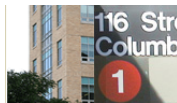


Bayesian inference and accurate approximation

Nancy Reid

November 24, 2008



Don Fraser, Ana-Maria Staicu, Ye Sun, Grace Yun-Yi



Bayesian asymptotics

Formulas

Example 1

Frequentist asymptotics

Normal circle

Example 2

Default priors

Well-calibrated priors

Approximate location models

Conclusions

Approximate posterior

$$\Pr_m(\Psi \leq \psi \mid \mathbf{y}) \doteq \Phi(r_B^*) = \Phi\left(r + \frac{1}{r} \log \frac{q_B}{r}\right)$$

$$Y \sim f(\mathbf{y}; \theta) \quad Y = (Y_1, \dots, Y_n) \quad \text{model}$$

$$\ell(\theta) = \ell(\psi, \lambda) = \log f(\mathbf{y}; \psi, \lambda), \quad \mathbf{y} \in R^n \quad \text{log-likelihood}$$

$$r = \pm[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2}, \quad \psi \in R \quad \text{likelihood root}$$

$$\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi) \quad \text{constrained m.l.e.}$$

$$\pi_m(\psi \mid \mathbf{y}) = \int \pi(\theta \mid \mathbf{y}) d\lambda$$

$$\propto \int \exp \ell(\theta) \pi(\theta) d\lambda \quad \text{marginal posterior}$$



... approximate posterior

$$\Pr_m(\Psi \leq \psi \mid \mathbf{y}) \doteq \Phi(r_B^*) = \Phi\left(r + \frac{1}{r} \log \frac{q_B}{r}\right)$$

$$q_B = -\ell'_p(\psi) j_p(\hat{\psi})^{-1/2} \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}$$

$$\ell_p(\psi) = \ell(\hat{\theta}_\psi)$$

profile log-likelihood

$$j(\theta) = -\ell''(\theta; \mathbf{y})$$

observed information

$$j(\theta) = \begin{bmatrix} j_{\psi\psi}(\theta) & j_{\psi\lambda}(\theta) \\ j_{\lambda\psi}(\theta) & j_{\lambda\lambda}(\theta) \end{bmatrix}$$

partitioned matrix



... approximate posterior

$$r_B^* \sim N(0, 1)$$

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

r is the likelihood root

q_B is an adjusted score statistic

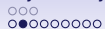
The approximation is very good!
Relative error $O(n^{-3/2})$

Example: Normal circle

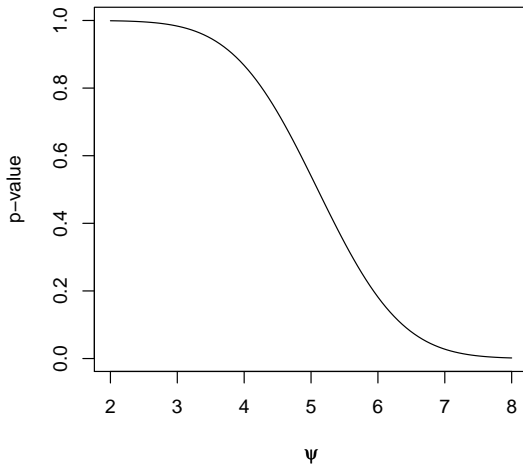
- ▶ $y_1 \sim N(\mu_1, 1/n), \dots, y_k \sim N(\mu_k, 1/n)$
- ▶ parameter of interest $\psi = (\mu_1^2 + \dots + \mu_k^2)^{1/2} = \|\mu\|$
- ▶ prior $\pi(\mu) = 1$
- ▶ Exact marginal posterior $\Pr\{\chi_k^2(n\|y\|^2) \geq n\psi^2\}$
- ▶ Third order

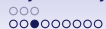
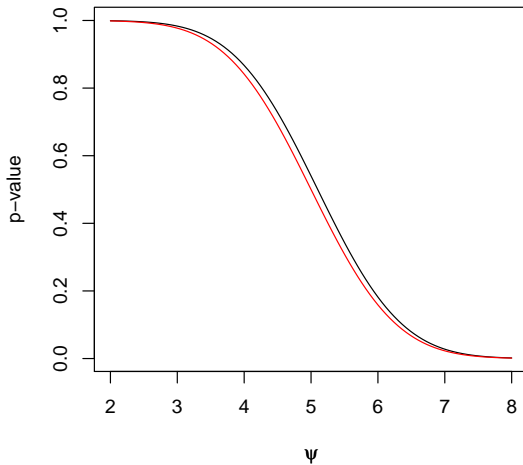
$$r_B^* = \sqrt{n}(\hat{\psi} - \psi) + \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\hat{\psi}}{\psi} \right)^{(k-1)/2} \right\} \sim N(0, 1)$$

