Journal of Statistical Planning and Inference 108 (2002) 71–83

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Likelihood inference for location, scale, and shape

George A Barnard[a,✠], John B. Copas[b,*]

[a] University of Essex, Colcheter CO4 3SQ, Essex, UK
[b] University of Warwick, Coventry CV4 7AL, UK

Abstract

If (μ, σ, Σ) denote the location, scale, and shape parameters of a continuous variate X, we show how the exact likelihood function $L(\mu, \sigma, \Sigma | x_1 \ldots, x_n)$ based on n independent observed values of X can be displayed and used to make frequency-interpretable inferences about any or all of the three parameters. When interest is confined to fewer than three parameters, "simplifying assumptions" may be needed to preserve accuracy in the frequency interpretation. Such simplifying assumptions *mathematically* resemble Bayesian priors but their logical status is quite different. The approach used leads towards a "Bayes–Frequentist" compromise.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Location; Scale; Shape; Likelihood principle; Bayes-non-Bayes compromise; Frequency interpretation

0. Sixty years of statistics

It is a particular pleasure for the senior author to welcome C.R. Rao to the over-eighties club. We were introduced to statistics via the early editions of Fisher's "Statistical Methods for Research Workers" through whose examples we worked with the help of tables of squares and machines which could add, subtract, multiply or divide two four-figure numbers in about 10 s per operation. We no longer need tables of squares nor any of a wide range of functions; operations take less than microseconds, and above all we have instant access memories for storing data and intermediate computed values. Statisticians now can and do deal with problems vastly more complicated than those discussed in Fisher's book. But increased complication is not the only change to

[*] Corresponding author.
E-mail address: jbc@stats.warwick.ac.uk (J.B. Copas).
[✠] Professor G.A. Barnard died on 30 July 2002.

PII: S0378-3758(02)00271-9

statistical practice which the PC has brought. It has made practicable procedures which our pioneers would doubtless have wished to follow but were prevented from doing so by computing limitations. We illustrate this by considering the effect of these changes on a very simple problem: the estimation of the location, scale, and shape of the distribution of a real-valued variate X on which we have made n independent observations.

1. Student's problem

Fisher's book was written with the primary aim of encouraging *experimenters* to use the "small sample" methods introduced in 1908 by Student. The data he quoted were requoted by Fisher, and the scientific background has been described in detail by Senn (1993). We therefore use Student's problem to illustrate the new *logical* possibilities opened up by the PC.

Prevented by confidentiality from using his own data, Student quoted work by two experimental psychologists, Cushny and Peebles (CP) interested in the brain centres involved in the action of sleep-prolonging drugs. The data (Fisher, 1958, p. 121) consisted of 10 independent observations:

$$x_i: \; 1.2 \; 2.4 \; 1.3 \; 1.3 \; 0.0 \; 1.0 \; 1.8 \; 0.8 \; 4.6 \; 1.4$$

of the additional hours of sleep gained by patients when taking drug A instead of drug B. Treatments A and B differed only in their chirality. Differences associated with chirality would imply chirality in the brain centres affected by the drugs. This was the issue with which CP were concerned. For the data given, Student found

$$t = (\bar{x} - \mu)\sqrt{n}/s_x = (1.58 - \mu)/0.3890 = 4.062,$$

giving a (one-sided) P-value against the null hypothesis $\mu = 0$ of about 0.002— strong evidence in favour of chirality in the brain centres.

Senn gives an excellent account of the mistakes Student and Fisher made in reading CPs paper, and the mistakes Cushny and Peebles made in their experimental design. For simplicity we shall discuss the CP data and the associated inference as if no such mistakes were made.

2. Scientific inference and decisions

Scientific inference from experimental data begins with a statistical model, usually parametric in form. A model is a collection of alternative hypotheses one of which is supposed true to sufficient approximation. No one who has lived through the past 80 years could fail to recognise that scientific models and the inferences made from them are always subject to revision in the light of further information. In 1938, Einstein, Podolsky, and Rosen (EPR) thought they had shown that the quantum-theoretical description of reality must be incomplete because the contrary assumption would require "action at a distance" of a kind then regarded as impossible. Within the past decade experiments have shown that the phenomenon regarded by EPR as impossible does in fact occur. Wegener's notion of continental drift, regarded all his life by no less an authority than Harold Jeffreys as clearly impossible, is now accepted, along with magnetic polar reversal, by all geophysicists as a major factor in the formation of continents.

