Statistical Inference

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA

1. Asymptotic Inference in Exponential Families

Let X_j be a sequence of independent, identically distributed random variables from a natural exponential family

$$f(x \mid \eta) = h(x)\, e^{\eta \cdot T(x) - A(\eta)}$$

and let q denote the dimension of T and η. We have seen that the log moment generating function for $T(X)$ is $A(\omega + \eta) - A(\eta)$ and hence that the mean and covariance for T are given respectively by

$$\mathsf{E}[T \mid \eta] = \nabla A(\eta) \qquad \mathsf{V}[T \mid \eta] = \nabla^2 A(\eta) = I(\eta), \tag{1}$$

where we recognize the Hessian of $A(\eta)$ as the Information matrix, both expected (Fisher) and observed. The likelihood function upon observing a sample of size n is

$$L_n(\eta) = \prod h(x_j)\, e^{\eta \cdot \sum T(x_j) - n\, A(\eta)},$$

so the Maximum Likelihood Estimator (MLE) $\hat{\eta}_n$ of η satisfies the equation

$$\nabla\, A(\hat{\eta}_n) = \bar{T}_n$$

Under suitable regularity conditions the Central Limit Theorem will ensure that $\bar{T}_n$ will have an asymptotically normal distribution with (by (1) mean $\nabla A(\eta)$ and covariance $I(\eta)/n$, so

$$\sqrt{n}\big[\nabla A(\hat{\eta}_n) - \nabla A(\eta)\big]$$

will have an asymptotical $\mathsf{No}(0, I(\eta))$ distribution. By Taylor's Theorem we can write

$$\begin{aligned}\big[\nabla A(\hat\eta_n) - \nabla A(\eta)\big] &= \nabla^2 A(\eta)(\hat\eta_n - \eta) + O(|\hat\eta_n - \eta|^2)\\ &= I(\eta)(\hat\eta_n - \eta) + O(1/n),\end{aligned}$$

from which we conclude

$$\sqrt{n}(\hat\eta_n - \eta) \sim \mathsf{No}\big(0, I(\eta)^{-1}\big)$$

or, more casually, that $\hat\eta_n$ has approximately a $\mathsf{No}\big(\eta, [n\, I(\eta)]^{-1}\big)$ distribution—so the Maximum Likelihood Estimator is consistent and efficient and asymptotically normal.

1.1. Unnatural Families

If we parametrize by some $\theta \in \Theta$ other than the natural parameter η, but still have a smooth mapping $\theta \mapsto \eta(\theta)$, we can note that $\hat\eta_n = \eta(\hat\theta_n)$, and (again, by Taylor)

$$\eta(\hat\theta_n) = \eta(\theta) + J'\,(\hat\theta_n - \theta) + O(|\hat\theta_n - \theta|^2),$$

where the Jacobian matrix J is given by $J_{ij} = \partial\eta_j/\partial\theta_i$, so

$$\sqrt{n}(\hat\theta_n - \theta) \sim \mathsf{No}\big(0, [J' I(\eta) J]^{-1}\big) = \mathsf{No}\big(0, I(\theta)^{-1}\big)$$

and again we have consistency, efficiency, and asymptotic normality. In one-dimensional families this leads to Frequentist confidence intervals of the form

$$1 - \alpha \approx \Pr\left[\theta \in \big(\hat\theta_n - \frac{Z_{\alpha/2}}{\sqrt{n\, I(\hat\theta_n)}},\quad \hat\theta_n + \frac{Z_{\alpha/2}}{\sqrt{n\, I(\hat\theta_n)}}\big) \mid \theta\right]$$

where the "$\approx$" is required both because the distribution of $\hat\theta_n$ is only approximately normal, and because $I(\hat\theta_n)$ is only approximately $I(\theta)$.

