Canad. Math. Bull. Vol. **40** (2), 1997 pp. 231–243

ASYMPTOTIC THEORY AND THE FOUNDATIONS OF STATISTICS

N. REID

ABSTRACT. Statistics in the 20th century has been enlivened by a passionate, occasionally bitter, and still vibrant debate on the foundations of statistics and in particular on Bayesian vs. frequentist approaches to inference. In 1975 D. V. Lindley predicted a Bayesian 21st century for statistics. This prediction has often been discussed since, but there is still no consensus on the probability of its correctness. Recent developments in the asymptotic theory of statistics are, surprisingly, shedding new light on this debate, and may have the potential to provide a common middle ground.

1. **Introduction.** I am very honoured to have been awarded this prize lectureship, and very humbled, in view of the many highly deserving candidates among our colleagues. I would like to thank the Research Committee, the Committee on Women in Mathematics and the CMS for their courage and leadership in establishing this prize lectureship. I also hope I will see the day when there are so many women in mathematics that such awards are viewed as quaint relics of an earlier time.

I was asked to present a lecture on my research for a general mathematical audience, and not having been faced with such a tall order before, I found it a little difficult to know where to start. The type of research that I've been involved in over the past ten or fifteen years would probably be broadly described by my statistical colleagues as "theoretical statistics". The particular area of theoretical statistics that I study is parametric inference. At the risk of boring you with some background material, I will describe this briefly, as an introduction and to fix some notation.

We will start from the assumption that a random variable y takes values on (some subset of) the real line according to a probability density function $f(y)$. The probability that y takes values in the interval $[a, b]$ is given by $\int_a^b f(y)\,dy$. The function f is non-negative and integrates to 1. In statistics probability density functions are typically defined relative to Lebesgue measure (as I have done here) or counting measure: the former for describing variables that take values on $\mathbb{R}$ or some sub-interval of $\mathbb{R}$, and the latter for variables taking values on a lattice such as Z^+. In the lattice case we would write $\sum_{y\in Z^+} f(y) = 1$. A much more general treatment using Lebesgue-Stieltjes integrals is of course available. A straightforward extension to vector-valued random variables is also possible, but not needed here.

The first Krieger-Nelson Prize Lectureship, given at the summer meeting of the Canadian Mathematical Society in Toronto, June 1995.

Received by the editors June 19, 1995.

AMS subject classification: Primary: 62-02; Secondary: 62A10, 62A15.

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

The density function will be further indexed by a parameter θ, taking values in a parameter space Θ which is most often (an interval of) $\mathbb{R}$ or a connected region of $\mathbb{R}^k$. Thus the probability model is really a family of densities $\{f(y;\theta) : \theta \in \Theta\}$, assumed to be non-negative and normed for every $\theta \in \Theta$.

Parametric inference is the study of the 'inverse problem'[1] of reasoning from the random variable y to the parameter θ. Since we can only describe the behaviour of the system probabilistically, we are interested in the general or average case and thus typically assume that we have available a series of observations $y_1, \ldots, y_n$ each taking values according the the family of densities $\{f(y;\theta)\}$. In the simplest case, and the only one I will discuss here, we assume that the observations $\mathbf{y} = (y_1, \ldots y_n)$ are taken independently, so that the model for the complete set of observations is, with a slight abuse of notation,

$$f(\mathbf{y};\theta) = \prod_1^n f(y_i;\theta).$$

Examples of the types of inferences that statisticians make and study are

- "The value $\theta = \theta_0$ is not consistent with $\mathbf{y}$, at level 0.03."
- "A 95% confidence interval for θ is $\big(\theta_L(\mathbf{y}), \theta_U(\mathbf{y})\big)$."
- "The probability is 0.90 that $\theta \in \big(\theta_{LB}(\mathbf{y}), \theta_{UB}(\mathbf{y})\big)$."

There is also an important body of work in nonparametric inference, in which the density function f is not parameterized, but assumed only to belong to a suitable class of smooth functions. As well, inference in models allowing dependence among the y_is is an important aspect of several areas in statistics, including time series analysis, multivariate analysis, and inference in stochastic processes.

To emphasize the goal of the inverse problem, working from $(y_1, \ldots, y_n)$ back to θ, we often write the density function $f(\mathbf{y};\theta)$ as a function of θ, and give it a new name, the *likelihood function*:

$$L(\theta) = L(\theta;\mathbf{y}) = f(\mathbf{y};\theta)c(\mathbf{y}).$$

In fact the likelihood function is an equivalence class of functions proportional to the joint density, since observation of a one-to-one function of $\mathbf{y}$, for example, should give us the same inference statement about θ. This means in particular that values of $L(\theta)$ can only be assessed relative to each other.

