Natural Sciences and Engineering Research Council of Canada | Conseil de recherches en sciences naturelles et en génie du Canada

1508

Institutional Identifier

System-ID (for NSERC use only): 125521286

FORM 101
Application for a Grant
PART I

Date: 2009/10/13

Family name of applicant	Given name	Initial(s) of all given names	Personal identification no. (PIN)
Fraser	Donald	DAS	**Valid** 1363

Institution that will administer the grant: Toronto

Language of application: [X] English [ ] French

Time (in hours per month) to be devoted to the proposed research / activity: 150

Type of grant applied for: Discovery Grants - Individual

For Strategic Projects, indicate the Target Area and the Research Topic; for Strategic Networks and Strategic Workshops indicate the Target Area.

Title of proposal: Likelihood Based Theory for Applications

Provide a maximum of 10 key words that describe this proposal. Use commas to separate them.

statistics, p-values, likelihood, conditioning, discrete data, ancillarity, confidence, inference, interest parameters, nuisance parameters

Research subject code(s)		Area of application code(s)	
Primary	Secondary	Primary	Secondary
3009	3001	1205	1211

CERTIFICATION/REQUIREMENTS

If this proposal involves any of the following, check the box(es) and submit the protocol to the university or college's certification committee.

Research involving : Humans [ ] Human pluripotent stem cells [ ] Animals [ ] Biohazards [ ]

Does any phase of the research described in this proposal a) take place outside an office or laboratory, or b) involve an undertaking as described in Part 1 of Appendix B?

[X] NO [ ] If YES to either question a) or b) – Appendices A and B must be completed

TOTAL AMOUNT REQUESTED FROM NSERC

Year 1	Year 2	Year 3	Year 4	Year 5
94,700	91,200	91,200	91,200	91,200

SIGNATURES (Refer to instructions "What do signatures mean?")

It is agreed that the general conditions governing grants as outlined in the NSERC *Program Guide for Professors* apply to any grant made pursuant to this application and are hereby accepted by the applicant and the applicant's employing institution.

Applicant

Applicant's department, institution, tel. and fax nos., and e-mail

Statistics
Toronto
Tel.: (416) 978 4448
FAX: (416) 978 5133
dfraser@utstat.toronto.edu

Head of department

Dean of faculty

President of institution (or representative)

The information collected on this form and appendices will be stored in the Personal Information Bank for the appropriate program.

PROTECTED WHEN COMPLETED

Version française disponible

Canada

Personal identification no. (PIN)	Family name of applicant
Valid 1363	Fraser

SUMMARY OF PROPOSAL FOR PUBLIC RELEASE (Use plain language.)

This plain language summary will be available to the public if your proposal is funded. Although it is not mandatory, you may choose to include your business telephone number and/or your e-mail address to facilitate contact with the public and the media about your research.

Business telephone no. (optional): 416 (978) 3452

E-mail address (optional): dfraser@utstat.toronto.edu

Statistical theory has two primary methodologies for analyzing a statistical model with data, the default Bayesian and the frequentist; and these can give widely different answers with demonstrably serious risks. Both take likelihood as their primary ingredient but differ on other input. Discrete data represent a very large area of need for methodology that goes beyond the use of just likelihood. This research continues present work that demonstrates that likelihood gradient or mean likelihood gradient can give the additional substance needed to separate an interest parameter from nuisance parameters and then give highly accurate p-values for the interest parameter. The methodology covers the continous parameter context but its development for the discrete data and the contingency table context represents an urgent need for large areas of application.

Other Language Version of Summary (optional).

Personal identification no. (PIN)	Family name of applicant
Valid 1363	Fraser

Before completing this section, **read the instructions** and consult the *Use of Grant Funds* section of the NSERC Program Guide for Professors concerning the eligibility of expenditures for the direct costs of research and the regulations governing the use of grant funds.