It is the role of the P-values given by statistical tests to indicate when the time has come to revise our *model*. Computer limitations in the past meant that P-values came to be used for comparing the plausibilities of the various *hypotheses* covered by our model — a purpose for which they are unsuited.

Given a specified model, the *relative* credibilities of the various hypotheses covered by the model are measured by the likelihood function. It is now possible, using spreadsheets, to display the likelihood function so that it is easy to *see* these relative credibilities.

That Fisher would have used spreadsheets had he been able to do so is shown by this passage, quoted in full from the first edition of his book: "Inferences respecting populations, from which known samples have been drawn, cannot be expressed in terms of probability, except in the trivial case when the population is itself a sample of a super-population the specification of which is known with accuracy.

"This is not to say that we cannot draw, from knowledge of a sample, inferences respecting the population from which the sample was drawn, but that the mathematical concept of probability is inadequate to express our mental confidence or diffidence in making such inferences and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term "**Likelihood**" to designate this quantity; since both the words "likelihood" and "probability" are loosely used in common speech to cover both kinds of relationship."

Fisher's habit of revising later editions of his book from interleaved copies meant that the crystal clear and correct passage quoted became overlaid with interpolations, particularly those arising from his 1931 discovery of "fiducial probability". But the original text remained uncut through all the many later editions. It was part of Fisher's tragedy that he did not live to see the practical possibility of directly exploring the whole likelihood function $L(\cdot|E)$. The aim of this paper is to show, by reference to CP's data, how Fisher and Student might have used spreadsheets to deal with Student's problem.

In 1946, the senior author was among a minority in judging that the statistical theory of sampling inspection should be approached via Bayes' Theorem. Fisher himself later acknowledged that the term "trivial" in the passage quoted was not altogether justified. It is now generally accepted that any *decision* that needs to be taken at a given time should receive similar treatment, using a prior coupled with the data-based likelihood. But it is the difference between inference and decision that unless decisive action is called for, the appropriate attitude to alternative hypotheses H and H' contained in a given model is best described as weighing the evidence for H as against H', using the likelihood ratio $L(\mathrm{H}/\mathrm{H}'|E)$ alone, and allowing for the possibility that the evidence available may not be sufficient to allow a reliable choice of hypothesis to be made. It is time we gave up using P-values for purposes for which they are not well suited.

3. P-values, L-values and W-values

In the passage quoted Fisher wrote *measuring*, not merely ranking. Partial justification for treating L as a measure derives from the fact that if data E make the (vector) parameter value α W times as likely as the value α', further independent evidence making α' at least W times as likely as α will be needed to reverse the order of credibility. But the full long run *frequency* justification of the term measure derives from the fact that when faced with choosing between two hypotheses H and H' on the basis of data E, if we choose H when $L(\mathrm{H}/\mathrm{H}'|E)$ exceeds W, and we choose H$'$ when $L(\mathrm{H}/\mathrm{H}'|E)$ is less than $1/W$, *making no choice if* $1/W < L(\mathrm{H}/\mathrm{H}'|E) < W$, then in a long run of choices correct choices will outnumber incorrect choices by more than W to 1. Assuming that the set of hypotheses considered includes the correct one, the long run ratio W of correct to incorrect choices may be called the long run odds against error. The long run involved here is the long run of our experience in dealing with *different* problems. Neyman's concept of repeated sampling from the *same* population is not involved.

The proof of our assertion is simple: The fact that we are choosing on the basis of given data E, *and nothing else*, means that our labelling of H and H$'$ can be done at random. Then the prior probabilities of H and H$'$ are equal, and by Bayes' theorem the posterior odds *when a choice is made* are either greater than W or less than $1/W$.

We use W to denote critical values of the likelihood ratio in honour of Abraham Wald, to whom the frequency interpretation of the L-ratio is due, though he called it the probability ratio. Its general applicability was pointed out in Barnard (1947), but the general practical application has had to wait for the coming of the PC.