2. Bayesian Asymptotic Inference

The *observed information* for a single observation $X = x$ from the model $X \sim f(x|\theta)$ is

$$i(\theta, x) \quad = \quad -\nabla^2_\theta \log f(x \mid \theta);$$

evidently the Fisher (expected) information is related to this by $I(\theta) = \mathsf{E}[i(\theta, X) \mid \theta]$. The likelihood for a sample of size n is just the product of the individual likelihoods, leading to a *sum* for the *log* likelihoods, and observed information

$$i(\theta, x) \quad = \quad \sum_{j=1}^{n} i(\theta, x_j).$$

If the log likelihood $\log f_n(x|\theta)$ is differentiable throughout Θ and attains a unique maximum at an interior point $\hat{\theta}_n(x) \in \Theta$, then we can expand $\log f_n(x|\theta)$ in a second-order Taylor series for $\theta = \hat{\theta}_n(x) + \epsilon/\sqrt{n}$ close to $\hat{\theta}_n(x)$ to find (for $q = 1$ dimensional θ)

$$\begin{aligned}
\log f_n(x|\theta) \quad &= \quad \log f_n(x|\hat{\theta}_n) + \frac{(\epsilon/\sqrt{n})^1}{1!} \nabla_\theta \log f_n(x|\hat{\theta}_n) + \frac{(\epsilon/\sqrt{n})^2}{2!} \nabla_\theta^2 \log f_n(x|\hat{\theta}_n) \\
&\qquad + o\big((\epsilon/\sqrt{n})^2 |\nabla_\theta^2 \log f_n(x|\hat{\theta}_n)|\big) \\
&= \quad \log f_n(x|\hat{\theta}_n) + 0 - \frac{\epsilon^2}{2} \frac{1}{n} \sum_{j=1}^{n} i(\hat{\theta}_n, x_j) + o(1) \\
&\to \quad \log f_n(x|\hat{\theta}_n) - \frac{\epsilon^2}{2} \mathsf{E}[i(\hat{\theta}_n, X)] \\
&= \quad \log f_n(x|\hat{\theta}_n) - \frac{\epsilon^2}{2} I(\hat{\theta}_n) \\
&= \quad \log f_n(x|\hat{\theta}_n) - \frac{1}{2} n\, I(\hat{\theta}_n)(\theta - \hat{\theta}_n)^2,
\end{aligned}$$

where we have used the consistency of $\hat{\theta}_n$ and have applied the strong law of large numbers for $i(\theta, X)$. Thus we have the likelihood approximation $f_n(x|\theta) \approx \mathsf{No}(\hat{\theta}_n(x), nI(\hat{\theta}_n))$, normal with mean the MLE $\hat{\theta}_n(x)$ and *precision* $nI(\hat{\theta}_n)$ (or covariance $\frac{1}{n} I(\hat{\theta}_n)^{-1}$).

Note that the prior distribution is irrelevent, asymptotically, so long as it is smooth and doesn't vanish in a neighborhood of $\hat{\theta}_n$; thus we have the *Bayesian Central Limit Theorem*,

$$\pi_n(\theta \mid x) \approx \mathsf{No}(\hat{\theta}_n, [n\, I(\hat{\theta}_n)]^{-1}),$$

leading in the $q = 1$-dimensional case to Bayesian credible intervals of the form

$$1 - \alpha \approx \Pr\Big[\theta \in \big(\hat{\theta}_n - \frac{Z_{\alpha/2}}{\sqrt{n\, I(\hat{\theta}_n)}}, \quad \hat{\theta}_n + \frac{Z_{\alpha/2}}{\sqrt{n\, I(\hat{\theta}_n)}}\big) \mid x\Big],$$

just as before, but now with a completely different interpretation.

3. An Example

Let X_i be Bernoulli random variables with $\mathsf{P}[X_i = 1] = \theta$ for some $\theta \in \Theta = (0, 1)$; this is an exponential family with natural parameter $\eta = \eta(\theta) = \log \frac{\theta}{1-\theta}$ and natural sufficient statistic (for a sample of size n) $T_n = \Sigma X_j$, hence with log normalizing factor $A(\eta) = \log(1 + e^\eta) = B(\theta) = -\log(1 - \theta)$, i.e. with likelihood

$$L(\theta) = \prod p^{T_n}(1-p)^{n-T_n} = e^{\eta T_n - n\log(1+e^\eta)}$$

and hence with MLE's and (1×1) information matrices

$$\begin{aligned} \hat\theta_n &= T_n/n = \bar X_n & I^\theta &= \tfrac{1}{\theta(1-\theta)} \\ \hat\eta_n &= \log \tfrac{T_n}{n-T_n} = \log \tfrac{\bar X_n}{1-\bar X_n} & I^\eta &= \tfrac{e^\eta}{(1+e^\eta)^2}, \end{aligned}$$

so $I^\eta(\hat\eta_n) = T_n(n - T_n)/n^2$, $I^\theta(\hat\theta_n) = n^2/(T_n(n - T_n))$, and 95% confidence intervals would be

$$\begin{aligned} 0.95 &\approx \Pr[\hat\eta_n - 1.96/\sqrt{nI(\hat\eta_n)} < \eta < \hat\eta_n + 1.96/\sqrt{nI(\hat\eta_n)}] \\ &= \Pr[\log \frac{T_n}{n-T_n} - \frac{1.96}{\sqrt{T_n(n-T_n)/n}} < \eta < \log \frac{T_n}{n-T_n} + \frac{1.96}{\sqrt{T_n(n-T_n)/n}}] \end{aligned}$$