The likelihood function, like most other concepts in modern statistics, was introduced into mathematical statistics by Sir Ronald Fisher [9,10,11]. He showed that the likelihood function contains all the information in the observations $\mathbf{y}$ about the parameter θ, and described the likelihood function as providing a "measure of the degree of our rational belief in a conclusion [about θ]". The problem of parametric inference is then to find ways to calibrate the likelihood function, in order to draw inferential conclusions like the ones quoted above.

[1] I am indebted to Bradley Efron for this description of inference.

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

2. **Examples.**

2.1. *Example 1: ulcer data.* To give a flavour of the uses to which parametric inference might be put, we illustrate with three fairly simple examples. The first, taken from Efron [6,8], concerns investigation of a new treatment for stomach ulcers. Table 1 shows the data from an experiment conducted to evaluate the new treatment relative to the standard treatment.

	success	failure
new treatment	2	36
old treatment	12	20

TABLE 1: A 2 × 2 table comparing two treatments for stomach ulcers. From Efron [8].

A reasonable and simple starting point for modelling the data (given the details of the experimental protocol, which are omitted here), is to assume that y_1, the number of successes on the new treatment, follows a Binomial (38, p_1) density: *i.e.*

$$f(y_1;p_1) = \binom{38}{y_1} p_1^{y_1}(1-p_1)^{38-y_1} \quad y_1 = 0,\ldots,38; \quad p_1 \in [0,1],$$

and that y_2, the number of successes on the standard treatment, follows a Binomial (32, p_2) density, independently of y_1. The likelihood function for (p_1, p_2) is

$$L(p_1,p_2;y_1,y_2) \propto p_1^{y_1}(1-p_1)^{38-y_1} p_2^{y_2}(1-p_2)^{32-y_2} \quad p_1,p_2 \in [0,1]^2.$$

To evaluate the new treatment relative to the standard means assessing the magnitude of the difference between p_1 and p_2. For various reasons it is convenient to measure this difference on the log-odds scale, so we define

$$\theta_1 = \log[\{p_1/(1-p_1)\}/\{p_2/(1-p_2)\}]$$
$$\theta_2 = \log\{p_2/(1-p_2)\}.$$

Then we can write

$$L(\theta_1,\theta_2;y_1,y_2) \propto \exp\{\theta_1 y_1 + \theta_2(y_1+y_2) - k(\theta_1,\theta_2)\}$$

which is in a mathematically more attractive form, even if the parametrization is a little more difficult to interpret. This likelihood function is an example of a likelihood function for an *exponential family* model: these models play a special role in parametric inference, because of their nice mathematical properties. They also provide a reasonably broad class of models for use in applications.

In fact Efron [8] concerns a set of 41 such tables of data from 41 different published studies of the new ulcer treatment. This "study of studies" is becoming very important in medical statistics, where it is variously called "outcomes research" or "meta-analysis".

Note that θ_1 measures directly how beneficial the new treatment is, relative to the standard. The parameter θ_2 measures the success rate of the standard treatment, and is not of direct interest in evaluating the new treatment. This splitting of a model's vector

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

parameter into a *parameter of interest* and a *nuisance parameter* is a characteristic feature of most recent work in parametric inference.

2.2. *Example 2: Shakespeare.* A similar model seemed appropriate for a recent problem that our statistical consulting service tackled this spring. In this case the data were obtained from a study of spelling variants in early published volumes of Shakespeare's sonnets. After some distillation, we summarized one piece of the data as shown in Table 2.

	page 1	page 2	page 3	page 4
far	0	4	1	2
farre	2	0	3	1

TABLE 2: Two spelling variants of *far* used on different pages in a volume of Shakespeare's sonnets

Of particular interest in this case was whether the two spelling variants were essentially segregated by page, which would in turn suggest that two different typesetters were involved in the printing. The model used was an extension from two independent binomials to $k(=4)$, and the question of interest is whether or not the observed data support the hypothesis $p_1 = \cdots = p_k$. There were approximately 100 such tables, one for each pair of spelling variants on about 100 different words. A special feature of this data is the presence of large numbers of zeroes in the cells of the table.