4. Definitions of location, scale and shape

The probability distribution of any real-valued variate X is specified by its cumulative distribution function (cdf)

$$F(x) = Pr\{X \leqslant x\} \text{ for all real } x.$$

For any interval $(a,b]$, the probability that X falls in $(a,b]$ is

$$Pr\{a < X \leqslant b\} = F(b) - F(a).$$

We define X to be *continuous* iff $F(x)$ is continuous and strictly increasing for all x for which $0 < F(x) < 1$. For any continuous X we define its *location parameter* μ by the condition $F(\mu) = 1/2$ and its *scale parameter* σ by the condition $F(\mu+\sigma) = 3/4$. Thus μ is the median and $\mu+\sigma$ is the upper quartile of X. The *shape parameter* Σ of X is then defined as the cdf of $U = (X-\mu)/\sigma$. Any cdf of a continuous U with median 0 and upper quartile 1 can be a shape parameter.

The Cauchy Shape parameter is

$$C(u) = 1/2 + (1/\pi)\tan^{-1} u, \tag{1}$$

the Normal Shape parameter is

$$N(u) = \Phi(u/0.67449), \tag{2}$$

where $\Phi(x)$ is the cdf of a standard normal deviate, and the Rectangular Shape parameter is

$$R(u) = \begin{cases} 0, & u < -2, \\ (u+2)/4, & -2 \leqslant u \leqslant 2, \\ 1, & u > 2. \end{cases} \tag{3}$$

If $S_i(u), i=1,2,\ldots,k$ are distinct shape parameters, the “mixture” $\Sigma_i v_i S_i(u)$ with $v_i \geqslant 0$ and $\Sigma_i v_i = 1$ is also a shape parameter. It is rarely the case that the shape parameter Σ is exactly known. It is usually best taken to be a mixture of standard shapes in proportions which can themselves be estimated along with μ and σ from the available data.

5. Exact calculation of the likelihood function $L(u, \sigma, \Sigma | x, \ldots, x_n)$

If X is continuous with median μ, scale σ, and shape $\Sigma(u)$, its cdf is

$$Pr\{X < x | \mu, \sigma, \Sigma\} = \Sigma((x-\mu)/\sigma) \tag{4}$$

and the probability that X lies in the interval $x_i \pm \delta_i$ is the central difference

$$\Sigma((x_i + \delta_i - \mu)/\sigma) - \Sigma((x_i - \delta_i - \mu)/\sigma) = \Delta\Sigma(u_i, \pm\varepsilon_i). \tag{5}$$

where $u_i = (x_i - \mu)/\sigma$ and $\varepsilon_i = \delta_i/\sigma$. Given a sample of independent observations x_i, $i = 1, 2, \ldots .n$, with x_i accurate to within $\pm\delta_i$, the likelihood $L(\mu, \sigma, \Sigma \| x_1, \ldots x_n)$ is proportional to the product of central differences

$$K\, \Pi_i\, \Delta\Sigma(u_i, \pm\varepsilon_i) \tag{6}$$

with K an arbitrary factor of proportionality not involving any of the three parameters μ, σ, or Σ.

Provided that all the observations are of equal high accuracy and Σ has a smooth derivative $\Psi(u) = \Sigma'(u)$, we can approximate (6) by the product of *densities*

$$K\, \Pi_i\, 2\varepsilon\Psi(u_i) \tag{7}$$

and the common factor 2ε can be absorbed into K to give

$$K\, \Pi_i\, \Psi(u_i). \tag{8}$$

In a startling proportion of modern texts one finds the likelihood *defined* as (8). But (8) is only an approximation to the exact (6) and is quite unsuitable as a general definition. Fifty years ago measurements were less accurate than now, and were often of varying accuracy. The need to use (6) was then universally recognised. The main reason now for distinguishing between (6) and (8) arises with variate transformations such as $T{:}X \Rightarrow 1/X$ under which a small x-interval of length 2δ independent of x transforms to an interval of length varying with x and becoming arbitrarily large as x approaches the origin.

Given a catalogue $\{S_i\}\, i = 1, 2, \ldots, k$ of shape functions, the shape-parameter space they generate is the $(k-1)$-dimensional set $\{S\} = \{\Sigma_i\, v_i S_i\}$ with $v_i \geqslant 0$ and $\Sigma_i\, v_i = 1$.

The complete parameter space of (μ, σ, Σ) is then the $(k+1)$-dimensional set product $R \times R^+ \times \{S\}$, and the three parameters are variation independent in Barndorff-Nielsen's (1978) sense, i.e. taking any single one, or any two of the three parameters as fixed and known does not restrict in any way the possible values of the remaining unknown parameter(s).