This can be written as an interval $0.95 = \Pr[L_n < \theta < R_n]$ for $\theta = e^\eta/(1 + e^\eta)$, with left and right endpoints

$$\begin{aligned} L_n(T_n) &= T_n/\big[T_n + (n - T_n)\exp\big(+1.96/\sqrt{T_n(n-T_n)/n}\big)\big] \\ R_n(T_n) &= T_n/\big[T_n + (n - T_n)\exp\big(-1.96/\sqrt{T_n(n-T_n)/n}\big)\big]; \end{aligned}$$

for examle, with $T_{100} = 10$ successes in $n = 100$ tries, the endpoints are $L_{100}(10) = 10/[10 + 90\exp(1.96/\sqrt{9})] = 1/[1 + 9\exp(0.6533)] = 0.05465$ and $R_{100}(10) = 1/[1 + 9\exp(-0.6533)] = 0.17597$, while with $T_{100} = 50$ the interval endpoints would be are $L_{100}(50) = 50/[50 + 50\exp(1.96/\sqrt{25})] = 1/[1+\exp(0.392)] = 0.40324$ and $R_{100}(50) = 1/[1+\exp(-0.392)] = 0.59676$. Intervals for θ can be made directly using

$$\begin{aligned} 0.95 &\approx \Pr[\hat\theta_n - 1.96/\sqrt{nI(\hat\theta_n)} < \theta < \hat\theta_n + 1.96/\sqrt{nI(\hat\theta_n)}] \\ &= \Pr[\hat\theta_n - 1.96\sqrt{\hat\theta_n(1-\hat\theta_n)/n} < \theta < \hat\theta_n + 1.96\sqrt{\hat\theta_n(1-\hat\theta_n)/n}], \end{aligned}$$

an interval with endpoints $L_{100}(10) = 0.10 - 1.96\sqrt{0.09/100} = 0.10 - 1.96 \times 0.30 = 0.0412$, $R_{100}(10) = 0.10 + 1.96 \times 0.03 = 0.1588$ for $T_{100} = 10$, and $L_{100}(50) = 0.50 - 1.96\sqrt{0.25/100} = 0.10 - 1.96 \times 0.05 = 0.402$, $R_{100}(50) = 0.50 + 1.96 \times 0.05 = 0.598$ for $T_{100} = 50$. Recall that the *exact* 95% confidence intervals for θ are `qbeta(0.025,` T_n`,n-`T_n`+1), qbeta(0.975,` T_n`+1,n-`T_n`)` in general, or `[qbeta(0.025, 10,91), qbeta(0.975, 11,90)] =[0.0490, 0.1762]` for $T_{100} = 10$ and `[qbeta(0.025, 50,51), qbeta(0.975, 51, 50)] = [0.3983, 0.6017]` for $T_{100} = 50$. In summary,

	$L_{100}(10)$	$R_{100}(10)$	$L_{100}(50)$	$R_{100}(50)$
Normal, based on η:	[0.05465,	0.17597]	[0.40324,	0.59676]
Normal, based on θ:	[0.04120,	0.15880]	[0.40200,	0.59800]
Exact Frequentist:	[0.04900,	0.17622]	[0.39832,	0.60168]
Exact Bayes (1.0,1.0):	[0.05564,	0.17456]	[0.40364,	0.59636]
Exact Bayes (0.5,0.5):	[0.05258,	0.17012]	[0.40317,	0.59683]
Exact Bayes (0.0,0.0):	[0.04951,	0.16557]	[0.40270,	0.59730]

Evidently the normal approximations to the Frequentist intervals are both rather close, but a bit too narrow (hence cover θ with a bit less than the promissed 95% probability). Of course the approximations improve with increasing n and become worse for smaller n. The intervals based on the natural parameter are very close to the Bayesian credible intervals. All the approximations are better near the middle of the range ($\theta \approx 1/2$) than near the endpoints.