} = \hat{\theta}$, $b = a$, and the constants of integration in (4) and (11) are equal. Curve (a) in Fig. 4 shows the confidence density or fiducial density for θ_1 based on the pivot Z_1 conditionally on $A = a$. The lower and upper 5% points in this density are 0·1363 and 2·1923 respectively, which are identical to the confidence limits obtained above. In the conditional distribution given $B = b$, however, there appears to be no exact pivot for θ_1, so no exact conditional confidence limits for θ_1 can be obtained. However, the joint fiducial density for ψ is proportional to $\psi_2^{-1} \Pi f(r_i; \psi)$. On treating this as an ordinary bivariate density, we may transform from ψ to θ and integrate out θ_2 to obtain the marginal fiducial density of θ_1 conditionally

on $B = b$ as shown in Fig. 4, curve (b). In density (b), however, the interval $(-\infty, 0{\cdot}1363)$ has probability 7·8%, whereas $(2{\cdot}1923, \infty)$ has probability approximately 0·5%. Intervals based on density (b), unlike those obtained from (a), do not have coverage probability equal to the nominal value. Density (b) has shorter tails than (a).

The two densities in Fig. 4 have the alternative Bayesian interpretation as marginal posteriors for θ_1, density (a) based on the prior $d\theta_1\, d\theta_2/\theta_2$, and (b) based on the prior $d\psi_1\, d\psi_2/\psi_2$.

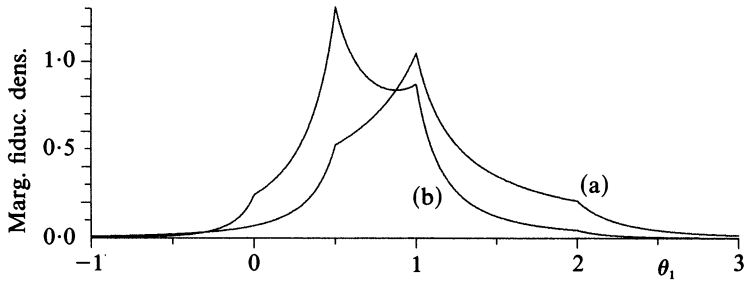

Fig. 4. Fiducial densities for θ_1 based on sample $y = (0{\cdot}5, 1{\cdot}0, 2{\cdot}0)$. Curve (a), confidence density based on the pivot Z_1; (b), marginal fiducial density based on the pivot (Z_1^*, Z_2^*). Density (b) has an extra cusp at the origin.

3·7. *Asymptotic analysis*

Consider now two statisticians, both conditional frequentists using Fisher's (1934) argument, one using an analysis based on Y conditional on the configuration A, and the other using an analysis based on $R = 1/Y$ conditional on the reciprocal configuration B. Suppose that the first statistician obtains his confidence interval or fiducial interval from the pivot Z_1 given A, and the second statistician uses the marginal fiducial distribution derived from the bivariate pivot Z^* given B. Their computations are then identical to those of two Bayesians. Conditioning on A is equivalent to using the prior $d\theta_1\, d\theta_2/\theta_2$: conditioning on B is equivalent to using the prior $d\psi_1\, d\psi_2/\psi_2 = d\theta_1\, d\theta_2/(\theta_2|\theta|^2)$. The effect on probability calculations, either frequency-theory coverage probability or Bayesian posterior calculations, is therefore $O(n^{-1})$ in large samples. In an asymptotic sense, therefore, the discrepancy between our two conditional frequentists, one orthodox and one fiducialist, is of the same order of magnitude as the discrepancy between two Bayesians using different noninvariant priors. By contrast, the difference between conditioning and not conditioning is typically $O(n^{-\frac{1}{2}})$.

The preceding comparison is not entirely satisfactory because it presupposes acceptance of the fiducial or joint pivotal argument. A more direct comparison of the effect of conditioning is achieved by comparing the two probability measures whose densities are given by (4) and (11). One way to compare the two densities is to choose an arbitrary set S in the parameter space and to compute the arithmetic difference

$$D_S(a, b) = \text{pr}\,(T \in S \mid A = a) - \text{pr}\,(T \in S \mid B = b). \tag{13}$$

If the random variable $D_S(A, B)$ is $O_p(n^{-k})$ for almost all sets S not depending on a, b, we say that the effect of conditioning appears in the $O(n^{-k})$ term in large samples. It is usual to choose sets S satisfying $\text{pr}\,(T \in S) = O(1)$ as $n \to \infty$, though this restriction is not necessary.

