The Canadian Journal of Statistics
Vol. 50, No. 4, 2022, Pages 1104–1115
La revue canadienne de statistique

D. A. S. Fraser: From structural inference to asymptotics

Nancy REID*

Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Key words and phrases: Foundations; inference; likelihood; saddlepoint approximation; tangent exponential model.

MSC 2020: Primary 6202; secondary 62F99.

Abstract: Don Fraser was my collaborator and life partner, so I had a uniquely close view of his life in research. This note describes how his early work in the structure of models informed our work in asymptotic theory.

Résumé: Don Fraser était mon collaborateur et mon compagnon de vie; j'ai donc été un témoin privilégié de sa vie de chercheur. Cette note décrit comment ses premiers travaux sur la structure des modèles ont guidé notre travail sur la théorie asymptotique.

1. INTRODUCTION

This note describes the background to our work on approximate inference derived from higher-order asymptotic expansions, developed over a number of years and in collaboration with many students and colleagues. The building block for this is an approximating density that Don referred to as the *tangent exponential model*. He was confident in his view that this approach leads to an essentially unique prescription for inference, at least to the order of approximation in which terms of $O(n^{-3/2})$ are ignored. Whether this confidence will be borne out, I cannot predict. Advances in statistical theory need to be repeatedly tested against statistical practice, and the adoption of new techniques depends on many factors, including their ease of use and their perceived relevance for applied problems.

2. EXPONENTIAL MODELS

Suppose $y_1, \dots, y_n$ is a sample from an exponential family model, and the ith variable has density

$$f(y_i; \varphi) = \exp\left\{\varphi^{\mathrm{T}} s(y_i) - \kappa(\varphi)\right\} h(y_i), \quad y_i \in \mathbb{R}, \varphi \in \mathbb{R}^p; \tag{1}$$

φ is the canonical parameter and $s(y_i) = \{s_1(y_i), \dots, s_p(y_i)\}$ is the sufficient statistic for φ. The joint density of the sample has the same form, as does the marginal density of $s = \Sigma_i s(y_i)$:

$$f(s; \varphi) = \exp\left\{\varphi^{\mathrm{T}} s - n\kappa(\varphi)\right\} \tilde{h}(s), \quad s, \varphi \in \mathbb{R}^p. \tag{2}$$

With $\mathcal{A} = \{(y_1, \dots, y_n) \in \mathbb{R}^n : s(y) = s\}$, $\tilde{h}(s) = \int_{\mathcal{A}} h(y) dy$, although an explicit expression for $\tilde{h}(s)$ may not be available.

* *Corresponding author: nancym.reid@utoronto.ca*

Many familiar distributions have densities in the exponential family, including the normal, multivariate normal, binomial, multinomial, gamma, inverse Gaussian and Poisson. Often the more conventional parametrization is not the canonical parameterization; for example, the normal distribution is typically parameterized by (μ, σ^2), the expected value and variance of y_i, but the canonical parameter is $(\mu/\sigma^2, -1/2\sigma^2)$. The sufficient statistic for a sample from a $N(\mu, \sigma^2)$ density is $\left(\Sigma y_i, \Sigma y_i^2\right)$, which is a one-to-one function of $\{\bar{y}, \Sigma(y_i - \bar{y})^2\}$.

Exponential family models are particularly easy to work with; their log-likelihood functions are typically concave, and the maximum likelihood estimator $\hat{\varphi}$ is uniquely determined by the solution of the score equation $\partial \log f(y; \varphi)/\partial\varphi = 0$. The asymptotic variance of the maximum likelihood estimator is the inverse of the expected Fisher information,

$$i(\varphi) = \mathrm{var}_{\varphi}\{\partial \log f(y; \varphi)/\partial\varphi\}, \tag{3}$$

evaluated at $\hat{\varphi}$, and in exponential families $i(\varphi) = n\partial^2\kappa(\varphi)/\partial\varphi\partial\varphi^{\mathrm{T}}$ is also the observed Fisher information function. Another convenient property of exponential families, less-widely remarked upon, is that the canonical parameter $\varphi = \partial\ell(\varphi; s)/\partial s$ is obtained (up to affine transformation) by differentiation on the sample space: this will be used in Section 4 to define the parameter of the tangent exponential model.

Extensions to independent, but not identically distributed observations, such as arise in generalized linear models, are straightforward, as long as the Fisher information increases with the sample size. To keep notation simpler, I will restrict attention to the i.i.d. case.

The expectation parameter is $\eta(\varphi) = \mathbb{E}_{\varphi}(s) = \kappa'(\varphi)$, and its maximum likelihood estimate $\hat{\eta} = s = \kappa'(\hat{\varphi})$. This directly links the sample space to the parameter space (Efron, 1978). As there is a one-to-one relationship between the maximum likelihood estimator and the sufficient statistic, the density of the maximum likelihood estimator $f(\hat{\varphi}; \varphi)$ can be computed by transformation of Equation (2), and captures all the information about φ contained in the data. If φ is a scalar parameter, we could imagine finding two values of φ, say φ_L and φ_U, for which $F(\hat{\varphi}^{\mathrm{o}}; \varphi_L) = 1 - F(\hat{\varphi}^{\mathrm{o}}; \varphi_U) = 0.025$, where $\hat{\varphi}^{\mathrm{o}}$ is the observed value of the maximum likelihood estimate from a given sample. This defines a 95% confidence interval for φ and could be equally expressed in terms of the cumulative distribution function $F(s; \varphi)$ obtained by integrating Equation (2).

