Modern Likelihood-Frequentist Inference

Donald A Pierce, OHSU

and

Ruggero Bellio, Univ of Udine

- Shortly before 1980, important developments in frequency theory of inference were “in the air”.

- Strictly, this was about new asymptotic methods, but with the capacity leading to what has been called “Neo-Fisherian” theory of inference.

- A complement to the Neyman-Pearson theory, emphasizing likelihood and conditioning for the reduction of data for inference, rather than direct focus on optimality, e.g. UMP tests

BIOMETRIKA

Table of Contents

Volume 67 Issue 2 1980

For checked items
view abstracts download to citation manager Go Clear

Articles

A few years after that, this pathbreaking paper led the way to remarkable further development of *MODERN LIKELIHOOD ASYMPTOTICS*

Article
O. E. BARNDORFF-NIELSEN
Infereni on full or partial parameters based on the standardized signed log likelihood ratio
Biometrika (1986) 73 (2): 307-322 doi:10.1093/biomet/73.2.307
...Trust August 1986 research-article...Infereni on full or partial...ratio O. E. BARNDORFF-NIELSEN...of drawing inference on the complementary...ancillary (Barndorff-Nielsen...Biometrika (1986), 73...Great Britain Inference on full or partial...BY O. E. BARNDORFF-NIELSEN...
» Abstract » Full Text (PDF)

That paper was difficult, so Dawn Peters and I had some success interpreting it in an invited RSS discussion paper

Practical use of higher order asymptotics for multiparameter exponential families
DA **Pierce**, D Peters - Journal of the Royal Statistical Society. Series B (..., 1992 - JSTOR
... **inferences**-tests and confidence intervals-about a single parametric function which ... applications of saddlepoint methods, mainly from the viewpoint of improving on the central imit theorem and direct Edgeworth expansions, without explicit attention to **inference**. **Barndorff**-Nielsen ...
Cited by 145 Related articles All 2 versions Cite Save More

HIGHLY ABRIDGED REFERENCE LIST

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed likelihood ratio. *Biometrika* **73**, 307-322.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*, Chapman and Hall, London.

Brazzale, A. R., Davison, A. C. and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics.* Cambridge Univ. Press, Cambridge.

Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika* **67**, 311-333.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457-482.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective.* World Scientific Publishing Co. , Singapore.

Pierce, D. A. and Peters, D. (1992). Practical use of higher-order asymptotics for multiparameter exponential families. *J. Roy. Statist. Soc. B.* ***54**,* 701-737.

Pierce, D. A. and Bellio, R. (2006). Effects of the reference set on frequentist inferences. *Biometrika* ***93**,* 425-438.

Pierce, D.A. and Bellio, R. (in prep). Modern likelihood-frequentist inference. (*with more complete reference list).*

Reid, N. (1996). Likelihood and higher-order approximations to tail areas: A review and annotated bibliography. *Canad. J. Statist.* **24,** 141-166.

Reid, N. (2003). The 2000 Wald Lectures: Asymptotics and the theory of inference. *Ann. Statist.* **31**, 1695-1731.

Severini, T. A. (2000). *Likelihood Methods in Statistics.* Oxford Univ. Press, Oxford.

Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145-165.

Skovgaard, I. M. (2001). Likelihood asymptotics. *Scand. J. Statist.* **28**, 3-32.

Skovgaard, I. M. (1985). A second-order investigation of asymptotic ancillarity. *Ann. Statist.* **13**, 534-551.

THIS LIST CONTAINS SEVERAL MAJOR BOOKS

Inference and Asymptotics (1994) Barndorff-Nielsen & Cox

Principles of Statistical Inference from a Neo-Fisherian Perspective (1997) Pace & Salvan

Likelihood Methods in Statistics (2000) Severini

- Salvan (Univ Padua) and Pace & Bellio (Univ Udine) made it possible for me to visit 2-4 months/year from 2000 to 2014 to study Likelihood Asymptotics

- In 2012 they arranged for me a Fellowship at Padua, work under which led to the paper in progress discussed today

- This is based on the idea that the future of Likelihood Asymptotics will depend on: (a) development of generic computational tools and (b) concise and transparent exposition amenable to statistical theory courses.

- Starting point is a simple and accurate ‘likelihood ratio approximation’ to the distribution of the (multidimensional) maximum likelihood estimator

- Next step is to transform & marginalize from this to the distribution of the signed LR statistic (sqrt of usual χ_1^2 statistic) --- requiring only a Jacobian and Laplace approximation to the integration

- This result is expressed as an adjustment to the first-order *N(0,1)* distribution of the LR: “If that approximation is poor but not terrible this mops up most of the error” (Rob Kass)

- This is not hard to fathom---accessible to a graduate level theory course---if one need not be distracted by arcane details

- A central concept in what follows involves *observed* and *expected* (Fisher) information.

- The *observed* information is defined as minus the second derivative of the loglikelihood at its maximum

$$\hat{j} = -\ddot{l}(\theta; y)\,|_{\theta=\hat{\theta}}$$

- The *expected* information (more usual Fisher info) is defined as

$$i(\theta) = E\{-\ddot{l}(\theta; Y)\}$$

- And we will write $\hat{i} = i(\hat{\theta})$

- The MLE is sufficient if and only if $\hat{i} = \hat{j}$, and under regularity this occurs only for exponential families without restriction on the parameter (full rank case)

- Inferentially it is unwise and not really necessary to use the average information

- With methods indicated here, it is feasible to condition on an *ancillary statistic* such as

$$a = \hat{j} / \hat{i} \qquad (\text{meaning actually } \hat{i}^{-1}\hat{j})$$

- This is key part of what is called *Neo-Fisherian Inference*

- Remarks on ancillary conditioning: Neo-Fisherian Inference

- To Fisher, “optimality” of inference involved sufficiency, more strongly than in the Neyman-Pearson theory

- But generally the MLE is not a sufficient statistic

- Thus to Fisher, and many others, the resolution of that was conditioning on something like $a = \hat{j} / \hat{i}$ to render the MLE sufficient beyond 1st order.