When $\Sigma_1 = N$ and $\Sigma_2 = C$ we have a normal density "contaminated" with Cauchy observations which might appear as outliers. Much effort has gone into devising ways of dealing with outliers, using concepts such as "Winsorising" and "trimming". But the spreadsheet of the 3-parameter likelihood $L(\mu, \sigma, \Sigma)$ can be exhibited as two-way tables in (μ, σ), one giving $L(\mu, \sigma, \Sigma = N)$, the other giving $L(\mu, \sigma, \Sigma = C)$. The likelihood of various values of v can then be estimated by interpolation and used to improve the estimates of λ and σ. If, as v varies from 0 to 1, the spreadsheet values for shape functions $S(v) = vC(u) + (1-v)N(u)$ vary little, this means that the data to hand are robust to Cauchy contamination when estimating μ and σ, and correspondingly insensitive to variations in v. But if the spreadsheet values differ seriously, then estimation of v is both easier and more important.

For reasons given below we use the terms N-map, C-map and R-map for the likelihood spreadsheets for the CP data based on $\Sigma = N$, $\Sigma = C$ and $\Sigma = R$. Abbreviated versions of these are presented below.

6. The likelihood ratio for "A or B" versus "A' or B'"

CP could have known from the randomisation they (should have) carried out that the distribution of their data was symmetric. Denoting by μ the centre of symmetry of the X distribution, CP were interested in the plausibility of $\mu = 0$ versus $\mu = a$ for some $a \neq 0$. In Neyman's terminology, they were interested in the relative credibility of the *composite* hypotheses $\mathrm{H}_0 : \mu = 0,\ 0 < \sigma < \infty,\ \Sigma$ symmetric but otherwise unspecified and $\mathrm{H}_1 : \mu = a \neq 0,\ 0 < \sigma < \infty, \Sigma$ symmetric but otherwise unspecified. Likelihood ratios are usually defined for simple hypotheses only. Can their frequency interpretations be extended to cover composite hypotheses?

In the final edition of Fisher's "Statistical Methods and Scientific Inference" (Fisher, 1959) we find "Whereas such a phrase as "the probability of A or B" has a simple meaning, where A and B are mutually exclusive possibilities, the phrase "the likelihood of A or B" is more parallel with "the income of Peter or Paul"— you cannot know what it is until you know which is meant". And the same statement is to be found elsewhere in Fisher's writings.

This has led many to deny the possibility of meaningfully adding two likelihoods. But CP were not in the least interested in the various possible values of σ and their situation with regard to σ corresponds to one in which "the income of Peter or Paul" could be taken to mean the income of either one of Peter or Paul, any possible *differences* between *their* incomes being irrelevant to the issues under discussion. In cases such as that with which CP were concerned it seems reasonable, when the H_i are simple hypotheses between which we are indifferent, to *define* "H_1 or H_2 or $\ldots, \mathrm{H}_k$" to mean the simple mixture $(v_1\mathrm{H}_1 + v_2\mathrm{H}_2 + \cdots + v_k\mathrm{H}_k)$, with $\Sigma_i v_i = 1$. And unless available data indicate otherwise, we may suppose that the H_i are exchangeable so that we may make the "simplifying assumption" that $v_i = 1/k$ for all i. The unweighted addition of the tabulated likelihoods in the likelihood maps given below corresponds to adopting this "simplifying assumption" which implies taking the distribution of $\lambda = \log \sigma$ to have a density *approximately* proportional to the row totals — i.e. that non uniformities in the density of λ are negligible given the observed data. Such a "simplifying assumption" is much weaker than the assumption that λ is uniformly distributed a priori. And the likelihood maps enable us to see how big the changes in the "prior" distributions of the nuisance parameters would have to be to make our likelihoods of the parameters of interest seriously in error. In the CP case we make two simplifying assumptions (i) that the shape parameter is a mixture of N, R, and C and (ii) that the scale parameter λ is uniformly distributed a priori. As will be seen below, given the CP data it is clear that both are justified.

Mathematically, our simplifying assumption about the scale parameter produces the same result as using the Jeffreys prior element $d\sigma/\sigma$ for the unknown σ. But the rationale for doing this is not the same as Jeffreys'. The validity or otherwise of his argument is not affected by the sample size nor is it affected by the data actually to hand. But in the CP case the data show that the information that would have been needed to have a serious effect on their inference was quite unavailable; indeed it remains so today.