In the effort to use the convenient properties of an exponential family in more general models, the *curved* exponential family plays a central role. The density for a single observation is

$$f(y_i; \theta) = \exp\left\{\varphi^{\mathrm{T}}(\theta)s(y_i) - \kappa(\theta)\right\}h(y_i), \tag{4}$$

where $\theta \in \mathbb{R}^q$, say, with $q < p$. The density for a sample of size n again depends only on the data through s;

$$f(s; \theta) = \exp\left\{\varphi^{\mathrm{T}}(\theta)s - n\kappa(\theta)\right\}\tilde{h}(s), \tag{5}$$

as at Equation (2). The normal distribution under the constraint $\sigma^2 = \mu^2$ is an example of a curved exponential family.

An arbitrary continuous density with sufficiently smooth dependence on its parameter could be written, for example, as

$$f(y; \theta) = \exp\left\{\ell(\theta_0) + (\theta - \theta_0)\ell'(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2\ell''(\theta_0) + \cdots\right\}, \tag{6}$$

where $\ell(\theta) = \log f(y; \theta)$ is the log-likelihood function and θ_0 is some reference point. If θ is a vector of length $q > 1$, then ℓ' is a vector, ℓ'' is a matrix, and higher-order derivatives

are multidimensional arrays. If we truncate the series at, for example, the kth term, and renormalize, the resulting density has the form of a curved exponential family, with $\varphi(\vartheta) = (\vartheta, \vartheta^2, \ldots, \vartheta^k)$, writing ϑ for $\theta - \theta_0$. The sufficient statistic in this notional model is $s = \{\ell'(\theta_0), \ell''(\theta_0), \ldots, \ell^{(k)}(\theta_0)\}$.

In curved exponential families, information is lost in summarizing the data by $\hat{\theta}$, since the sufficient statistic is of dimension $p > q$. One way to recover this information is to consider instead the conditional density of $\hat{\theta}$ given an auxiliary statistic a for which the transformation from s to $(\hat{\theta}, a)$ is one-to-one. If the marginal density of a is free of θ, then no information is lost in restricting attention to the conditional distribution. Such a function a is said to be ancillary for θ. Fisher (1934) suggested in passing that such a statistic could be constructed from the higher-order derivatives of the log-likelihood function. Cox (1980) detailed the use of the expansion identified in Equation (6), stopping at $k = 2$, to construct a function of ℓ' and ℓ'' that was approximately ancillary in a specified sense. He then derived an approximation to the conditional distribution of the maximum likelihood estimator, conditional on this approximate ancillary statistic. Several other papers in the same issue of *Biometrika* tackled related problems (Barndorff-Nielsen, 1980; Durbin, 1980; Hinkley, 1980), and what emerged as a common thread was an approximation to the conditional distribution of the maximum likelihood estimator later known as the p^*-approximation. Reid (1988) discussed this p^*-approximation and its relation to saddlepoint approximation in exponential families, and Barndorff-Nielsen (1990) showed its connection to the Laplace approximation.

In about 1987, David Cox and I were at the blackboard in my office puzzling over how expansions like the one used in Equation (6) might provide insight into aspects of inference. It is natural to choose as the expansion point $\theta_0 = \hat{\theta}^{\mathrm{o}}$, the observed maximum likelihood estimate, in which case the terms in the exponent would be expected to decrease in powers of n. (A careful treatment of this was given in Skovgaard (1990); see also Skovgaard (1989).) Don wandered in to see what we were doing, seemed uneasy because the board was filled with algebra and no pictures, said little, but left for his regular swim looking thoughtful. A few weeks later, he was writing out long expansions himself, but with a key difference. To my surprise, he was using Taylor series expansions in both the parameter and the data. I think this may have seemed natural to him because of his work in structural inference, but it is most directly related to his 1964 paper on local location models (Fraser, 1964).

3. LOCATION MODELS

The simplest location model has a family of density functions $\{f(y - \theta); \theta \in \mathbb{R}, y \in \mathbb{R}\}$, with f known, and both θ and y taking values in $\mathbb{R}$. Exact inference for θ is straightforward: by pivoting on θ we can find, for a given value y^{o}, the location density in the family that has exactly $\alpha/2$ probability in each tail, and this defines an exact confidence interval for θ with confidence $1 - \alpha$. If the confidence intervals at different values of α are nested then we could summarize them via a *confidence distribution* for θ, as pointed out in Cox (1958) and discussed further in Efron (1993). The location model can be expressed as $y = \theta + e, e \sim f(e)$; once $y = y^{\mathrm{o}}$ has been observed, we can write $\theta = y^{\mathrm{o}} - e$. Fisher's *fiducial distribution* for θ is induced by the randomness in e. In this simple model, this has location form and is in fact the same as Cox's confidence distribution. Fraser (1966) generalized Fisher's fiducial argument to a class of transformation models generated by a group, and in this context defined what he called a *structural distribution* for θ. These so-called "distributions" do not obey the rules of probability calculus, so generalizing the arguments to models with vector parameters has proved to be elusive. It is a subject of ongoing discussion, for example, in the series of workshops on Bayes, Frequentist and Fiducial inference.