TOTAL PROPOSED EXPENDITURES (Include cash expenditures only)

	Year 1	Year 2	Year 3	Year 4	Year 5
1) Salaries and benefits					
a) Students	22,200	22,200	22,200	22,200	22,200
b) Postdoctoral fellows	40,000	40,000	40,000	40,000	40,000
c) Technical/professional assistants	1,500	1,500	1,500	1,500	1,500
d)	0	0	0	0	0
2) Equipment or facility					
a) Purchase or rental	5,500	2,000	2,000	2,000	2,000
b) Operation and maintenance costs	1,000	1,000	1,000	1,000	1,000
c) User fees	4,000	4,000	4,000	4,000	4,000
3) Materials and supplies	4,000	4,000	4,000	4,000	4,000
4) Travel					
a) Conferences	10,500	10,500	10,500	10,500	10,500
b) Field work	3,000	3,000	3,000	3,000	3,000
c) Collaboration/consultation	2,000	2,000	2,000	2,000	2,000
5) Dissemination costs					
a) Publication costs	1,000	1,000	1,000	1,000	1,000
b)	0	0	0	0	0
6) Other (specify)					
a)	0	0	0	0	0
b)	0	0	0	0	0
TOTAL PROPOSED EXPENDITURES	94,700	91,200	91,200	91,200	91,200
Total cash contribution from industry (if applicable)					
Total cash contribution from university (if applicable)					
Total cash contribution from other sources (if applicable)	0	0	0	0	0
TOTAL AMOUNT REQUESTED FROM NSERC (transfer to page 1)	94,700	91,200	91,200	91,200	91,200

Explanation of Proposed Expenditures (per annum):

1. Salaries and Benefits

a1. Doctoral students: $16,000

Two PhD candidates @ $8,000. Currently: Ramya Thinniyam, Muzzaffar Mallo.

a2. Undergraduates: $6,2000

Undergraduate summer students: One with NSERC summer support @$1,200. and one without NSERC summer support @$5,000. Recent: Kexin Ji and Yanling Cai

b. Postdoctoral Fellows: $40,000

Two half-support as shared or one full-support. Recent: J.-F. Plante and Ye Sun; This has been abundantly fruitful in the last two years with active involvement in research activities and shared research publications and presentations; see Form 100 for student entries.

c. Technical and Professional; $1,500

Programming and computing assistance by undergraduates in advanced classes. This gets starting students involved in the realities of research development. Recent: Kexin Ji and Yanling Cai.

2. Equipment and User Fees.

a. Purchases: $2,000 ($5,500 in first year):

Software and hardware support materials for laptop, presentation tablet and desktop computers (plus $3500 for laptop tablet in year 1).

b. Maintenance: $1,000:

Technical repairs not supplied by the Department.

c. User fees: $4,000:

Department of Statistics user fees for Applicant and students (Thinniyam, Mallo. and Ji)

3. Material and supplies

Copying, printing, fax $2,000. Desk: paper and materials $2,000

4. Travel.

a. Conferences:

One international conference: ISBA World, Benidorm, June 3-8, 2010, $3,000. One North American conference: JSM, Vancouver, July 31-August 5, 2010, $2,500. One Domestic conference: SSC Annual meeting, Quebec, May 23-26, 2010, $2,000. Two Doctoral students to SSC Annual meeting, $3,000.

b. Field work:

One international workshop: $3,000.

c. Collaboration:

Travel expenses for collaborator: Judith Rousseau, U of Paris, 2010: $2,000.

5. Dissemination costs:

Page charges: two papers @ $500: $1,000.

Proposal

1. Recent Progress related to the proposal and the present grant

Statistical methodology tends to cluster around two extreme approaches, methods that are fully conditional on available data as in the Bayesian methodology, and methods that are fully marginal for some chosen evaluation statistic as with typical large-sample, bootstrap or simulation approaches; and then often with the same model and data the two extremes give quite different information summaries. Accordingly, from say an outside viewpoint, it would then seem unclear how such contradictory results can be viewed as acceptable, particularly within the major discipline of statistics. The recent overview [1] surveys such contradictions and attributes a major cause to curvature in the parameters being considered. The concern for the contradictions and their role in the process of statistical inference is a primary focus for the grant, both for the present grant and the proposed renewal. As further background for this concern, recent higher-order likelihood methodology does condition on model characteristics in the neighbourhood of observed data and in doing so presents a direct compromise between the two extremes. Thus while recognizing the breadth and richness of the Bayesian approach for exploratory and preliminary analysis, it does seem essential that the methods have wider calibration and the reliability be assessed outside the paradigm.