The status of our simplifying assumption is the same as in other branches of applied mathematics. In many problems of fluid mechanics it is customary to assume, for example that viscosity is negligible except in the boundary layer, or that flow is laminar. In both cases the experienced professional will usually be able to judge whether the assumption can safely be made. If there is serious doubt, then further experimentation may be required.

7. The *N, C, R* maps for the CP data

Table 1 shows three tables of values of (6) for the CP data, taking the shape Σ to be N, C and R. We take $\delta = \pm 0.05$, recognising the one decimal place accuracy of the data. For convenience we choose the value of K so that the maximum value of (6) for the normal model is 100. (Note that the *same* value of K is used for all three choices of Σ). These tables are called maps because they enable us to "explore" the whole likelihood function. They show the "heights" of the likelihood at the various parameter points. The odds on any point in a mountain region where $L > M$ as against any point in a valley region where $L < m$ are at least M/m.

The maps given in Table 1 are abbreviated from fuller maps printed directly by computer using (6) above. Their immediate interpretation is illustrated by comparing the entry 61.4 at $(1.2, -0.2)$ on the N-map with the entry 893.4 at $(1.2, -1.0)$ on the C-map. These tell us that the likelihood ratio for $(\mu, \lambda, \Sigma) = (1.2, -1.0, C)$ as against $(\mu, \lambda, \Sigma) = (1.2, -0.2, N)$ is $893.4/61.4 = 14.55$. If presented with these two hypotheses as alternatives, one of which can be taken as true, a W value of 10 would lead us to choose $(1.2, -1.0, C)$ as against $(\mu, \lambda, \Sigma) = (1.2, -0.2, N)$. But with a more cautious $W = 20$ we would refrain from making a choice.

Table 1
Likelihood maps for the CP data

	λ	μ								
		0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2
N-map	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.8	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0
	0.6	0.2	0.5	0.9	1.2	1.3	1.2	0.8	0.5	0.2
	0.4	0.5	1.5	3.4	5.4	6.3	5.2	3.2	1.4	0.4
	0.2	0.5	2.8	9.2	18.7	23.3	17.8	8.4	2.4	0.4
	0.0	0.2	2.6	15.6	44.7	62.1	41.6	13.5	2.1	0.2
	−0.2	0.0	0.9	12.7	61.4	100.0	55.1	10.2	0.6	0.0
	−0.4	0.0	0.1	3.5	36.8	76.2	31.3	2.5	0.0	0.0
	−0.6	0.0	0.0	0.2	6.4	19.1	5.1	0.1	0.0	0.0
	−0.8	0.0	0.0	0.0	0.2	0.9	0.1	0.0	0.0	0.0
	−1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C-map										
	1.0	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.0
	0.8	0.0	0.1	0.2	0.4	0.4	0.2	0.1	0.0	0.0
	0.6	0.1	0.3	0.9	1.7	1.6	0.8	0.3	0.1	0.0
	0.4	0.1	0.6	2.9	6.2	5.4	2.1	0.4	0.1	0.0
	0.2	0.1	1.1	7.5	20.1	15.7	4.2	0.6	0.1	0.0
	0.0	0.1	1.4	16.0	56.6	38.0	6.5	0.5	0.0	0.0
	−0.2	0.0	1.3	27.1	136.5	74.6	7.6	0.4	0.0	0.0
	−0.4	0.0	0.9	36.1	280.6	116.8	6.7	0.2	0.0	0.0
	−0.6	0.0	0.5	37.4	489.0	143.4	4.4	0.1	0.0	0.0
	−0.8	0.0	0.2	30.1	720.1	136.8	2.2	0.0	0.0	0.0
	−1.0	0.0	0.1	19.0	893.4	100.8	0.9	0.0	0.0	0.0
	−1.2	0.0	0.0	9.7	931.8	57.5	0.3	0.0	0.0	0.0
	−1.4	0.0	0.0	4.0	815.6	25.6	0.1	0.0	0.0	0.0
	−1.6	0.0	0.0	1.4	598.8	9.1	0.0	0.0	0.0	0.0
	−1.8	0.0	0.0	0.4	369.6	2.6	0.0	0.0	0.0	0.0
	−2.0	0.0	0.0	0.1	192.7	0.6	0.0	0.0	0.0	0.0
	−2.2	0.0	0.0	0.0	85.6	0.1	0.0	0.0	0.0	0.0
	−2.4	0.0	0.0	0.0	32.7	0.0	0.0	0.0	0.0	0.0
R-map										
	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	0.6	0.0	0.0	0.0	1.6	1.6	1.6	1.6	1.6	1.6
	0.4	0.0	0.0	0.0	0.0	4.1	12.1	12.1	12.1	0.0
	0.2	0.0	0.0	0.0	0.0	0.0	0.0	82.8	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The likelihood maps in Table 1 are shown as contour plots in Fig. 1. These use the same scale factor K as in the tables, but label the contours on the log scale from $L = \exp(0) = 1$, $L = \exp(1)$, $L = \exp(2)$ and so on. Although the parameter $\lambda = \log \sigma$ is defined consistently in the sense that σ is the difference between the upper quartile