Don viewed this argument as developing a direct link between y and θ, or more generally between the sample space and the parameter space. For example in Fraser (1964), he wrote "Transformation-parameter models … provide a position relationship between variable and parameter and this seems to lead to stronger inference statements".

In i.i.d. sampling from the location model, the structural form of the model is now $y_i = \theta + e_i, i = 1, \dots, n$, with e_i independently distributed as $f(\cdot)$. Once the vector of observations has been observed, Fraser (1966) noted that the vector $a = (e_2 - e_1, \dots, e_n - e_1)$ has also been observed, since it is identically equal to the differences between the observations. On the principle that one should condition on what is known, he argued that all the information about θ is contained in the conditional distribution of y_1, say, given a. In more conventional terms, we say that a is ancillary for θ, i.e., its marginal distribution is free of θ, so the model can be expressed as

$$f(y;\theta) = f_1(y_1|a;\theta)f_2(a)$$

and the likelihood function depends on a single variable, here y_1. Again the structural or fiducial or confidence distribution for θ follows directly, as $f_1(y_1;\theta|a)$ is a location model on $\mathbb{R}$. The choice of (y_1, a) above is not unique; y_1 could be replaced by any single y_i, or the average $\bar{y}$, or the maximum likelihood estimator $\hat{\theta}$, with a corresponding change to a. In the last form, we would write

$$f(\hat{\theta}, a;\theta) = f(\hat{\theta}|a;\theta)f(a), \tag{7}$$

where now $a = (y_1 - \hat{\theta}, \dots, y_n - \hat{\theta})$, a vector in $\mathbb{R}^{n-1}$, as it is constrained by the maximum likelihood equation $\partial\ell(\hat{\theta};a)/\partial\theta = 0$. Fisher (1934) showed that the conditional density in Equation (7) is exactly

$$f(\hat{\theta}|a;\theta) = \frac{L(\theta;y)}{\int L(\theta;y)d\theta} = \frac{L(\theta;\hat{\theta},a)}{\int L(\theta;\hat{\theta},a)d\theta},$$

where $L(\theta;y) = L(\theta;\hat{\theta},a)$ is the likelihood function for θ based on the sample $y = (y_1, \dots, y_n)$.

In the location model, the "position relationship" of Fraser (1964) can be expressed by considering a fixed quantile of the model for a single observation: $F(y - \theta) = q$, say, where $F(\cdot)$ is the cumulative distribution function for the model. As $q = F(y + a - \theta - a) = F(y' - \theta')$, change in the parameter is directly linked to change in the observation. In Fraser (2004), this is described equivalently as $dy/d\theta = 1$.

In a general model, we can impose this linking by fixing a quantile and taking the total derivative of the cumulative distribution function at that quantile:

$$0 = dF(y;\theta) = \frac{\partial F(y;\theta)}{\partial\theta}d\theta + \frac{\partial F(y;\theta)}{\partial y}dy,$$

leading to

$$\frac{dy}{d\theta} = -\frac{\partial F(y;\theta)/\partial\theta}{f(y;\theta)}. \tag{8}$$

Fraser (1964) established Equation (8) by defining a new variable (which he called l):

$$x = \int^{y} -\frac{f(y;\theta_0)}{\partial F(y;\theta_0)/\partial\theta}dy,$$

where θ_0 is some fixed point in the parameter space. He then showed that the density, say $g(x;\theta)$, of the new variable is of location model form near θ_0; for example, it satisfies $\partial g(x;\theta_0)/\partial x = -\partial g(x;\theta_0)/\partial\theta$, but only at θ_0.

A sample $(y_1,\dots,y_n)$ from f then leads to a transformed sample $(x_1,\dots,x_n)$, with the density of each observation by construction having location-model dependence on θ to first derivative at θ_0. Fraser (1964, Section 3) also showed that any ancillary statistic for a sample from a location model, such as $a = (x_2 - x_1, \dots x_n - x_1)$ or $(x_1 - \bar{x}, \dots, x_n - \bar{x})$, is an ancillary statistic for the original density locally at θ_0. The distribution of a is not completely free of θ, but its derivative with respect to θ is 0, at θ_0. In Sections 4 and 5, he describes how to use the conditional distribution, given the ancillary, for inference about θ, and calls the variable for that conditional distribution *conditionally sufficient* for θ.

In Fraser & Reid (1995, Section 5), this construction was used to approximate the distribution of this conditionally sufficient statistic, via Taylor series approximations in a neighbourhood of $y^o, \hat{\theta}^o = \hat{\theta}(y^o)$, where y^o is the observed value of the sample and $\hat{\theta}^o$ is the corresponding maximum likelihood estimate. By assuming that we are working in a $\sqrt{n}$-neighbourhood of the maximum likelihood estimate, we can control the order in n of the successive terms in the series expansion. The notation v for the local ancillary statistic evaluated at the observed sample point y^o was introduced here:

$$v = v\big(\hat{\theta}^o; y^o\big) = \left(-\frac{\partial F\big(y_1^o;\hat{\theta}^o\big)/\partial\theta}{\partial F\big(y_1^o;\hat{\theta}^o\big)/\partial y_1}, \dots, -\frac{\partial F\big(y_n^o;\hat{\theta}^o\big)/\partial\theta}{\partial F\big(y_n^o;\hat{\theta}^o\big)/\partial y_n} \right)^{\mathrm{T}}. \tag{9}$$

If θ is a (column) vector of length p, then Equation (8) generalizes to

$$\frac{dy}{d\theta^{\mathrm{T}}} = -\frac{\partial F(y;\theta)/\partial\theta^{\mathrm{T}}}{f(y;\theta)} = V(\theta), \tag{10}$$

where $V(\theta;y)$ is a $1\times p$ vector reflecting perturbations in θ linked to local perturbations in y. Fraser (2004) calls $V(\theta;y^o)$ the sensitivity of y relative to θ, and it was used in Fraser et al. (2010) to define a data-dependent prior.

