SUMMARY OF SOME STATISTICAL ISSUES

N. REID

Department of Statistics, University of Toronto, 100 St. George St., Toronto Canada M53 3G3
E-mail: reid@utstat.utoronto.ca

A brief summary of some statistical issues that arose during the conference is presented.

In terms of statistical ideas, I would make a very broad distinction between two prominent sets of problems at this conference. First there are a number of problems in which the main feature is a very large amount of data, requiring new methods and considerable computing power. An example that has already been used with success in astronomy is the use of false discovery rates in problems involving a great number of tests, and we heard here about new adaptations of wavelet and ridgelet techniques for identifying structure in images, about smoothing methods in multi-dimensional image processing, and new methods for on-line data mining. I won't attempt to summarize this class of problems, although it is clearly very important, not only in physics and astronomy but in a number of scientific problems, especially including genomics, where there is very active development of statistical techniques.

Another class of problems seem simpler (to a statistician) on a first reading. An example is independent Poisson counts from background events and possible signal events. We should not forget, though, that elaborate experimental techniques and considerable ingenuity in data processing, have preceded the presentation of a small amount of data. For this setting one would expect that standard statistical methods would provide a simple, and even a best, answer, but as we have seen even in this context this is not always the case. Certainly inference about the ratio of Poisson mean parameters is satisfactorily solved using the binomial likelihood. Statistical inference for the difference between two Poisson means is somewhat more difficult, as we have to rely on some approximate argument, and with small counts the usual normal approximations will not be reliable. in As Sir David Cox stressed in the panel discussion, the science of statistics develops most fruitfully in close collaboration with applications, and this problem is a good example of something that is indeed sufficiently specialized to the HEP context that it is not in the repertoire of 'off-the-shelf' statistical methods.

Some general ideas which should inform the solution include the very important notion that confidence intervals, however developed, should have good properties in repeated observation of the same experimental system, even if these repetitions are hypothetical. In my view the definition of 'same experimental system' needs great care, in order to avoid difficulties similar to, but more subtle than, the problem of two measuring instruments discussed in Cox[4] and mentioned in Cousins [this volume]. Unfortunately it seems extremely difficult to 'mathematize' this notion; statisticians have spent many years of effort on the topic, and a single widely accepted solution has not emerged. At this time the best we can advise is to look at problems on a case by case basis.

Likelihood methods are well accepted in the HEP community, but not always used in quite the same manner as used by statisticians. To clarify, suppose we have a single parameter model $f(x;\theta)$ and observe a sample $\underline{x} = (x_1, \ldots, x_n)$ of independent observations from this model. The log-likelihood function $\ell(\theta;\underline{x}) = \log \Pi f(x_i;\theta)$ is a sum of n terms, and we can apply the central limit theorem to $\partial\ell(\theta;\underline{x})/\partial\theta$ to derive the following approximations:

$$\begin{aligned} (\hat\theta - \theta) i^{1/2}(\theta) &\;\dot\sim\; N(0,1) \\ \ell'(\theta) i^{-1/2}(\theta) &\;\dot\sim\; N(0,1) \\ \pm\sqrt{[2\{\ell(\hat\theta) - \ell(\theta)\}]} &\;\dot\sim\; N(0,1) \end{aligned}$$

where $\hat\theta$ is the maximum likelihood estimate and $i(\theta) = E\{-\partial^2\ell(\theta)/\partial\theta^2\}$ is the expected information. Barlow [this volume, 2nd talk] described the second of these as Bartlett's statistic and the third as $-2\ln L$ (although I have here taken the square root, since the parameter is scalar). Each approximation provides a different way to compare the expected value to the observed value, but each is a so-called 'first order approximation', because the error in the approximation is $O(n^{-1/2})$. In the limit when the log-likelihood function becomes quadratic, with second derivative equal in the limit to its expectation, they all lead to the same measure. To these three approximations we can further confuse things by adding standardization by observed information:

$$\begin{aligned}(\hat{\theta}-\theta)j^{1/2}(\hat{\theta}) &\;\dot{\sim}\; N(0,1)\\ \ell'(\theta)j^{-1/2}(\hat{\theta}) &\;\dot{\sim}\; N(0,1)\end{aligned}$$

where $j(\hat{\theta}) = -\ell''(\hat{\theta})$ is the curvature of the log-likelihood function at the maximum.

