

Asymptotics and the theory of inference

Outline

1. Introduction
2. Asymptotics and Wald
3. First order asymptotics
4. Higher order likelihood asymptotics
5. Some other applications of asymptotic theory
6. Lectures 2 and 3

1. Introduction

- Asymptotic arguments are a natural/inevitable consequence of a frequency based theory of probability, even in a Bayesian context
- Provide approximations that have proved to be relatively robust
- Provide insight into inference:
 - verifying that we're being moderately sensible
 - comparing competing methods
 - understanding the structure of models

re 1: David Andrews' 'thought experiment' in statistical consulting

re 2: early example: Fisher's comparison of standard deviation and mean deviation discussed in Stigler (Bka, 1973) (I learned about it from Bickel and Doksum's book)

$s = \sqrt{(n-1)^{-1} \sum (x_i - \bar{x})^2}$ compared to scaled version of $n^{-1} \sum |x_i - \bar{x}|$; s^2 has smaller asymptotic variance (this also led Fisher on to discover sufficiency!)

later example: Results of Stone (1980, 82, Annals), Fan (1993 Annals), minimax efficiency of local linear smoothers; see Hastie and Loader (1993, Statist. Sci) for overview

re 3: early example: Neyman and Scott (1948, Econometrica) on inconsistency of maximum likelihood estimate in problems with increasing number of nuisance parameters

recent example: Smith (Bka, 1985) on asymptotic theory for non-regular models, e.g. endpoint problems $f(x; \psi, \lambda) = (x - \psi)^{\lambda-1} g(x - \psi; \phi), x > \psi$; asymptotics for m.l.e. of ψ depends crucially on $\lambda, > 2, = 2, \in (1, 2), \in (0, 1]$

$$s_2 = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

$$s_1 = \frac{1}{n \cdot c} \sum |x_i - \bar{x}|$$

2. Previous Wald lectures

- 41 lectures, 11 on “pure probability”, 21 on “inference”, design of experiments, boundary crossing, linear models, dynamic programming, statistical methods (applications) (1!)
- inference: fiducial, estimation, decision theory, nonparametric (infinite-dimensional parameters), testing, likelihood

all use asymptotic theory for insight, to suggest approximations, as a ‘reality check’

| | |
|-------------------|--|
| Tukey, 1958 | Fiducial inference |
| Bose, 1959 | Design of experiments |
| Stein, 1961 | Estimation of many parameters |
| Kiefer, 1962 | Optimal experimental design |
| LeCam, 1963 | Asymptotic models in decision theory |
| Lehmann, 1964 | Topics in nonparametric statistics |
| Wolfowitz, 1965 | Efficiency, information theory, design of ex |
| Hoeffding, 1967 | Parametric large sample theory |
| Chernoff, 1968 | Optimal stochastic control |
| Robbins, 1969 | Boundary crossing problems |
| Rosenblatt, 1970 | Estimation for stationary processes |
| Burkholder, 1971 | Martingale theory and its applications |
| Huber, 1972 | Robustness |
| Billingsley, 1973 | Additive arithmetic functions |
| Bahadur, 1974 | Testing and estimation |
| Rao, 1975 | Estimation |
| Blackwell, 1976 | Dynamic programming |
| Kingman, 1977 | Exchangeability in population genetics |
| Kac, 1978 | Statistical physics |
| Spitzer, 1979 | Interacting particle systems |
| Bickel, 1980 | Robustness and adaptation |
| Efron, 1981 | Bootstrap, transformation theory, m.l.e.s |

3. First order asymptotics

(a) Maximum likelihood estimate is consistent, asymptotically normal, efficient (Wald)