With a sample of size n, the same construction gives an $n\times p$ matrix V, and in the asymptotic theory of Section 4, this matrix is evaluated at $\hat{\theta}^o, y^o$, as at Equation (9); the jth column v_j corresponding to θ_j.

Having established an approximate conditioning, we have a factorization analogous to Equation (7), where the dependence on θ only appears in the density of the conditionally sufficient statistic, which is a density on $\mathbb{R}^p$. This result provides a basis for the tangent exponential model, first introduced in Fraser (1990).

4. TANGENT EXPONENTIAL MODEL

The tangent exponential model is a density approximation to a general statistical model $\{f(y;\theta), \theta\in\mathbb{R}^p\}$, where for simplicity here we write y for $(y_1,\dots,y_n)$. It combines the local ancillary construction of the previous section with the observation in Section 2 that the canonical parameter of an exponential model can be obtained by differentiating the log-likelihood function with respect to the sufficient statistic. The density of the tangent exponential model is

$$\begin{aligned} f_{\mathrm{TEM}}(s|a;\theta) &= \exp\left[s^{\mathrm{T}}\varphi(\theta) + \ell\{\theta(\varphi);y^o)\}\right] h(s) \\ &\doteq c|j(\hat{\varphi})|^{-\frac{1}{2}} \exp\left[s^{\mathrm{T}}\{\varphi(\theta) - \varphi(\hat{\theta}^o)\} + \ell(\theta;y^o) - \ell(\hat{\theta}^o;y^o)\right], \end{aligned} \tag{11}$$

where $\ell(\theta; y^o) = \ell\{\theta(\varphi); y^o\}$ and $j(\varphi) = -\partial^2 \ell(\theta; y^o)/\partial\varphi\partial\varphi^T$ are the log-likelihood function and the observed Fisher information function computed in the φ parametrization, defined below. The second expression in Equation (11) is the saddlepoint approximation to the first and has relative error $O(n^{-3/2})$ when the original model is continuous and the information in the sample is $O(n)$.

The canonical parameter $\varphi(\theta)$ is constructed using the ancillary direction vectors V:

$$\varphi(\theta; y^o) = \ell_{;V}(\theta; y^o) := \frac{d}{dt}\ell(\theta; y^o + Vt)\Big|_{t=0} = V^T \frac{\partial \ell(\theta; y^o)}{\partial y}\Big|_{y=y^o}, \tag{12}$$

where V is defined at Equation (10), but now evaluated at $\left(\hat{\theta}^o, y^o\right)$. If the observations $y_1, \ldots, y_n$ are independent, then this is

$$\varphi^T(\theta) = \sum_{i=1}^{n} V_i^T \frac{\partial \ell(\theta; y^o)}{\partial y_i};$$

i.e., $\varphi(\theta)$ is a linear combination of sample space derivatives of $\ell(\theta; y)$. In exponential families, the canonical parameter is only defined up to linear transformations (not depending on θ), so there are other ways to calculate φ. One alternative is described in Davison & Reid (2022): they assume there is a one-to-one transformation from y to (s, a), where $s \in \mathbb{R}^p$ is sufficient for θ and a is approximately ancillary, and show that the rows of V are equivalent to $\partial y/\partial s$, with $a = a^o$ fixed, and are there called "sufficient directions".

The tangent exponential model is only useful if valid inference can be obtained from this approximation to the original model. Establishing that this is the case involves a combination of both numerical examples and theoretical results. Although the theory was set out in Fraser & Reid (1993, 1995), the explanations were relatively complicated, and a simpler picture emerged slowly over a series of related papers. Throughout this development, the empirical observation that the resulting inferential approximations seemed to be extremely accurate in example after example provided inspiration for clarifying the theory.

An important first step was to study in detail the tangent exponential model for s and $\theta \in \mathbb{R}$. This was developed in Fraser & Reid (1993), but the most complete treatment is Andrews, Fraser & Wong (2005), where fourth-order Taylor series expansions of the density and distribution function are given explicitly, with an assumption that successive terms in the Taylor series are decreasing in powers of n, and ignoring terms of $O(n^{-3/2})$. The paper shows (see, e.g., Eq. 2.6) that such an expansion has a single term of $O(1/n)$ that captures the nonexponentiality of the model, and further that the distribution function evaluated at s^o, the observed value of s, does not depend on this single term. Thus, for the calculation of a P-value, i.e., the cumulative distribution function at s^o, the nonexponential term is not needed. The tangent exponential model results from setting this term to 0 and using the rest of the expansion as the approximation to the original model. This paper and several earlier ones on which it builds use y for s, and at first read seem to assume that we only observe a single observation, which seems quite specialized. Don wrote "assuming that the asymptotic properties come from some antecedent sample size n", which was shorthand for the assumption that the preliminary work of reducing the sample $y_1, \ldots, y_n$ to a one-dimensional sufficient statistic, as in Section 2, or a one-dimensional conditionally sufficient statistic, as in Section 3, has already been done.