A very natural question is which of these approximations is to be preferred in finite samples, and some reasons for expecting the log-likelihood ratio to be preferred are that it is invariant to reparametrization, and that it preserves the asymmetry in the log-likelihood function. It is also the leading term in a higher order expansion, the correction term of which uses one or other of the two j-standardized statistics. Indeed the statistical literature has since Efron & Hinkley[8] preferred the j-standardization for $\hat{\theta}$, and later somewhat technical development of improved approximations to the distribution of $\hat{\theta}$ have confirmed this preference. It is related to conditioning on ancillary statistics, i.e. functions of the data that have a distribution exactly or approximately free of θ.

Unfortunately however there are no general results on rates of convergence or other properties that could lead to a definitive conclusion about which departure measure to use, and case by case studies are thus needed. Barlow [this volume, 2nd talk] showed that for the exponential mean, Bartlett's statistic, i.e. the score function using the i-standardization (which coincidentally is the same in this example as the i-standardized maximum likelihood estimate), is better approximated by a standard normal than the log-likelihood ratio. This I found quite surprising, given my 'prior belief' in the log-likelihood ratio statistic. The explanation is that Bartlett's statistic has exact mean 0 and exact variance 1, these moments coinciding with those of the normal approximation, which is therefore reasonably accurate for moderate deviations. However if we move out to the tails the likelihood ratio statistic is more accurately approximated by a standard normal than Bartlett's statistic. Figure 1 compares the p-values, as functions of the mean parameter, to the exact p-value based on the gamma distribution, for a sample of size 5 and an observed sample mean of 1, first in the '1-sigma' range and then in the '4-sigma' range. This example is also treated in Barndorff-Nielsen & Cox[3].

It does seem very difficult to draw any general conclusions about the first order approximations, although for most examples I have looked at the normal approximation to the square root of the likelihood ratio has been the most accurate in the tails. A relatively simple combination of this with the Bartlett score statistic, as outlined in Reid & Fraser[11] gives essentially exact results for the exponential example.

As has been mentioned several times during this workshop, adding nuisance parameters further complicates the issues. There are a number of somewhat different lines of argument in the statistical literature leading to the idea of improving the profile likelihood by adding a term to allow for the estimation of the nuisance parameters. The simplest motivation is from a Bayesian argument. We can get an approximation for the marginal posterior distribution of the parameter of interest as follows:

$$\begin{aligned}\pi_m(\psi \mid \underline{x}) &= \int \pi(\psi,\nu \mid \underline{x})d\nu\\ &\propto \int \exp\{\ell(\psi,\nu)\}\pi(\psi,\nu)d\nu\\ &= \int \exp\{\ell(\psi,\nu)\}\pi(\nu \mid \psi)d\nu\pi(\psi)\\ &\doteq \exp\{\ell(\psi,\hat{\nu}_\psi)\}|j_{\nu\nu}(\psi,\hat{\nu}_\psi)|^{-1/2}\cdot\\ &\qquad \pi(\hat{\nu}_\psi \mid \psi)\sqrt{(2\pi)}^{k-1}\pi(\psi)\end{aligned}$$

where $\theta = (\psi,\nu)$ has been partitioned into a parameter of interest ψ and a $k-1$-dimensional nuisance parameter ν, and $j_{\nu\nu}(\psi,\hat{\nu}_\psi) = -\partial^2\ell(\psi,\nu)/\partial\nu\partial\nu^T$ is the portion of the observed information matrix related to the nuisance parameter. The last approximation comes from a Laplace approximation of the integral defining the marginal posterior.

Now it can be shown that if ψ and ν are orthogonal parameters, in the sense that the (ψ,ν) components of the expected Fisher information matrix are 0, then $\hat{\nu}_\psi = \hat{\nu} + O_p(1/n)$; in the absence of parameter orthogonality the error would be $O_p(1/\sqrt{n})$. Sweeting[13] in the discussion of Cox & Reid[6] argued that if ψ and ν are orthogonal then it would make sense to assign independent priors to them, in which case the term involving the prior on ν vanishes (to $O(n^{-1})$) and the log of the posterior marginal den-