(b) Score statistic has mean zero, asymptotically normal

(c) Likelihood ratio statistic has mean k , asymptotically distributed χ^2

(d) posterior distribution is asymptotically normal

$$(\hat{\theta} - \theta)j^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, I)$$

$$U(\theta)\{j(\hat{\theta})\}^{-1/2} \xrightarrow{d} N(0, I)$$

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi_p^2$$

$$\int_{a_n}^{b_n} \pi(\theta|y) d\theta \longrightarrow \Phi(b) - \Phi(a)$$

$$a_n = \hat{j}^{-1/2}a + \hat{\theta}$$

,

Notation

| | |
|-----------------------------|---|
| model | $f(y; \theta), \quad \theta \in \mathfrak{R}^k$ |
| data | $y = (y_1, \dots, y_n)$ |
| likelihood | $L(\theta) = L(\theta; y) = c(y)f(y; \theta)$ |
| log-likelihood | $\ell(\theta) = \ell(\theta; y) = \log f(y; \theta)$ |
| maximum likelihood estimate | $\hat{\theta} = \hat{\theta}(y) = \arg \sup_{\theta} L(\theta)$ |
| score function | $U(\theta) = \ell'(\theta)$ |
| observed information | $j(\theta) = -\ell''(\theta)$ |
| expected information | $i(\theta) = \text{cov}_{\theta}\{U(\theta)\}$ |

...3 (a) Maximum likelihood estimate is consistent, asymptotically normal, efficient (Wald)

except (e.g.)

- infinitely many nuisance parameters; Neyman-Scott; nonparametric

$$f(y_{ij}; \psi, \lambda_j), \quad j = 1, \dots, J; i = 1, \dots, I \quad J \rightarrow \infty$$

- non-regular problems; true parameter on boundary, endpoint problems, changepoints, mixtures (e.g. Smith, 1989)

$$f(x; \psi, \lambda) = (x - \psi)^{\lambda-1} g(x - \psi; \phi), \quad x > \psi$$

| | |
|----------------------|---------------------------------------|
| $\lambda > 2$ | m.l.e. asymptotically normal |
| $\lambda = 2$ | m.l.e. asympt. normal, not $n^{-1/2}$ |
| $\lambda \in (1, 2)$ | not efficient (Woodroffe, ...) |
| $\lambda \in (0, 1)$ | not easily summarized |

- Long range dependence (Beran, 1994)

...3 (b) Score statistic has mean zero, asymptotically normal

- starting point for generalizations to complex models, e.g. proportional hazards model for survival data (Cox, 1972)
- most problems with (a) when m.l.e. not a root of the score equation
- e.g. Speed (Biostatistics, 1,1): score test for genetic linkage based on second derivative of loglikelihood
- Rotnitzky et al. (Bernoulli, 2000) on asymptotic theory for such
- Use $j(\hat{\theta})$ not $i(\theta)$ for variance estimate (Efron-Hinkley, Sorensen)

(c) Likelihood ratio statistic $2\{\ell(\hat{\theta}) - \ell(\theta)\}$
asymptotically distributed χ_k^2

$$\begin{aligned} &\sim U(\theta)^T j(\hat{\theta})^{-1} U(\theta) \\ &\sim (\hat{\theta} - \theta)^T j(\hat{\theta})(\hat{\theta} - \theta) \end{aligned}$$

(d) posterior distribution is asymptotically

$$N(\hat{\theta}, j^{-1}(\hat{\theta}))$$

not true in nonparametric models; Freedman
(Annals, 1999)

Insight?

- three (frequentist) test statistics asymptotically equivalent
- influence of prior “washed out”
- to compare frequentist statistics use
 - power
 - rate of convergence to null limiting distribution
 - other properties, e.g. similar on boundary
- no uniform domination, but likelihood ratio statistic generally seems preferable

4. Higher order asymptotics for likelihood

Main results

(a) p^* (density for m.l.e.)

(b) r^* (distribution function for likelihood ratio statistic)

(c) adjusted likelihood (elimination of nuisance parameters)

(d) posterior expansions (and matching priors)

...4

(a) Density of maximum likelihood estimate
(Barndorff-Nielsen, 1980, 1983; Barndorff-Nielsen
and Cox, 1994)

$$f(\hat{\theta}; \theta | a) \doteq \frac{c}{(2\pi)^{p/2}} |j(\hat{\theta})|^{1/2} \exp\{\ell(\theta; a) - \ell(\hat{\theta}; a)\}$$

$$\begin{aligned} f(\hat{\theta}; \theta|a) &\doteq \frac{c}{(2\pi)^{p/2}} \frac{L(\theta)}{L(\hat{\theta})|j(\hat{\theta})|^{-1/2}} \\ &= c \exp\{\ell(\theta)\} \frac{1}{(2\pi)^{p/2} \exp\{\ell(\hat{\theta})\}|j(\hat{\theta})|^{-1/2}} \end{aligned}$$