- Indeed, Skovgaard (1985) showed that in general $(\hat{\theta}, a)$ is to $O_P(1/n)$ sufficient, and conditioning on $a = \hat{j}/\hat{i}$ (among other choices) leads in that order to: (a) no loss of "information", (b) the MLE being sufficient

- The LR approximation to the distribution of the MLE (usually but less usefully called the p^* formula, or the "magic formula" as by Efron in his Fisher Lecture) is then

$$pr^*(\hat{\theta} \mid a;\theta) = \frac{|\, j(\hat{\theta})\,|^{1/2}}{(2\pi)^{p/2}} \frac{pr(y;\theta)}{pr(y;\hat{\theta})}$$

$$= pr(\hat{\theta} \mid a;\theta)\left\{1 + O(n^{-1})\right\}$$

- Though this took some years to emerge, in restrospect it becomes fairly simple:

$$p(\hat{\theta}|a;\theta) \equiv \frac{p(\hat{\theta}|a;\theta)}{p(\hat{\theta}|a;\hat{\theta})} p(\hat{\theta}|a;\hat{\theta})$$

$$\doteq \frac{p(y|a;\theta)}{p(y|a;\hat{\theta})} \; \frac{p(\hat{\theta}|a;\theta)}{p(\hat{\theta}|a;\hat{\theta})} \; p(\hat{\theta}|a;\hat{\theta}) \quad \text{since the first term is nearly unity}$$

$$= \frac{p(y;\theta)}{p(y;\hat{\theta})} \; p(\hat{\theta}|a;\hat{\theta}) \quad \text{and with Edgeworth expansion to the final term}$$

$$\doteq \frac{p(y;\theta)}{p(y;\hat{\theta})} \frac{|j(\hat{\theta})|^{1/2}}{(2\pi)^{p/2}} \quad \text{this having relative error } O(1/n) \text{ for all } \theta = \hat{\theta} + O(n^{-1/2})$$

$$= p^{*}(\hat{\theta}|a;\theta)$$

- The Jacobian and marginalization to be applied to $p^*(\hat{\theta})$ involve rather arcane sample space derivatives

$$C_\psi = \left| \frac{\partial^2 l(\hat{\theta}_\psi)}{\partial\lambda\,\partial\hat{\lambda}^T} \right| \left\{ |\,\hat{j}_{\lambda\lambda}\,|\,|\,\tilde{j}_{\lambda\lambda}\,| \right\}^{-1/2} \quad , \qquad \tilde{u}_\psi = |\,\partial\{l_P(\hat{\theta};\hat{\theta},a) - l_P(\theta;\hat{\theta},a)\} / \partial\hat{\psi}\,|\,|\,\tilde{j}_{\psi|\lambda}\,|^{-1/2}$$

approximations* to which are taken care of by the software we provide.

- The result is an inferential quantity that is standard normal to 2nd order

$$r_\psi^* = r_\psi + r_\psi^{-1}\log(C_\psi) + r_\psi^{-1}\log\left\{\tilde{u}_\psi / r_\psi\right\} = r_\psi + NP + INF$$

modifying the usual 1st order standard normal LR quantity

$$r_\psi = sign(\hat{\psi} - \psi)\sqrt{2\{l(\hat{\theta};y) - l(\tilde{\theta};y)\}}$$

- It was almost prohibitively difficult to differentiate the likelihood with respect to MLEs while holding fixed an ancillary statistic

- The approximations* to sample space derivatives referred to came in a breakthrough by Skovgaard, making the theory practical

- Skovgaard's approximation uses projections involving covariances of likelihood quantities computed without holding fixed an ancillary

- Our software uses simulation for these covariances, NOT involving model fitting in simulation trials

- To use the generic software, the user specifies an R-function for computing the likelihood. The choices made render the routines fairly generally applicable.

- Since higher-order inference depends on more than the likelihood function, one defines the extra-likelihood aspects of the model by providing another R-function that generates a dataset.

- The interest parameter is defined by one further R-function.

- We illustrate this with a Weibull example, and interest parameter the survival function at a given time and covariate

```
loglik.Wbl <- function(theta, data)
 {
 logy <- log(data$y)
 X <- data$X
 loggam <- theta[1]
 beta <- theta[-1]
 gam <- exp(loggam)
 H <- exp(gam * logy + X %*% beta)
 out <- sum(X %*% beta + loggam + (gam-1) * logy - H)
 return(out)
 }


gendat.Wbl <- function(theta, data)
 {
 X <- data$X
 n <- nrow(X)
 beta <- theta[-1]
 gam <- exp(theta[1])
 data$y <- (rexp(n) / exp(X %*% beta)) ^ (1 / gam)
 return(data)
 }

psifcn.Wbl <- function(theta)
 {
 beta <- theta[-1]
 gam <- exp(theta[1])
 y0 <- 130
 x0 <- 4
 psi <- -(y0 ^ gam) * exp(beta[1] + x0 * beta[2])
 return(psi)
 }
```

- For testing this ψ at the Wald-based 95% lower confidence limit, the results are

$$r_\psi = 1.66 \ (P = 0.048)$$

$$r_\psi^* = 2.10 \ \left(P = 0.018\right)$$

$$Wald = 1.95 \ (P = 0.025)$$

- This is typical for settings with few nuisance parameters, when there are several the adjustment can be much larger