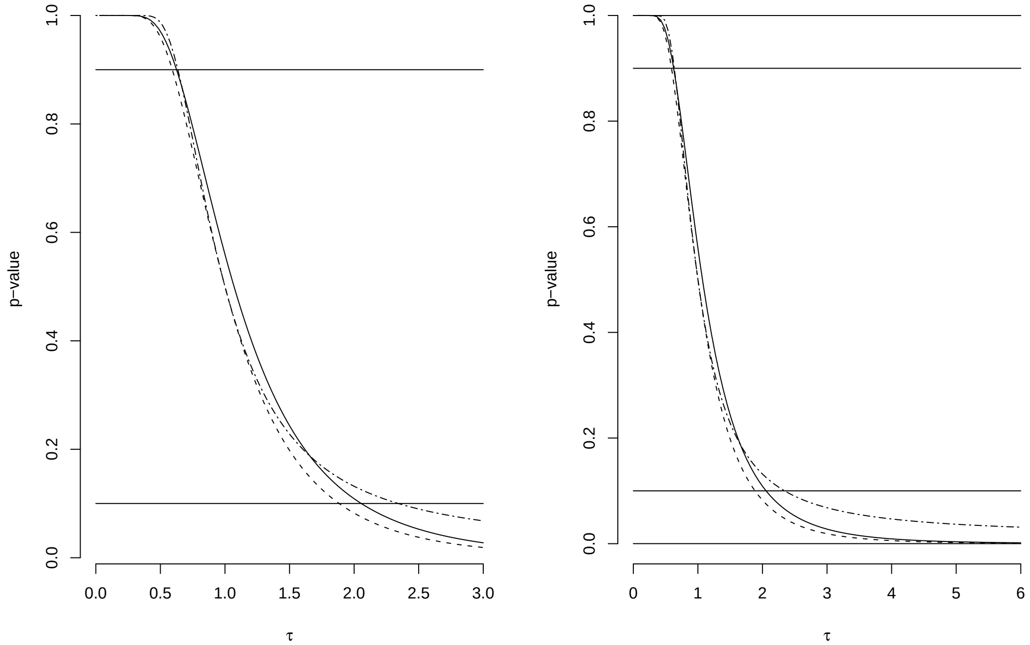

Fig. 1. Plot of p-value functions for exponential mean parameter τ, computing using the exact distribution (solid), the normal approximation to the square root of the log-likelihood difference (dashed), and the normal approximation to the standardized score (dot-dash). Horizontal lines show the 0.10 and 0.90 limits (left), as well as the 0.0001 and 0.9999 limits (right). The sample size n is 5 and the sample mean is 1. Very similar results are obtained for both smaller and larger values of n.

sity is

$$\log \pi_m(\psi \mid \underline{x}) \doteq \ell(\psi, \hat{\nu}_\psi) - \frac{1}{2}\log|j_{\nu\nu}(\psi, \hat{\nu})| + \log \pi(\psi);$$

this is one way to motivate the so-called "adjusted" or "modified" profile log-likelihood

$$\ell_a(\psi) = \ell(\psi, \hat{\nu}_\psi) - \frac{1}{2}\log|j_{\nu\nu}(\psi, \hat{\nu})|.$$

If ψ is scalar then a transformation from some original parameterization (ψ, ϕ) to (ψ, ν) where ν is orthogonal to ψ can always be found; Cox & Reid[7] indicate how to compute the adjusted profile without explicitly reparameterizing the model. The term "modified profile likelihood" is usually used for one of a family of adjusted profile log-likelihoods of the form

$$\ell(\psi, \hat{\nu}_\psi) - \frac{1}{2}\log|j_{\nu\nu}(\psi, \hat{\nu})| + B(\psi)$$

where $B(\psi)$ is to be specified, but is always $O(1)$, i.e. the same order as the $\log j$ term, and serves among other things to make the result parameterization invariant, which the simple version ℓ_a is not.

Although motivated by higher order asymptotic arguments, only first order asymptotics apply to ℓ_a and its variants. In particular we have, in analogy to the results for a scalar parameter

$$\begin{aligned}
(\hat{\psi}_a - \psi)\{-\ell_a''(\hat{\psi}_a)\}^{1/2} &\mathrel{\dot{\sim}} N(0,1) \\
\ell_a'(\psi)\{-\ell_a''(\hat{\psi}_a)\}^{-1/2} &\mathrel{\dot{\sim}} N(0,1) \\
\pm\sqrt{2\{\ell_a(\hat{\psi}_a) - \ell_a(\psi)\}} &\mathrel{\dot{\sim}} N(0,1)
\end{aligned}$$

where $\hat{\psi}_a$ is the maximum likelihood estimate from $\ell_a(\psi)$. These approximations are no more accurate in asymptotic theory than those based on the profile likelihood but in practice the adjustment for nuisance parameters seems to lead to better approximations, especially when the number of nuisance parameters is large.