- renormalized likelihood function gives the density: extension of Fisher's (1934) result for location models
- saddlepoint approximation to density of minimal sufficient statistic in exponential families (easiest derivation for general models involves embedding in an exponential family)
- nearly trivial if we regard RHS as instead approximating the posterior density for θ
- conditioning on an ancillary or approximate ancillary statistic essential part of dimension reduction in frequency theory of inference

- not symmetric
- $c = c(\theta, a) = 1 + d(\theta, a)/n + O(n^{-3/2})$
- relative error $O(n^{-3/2})$
- $y \leftrightarrow (\hat{\theta}, a)$
- uses likelihood ratio as key component
- a "Laplace-type" approximation (Skovgaard, 1999; Dinges, 1986; "Wiener germs")

...4

(b) Approximation to p -value
(Barndorff-Nielsen, 1986, 1988, 1991)

$$F(r^*|a; \theta) \doteq \Phi(r^*) =$$

$$r^* = r +$$

$$-\frac{1}{2}r^2 =$$

$$q = \{\ell_{;\hat{\theta}}(\theta) - \ell_{;\hat{\theta}}(\hat{\theta})\} \{j(\hat{\theta})\}^{-1/2}$$

$$\ell_{;\hat{\theta}}(\theta) = \partial \ell(\theta; \hat{\theta}, a) / \partial \hat{\theta}$$

...4

(b) Approximation to p -value

$$F(r^*|a; \theta) \doteq \Phi(r^*) \doteq \Phi(r) + \phi(r)(1/r - 1/q)$$

$$r^* = r + \frac{1}{r} \log \frac{q}{r}$$

$$r = \pm \sqrt{[2\{\ell(\hat{\theta}) - \ell(\theta)\}]}$$

$$q = \{\ell_{;\hat{\theta}}(\theta) - \ell_{;\hat{\theta}}(\hat{\theta})\} \{j(\hat{\theta})\}^{-1/2}$$

$$\ell_{;\hat{\theta}}(\theta) = \partial \ell(\theta; \hat{\theta}, a) / \partial \hat{\theta}$$

...4

- likelihood root is the 'natural' statistic
- r is nearly normal; small adjustment makes it much closer (similar to mean and variance correction of r^2)
- 'small adjustment' $q = r + \frac{A}{\sqrt{n}}r^2 + \frac{B}{n}r^3 + \dots$
- p^* to r^* requires calculation of $\partial\ell/\partial\hat{\theta}$ with ancillary statistic held fixed
- Bayesian version will involve $\partial\ell/\partial\theta$ instead

...4

(b) Approximation to p -value
(Fraser, 1988, 1990)

$$F(r^*|a; \theta) \doteq \Phi(r^*) = \Phi(r) + \phi(r)(1/r - 1/q)$$

$$r^* = r + \frac{1}{r} \log \frac{q}{r}$$

$$r = \pm \sqrt{[2\{\ell(\hat{\theta}) - \ell(\theta)\}]}$$

$$q = \{\ell_{;V}(\theta) - \ell_{;V}(\hat{\theta})\} \{j(\hat{\theta})\}^{1/2} |\ell_{\theta;V}(\hat{\theta})|^{-1}$$

$$\begin{aligned} \ell_{;V}(\theta) &= \partial \ell(\theta; y) / \partial V(y) \\ &= \left. \frac{d}{dt} \ell(\theta; y^0 + tV) \right|_{t=0} \end{aligned}$$

...4(c) Nuisance parameters $\theta = (\psi, \lambda)$

Posterior density (marginal)

$$\pi_m(\psi|y) \doteq \frac{1}{\sqrt{(2\pi)}} \exp\{\ell_a(\psi) - \ell_a(\hat{\psi})\} \{j_a(\hat{\psi})\}^{1/2} \frac{\pi(\psi, \hat{\lambda})}{\pi(\hat{\psi}, \hat{\lambda})}$$