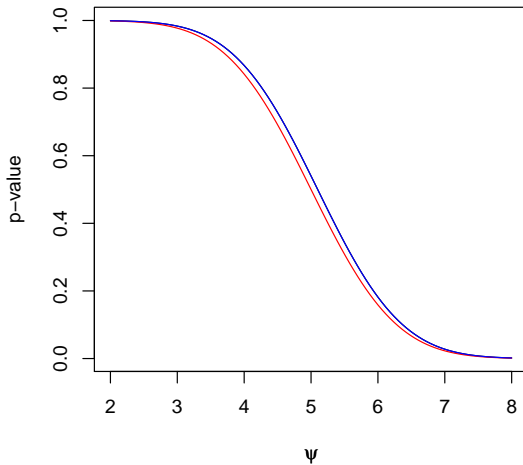
- ▶ Normal approximation to posterior $\sqrt{n}(\hat{\psi} - \psi) \sim N(0, 1)$

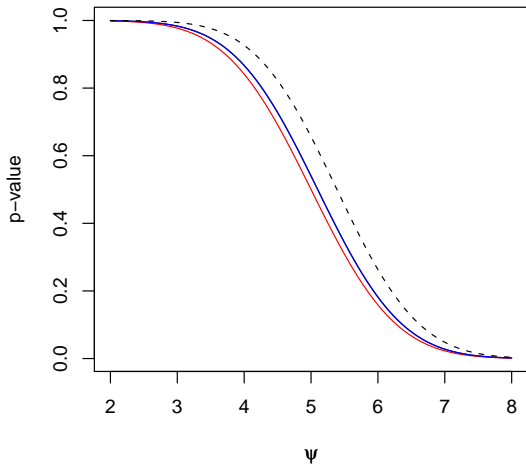


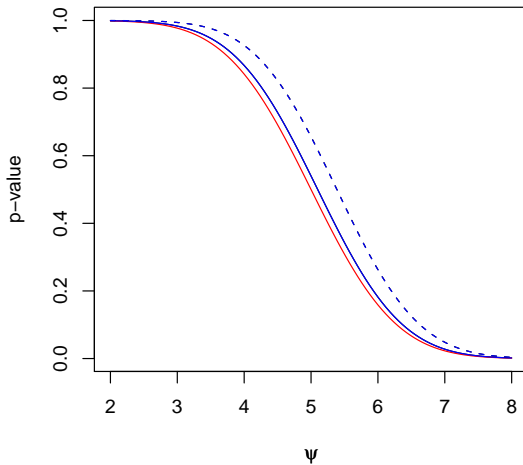
normal circle, $k=2$

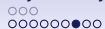
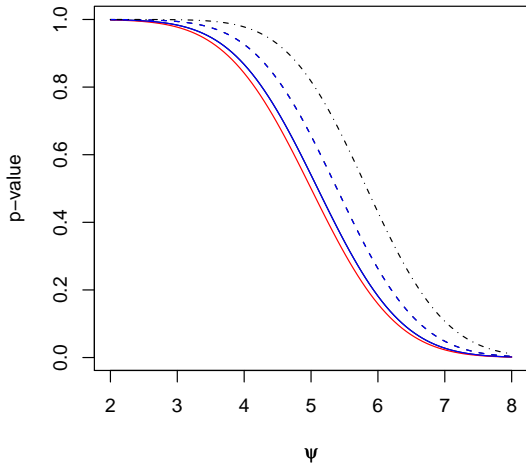