Default priors from asymptotics and from continuity: The original Bayes (1763) proposal examined what we now call a location model and recommended a constant or flat prior as indicated by location invariance. In doing this it produced a posterior density that corresponds exactly to what is now called likelihod (Fisher,1922) and gave probabilities that agree precisely with what is now called confidence (Fisher, 1930; Neyman, 1937). The original derivation however used just location symmetry with an appeal to conditional probability. That derivation was not acceptable to some observers, typically because the prior was not a frequency distribution as is proper for conditional probability calculations. The results were acceptable however to others, perhaps because of good performance, performance that might be attributable to the underlying confidence basis. Various extensions of Bayes original procedure were examined by Jeffreys (1939, 1946), Bernardo (1979) and many others, seeking to avoid limitations found with the preceding Bayes priors. In [2] Fraser & Reid used strong matching and likelihood asymptotics to examine repeatability for Bayes procedures and obtained priors that extended the Jeffreys approach. And in current work now in requested revision, Fraser & Reid [3] with two former students use parameter-variable continuity to develop general default priors that do have the repeatability to second and third order; this development is presently in the context of independent scalar coordinates but the wider extension is part of the present proposal; this will make use of characteristics of parameters as examined in [19]. The two approaches use different input but are supportive on areas of overlap.

Large sample conditioning: The parameter-variable continuity has also been used to develop conditioning that operates in the neighbourhood of observed data; this uses a tangent direction approach initiated in Fraser & Reid [4]; the higher derivative development of this conditioning has proceeded with colleagues [5]. The tangent direction approach has also extended the applicability of the conditioning approach into the discrete data context [6] making use of mean likelihood [7]. This with related developments also gives access to simulation procedures.

***p*-values and likelihood**: Higher order likelihood has focussed on a p-value as recording just the percentage position of data with respect to the parameter value, and on the related use of

corresponding confidence quantiles, in place of intervals and other familiar forms of processed information concerning parameters. This p-value usage was implicit in Barndorff-Nielsen (1986) and Daniels (1987) and formalized for large and general data contexts in [8]. As a consequence, definitive p-values are widely available for scalar parameters. In a somewhat parallel development, marginal likelihoods have also been derived with third order accuracy, for both scalar and vector parameters of interest; see Fraser [9]; this clarified various second order developments in the literature that were often in mutual disagreement.

Inference: The preceding contributions offer methodology that has consistency and uses conditioning that is intermediate to the Bayesian and the familiar frequentist approaches. And in addition for the Bayesian approach it makes use of local model form as part of determining the default priors; this is in addition to conditioning on the data. And for the frequentist it uses local model form for the conditioning that is essential to the third order p-value calculations. In some generality these two directions are in close agreement.

2. Objectives

This proposal is targetted on a more unified approach to statistical inference linking Bayesian and frequentist methods, on providing calibration for Bayesian methods but within a broad exploratory framework, and on the related development of the appropriate interface with applications.

Directional assessment of vector interest parameters: The assessment of vector interest parameters is widely based on the log-likelihood ratio quantity where the related null distribution theory is available to high order as part of the Bartlett corrections. But this distributional accuracy is at the price of an overall averaging that neglects informative directional effects that can be of first-order magnitude; and it overlooks conditioning details that are central to the third-order accuracy available for scalar parameters. The resolution of this anomaly is of prime importance for the continuing clarification statistical theory. We continue the development of the appropriate directional tests that build on widely available continuity in the statistical model; these can lead to second and third order directional assessments of parameters in some generality.

Discrete and qualitative data: The development of refined likelihood procedures for the analysis of quantitative data where the parameter of interest is typically vector valued. This builds on the joint work in [6] and the subsequent exploration of many applied problems jointly with colleagues Davison, Reid and Sartori; and it targets on an appropriate directional correction to the log-likelihood ratio p-value, together with an included allowance for boundary effects uncovered by the directional approach.

Parameter curvature from continuity: Parameter curvature can be seen as a prime source for the difference between Bayesian and frequentist results [1]: a preliminary report on this has determined the feasibility of using a curvature measure to correct Bayesian assessments; we seek an expression for a modified Efron curvature measure that targets the curvature underlying the Bayesian-frequentist difference; feasibility has been affirmed through the exploration [10] with **Sun** that uses continuity in the model .