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

$$j_a(\psi) = -\ell''_a(\psi)$$

Posterior distribution (marginal)

$$\Pi_m(\psi|y) \doteq \Phi(r^*) \doteq \Phi\{r + r^{-1} \log(q/r)\}$$

$$= \Phi(r) + \phi(r)(1/r - 1/q)$$

$$r = \pm \sqrt{[2\{\ell_a(\hat{\psi}) - \ell_a(\psi)\}]}$$

$$q = \ell'_a(\psi) \{j_a(\hat{\psi})\}^{-1/2} \frac{\pi(\psi, \hat{\lambda}_\psi)}{\pi(\hat{\psi}, \hat{\lambda})}$$

More notation: $\theta = (\psi, \lambda)$

profile log-likelihood $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$

restricted m.l.e. $\partial \ell(\psi, \hat{\lambda}_\psi) / \partial \lambda = 0$

partitioned information $j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$

partitioned inverse $j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}$

profile information $j_p(\psi) = -\ell_p''(\psi)$

$$|j_p| = j / |j_{\lambda\lambda}|$$

adjusted log-likelihood $\ell_a(\psi) =$

$$\ell_p(\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|$$

- Alternatives to r and q are

$$r = \pm \sqrt{[2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]}$$

$$q = \ell'_p(\psi) \{j_p(\hat{\psi})\}^{-1/2} \cdot \frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{1/2}}$$

- Frequentist versions exist, derived from either marginal or conditional p^* type densities
- Frequentist versions involve sample space derivatives, as in one-dimensional case
- dimension reduction from n to k achieved by conditioning
- dimension reduction from k to 1 achieved by marginalizing

...4(c)

- adjustment to loglikelihood function to accomodate nuisance parameters
- calculations otherwise essentially the same
- need p^* like formula as starting point (dimension reduction by conditioning)
- eliminate nuisance parameters by integration
- nuisance parameter adjustment more important than distributional approximation (Pierce and Peters, 1992, 1994; Barndorff-Nielsen, 1994)

...4 (d) Posterior expansions

Matching of Bayes and frequentist inferences by choice of prior, leads to:

$$i(\theta)^{1/2}$$

when θ scalar (Welch and Peers, 1963), and to

$$i_{\lambda\lambda}^{1/2}(\psi, \lambda)g(\lambda)$$

for orthogonal nuisance parameters (Peers, 1965; Stein, 1970; Tibshirani, 1989; Nicolau, 1993)

Matching: If

$$\Pr_{\theta|Y}\{\theta \leq \theta^{(1-\alpha)}|y\} = 1 - \alpha$$

then

$$\Pr_{Y|\theta}\{\theta \leq \theta^{(1-\alpha)}(\pi, Y)\} = 1 - \alpha + O(n^{-j})$$

Alternative approach is to set $q_f = q_B$, leads to data dependent prior.

In less regular problems (normal mixtures) matching *requires* data dependent priors (Wasserman, JRSSB, 2000)

Bayesian computations can lead to frequentist results via J.K. Ghosh's *shrinkage argument*, also Dawid (1991): do expansions in the parameter space, retrieve frequentist result by letting prior shrink to point mass

5. Other applications of higher order asymptotics

- bootstrap is second order correct (DiCiccio and Efron, Stat. Sci., 1996; Davison and Hinkley, 1998)
- Edgeworth expansions for U -statistics (Bickel, Van Zwet)
- Bayesian Bartlett correction (Bickel and Ghosh, 1980, Annals)
- Bartlett identities for martingales and dual likelihood (Mykland, 1999, Annals)
- expansions of power functions (Amari; Pfanzagl)

- developments in matching priors (Mukerjee; Severini 1999)
- combining information (Fisher; Akahira and Takeuchi; Cox)
- empirical likelihood (Owen; DiCiccio, Hall and Romano)

6. Next lectures

2.

$$\begin{aligned} F(r; \theta) &\doteq \Phi(r) + \phi(r) \left(\frac{1}{r} - \frac{1}{q} \right) \\ &\doteq \Phi\left(r + \frac{1}{r} \log \frac{q}{r}\right) \end{aligned}$$

- discussion and derivation of general formula for q (Fraser/Reid; Barndorff-Nielsen)
- role in Bayesian inference and matching priors

3. (evolving)

Some more practical aspects:

- applications and implementation
- roles in statistical methods
- new developments needed