2.3. *Example 3: salaries.* The third example is of some (notionally continuous) data on faculty salaries, carried out as one part of a large university's pay equity exercise a few years ago. Figure 1 shows a typical plot of salary vs. years of experience for males and females, across the university. A possible model for this data is the linear regression model

$$y_i = \alpha + \beta(\text{years}_i) + \gamma(\text{gender}_i) + \sigma e_i, \quad i = 1, \ldots n$$

where y_i is the salary of the i-th employee, and e_i is a random variable with a density $f_0(e)$ centered at 0. The likelihood function is thus of the form

$$L(\alpha, \beta, \gamma, \sigma; \mathbf{y}) \propto \prod_{i=1}^{n} \sigma^{-1} f_0\Big\{\big(y_i - \alpha - \beta(\text{years}_i) - \gamma(\text{gender}_i)\big)\sigma^{-1}\Big\}.$$

Of particular interest is inference about the 'gender effect', γ, although this is a substantial abstraction of the real applied problem. In fact the original analysis was carried out department by department. As well, there are clearly questions about the suitability of a model linear in years, and the assumption of constant variance.

3. **Inference from likelihood.** I will now abstract from the previous examples, and assume that our starting point is a vector of observations $\mathbf{y} = (y_1, \ldots y_n)$, a family of densities $\{f(y;\theta); \theta \in \Theta\}$, and a likelihood function $L(\theta) = L(\theta; \mathbf{y}) \propto \prod f(y_i; \theta)$. Much of parametric inference is devoted to the study of how $L(\theta)$ can be used to provide 'good'

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

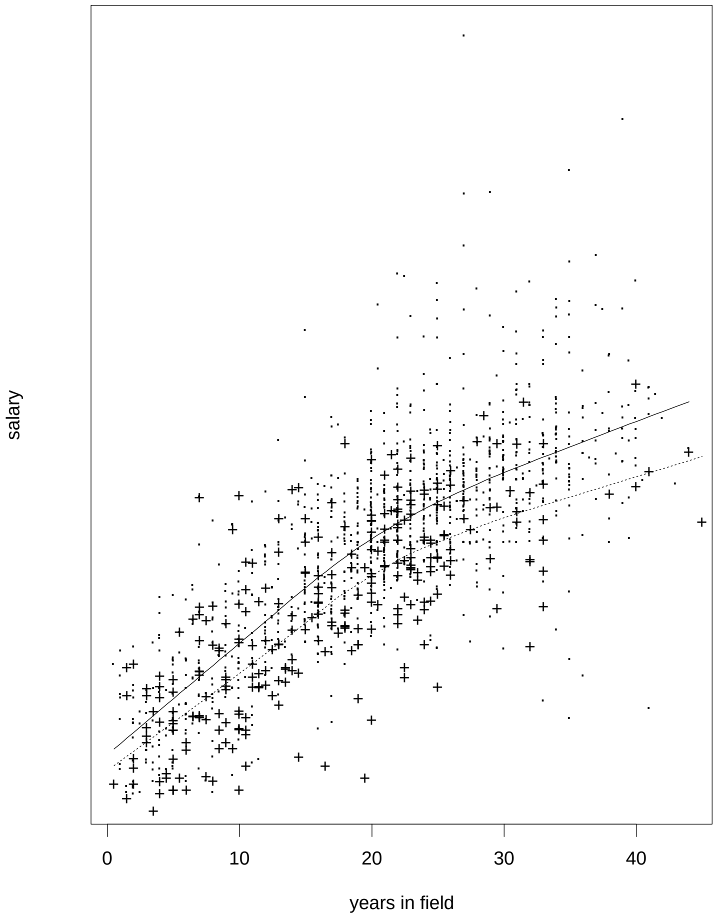

FIGURE 1. Salary plotted against years of experience, by gender. (+: female, .: male) The two curves are locally determined average salaries as a function of years, and are not entirely consistent with the linear model suggested in the text.

inference for θ, where ‘good’ is an operational concept that needs to be decided on first. When θ takes values on the real line, or even in $\mathbb{R}^2$, it is possible to plot $L(\theta)$ for the observed data $\mathbf{y}$. Figure 2 shows such a plot for Efron’s ulcer data example, and Figure 3 shows a more informative ‘profile’ curve for the parameter θ_1 (incorporating an adjustment for the nuisance parameter described at (10)) that measures the improvement of the new treatment. In fact what is plotted is the log of the (adjusted) profile likelihood $\ell(\theta_1) = \log L(\theta_1)$, which is more convenient to analyse theoretically.