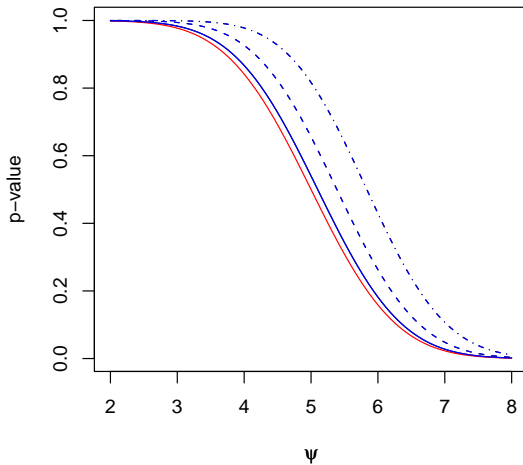
normal circle, $k=2$ 

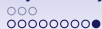
normal circle, $k=2$ 

normal circle, $k = 2, 5, 10$ 

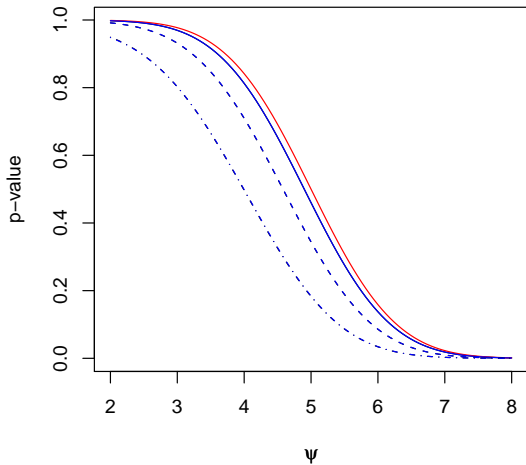
normal circle, $k = 2, 5, 10$ 

normal circle, $k = 2, 5, 10$ 

normal circle, $k = 2, 5, 10$ 



normal circle, $k = 2, 5, 10$



Normal circle

- ▶ exact:

$$\text{Pvalue}(\psi) = \Pr\{\chi_k^2(n\psi^2) \geq n\|y\|^2\}$$

- ▶ approx:

$$\text{Pvalue}(\psi) \doteq \Phi(r_F^*) = \Phi\left(r + \frac{1}{r} \log \frac{q_F}{r}\right)$$

- ▶

$$r_F^* = \sqrt{n}(\hat{\psi} - \psi) - \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\hat{\psi}}{\psi} \right)^{(k-1)/2} \right\} \sim N(0, 1)$$

- ▶ Bayes:

$$r_B^* = \sqrt{n}(\hat{\psi} - \psi) + \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\hat{\psi}}{\psi} \right)^{(k-1)/2} \right\} \sim N(0, 1)$$

- ▶ $r_B^* - r_F^* \sim \frac{k-1}{\psi\sqrt{n}}$

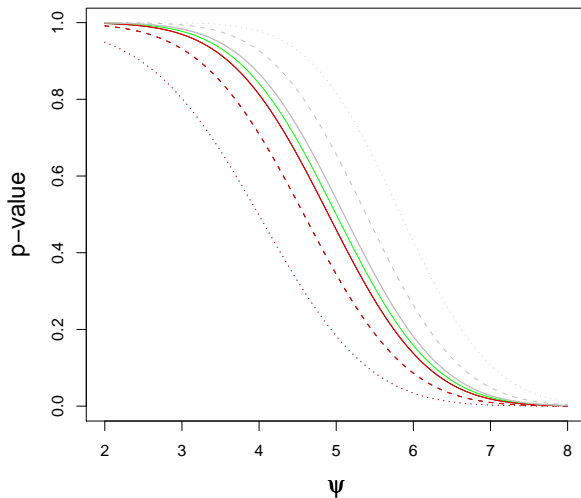
Frequentist P-value

$$\text{Pvalue}(\psi) \doteq \Phi(r_F^*) = \Phi\left(r + \frac{1}{r} \log \frac{q_F}{r}\right)$$

$$r = \pm [2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2}$$

$$q_F = \frac{|\ell_{;\nu}(\hat{\theta}) - \ell_{;\nu}(\theta) \quad \ell_{\lambda;\nu}(\hat{\theta}_\psi)|}{|\ell_{\theta;\nu}(\hat{\theta})|} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}$$

$$q_B = -\ell'_p(\psi) j_p(\hat{\psi})^{-1/2} \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}$$

Normal Circle, $k=2, 5, 10$ 

Normal circle

- ▶ frequentist:

$$r_F^* = \sqrt{n}(\hat{\psi} - \psi) - \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\hat{\psi}}{\psi} \right)^{(k-1)/2} \right\} \sim N(0, 1)$$

- ▶ Bayesian:

$$r_B^* = \sqrt{n}(\hat{\psi} - \psi) + \frac{1}{\sqrt{n}(\hat{\psi} - \psi)} \log \left\{ \left(\frac{\hat{\psi}}{\psi} \right)^{(k-1)/2} \right\} \sim N(0, 1)$$