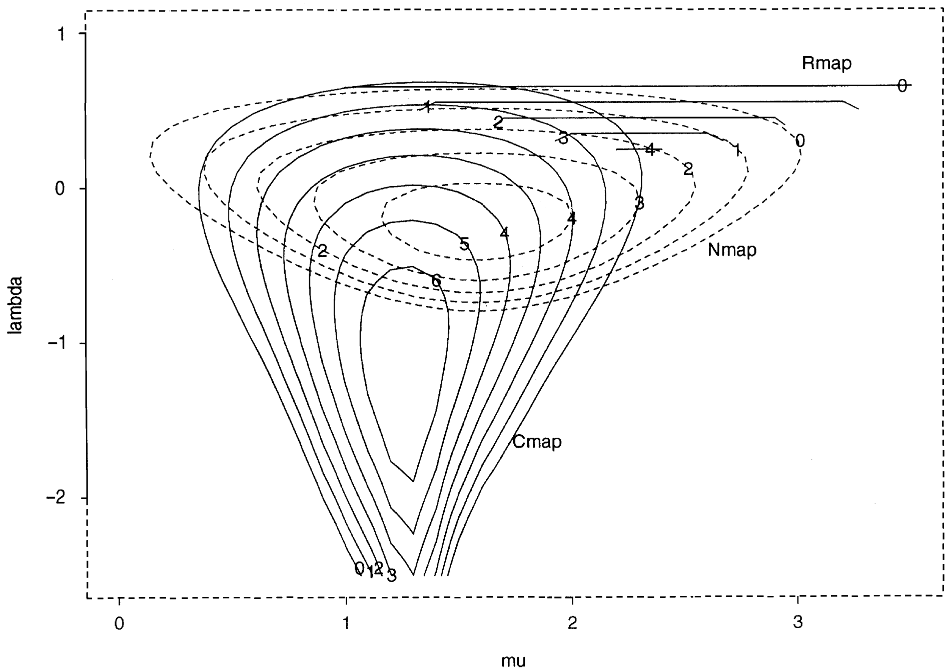

Fig. 1. Likelihood contours.

and the median for each of the shape models, the ranges of λ given most weight by the likelihood maps show differences owing to the different tail behaviour of each distribution.

There is some evidence that the true shape might be nearer to C rather than to N; if it were at all important it might be worth checking for—one explanation might be failure of the independence assumption due to the arousal of one patient interfering with the sleep of another. But with $W = 10$ the possibility need not be taken very seriously. And anyway the answer to this question is scientifically unimportant since CP were interested only in whether the centre of symmetry μ was or was not zero. In the light of the discussion in Sections 1–6 and the maps in Table 1, it is reasonable to make each of the three simplifying assumptions needed to derive the inferences that the likelihood ratios for the various possible values of μ are given approximately by column totals shown in Table 2. Column totals of these likelihood maps are also illustrated in Fig. 2, using the log scale. Note that both Table 2 and Fig 2 use more detailed spreadsheets than the summary tables presented in Table 1. The unusual appearance of the curves for the R model in Figs. 1 and 2 reflects the lack of mathematical regularity of the rectangular distribution.

When each of the three upper rows in Table 2 is divided by its first element we have the figures in Table 3, showing that we could offer 100 to 1 against $\mu = 0$ and reasonably expect to win for any mixture of shapes $\Sigma = \xi N + \eta C + \zeta R$ with $\xi + \eta + \zeta = 1$. Cushny and Peebles were fully justified in pursuing the source of the brain chirality.