There are two classes of models where, at least for some of their parametrizations, exact elimination of nuisance parameters is possible: exponential family models and non-normal linear regression models. Some examples are given in Reid & Fraser[11]. In these two classes the adjusted profile likelihood ℓ_a arises quite naturally as a kind of 'leading term'.

In models with a single scalar parameter, there is a uniquely determined, albeit improper, prior for which Bayesian posterior upper limits are guaranteed to have frequentist coverage to high accuracy: more precisely we have

$$\Pr(\theta \leq \theta^{(1-\alpha)}(\underline{x}) \mid \underline{x}) = \Pr_\theta(\theta^{(1-\alpha)}(\underline{X}) \geq \theta) + O(1/n)$$

if and only if the prior is proportional to $i^{1/2}(\theta)$; the first probability above is calculated under the posterior distribution, and defines $\theta^{(1-\alpha)}(x)$ by the requirement that this probability equal α, and the second probability is calculated under the sampling model $f(\underline{x};\theta)$. In multiparameter problems, matching priors do not exist in general, but there is an important exception. In statistical models whose mathematical structure is generated by a group of transformations, then it is possible to obtain the exact distribution of the maximum likelihood estimator by conditioning, and this is identical to the Bayesian posterior distribution for a special choice of prior measure related to the group structure; see Fraser[9], Barndorff-Nielsen[2] and also Podobnik & Zivko (2005, this volume). These arguments do not apply however to models for discrete data.

A recurring theme in this meeting has been the possible dangers in using flat priors in multiparameter problems. An early and compelling example is described in Example 10.6 of Cox & Hinkley[5]. Suppose $X_1, \ldots, X_n$ are independent normal random variables with mean μ_i and variance σ^2, and that

$$\mu_i = EX_i = \gamma + \beta\rho^{x_0+ia}, \quad 0 \leq \rho \leq 1$$

where x_0 and a are known, and $\theta = (\gamma, \beta, \rho, \sigma)$. In a linear regression model, the matching prior and most usual prior is proportional to $d\beta d\sigma/\sigma$, so a very natural 'flat' extension of this is to choose the prior

$$\pi(\theta) \propto d\gamma d\beta d\sigma/\sigma d\rho, \quad 0 \leq \rho \leq 1;$$

however the marginal posterior for ρ concentrates on the points $\rho = 0$ and $\rho = 1$. I don't know if this phenomenon is widespread or not, but the fact that one can so easily get into trouble in a relatively simple with a seemingly vague choice of prior is somewhat worrying. Heinrich [this volume] also raises several issues with flat priors. There is an active research effort in the statistics community to investigate what have come to be called 'objective' priors; the most recent conference was 'OBayes5', held in June, 2005.

Speaking as a statistician who has been largely involved with theoretical issues, it is exciting to discuss these issues in the context of applications to high energy physics, and I look forward to further fruitful collaborations between the two disciplines.

References

1. R. Barlow, *this volume* (A note on Delta ln L = -1/2 errors), 2005.
2. O.E. Barndorff-Nielsen, *Biometrika* **67**, 293 (1980).
3. O.E. Barndorff-Nielsen and D.R. Cox, *Inference and Asymptotics.* (Chapman & Hall, London, p, 83, 1994).
4. D.R. Cox, *Ann. Math. Statist.* **29**, 257 (1958).
5. D.R. Cox and D.V. Hinkley, *Theoretical Statistics.* (Chapman & Hall, London, 1974).
6. D.R. Cox and N. Reid, *J. R. Statist. Soc.* **B**, 49 (1)1987.
7. D.R. Cox and N. Reid, *J. R. Statist. Soc., B* **55**, 467 (1993).
8. B. Efron and D.V. Hinkley, *Biometrika* **65**, 457 (1978).
9. D.A.S. Fraser, *Inference and Linear Models.*, Ch. 7, McGraw-Hill, New York, 1979.
10. J. Heinrich, *this volume* (The Bayesian approach to setting limits: what to avoid), 2005.
11. N. Reid and D.A.S. Fraser, in *Proceedings of PHYSTAT2003*, L. Lyons, R. Mount, R. Reitmeyer, eds. SLAC e-Conf C030908, 265 (2003).
12. T. Podobnik and T. Zivko, *this volume* (On consistent and calibrated inference about the parameters of sampling distributions), 2005.
13. T.J. Sweeting, *J. R. Statist. Soc.* **B**, 49 (20)1987.