There are, broadly speaking, two approaches to inference for θ_1 from Figure 3: the frequentist and the Bayesian. The frequentist approach considers the probability distribution of $\ell(\theta) = \ell(\theta; \mathbf{y}) = \log L(\theta)$, and of summary quantities computed from $\ell(\theta)$,

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

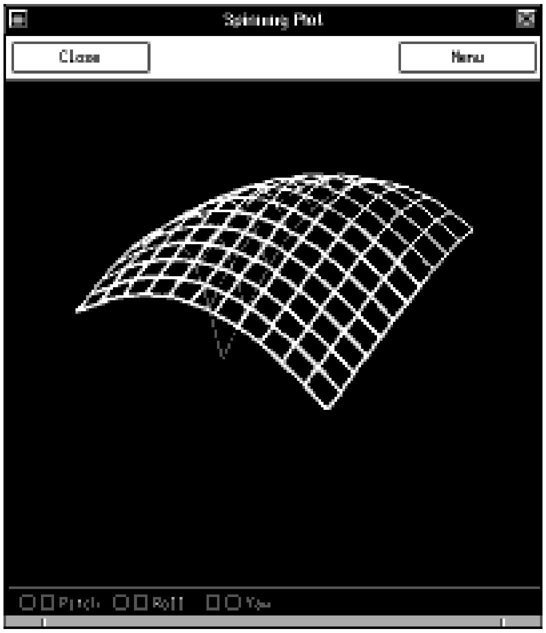

FIGURE 2. A plot of the likelihood function vs. θ_1 and θ_2 for the ulcer example.

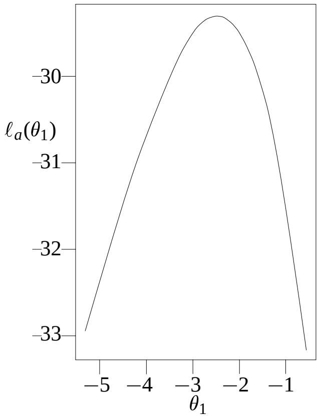

FIGURE 3. A plot of the profile log-likelihood function against θ_1.

under the model $f(\mathbf{y};\theta)$. For example, defining

$$\hat\theta = \arg\sup_\theta \ell(\theta;\mathbf{y}), \quad \hat\sigma^2 = \{-\ell''(\hat\theta)\}^{-1}$$

it can be shown as an application of the central limit theorem that

$$(\hat\theta - \theta)/\hat\sigma \xrightarrow{d} N(0,1) \tag{1}$$

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

i.e. converges in distribution to a standard normal random variable as $n \longrightarrow \infty$. This means, for example, that $\Pr\{|(\hat\theta - \theta|/\hat\sigma \leq 1.96\} \doteq 0.95$. In our example $\hat\theta = -2.38$ and $\hat\sigma = 0.75$, so the approximate 95% confidence limits for θ (*i.e.* the limits implied by (1)) are (-3.85, -0.91), corresponding to an interval for the odds ratio of (0.02, 0.40).

An asymptotically equivalent result is that

$$\text{sign}(\hat\theta - \theta)[2\{\ell(\hat\theta) - \ell(\theta)\}]^{1/2} \xrightarrow{d} N(0,1). \tag{2}$$

For the ulcer example, an approximate 95% confidence interval based on this result is (-4.20, -1.22). The two statistics are illustrated graphically in Figure 4. Note that the confidence interval based on (2) can capture the asymmetry of the log-likelihood function, which suggests that the approximation suggested by (2) is better than is that suggested by (1), which is indeed the case.

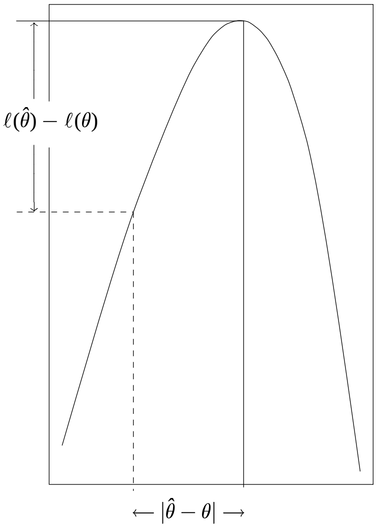

FIGURE 4. Two summary statistics derived from the log-likelihood function.