Table 2
Column totals from detailed tables of $L(\mu, \lambda, \Sigma)$

	μ								
	0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2
N	3.1	17.1	91.5	349.9	578.9	315.3	77.9	14.5	2.6
C	0.8	12.8	385.7	11,270.0	1458.3	71.9	5.3	0.5	0.0
R	0.1	0.3	0.9	2.5	11.0	51.8	134.6	19.0	6.9
$N+C+R$	4.0	30.2	478.1	11,622.4	2048.2	439.0	217.8	34.0	9.5

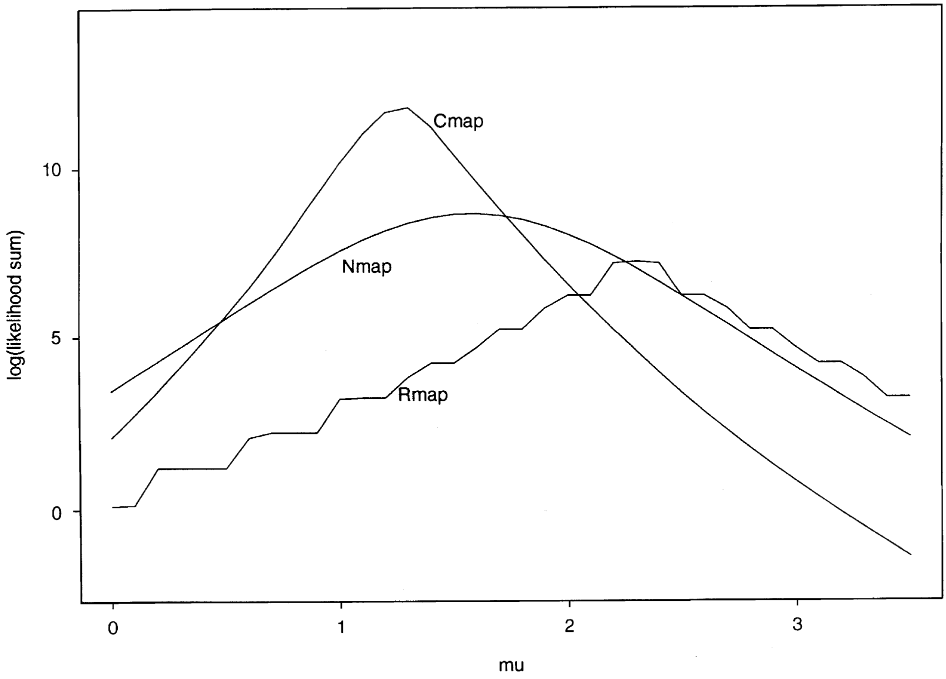

Fig. 2. Likelihood column totals (log scale).

Table 3
Rescaled column totals

	μ								
	0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2
N	1.0	5.5	29.5	112.9	186.7	101.7	25.1	4.7	0.8
C	1.0	16.0	482.1	14,087.5	1822.9	89.9	6.6	0.6	0.0
R	1	3	9	25	110	518	1346	190	69

8. Another approach to the elimination of nuisance parameters

A very important paper (Berger et al., 1999—BLW) came to hand after much of the present paper was drafted. The authors consider cases like that considered here, where the model parameters can be separated into two sets (α, β), variation independent of each other, with α the parameter(s) of interest and β as nuisance parameter(s). Their treatment is also based on the likelihood principle. The authors accept that likelihoods may also provide suitable answers to inferential problems though their paper leans towards a Bayesian treatment rather more than that developed here.

The CP problem discussed here in detail would seem to provide a case where BLW might agree that the inference as given in Section 6 is more appropriate than any alternative. Recalling the final paragraph of Section 2 above, a posterior *distribution* for the parameter of interest will strictly be needed if and only if we have a loss function $D(A, \alpha)$ measuring the damage done by taking action A when the true parameter value is α. In such a case, the likelihood needs to be supplemented by a prior to produce a posterior distribution allowing the expected loss from taking action A to be calculated.

Cushny and Peebles did go on to look for chiral molecules which could be the seat of their drug action. Had μ in fact been zero, this would have been a wild goose chase whose cost could be assessed *after* the event, but not before. Besides clearly indicating chirality, the odds against $\mu = 0$ for the various positive values of μ would have indicated the size of the chirality effect they needed to search for. Incidentally, the fact that μ is a time implied that there could be no ambiguity as to whether the likelihoods should be given as functions of μ or whether, for example, likelihoods of μ^3 should be used. Wald's first use of the likelihood ratio was in connection with sampling inspection problems in which both the costs of wrong decisions and the costs of further experimentation could be assessed with reasonable accuracy. This allowed a fully Bayesian approach to the problems of whether or not to continue sampling, and, if not, whether to accept or reject. Real life produces a host of different situations, some approximating to the CP case, others approximating the Wald case. Recognition of this fact should assist in reducing the apparent antagonism between "Bayesians" and "Frequentists".