Although the regularity conditions required by McCullagh (1984) are not satisfied here because the parameter is two-dimensional and A does not have an Edgeworth expansion, the result given there suggests that the difference in (13) should be not greater than $O(n^{-1})$. For the present example it is possible to show that the Taylor expansion of $\log Q(t;\theta)$ about $t=\theta$ has the form

$$\log Q(t;\theta)=\log Q(\theta;\theta)+Q_2(t-\theta)+Q_3(t-\theta)+O(|t-\theta|^4),$$

where $Q_r(.)$ is homogeneous of degree r. The coefficients, which depend on a, b, and θ, are $O_p(1)$ for $Q_2(.)$, $O_p(n^{\frac{1}{2}})$ or smaller for $Q_3(.)$, and $O_p(n)$ or smaller for the higher-order terms. I have obtained an explicit expression for $Q_2(.)$ in terms of certain Fourier coefficients T_r, which are functions of the configuration and are $O_p(n^{\frac{1}{2}})$ with zero mean provided that the configurations are marginally typical. For example, the coefficient of $(t_1-\theta_1)^2$ is

$$\{T_0(\cos 4\alpha-\cos 4\hat{\alpha})+T_1(\sin 4\alpha-\sin 4\hat{\alpha})\}/(4\theta_2^2),$$

where $\alpha=\arg(\theta)$, and $\hat{\alpha}=\arg(\hat{\theta})$ is expressed as a function of a and b. In the moderate deviation range, for which $t=\theta+O(n^{-\frac{1}{2}})$, it follows that

$$\log Q(t;\theta)=\log Q(\theta;\theta)+O_p(n^{-1}).$$

The effect on probability calculations of choosing ancillary A rather than B is therefore $O_p(n^{-1})$ as measured by (13).

In the case of large deviations with $t-\theta=O(1)$, however, the logarithmic or relative difference

$$D_S^*(a,b)=\log \text{pr}\,(T\in S\,|\,A=a)-\log \text{pr}\,(T\in S\,|\,B=b) \tag{14}$$

is of order $O(T_0)O(\hat{\alpha}-\alpha)$ at least, as evidenced by the quadratic term in the preceding expansion. In the large-deviation regime, we have $\hat{\alpha}-\alpha=O(1)$, so that the relative difference is $O_p(n^{\frac{1}{2}})$. In the case of intermediate deviations with $t-\theta=O(n^{-\frac{1}{2}+q/2})$, it is possible to show that the relative difference in (14) is $O(n^{3q/2-1})$, providing a smooth transition between the moderate-deviation region ($q=0$) and the large-deviation region ($q=1$).

3·8. *The case $n=2$*

This apparently trivial case is of independent theoretical interest for the following reasons.

(i) The minimal sufficient statistic, which can be written in the form $S=(\bar{Y}, r=\frac{1}{2}|Y_1-Y_2|)$, is complete, so there is no nontrivial ancillary statistic.

(ii) The maximum-likelihood estimate of θ is not a single point, but coincides with the circle having (y_1, y_2) as diameter, i.e. the set $\{\bar{y}+re^{\pm i\phi}: 0<\phi<\pi\}$. This is the unique equivariant estimator under the real Möbius group.

(iii) The estimator $\hat{\theta}=\bar{y}+ir$ maximizes the likelihood, and is the unique single-point equivariant estimator under the location-scale group $y\to a+by$ with $b\neq 0$. Under the location-scale group with $b>0$, there exist other single-point equivariant estimators such as $\bar{y}+re^{i\pi/4}$.

(iv) With $\hat{\theta}$ taken as $\bar{y}+ir$, the joint density of the pivots Z_1, Z_2 exists and is given by (6). Density (8) also exists both as a joint fiducial density and as a Bayesian posterior density for θ. The marginal density of the pivot Z_1 is

$$f_1(z_1)=\frac{1}{2\pi^2 z_1}\log\left|\frac{z_1+1}{z_1-1}\right|,$$

which is symmetric with singularities at $z_1=\pm 1$. The implied confidence density for θ_1 is symmetric about $\bar{y}$, with singularities at y_1 and y_2. Despite its startling appearance, this is by no means an outrageous conclusion.

It is perhaps worthwhile clarifying point (ii), which, though technically correct, does not tell the full story. There exists a one-dimensional sub-group of Möbius transforms with real coefficients that keeps the sample points $\{y_1, y_2\}$ invariant. Under the action of this group, parameter points move in distinct nonoverlapping orbits. The equivariant sets in the complex plane are these orbits. With one exception, each orbit is a complex conjugate pair of circles having (y_1, y_2) as chord, but excluding the points y_1, y_2; the real axis is a limiting circle that coincides with its conjugate image. The only exceptional orbit is the set $\{y_1, y_2\}$. The likelihood function is constant on orbits, and takes distinct values on different orbits: it is not defined on the exceptional orbit. In this sense, the value assumed by the likelihood function uniquely identifies an orbit, and hence an equivariant estimator. The orbit of maximum likelihood is the circle identified in (ii), but excluding the exceptional orbit.

Thus, if the full implications of invariance under the Möbius group are accepted at face value, the only equivariant statements that can be made concerning θ are those based on level sets of the likelihood function. Statements based on density (8) or on the pivot Z_1 are typically not equivariant.

4. DISCUSSION

From a purely mathematical viewpoint, given that the two parameters of the Cauchy model are treated on an equal footing, the choice between what we call Y and what we call the reciprocal scale $R=1/Y$ is an arbitrary matter of labelling. For internal consistency, therefore, the conclusions from an analysis based on Y should be identical to those based on R. However, standard methods of conditional inference, and of Bayesian inference using the Pitman prior $d\mu\, d\sigma/\sigma$, do not have this property. The defect is curable in a Bayesian analysis by switching to the invariant prior $d\mu\, d\sigma/\sigma^2$, but no equally simple remedy is available in a conditional frequency-theory analysis.