The Bayesian approach to inference treats θ as a random variable with a density $\pi(\theta)$ called a "prior" density, as it is assumed to be available before the random variable y is observed. An application of Bayes theorem[2] then gives

$$\begin{aligned} \pi(\theta|\mathbf{y}) &\propto L(\theta;\mathbf{y})\pi(\theta) \\ &= \frac{L(\theta;\mathbf{y})\pi(\theta)}{\int L(\theta;\mathbf{y})\pi(\theta)\,d\theta} \end{aligned} \tag{3}$$

and all inference statements are constructed from $\pi(\theta|\mathbf{y})$, which is called the posterior density for θ. Inference statements of the type given third in Section 1 can then be constructed from the posterior density. For example, a maximum posterior density interval can be identified that satisfies $\int_{\theta_{LB}}^{\theta_{UB}} \pi(\theta|\mathbf{y})\,d\theta = 0.90$.

[2] $\Pr(B|A) = \Pr(A|B)\Pr(B)/\Pr(A)$

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

Which type of inference is better? Statisticians are still not agreed, and the debate between the two approaches has enlivened our subject for more than 50 years[3]. Frequentists take the view that in most applications, particularly in science, θ is a fixed but unknown constant, and not a random variable. Even if it is agreed to model θ as a random variable, there is a substantial question in how to choose a prior density for θ that in my view is not yet solved. The problem is especially acute when θ is a high-dimensional parameter.

In the ulcer example, θ_2 measures the success rate of the standard treatment, and there might be available sufficient past experience to provide a reasonable assessment of its distribution, before the experiment in question was carried out. It seems less likely that it would be feasible to construct a prior distribution for the improvement represented by the new surgery[4]. Considerable effort has been expended on constructing priors that convey 'ignorance' or 'lack of information' about a parameter. We might for example assume that p_1, the probability of success on the new treatment, followed a uniform distribution on (0,1), which would seem to favour no particular value of p_1 over any other. This is a highly non-uniform distribution for $\log\{p_1/(1-p_1)\}$, so any theory of non-informative priors has to consider issues of parametrization: in particular whether or not it is a good idea for a prior to be parametrization invariant, and if so how such priors might be constructed. Sir Harold Jeffreys was a strong proponent of Bayesian inference with non-informative priors, and his book [15] states this position quite clearly[5].

Fisher was adamant in his rejection of the Bayesian method of inference, writing, for example in 1930:

> I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation.

A discussion of the published exchanges of Fisher and Jeffreys on the topic of Bayesian vs. non-Bayesian inference is given in Lane [16].

The Bayesian approach provides an axiomatic foundation for statistical inference, and non-Bayesian attempts to put the theory of statistics on a coherent mathematical foundation have not been very successful. A key foundational result is deFinetti's theorem, which in statistical terms states that any exchangeable distribution is obtained from a mixture, over a probability distribution on the parameter, of a family of parametric models. A helpful introduction to the Bayesian interpretation of deFinetti's theorem is Smith [23]. The so-called subjective Bayesian viewpoint of the foundations of inference has perhaps its clearest statement in Savage [22]. Subjective Bayesians hold that correct inference proceeds from a prior distribution which is obtained as the investigator's opinion

[3] An excellent introductory account is given in Efron [4].

[4] Efron [8] discusses combining the information from the other 40 published studies to estimate the prior for θ_1, θ_2, using a methodology called empirical Bayes.

[5] In the case that θ is a scalar parameter, and there are no nuisance parameters in the problem, Jeffreys proposed a parametrization invariant prior that is generally accepted as providing a suitable non-informative prior. It is $\pi(\theta) \propto i^{1/2}(\theta)$, where $i(\theta) = \int -\ell''(\theta; y) f(y; \theta)\, dy$ is called the expected Fisher information.

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

about the plausible values of θ, which can in theory be obtained by a series of betting arguments. The fact that two different investigators will have two different inferences about θ is not considered a drawback of the approach, but a natural consequence of the uncertainty in the problem.

Lindley has been a tireless proponent of (subjective) Bayesian inference throughout his career. In Lindley [17], from which I took the title of this paper, he wrote

> . . . the only good statistics is Bayesian statistics. Bayesian statistics is not just another technique to be added to our repertoire alongside, for example, multivariate analysis: it is the only method that can produce sound inferences and decisions in multivariate, or any other branch of, statistics. It is not just another chapter to add to that elementary text you are writing: it is that text.

Both Fisher's and Lindley's remarks are extremely mild expressions of their opinions, relative to some of the surprisingly acrimonious debate that has been published on this debate.

In 1986, Efron stated [4] that the "high ground of objectivity has been seized by the frequentists": he also commented vis-a-vis objectivity:

> "Scientific objectivity" is more than a catch-phrase. Strict objectivity is one of the crucial factors separating scientific thinking from wishful thinking. Complete objectivity about one's own work is a little much to expect from a human being, even a scientist, but it is not too much to expect from one's colleagues.