The resulting approximation to the P-value can be expressed using a formula due to Barndorff-Nielsen (1986) called the r^*-approximation, which results from integrating the saddlepoint approximation found in Equation (11).

For more usual and practical settings where $y \in \mathbb{R}^n$, a technical argument is needed to verify that the construction of Section 2 gives a valid model on $\mathbb{R}^p$ for inference about θ. The local

ancillary directions $V_1, \ldots, V_n$ define tangent vectors at $\left(y^{\mathrm{o}}, \hat{\theta}^{\mathrm{o}}\right)$, but it needs to be checked that they are consistent with an approximately ancillary statistic for θ. Fraser & Reid (1995, 2001) showed that the tangent vectors could be modified to be ancillary to a higher order than first derivative, but rigorous verification that this defines an ancillary surface for the original model to the needed order of approximation appeared later (Fraser, Fraser & Staicu, 2010).

To compute P-values, we need to work on $\mathbb{R}^1$, so when $p > 1$ some further reduction is needed. This is achieved by assuming that $\theta = (\psi, \lambda)$, with $\lambda \in \mathbb{R}^{p-1}$, where ψ is the scalar parameter of interest and λ is a nuisance parameter. A one-dimensional model is obtained by a second application of the tangent exponential model (Fraser & Reid, 1995, Section 6). We tried to clarify this somewhat elusive argument in Reid (2003, Section 3.3) and Reid & Fraser (2010, Appendix). Don later simplified the construction using a geometric argument (of course) in Fraser (2017, Section 4).

The notation is unfamiliar, and the verification of the analysis is difficult, but the resulting approximation is surprisingly easy to implement, as it depends entirely on the observed log-likelihood function $\ell(\theta; y^{\mathrm{o}})$ and the constructed canonical parameter $\varphi(\theta; y^{\mathrm{o}}) = \ell_{;V}(\theta; y^{\mathrm{o}})$, i.e., on the log-likelihood function and its first derivative in the sample space at y^{o}. The resulting approximation to the P-value for testing ψ is

$$p(\psi) \doteq \Phi\left(r_\psi^*\right) = \Phi\left\{ r_\psi + \frac{1}{r_\psi} \log\left(\frac{Q_\psi}{r_\psi}\right)\right\}, \tag{13}$$

where

$$r_\psi = \mathrm{sign}(\hat{\psi} - \psi)\left[2\left\{\ell(\hat{\theta}) - \ell\left(\hat{\theta}_\psi\right)\right\}\right]^{1/2}, \tag{14}$$

$$Q_\psi = \frac{\left|\varphi(\hat{\theta}) - \varphi\left(\hat{\theta}_\psi\right) \quad \partial\varphi\left(\hat{\theta}_\psi\right)/\partial\lambda^{\mathrm{T}}\right|}{\left|\partial\varphi(\hat{\theta})/\partial\theta^{\mathrm{T}}\right|} \left\{ \frac{|j(\hat{\theta})|}{\left|j_{\lambda\lambda}\left(\hat{\theta}_\psi\right)\right|} \right\}^{1/2}. \tag{15}$$

In Equations (14) and (15), $\hat{\theta}$ is the maximum likelihood estimate, $\hat{\theta}_\psi = \left(\psi, \hat{\lambda}_\psi\right)$ is the constrained maximum likelihood estimate, $j(\theta) = -\partial\ell^2(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$ is the observed Fisher information function, and $j_{\lambda\lambda}(\theta)$ is the $(p-1)\times(p-1)$ block of this matrix corresponding to the nuisance parameters λ. In Equation (15), the numerator is the determinant of a $p \times p$ matrix; the first entry is a column vector of length p, and the partial derivative is a $p \times (p-1)$ matrix. Many of the components in Equations (14) and (15) are obtained through usual maximum likelihood fitting of constrained and unconstrained models. The extra step needed is a function for evaluating $\varphi(\theta)$.

If the original model is a p-dimensional exponential family already, and the parameter of interest is a component of the canonical parameter, the inference from (13) is the same as that based on the (saddlepoint approximation to the) conditional distribution of the corresponding component of the sufficient statistic, given the remaining components, and Equation (15) is the Wald statistic based on the profile log-likelihood function, multiplied by a correction factor $\{|j_{\lambda\lambda}(\hat{\theta})|/|j_{\lambda\lambda}(\hat{\theta}_\psi)|\}^{1/2}$. A similar simplification occurs for linear regression models, where Q is the standardized score function times a similar correction factor (Pierce & Peters, 1992; Brazzale, Davison & Reid, 2007, Chs. 2,8).

In Fraser (1991), which was based on Don's Fisher Lecture to the Joint Statistical Meetings in 1990, he focused on the use of the function $p(\psi)$ in Equation (13) for all values of ψ; there he called it a significance function, but it is essentially equivalent to a confidence distribution

 DOI: 10.1002/cjs.11720

function as defined in Cox (1958). This view that the significance function, or *P*-value function, provides a full summary of the information in the data about the parameter was also emphasized in Fraser (2017).