From a more practical point of view, one can argue that it is rarely the case that one would be interested in treating the components of $\theta=(\mu, \sigma)$ on an equal footing. As a consequence, unless Y has a clear physical interpretation as a ratio or the tangent of an angle, any model that is closed under reciprocal transformation is unlikely to be appropriate. I interpret this as an eminently sensible objection to the use of Cauchy models in many, if not most, 'typical' applications. From this point of view, conditioning on the ancillary A might be more sensible than conditioning on the reciprocal configuration B. Pitman's (1939) arguments could be used to buttress such a claim. Nevertheless, given circumstances for which the Cauchy model is deemed appropriate, the theoretical problem remains.

One might view the Cauchy example as an argument against conditioning, but if anything that is the opposite of what is intended. In terms of p-values and coverage

probability assessments, fully conditional probability calculations differ by only $O(n^{-1})$. This is similar to the difference between Bayesian posterior calculations based on different priors. Frequency-theory probability assessments that are not fully conditional can differ by $O(n^{-\frac{1}{2}})$. Thus, conditioning guarantees uniqueness, but only as far as the $O(n^{-1})$ term.

Acknowledgements

I am grateful to D. R. Cox, D. A. S. Fraser, S. Morgenthaler, J. Pitman, M. Stein, D. L. Wallace, D. Williams and E. J. Williams for help with several aspects of the paper.

References

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika* **67**, 293-310.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343-65.

Barndorff-Nielsen, O. E. (1984). On conditionality resolution and the likelihood ratio for curved exponential models. *Scand. J. Statist.* **11**, 157-70. Corr. (1985) **12**, 191.

Basu, D. (1959). The family of ancillary statistics. *Sankhyā* A **21**, 247-56.

Basu, D. (1964). Recovery of ancillary information. *Sankhyā* A **26**, 3-16.

Buehler, R. J. (1982). Some ancillary statistics and their properties (with discussion). *J. Am. Statist. Assoc.* **77**, 581-94.

Copas, J. B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika* **62**, 701-4.

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357-72.

Cox, D. R. (1971). The choice between alternative ancillary statistics. *J. R. Statist. Soc.* B **33**, 251-5.

Cox, D. R. (1980). Local ancillarity. *Biometrika* **67**, 279-86.

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

Dunau, J.-L. & Senateur, H. (1987). An elementary proof of the Knight-Meyer characterization of the Cauchy distribution. *J. Mult. Anal.* **22**, 74-8.

Efron, B. & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika* **65**, 457-87.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. R. Soc.* A **144**, 285-307.

Fraser, D. A. S. (1968). *The Structure of Inference.* New York: Wiley.

Hinkley, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67**, 287-92.

Jeffreys, H. (1966). *Theory of Probability*, 3rd ed. Oxford: Clarendon Press.

Knight, F. B. (1976). A characterization of the Cauchy type. *Proc. Am. Math. Soc.* **55**, 130-5.

Knight, F. B. & Meyer, P. A. (1976). Charactérization de la loi de Cauchy: *Z. Wahrscheinlichskeitstheorie* **34**, 129-34.

Lindley, D. V. (1958). Fiducial distributions and Bayes's theorem. *J. R. Statist. Soc.* B **20**, 102-7.

McCullagh, P. (1984). Local sufficiency. *Biometrika* **71**, 233-44.

McCullagh, P. (1987). *Tensor Methods in Statistics.* London: Chapman and Hall.

Menon, M. V. (1962). A characterization of the Cauchy distribution. *Ann. Math. Statist.* **33**, 1267-71.

Menon, M. V. (1966). Another characteristic property of the Cauchy distribution. *Ann. Math. Statist.* **37**, 289-94.

Pitman, E. J. G. (1938). The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika* **30**, 391-421.

Pitman, E. J. G. (1939). Tests of hypothesis concerning location and scale parameters. *Biometrika* **31**, 200-15.

Pitman, E. J. G. & Williams, E. J. (1967). Cauchy-distributed functions of Cauchy variates. *Ann. Math. Statist.* **38**, 916-8.

Rogers, L. C. G. & Williams, D. (1986). *Diffusions, Markov Processes, and Martingales*, **2**, Chichester: Wiley.

Tukey, J. W. (1957). Some examples with fiducial relevance. *Ann. Math. Statist.* **28**, 687-95.

Tukey, J. W. (1990). The present state of fiducial probability. In *The Collected Papers of J. W. Tukey*, **4**, pp. 55-118. Pacific Grove, CA: Wadsworth.

Williams, E. J. (1969). Cauchy-distributed functions and a characterization of the Cauchy distribution. *Ann. Math. Statist.* **40**, 1083-5.

Zellner, A. (1984). Maximal data information prior distributions. In *Basic Issues in Econometrics*, Ed. A. Zellner, pp. 201-15. University of Chicago Press.

[*Received May* 1991. *Revised December* 1991]