We would be interested to learn from BLW whether they would accept the term "simplifying hypothesis" for their and our assumptions that the distributions of the nuisance parameters have densities proportional to their likelihoods. The deliberate ambiguity as to the sense of the word "simplifying" resembles that of the word "unbiased" as applied to a parameter estimate whose mean is equal to the true value.

9. Concluding remarks

It is an unfortunate accident of history that Fisher's dispute with Karl Pearson over the number of degrees of freedom in χ^2, coupled with his justified enthusiasm for Student's t led him to over-emphasise the virtues of "exactness" in P-values. He was well aware that exact normal distributions rarely, if ever, occur in nature as is evident from the encouragement he gave to O'Toole in his work on densities whose logarithms are quartic rather than quadratic. Another historical accident encouraging excessive emphasis on pure mathematical exactitude was the development in France and Russia of powerful schools working in mathematical probability culminating in Kolmogoroff's "Grundbegriffe der Warscheinlichkeitsrechnung". It has too often been forgotten that the connection of this work with statistics resembles that of the theory of perfect fluids with the dynamics of real fluids.

The role of mathematics in any field of application is to provide a language in which problems in the *real* world can be discussed in *approximate* quantitative terms. Our observations of the real world can never be mathematically exact. It is devoutly to be hoped that the quantification of the odds against error proposed above for assessing the weight of evidence for or against hypotheses will not lead anyone to elevate particular W values to the absurd status once held by the numbers 1/20 and 1/100 in assessing P-values. Had Fisher and Student possessed PCs there is no question but that they would have calculated reasonably accurate P-values for the data sets they considered. As it was, their correspondence vividly shows the labour they had to undertake to produce their 5% and 1% tables whose misinterpretations continue to plague applied statistics.

The fact, taken for granted and illustrated here, that robustness is a property *conditional* on the data was first stressed by Dempster (1975). It is not easy to find a recent paper in which the necessarily finite precision of any observation is noted. Matthews and Farewell (1985) may perhaps be the most recent one.

Nelder (1999), stressing the risks of "over-mathematising" statistical science, also came to hand when this paper was in an advanced state of drafting. Among very much else, Nelder criticises the excessive emphasis placed on P values in assessing the credibility of specific hypotheses. He points to the serious dangers of publication bias resulting from what he calls this "P-value culture".

Returning to Berger et al. (1999) we may remark that one way of relating the Bayesian and the likelihood approaches to inference is to bear in mind another distinction between natural science and decision making. An essential element of the former is the *indefinite repeatability* of experimental results. It was at one time thought that such repeatability must necessarily be deterministic. But Mendel's laws, having been repeatedly tested throughout the present century, are now universally accepted as basically true although they predict only the relative frequency with which results of a given type will occur in sufficiently large experiments. The first philosopher to point this out seems to have been E. Kolman.

References

Barnard, G.A., 1947. Sequential analysis (review). J. Am. Statist. Assoc. 42, 658–664.

Barndorff-Nielsen, O., 1978. Information and Exponential Families in Statistical Theory. Wiley, Chichester, pp. ix+238 (Chapter IV).

Berger, J.O., Liseo, B., Wolpert, R.L., 1999. Integrated likelihood methods for eliminating nuisance parameters. Statist. Sci. 14, 1–22.

Dempster, A.P., 1975. A subjectivist look at robustness. Bulletin of International Statistical Institute 46 (Book 1) 349–374.

Fisher, R.A., 1958. Statistical Methods for Research Workers, 3rd Edition. Oliver and Boyd, Edinburgh.

Fisher, R.A., 1959. Statistical Methods and Scientific Inference, 2nd Edition. Oliver and Boyd, Edinburgh.

Matthews, D.E., Farewell, V.T., 1985. On a singularity in the likelihood for a change-point hazard rate model. Biometrika 72, 703–704.

Nelder, J.A., 1999. Statistics for the Millenium. From statistics to statistical science. J. Roy. Statist. Soc. D 48, 257–267.

Senn, S., 1993. Cross-over Trials in Medical Research. Wiley, Chichester.