Brazzale, Davison & Reid (2007) apply Equation (13) to several practical settings, and many other examples are discussed in Davison & Brazzale (2008). Regression models are illustrated in Fraser, Wong & Wu (1999), Fraser, Rekkas & Wong (2005), and Fraser, Wong & Sun (2009); the latter paper considers the Box-Cox transformed-regression model as one of the examples. Don's PhD student Augustine Wong, with his collaborators and students, has published a wealth of examples; see, for example, Wong & Jiang (2019) and Qi, Rekkas & Wong (2018). Belzile & Davison (2022) apply the tangent exponential model to inference about extreme values and threshold exceedances.

It is probably fair to say, however, that Don was more interested in the theoretical aspects. He published several papers on the technical aspects connected with the tangent model approximation and the local ancillarity construction; for example, Fraser & Rousseau (2008), Fraser & Staicu (2010), Fraser, Fraser & Staicu (2010), Fraser, Fraser & Fraser (2010), and Reid & Fraser (2010).

The tangent exponential model, and in particular the sensitivity matrix $V(\theta)$, was used to examine the higher-order asymptotic properties of Bayesian inference in Fraser, Fraser & Fraser (2010) and Fraser (2011). Most recently, it is the basis for an approach to inference about a vector parameter called directional inference that was originally developed in Fraser & Massam (1985). Directional inference is illustrated in Davison et al. (2014), Fraser, Reid & Sartori (2016), Wong & Zhang (2017), and McCormack et al. (2019).

A more detailed overview of the tangent exponential model and the approximation to *P*-values is provided in Davison & Reid (2022).

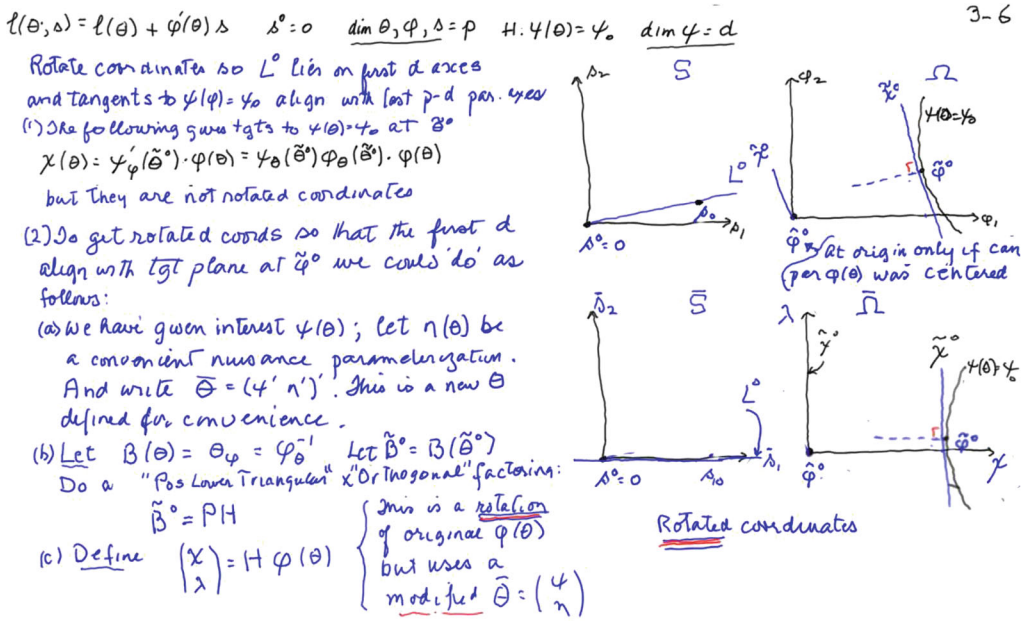

FIGURE 1: Don thought in pictures, although to me it always seemed to be the same picture! Slides for his talks were prepared using a sketching program and a large Wacom tablet, with many layers built up painstakingly, and each resulting slide then converted to pdf for presentation.

5. SOME PERSONAL REFLECTIONS

I have concentrated here on the work on higher-order asymptotics, as I was most directly involved in that, but Don had a long career and wrote on many different aspects of statistical inference. His early work on nonparametric methods, and in particular tolerance regions (Fraser, 1953; Fraser & Guttman, 1956), has echoes in modern work on conformal prediction. His development of structural inference grew out of efforts to put Fisher's fiducial inference on a more rigorous footing. His study of Cox (1958) and the role of the likelihood principle in inference led to several papers on foundations, e.g., Evans, Fraser & Monette (1985, 1986).

Don's approach to problems was geometric, he thought in pictures (Figure 1), and each time I asked him for an explanation of a difficult point, he started with a sketch, often the same sketch. Don also had a clear understanding of what mathematical arguments would be needed to turn the pictures into theorems, he just did not always see the point of filling all this in, since the solution was "obvious".

It was often remarked that he was very original, and his ability to "think sideways" about nearly everything was a continual surprise to me. He was extraordinarily willing to set aside everything that was known about a problem or set of problems and start with his own picture.

FIGURE 2: Don worked hard, but to him it was not work, it was fun.