Lindley's discussion of Efron [7] concluded with the words "Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly and was honest enough to admit that he might have been wrong."

In recent years, I think the high ground of practicality is being seized by the Bayesians[6]. As we will see in the next section, the Bayesian approach to inference can apply with equal ease to problems with or without nuisance parameters. Frequentist theory has more difficulty in deciding how to eliminate nuisance parameters in constructing an inference statement about a parameter of interest. A fairly usual Bayesian approach in practice is to use Lebesgue measure priors on at least a large subset of nuisance parameters and to assume (or sometimes check) that such a prior does not have a large effect on the inference.

4. **Asymptotic theory.** Recent developments in asymptotic theory have provided some hints on where the frequentist and Bayesian approaches diverge quantitatively, if not philosophically. Recall from Section 3 that a (relatively crude) approximation to the distribution of $\hat{\theta}$, the maximum likelihood estimator, is that of a normal distribution with mean θ and variance $\hat{\sigma}^2$, where $\hat{\sigma}^2 = j^{-1}(\hat{\theta}) = \{-\ell''(\hat{\theta})\}^{-1}$. (In regular problems the log-likelihood is concave, and has negative second derivative at the maximum, as in Figures 2 and 3.) This gives the inference statement "$\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}$ is an approximate $1-\alpha$

[6] This is partly due to advances in computing power, a development emphasized in Efron's [7] predictions for the next century.

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

confidence interval for θ", where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. We can also write

$$f(\hat{\theta};\theta) \doteq \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{1}{2\hat{\sigma}^2}(\hat{\theta}-\theta)^2}. \tag{4}$$

An approximation to the Bayesian posterior given by (3) can be obtained by expanding $\ell(\theta)$ in a Taylor series about $\hat{\theta}$:

$$L(\theta) = \exp\{\ell(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}) + \cdots\}, \tag{5}$$

leading to the approximation

$$\pi(\theta|y) \doteq \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{1}{2\hat{\sigma}^2}(\theta-\hat{\theta})^2} \tag{6}$$

which has the same form as the approximation derived from the frequentist argument, and does not depend on the prior. This gives the inference statement "a $1-\alpha$ posterior probability interval for θ is $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}$", and this is the same interval as the confidence interval described above.

Although this asymptotic coincidence may seem reassuring, it is a fairly crude result, both in order of accuracy (the relative errors in (4) and (6) are $O(n^{-1/2})$) and inferentially. As we have seen, neither interval makes any provision for asymmetry in the log likelihood function (or the posterior density), which is the most unsatisfactory aspect from a practical point of view. In addition, (6) in not depending on the prior would not be considered by Bayesians to be a very satisfying use of the Bayesian methodology.

If we go to the next order of asymptotic theory, the difference between the two approaches starts to emerge. A second order approximation to the posterior density is available by using the Taylor series expansion in (3) to approximate the denominator, but not the numerator. (Approximating the integral using (5) is an application of Laplace's method.) The result is

$$\pi(\theta|y) \doteq \frac{1}{\sqrt{2\pi}} |j(\hat{\theta})|^{1/2} \frac{L(\theta)}{L(\hat{\theta})} \frac{\pi(\theta)}{\pi(\hat{\theta})}. \tag{7}$$

A very similar result is available in exponential family distributions by using the saddlepoint approximation to the density of a sample mean. The result is

$$f(\hat{\theta};\theta) \doteq \frac{1}{\sqrt{2\pi}} |j(\hat{\theta})|^{1/2} \frac{L(\theta)}{L(\hat{\theta})}. \tag{8}$$

The same approximations hold for vector parameters $(\theta_1, \ldots \theta_k)$, with the factor $\sqrt{2\pi}$ replaced by $\sqrt{2\pi}^k$. Both approximations (7) and (8) have relative error $O(n^{-1})$. The effect of the prior is clearly isolated in the final factor in (7). If the prior for θ is flat near the maximum value $\hat{\theta}$, then the two formulae are the same, although their interpretations are quite different.