Second-order location-exponential equivalence: A second order asymptotic model with variable and parameter of the same dimension can be represented to second order as a location or an exponential model: see [11] for the scalar case and [12] for the vector case. Many things build on this including aspects of the default priors mentioned above, and the ancillary large-sample conditioning also mentioned above. The analysis however can be extended from location to location with shearing, and infinitessimal generators can be used within the asymptotic framework to generate a magnitude more general framework; this has implications for second order analysis but also for the

domain for the pivotals that are appropriate in a particular problem.

Combining model-data information: Third order asymptotic inference for scalar parameters is widely available from just the observed log-likelihood and an appropriate derivative of the log-model; this first derivative approximation to the statistical model is described in [6] and is represented in terms of those ingredients as a pair $\{\ell(\theta), \varphi(\theta)\}$ which summarizes the tangent exponential model $\exp\{\ell(\theta) + \varphi'(\theta)s\}h(s)$ with data $s^0 = 0$. With colleagues Saleh and Wong we investigate the combining of two or more such data summaries; this generalizes the combining of likelihoods or the combining of p-values to a full third-order accurate combining procedure

Parameter curvature from asymptotics: The application of marginal likelihood [9] to default priors [3] revealed an unusual additional factor, a standardized nuisance root information $|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}|$ that appears to reflect curvature characteristics of the interest parameter ψ; it arose as part of a sample space effect deriving from the Welch-Peers [13] transfer of integration on the parameter space to integration on the sample space and to be effectively absent when parameter linearity is present. We seek to calibrate this nuisance information term as a measure of curvature and to relate it to the curvature measure obtained above from the asymptotic approach.

Conditioning with independent vector variables: Continuity of the distribution function or the equivalent quantile function leads to the default priors and to the conditioning that enables the higher order approximations. For vector coordinates the direct effect of parameter change on the variable requires more. In some contexts the physical interpretation of parameters [19] can demonstrate how a change in the parameter can affect the variable. This will be investigated with high priority.

3. Pertinent literature

[1] Fraser, D.A.S. (2009). Is Bayes posterior just quick and dirty confidence? *Statistical Science*; in review.

[2] Fraser, D.A.S., and Reid, N. (2002) Strong matching of frequentist and Bayesian parametric inference. *Journal of Statistical Planning and Inference*. **103**, 263-285.

[3] Fraser, D.A.S., Reid, N., **Marras, E., and Yi, G.Y.** (2007) Default priors for Bayesian and frequentist inference. *J. Royal Statist. Soc. B*, revision requested.

[4] Fraser, D.A.S., and Reid, N. (1993). Third Order Asymptotic Models: Likelihood functions leading to accurate approximations for distribution functions. *Statist. Sinica* **3**, 67-82.

[5] Fraser, A.M., Fraser, D.A.S. and **Staicu, A.-M.** (2009). The second order ancillary: A differential view with continuity. *Bernoulli*, accepted, minor revision.

[6] Davison, A.C., Fraser, D.A.S. and Reid, N. (2006). Improved likelihood inference for discrete data. *J. Royal Statist. Soc.***B 68**, 495-508.

[7] Reid, N, and Fraser, D.A.S. (2009) Mean likelihood and higher order inference. *Biometrika*; accepted, September 2009.

[8] Fraser, D.A.S., Reid, N., and **Wu, J.** (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249-264.

[9] Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327-339.

[10] Fraser, D.A.S. and **Sun, Y.** (2009). Some corrections for Bayes curvature. *Parkistan J. of Statist.*; accepted: June 2009.

[11] **Cakmak, S.**, Fraser, D.A.S., McDunnough, P., Reid, N., and **Yuan, X.** (1998). Likelihood centered asymptotic model: exponential and location model versions. *J. Statist. Planning and Inference* **66**, 211-222.