 DOI: 10.1002/cjs.11720

At least while I knew him, he was not an avid reader of the statistical literature, although well-marked copies of old issues of the *Annals* and *JRSS B* in our basement confirm this was not always the case. One anecdote that illustrates his supreme confidence in swimming against the tide occurred during a reading week break (of course, we were both working) when he insisted for a few days that the central limit theorem was incorrect. Even though he accepted with a laugh that this was unlikely to be the case, he was quite willing to entertain the possibility until he had sorted things out to his own satisfaction.

While his single-mindedness was a great strength in tackling difficult problems, it was also sometimes a weakness, in that once he got going in a particular direction, it was hard for him to reset. Many of his published papers can seem repetitive, because he found one more piece of the puzzle to fit into the whole, whereas we might not have noticed that there were pieces missing. When he illustrated arguments with examples, they sometimes seemed to me to be overly simplistic: often this was deliberate as he felt they captured the essence of the point to be made. I think this made his published work seem less than practical for the wide range of applications that is today's world of statistics. But on the occasions when I watched him tackling a particular applied problem, for example, in a consulting setting, or helping me with something, he was very skilled at seeing what needed to be done. Students and colleagues who worked with him also commented on his uncanny sixth sense in describing exactly what needed to be computed and which examples could be used to highlight one property or another of a method.

In later years, Don became increasingly convinced that Bayesian inference was leading the field astray, and he became more strident about this in his talks and papers (e.g., Fraser, 2011; Fraser & Reid, 2015; Fraser et al., 2016). I think the basis of his concern was the widespread acceptance of posterior probabilities derived from priors of convenience; he felt the only possible justification of this could be that the resulting answer was close to one that would be obtained by non-Bayesian arguments. In that respect, his position seems to me to be close to that in Cox (2006, Ch. 5) and Reid & Cox (2015, Section 2.2); the latter paper suggests that procedures derived by Bayesian arguments be assessed by their properties in repeated sampling from the model.

Like many academics, Don viewed his work as fun (Figure 2) and was infectiously enthusiastic about whatever he was currently working on. His latest paper (Fraser & Bédard, 2022) involved many weekdays and weekends looking at the Lasso from his own unique viewpoint. It is a measure of his intellectual energy that he was eager to put his own stamp on such a well-studied methodology and willing to turn his research in a completely different direction at the end of his career.

A partnership that combines academic and personal life is both wonderful and challenging, as many academics know. It is enriching to share your work interests with the person closest to you, but it can sometimes be overwhelming, and needs some negotiation, whether spoken or unspoken. That the boundary between work and home was quite blurred was driven home to me in a conversation with our youngest daughter when she was about 13 or 14 years old. She asked (as we all did every year), "what do you think Daddy would like for his birthday?" and answered her own question with, "all the Bayesians to agree with him? I'll get on that right away".

ACKNOWLEDGEMENTS

I would like to thank Heather Battey, Mylène Bédard, Bruno Rémillard, Mary Thompson, and Yanbo Tang for helpful comments on an earlier draft. The work was supported in part by the Natural Sciences and Engineering Research Council.

REFERENCES

Andrews, D. A., Fraser, D. A. S., & Wong, A. C. M. (2005). Computation of distribution functions from likelihood information near observed data. *Journal of Statistical Planning and Inference*, 134, 180–193.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika*, 67, 293–310.
Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, 73, 307–322.
Barndorff-Nielsen, O. E. (1990). p^* and Laplace's method. *Brazilian Journal of Statistics*, 4, 89–103.
Belzile, L. R. & Davison, A. C. (2022). Improved inference on risk measures for univariate extremes. arXiv preprint, arXiv:2007.10780.
Brazzale, A. R., Davison, A. C., & Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*, Cambridge University Press, Cambridge.
Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29, 357–372.
Cox, D. R. (1980). Local ancillarity. *Biometrika*, 67, 279–286.
Cox, D. R. (2006). *Principles of Statistical Inference*, Cambridge University Press, Cambridge.
Davison, A. C. & Brazzale, A. R. (2008). Accurate parametric inference for small samples. *Statistical Science*, 23, 465–484.
Davison, A. C., Fraser, D. A. S., Reid, N., & Sartori, N. (2014). Accurate directional inference for vector parameters in linear exponential families. *Journal of the American Statistical Association*, 109, 302–314.
Davison, A. C. & Reid, N. (2022). The tangent exponential model. In Berger, J. O. et al. (Eds.), *Bayes, Frequentist, Fiducial*, Cambridge University Press, Cambridge. https://arxiv.org/abs/2106.10496
Durbin, J. (1980). Approximations for densities of sufficient statistics. *Biometrika*, 67, 311–333.
Efron, B. (1978). The geometry of exponential families. *Annals of Statistics*, 6, 362–376.
Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80, 3–26.
Evans, M. J., Fraser, D. A. S., & Monette, G. (1985). Mixtures, embedding, and ancillarity. *Canadian Journal of Statistics*, 13, 1–6.
Evans, M. J., Fraser, D. A. S., & Monette, G. (1986). On principles and arguments to likelihood. *Canadian Journal of Statistics*, 14, 181–199.
Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London A*, 144, 285–307.
Fraser, D. A. S. (1953). Nonparametric tolerance regions. *Annals of Mathematical Statistics*, 24, 44–55.
Fraser, D. A. S. (1964). Local conditional sufficiency. *Journal of the Royal Statistical Society B*, 26, 52–62.
Fraser, D. A. S. (1966). Structural probability and a generalization. *Biometrika*, 53, 1–9.
Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, 77, 65–76.
Fraser, D. A. S. (1991). Statistical inference: Likelihood to significance. *Journal of the American Statistical Association*, 86, 258–265.
Fraser, D. A. S. (2004). Ancillaries and conditional inference. *Statistical Science*, 19, 333–369.
Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? (with discussion). *Statistical Science*, 26, 299–316.
Fraser, D. A. S. (2017). The p-value function: The core concept of modern statistical inference. *Annual Review of Statistics and Its Application*, 4, 1–14.
Fraser, D. A. S. & Bédard, M. (2022). The linear Lasso: A location model approach. *Canadian Journal of Statistics*, 50, 437–453. https://doi.org/10.1002/cjs.11691
Fraser, D. A. S., Bédard, M., Wong, A. C. M., Lin, W., & Fraser, A. M. (2016). Bayes, reproducibility and the quest for truth. *Statistical Science*, 31, 578–590.
Fraser, D. A. S., Fraser, A. M., & Fraser, M. J. (2010). Parameter curvature revisited and the Bayesian frequentist divergence. *Journal of Statistical Research*, 44, 335–346.
Fraser, D. A. S., Fraser, A. M., & Staicu, A.-M. (2010). Second order ancillary: A differential view with continuity. *Bernoulli*, 16, 1208–1223.
Fraser, D. A. S. & Guttman, I. (1956). Tolerance regions. *Annals of Mathematical Statistics*, 27, 162–179.
Fraser, D. A. S. & Massam, H. (1985). Conical tests: Observed levels of significance and confidence regions. *Statistische Hefte*, 26, 1–17.
Fraser, D. A. S. & Reid, N. (1993). Simple asymptotic connections between densities and cumulant generating functions leading to accurate approximations for distribution functions. *Statistica Sinica*, 3, 67–82.
Fraser, D. A. S. & Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematicae*, 47, 33–53.