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

Asymptotic arguments are even more useful in the case of several nuisance parameters. We write the parameter vector as (θ, λ), where now θ is a vector of parameters of particular interest (in many applications scalar), and λ is a (possibly large) vector of nuisance parameters. As mentioned above, the Bayesian method of eliminating nuisance parameters is quite straightforward: they are integrated out. From the joint posterior $\pi(\theta, \lambda|y) = L(\theta, \lambda)\pi(\theta, \lambda) / \int L(\theta, \lambda)\pi(\theta, \lambda)\, d\theta$, we compute

$$\pi(\theta|y) = \int \pi(\theta, \lambda|y)\, d\lambda,$$

which is the ratio of two, possibly high-dimensional, integrals. Another application of Laplace approximation, this time to the numerator integral over λ, leads to the approximation

$$\begin{aligned}\pi(\theta|y) &= c\, L(\theta, \hat\lambda_\theta)\left|-\frac{\partial^2 \ell(\theta, \hat\lambda_\theta)}{\partial\lambda\partial\lambda'}\right|^{-1/2}\pi(\theta, \hat\lambda_\theta) \\ &= c\ \exp\left\{\ell(\theta, \hat\lambda_\theta) - \frac{1}{2}\log\left|\frac{\partial^2 \ell(\theta, \hat\lambda_\theta)}{\partial\lambda\partial\lambda'}\right|\right\}\pi(\theta, \hat\lambda_\theta).\end{aligned} \tag{9}$$

As it turns out, frequentist theory, using entirely different arguments related to eliminating nuisance parameters in exponential families by conditioning, leads to an adjusted log-likelihood for the parameter of interest given by

$$\ell_a(\theta) = \ell(\theta, \hat\lambda_\theta) - \frac{1}{2}\log\left|-\frac{\partial^2 \ell(\theta, \hat\lambda_\theta)}{\partial\lambda\partial\lambda'}\right| \tag{10}$$

which is exactly the 'likelihood' function appearing in (9) for the marginal posterior density of θ. Again the effect of the prior is isolated from the effect of the parameter, and it is in principle possible to compare inference statements from results (9) and (10) for various types of priors, particularly flat priors along the curve $(\theta, \hat\lambda_\theta)$.

The approximation (8) for the density of the maximum likelihood estimator holds in models more general than the exponential family models, although some additional concepts need to be added. It is known in the statistical literature as "Barndorff-Nielsen's formula" or the "p^* approximation", and is reviewed in Reid [20][7].

Expression (10) for the adjusted log-likelihood also can be used in more generally, as discussed in Cox and Reid [2]. The correspondence between the Bayesian and frequentist results outlined here are discussed in more detail in Reid [21].

5. **Further common ground.** The similarities of the expressions in the previous section are intriguing, but are difficult to translate into a statement about quantitative or analytical similarities between the inferences[8]. A different approach to comparing Bayesian and frequentist inferences from an asymptotic point of view is the theory of

[7] In fact the approximation can be improved to relative error $O(n^{-3/2})$ by the simple technique of renormalizing the approximation so that it integrates to 1 (with respect to $\hat\theta$).

[8] Some progress is made in DiCiccio and Martin [3].

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

'matching priors'. Matching priors are priors for which the resulting posterior probability intervals have the correct sampling behaviour under the model $f(y;\theta)$, at least to $O(n^{-1})$. In the case that θ is a scalar and there are no nuisance parameters, Welch and Peers [26] showed that the matching prior is given by Jeffreys' prior described in Section 4. Matching priors could be argued to be non-informative, in the sense that a frequentist would presumably be satisfied with the resulting inference statement, even though it was obtained by Bayesian arguments. On the other hand, many Bayesians do not believe that this is a compelling argument for using matching priors, and argue that the prior should reflect all available knowledge about the parameter. Most Bayesians do agree though that matching priors could be used as a reference point for discussion.

In the case that we have nuisance parameters as well, development of matching priors is considered by Peers [19], Stein [24] and Tibshirani [25]. Unfortunately in general matching priors are only determined up to arbitrary functions of the nuisance parameter, and thus do not provide a guideline for constructing an objective prior for the full parameter (θ, λ). Recent work has concentrated on adding some reasonable conditions to make the parameter well-defined, but these conditions are more than 'matching', of course.

Another point of contact developed from asymptotics is the use of a shrinking argument sketched in Stein [24] and in more detail in Bickel and Ghosh [1] and Ghosh [14]. Asymptotic results are established under the joint probability distribution of y and θ; for fixed y this gives the Bayesian asymptotic result, and by letting the prior shrink to a point mass the frequentist result is obtained from the same argument. This is used to establish an elegant result related to the distribution of (2) in Bickel and Ghosh [1].

6. **Conclusion.** The Bayesian/frequentist asymptotic connection is fairly recent, and although it might lead to a theory of non-informative priors, it might not: the jury is still out.

The arguments between Bayesians and frequentists have shifted ground somewhat, from philosophical issues to practical ones. In particular the recent development of techniques of integration in high-dimensional spaces and in particular Markov chain Monte Carlo methods, have given new impetus to the use of Bayesian methods in complex applied problems.