[12] **Cakmak, S.**, Fraser, D.A.S., and Reid, N. (1994). Multivariate asymptotic model: exponential

and location approximations. *Utilitas Mathematica* **46**, 21-31.
[13] Welch, B.L. and Peers, H.W. (1963). On formulae for confidence points based in intervals of weighted likelihoods. *J. Roy. Statist. Soc.* B **25**, 318–329.
[14] **Abebe, F., Cakmak, S., Cheah, P.K.**, Fraser, D.A.S., **Kuhn, J.**, McDunnough, P., Reid, N., and **Tapia, A.** (1995). Third order asymptotic model: Exponential and location type approximations. *Pari–sankhyan Samikkha*. **2**, 25-33.
[15] Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307-22.
[16] Stainforth, D. A., Allen, M. R., Tredger, E. R. and Smith, L. A. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Phil. Trans. Roy. Soc.* A, **365**, 2145-2162. See also: Gambling on tomorrow. Modelling the Earth's climate mathematically is hard already. Now a new difficulty is emerging. *Economist*. August 18, 2007, p69.
[17] Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327-339.
[18] Fraser, D.A.S. and Rousseau, J. (2008). Studentization and deriving accurate p-values. *Biometrika*, **95**, 1-16.
[19] McCullagh, P. (2002). What is a statistical model? (with discussion). *Annals of Statistics*, **30**, 1225–1310.

4. Methods and proposed approach

Asymptotics: Asymptotic methods in statistics examine how statistical models are modified by increasing amounts of data information measured initially by a nominal count n of data inputs. Typically a log-model grows as $O(n)$ and can be expanded by Taylor series in both sample variable and parameter. This expansion can be relative to given data y^0 with related observed maximum likelihood value $\hat{\theta}^0$ or it can be relative to a parameter value of prime interest and a related maximum density value; the two possibilities provide different information summaries concerning the model; see [11], [12] and [14]. The exceptional power of this approach has not been widely recognized. Another approach involves direct functional manipulation of the log-model [15] and provides different access to model asymptotic properties; an advantage when available is that concluding expressions are typically in a form invariant of the mode of expression of the input variables and parameters. This with recent ancillary developments has led to definitive p-values [8] for general scalar parameters. These techniques are supportive for the development of default priors as proposed here and also for the directional assessment of vector parameters also proposed.

Continuity: A powerful related tool involves the use of directly available continuity typically expressed in coordinate distribution functions or in the corresponding quantile functions. This is a case where of course continuity is wanted, but its direct use has been generally overlooked. This continuity leads to the ancillaries in [5] and the default priors in [3]. Together the asymptotics and the continuity are primary techniques for the proposed research.

The observed nuisance maximum-likelihood surface: A contour for assessing nuisance parameters intersects the observed nuisance maximum-likelihood surface in a single point; the points on the surface thus index the contours for the nuisance parameter and the corresponding marginal distribution recorded on that surface is third-order unique and records the distribution for assessing the interest parameter: this led to the marginal likelihood in [17] and the testing distribution that was free of the nuisance parameter in [18]. And in turn it provides the basic nuisance-free distribution for directionally assessing a vector parameter value. This plus the availability of the approximating exponential model provides the core theory and methodology for the vector interest problem and its further development for qualtitative data analysis.

5. Anticipated significance:

A widely available method of analyzing complex models in chemical engineering and more generally is to use a flat prior on the available range for input variables and then do a standard Bayesian analysis. As reported from chemical engineering and now from weather modelling [16] this can lead to conflicting results with differing models. A partial explanation is that a flat prior for a particular variable is not a flat prior for a reexpressed variable. But the overall effect with large models may be much magnified. The use of the Bayesian approach has had a rich effect in liberalizing statistics and giving directly implementable procedures. But the claim that probabilities are obtained is contradicted by [16]; and the implicated cause is curvature in the parameters being examined in addition to presumed linearity in the input variables; see [1]. It would seem imperative that suitable calibration be part of the process; this may not be easy but may be necessary at some level.

6. HQP training:

Undergraduates, summer students, graduate students and postdoctural students have been regularly involved in seminars and workshops on the themes of this grant, as indicated by their authorship in the papers cited here and on Form 100. Various computational and analytic issues arise frequently with log-models and measures of departure and students pursuing these find a welcome contrast to extensive literature surveys. They find excitement in the exploration and use of rather different techniques, and then eagerly pursue a broader contact with the discipline and its applications. The most recent two years with two shared PostDocs was very enriching: helpful for them and for the grant holders but also for the discipline. We hope to continue and expand this enriching activity.