Fraser, D. A. S. & Reid, N. (2001). Ancillary information for statistical inference. In Ahmed, S. E. & Reid, N. (Eds.), *Empirical Bayes and Likelihood Inference*, Springer-Verlag, New York, 185–209.

Fraser, D. A. S. & Reid, N. (2015). Crisis in science? Or crisis in statistics: Mixed messages in statistics with impact on science. *Journal of Statistical Research*, 47, 107–115.

Fraser, D. A. S., Reid, N., Marras, E., & Yi, G. Y. (2010). Default priors for Bayesian and frequentist inference. *Journal of the Royal Statistical Society B*, 75, 631–654.

Fraser, D. A. S., Reid, N., & Sartori, N. (2016). Accurate directional inference for vector parameters, with curvature. *Biometrika*, 103, 625–635.

Fraser, D. A. S., Rekkas, M., & Wong, A. C. M. (2005). Highly accurate likelihood analysis for the seemingly unrelated regression problem. *Journal of Econometrics*, 127, 17–33.

Fraser, D. A. S. & Rousseau, J. (2008). Studentization and deriving accurate *p*-values. *Biometrika*, 95, 1–16.

Fraser, D. A. S. & Staicu, A.-M. (2010). The second order ancillary is rotation based. *Journal of Statistical Planning and Inference*, 140, 831–836.

Fraser, D. A. S., Wong, A. C. M., & Sun, Y. (2009). Three enigmatic examples and inference from likelihood. *Canadian Journal of Statistics*, 37, 161–181.

Fraser, D. A. S., Wong, A. C. M., & Wu, J. (1999). Regression analysis, nonlinear or non-normal: Simple and accurate *p*-values from likelihood analysis. *Journal American Statistical Association*, 94, 1286–1295.

Hinkley, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika*, 67, 287–292.

McCormack, A., Reid, N., Sartori, N., & Theivendran, S. (2019). A directional look at *F*-tests. *Canadian Journal of Statistics*, 47, 619–627.

Pierce, D. A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *Journal of the Royal Statistical Society B*, 54, 701–737.

Qi, J., Rekkas, M., & Wong, A. C. M. (2018). Highly accurate inference on the Sharpe ratio for autocorrelated return data. *Journal of Statistical and Econometric Methods*, 7, 1–2.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, 3, 213–238.

Reid, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics*, 31, 1695–1731.

Reid, N. & Cox, D. R. (2015). On some principles of statistical inference. *International Statistical Review*, 83, 293–308.

Reid, N. & Fraser, D. A. S. (2010). Mean loglikelihood and higher-order approximations. *Biometrika*, 97, 159–170.

Skovgaard, I. M. (1989). *Analytic Statistical Models*, Institute of Mathematical Statistics, Hayward.

Skovgaard, I. M. (1990). On the density of minimum contrast estimators. *Annals of Statistics*, 18, 779–789.

Wong, A. C. M. & Jiang, L. (2019). Improved small sample inference on the ratio of two coefficients of variation of two independent lognormal distributions. *Journal of Probability and Statistics*, 2019, 7173416. https://doi.org/10.1155/2019/7173416

Wong, A. C. M. & Zhang, S. (2017). A directional approach for testing homogeneity of inverse Gaussian scale-like parameters. *Biostatistics Biometrics Open Access Journal*, 3, 555608. https://doi.org/10.19080/BBOAJ.2017.03.555608

Received 8 April 2022
Accepted 17 April 2022