More importantly, statistical modelling and inference can be applied in either version in a very wide variety of applications, and discussions of theoretical statistics help us to find the common threads in a variety of techniques, and provide a framework for comparing them.

References

[1] P. J. Bickel and J. K. Ghosh, *A decomposition for the likelihood ratio statistic and Bartlett correction—a Bayesian argument,* Ann. Statist. **18**(1990), 1070–1090.

[2] D. R. Cox and N. Reid, *Parameter orthogonality and approximate conditional inference*, J. R. Statist. Soc. B **49**(1987), 1–39.

[3] T. J. DiCiccio and M. A. Martin, *Simple modifications for signed roots of likelihood ratio statistics*, J. R. Statist. Soc. B **55**(1993), 305–316.

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press

[4] B. Efron, *Controversies in the foundations of statistics*, Amer. Math. Monthly **85**(1978), 231–246.

[5] ______, *Why isn't everyone a Bayesian?* Amer. Statist. **40**(1986), 1–11.

[6] ______, *Bayes and likelihood calculations from confidence intervals*, Biometrika **80**(1993a), 3–26.

[7] ______, *Statistics in the 21st century*, Statistics and Computing **3**(1993b), 188–190.

[8] ______, *Empirical Bayes methods for combining likelihoods*, J. Am. Statist. Assoc. **91**(1996), 538–565.

[9] R. A. Fisher, *On the 'Probable Error' of a coefficient of correlation deduced from a small sample*, Metron **I**(1921), 3–32, Reprinted in Fisher (1950).

[10] ______, *On the mathematical foundations of theoretical statistics*, Phil. Trans. Roy. Soc. A **222**(1922), 309–368, Reprinted in Fisher (1950).

[11] ______, *Theory of statistical estimation*, Proc. Camb. Phil. Soc. **22**(1925), 700–725, Reprinted in Fisher (1950).

[12] ______, *Inverse probability*, Proc. Camb. Phil. Soc. **26**(1930), 528–535, Reprinted in Fisher (1950).

[13] ______, *Contributions to Mathematical Statistics*, J. Wiley & Sons, New York, 1950.

[14] J. K. Ghosh, *Higher order asymptotics*, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 4, Institute of Mathematical Statistics, Hayward, 1994.

[15] H. Jeffreys, *Scientific Inference*, 2nd edition 1957, Cambridge University Press, 1931.

[16] D. A. Lane, *Fisher, Jeffreys, and the nature of probability*. In: R. A. Fisher, an appreciation, (eds. D. V. Hinkley and S. E. Fienberg), Springer-Verlag, New York, 1980.

[17] D. V. Lindley, *The future of statistics—a Bayesian 21st century*. In: Proceedings of the Conference on Directions for Mathematical Statistics, (ed. S. G. Ghurye), Special Supplement to Adv. Appl. Probab., 1975.

[18] ______, *Contribution to the discussion of Efron (1986)*, Amer. Statist. **40**(1986), 6–7.

[19] H. W. Peers, *On confidence points and Bayesian probability points in the case of several parameters*, J. R. Statist. Soc. B **27**(1965), 16–27.

[20] N. Reid, *Saddlepoint methods and statistical inference*, Statist. Science **3**(1988), 213–238.

[21] ______, *Likelihood and Bayesian approximation methods*. In: Bayesian Statistics 351–368, (eds. J. O. Berger, J. M. Bernardo, D. V. Lindley and A. F. M. Smith), Oxford University Press, 1995.

[22] L. J. Savage, *The Foundations of Statistics*, J. Wiley & Sons, New York, 1954.

[23] A. F. M. Smith, *Present position and potential developments: some personal views. Bayesian statistics*, J. R. Statist. Soc. A (2)**147**(1984), 245–259.

[24] C. Stein, *On the coverage probability of confidence sets based on a prior distribution*. In: Sequential Methods in Statistics, Banach Center publications, 16, PWN-Polish Scientific Publishers, Warsaw, 1985.

[25] R. J. Tibshirani, *Non-informative priors for one parameter of many*, Biometrika **76**(1989), 604–608.

[26] B. L. Welch and H. W. Peers, *On formulae for confidence points based on integrals of weighted likelihoods*, J. R. Statist. Soc. B **25**(1963), 318–329.

Department of Statistics
University of Toronto
Toronto, ON
M5S 3G3
e-mail: reid@utstat.toronto.edu

https://doi.org/10.4153/CMB-1997-028-x Published online by Cambridge University Press