- ▶

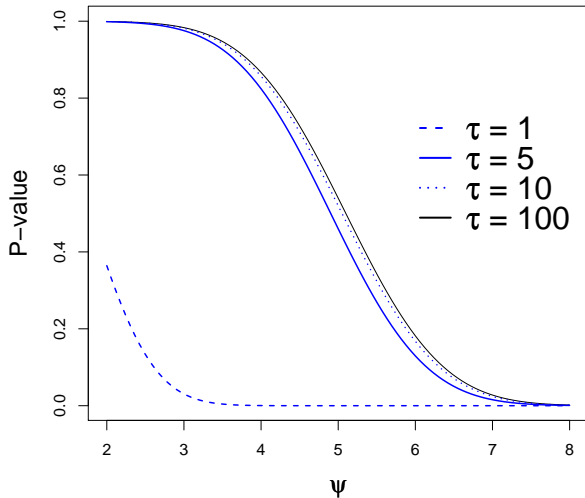
$$r_B^* = r_F^* \iff \pi(\mu) d\mu \propto \|\mu\|^{-(k-1)} d\mu \quad (\psi = \|\mu\|)$$

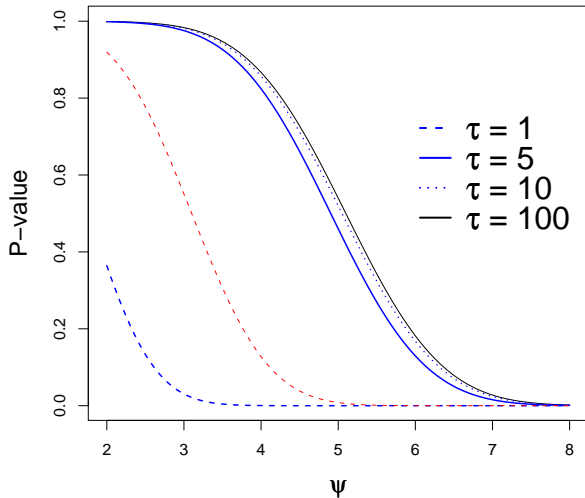
- ▶ simple hierarchical priors on μ don't fix this: R & Sun/08

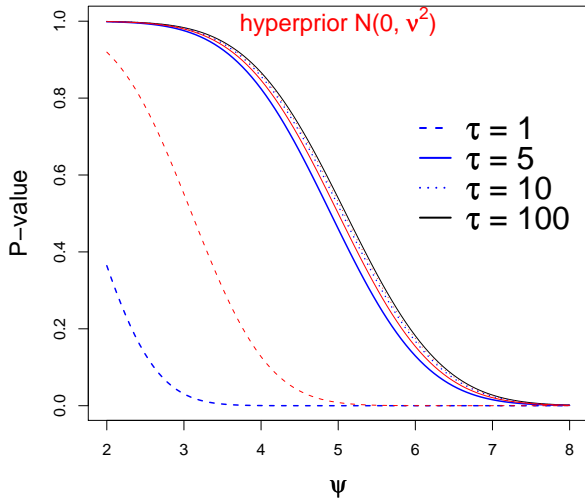
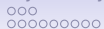


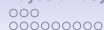
Normal circle: hierarchical priors

- ▶ (i) : $\mu_j \sim N(0, \tau^2)$
- ▶ (ii): $\mu_j \sim N(a, \tau^2), \quad a \sim N(0, \nu^2)$









Logistic regression

- ▶ $\log\{\Pr(Y_i = 1) / \Pr(Y_i = 0)\} = \alpha + \beta x_i$
- ▶ $\pi(\theta) d\theta \propto d\alpha d\beta$
- ▶ $\psi = -\alpha/\beta = ED_{50}$

method	lower	central	upper
$\Phi(r)$	0.0315	0.9254	0.0431
$\Phi(r_F^*)$	0.0232	0.9461	0.0306
$\Phi(r_B^*)$	0.0471	0.8992	0.0674

$n=15, \psi = 1$
10,000 simulations

Example: logistic regression

$$y_i \sim \text{Bernoulli}(p_i), \quad \text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_6 x_{6i}$$

$$\psi = \Pr(Y = 1 \mid x^*)$$

method	lower	central	upper
$\Phi(r)$	0.0228	0.9278	0.0494
$\Phi(r_F^*)$	0.0296	0.9511	0.0192
$\Phi(r_{B1}^*)$	0.0177	0.8722	0.1101
$\Phi(r_{B2}^*)$	0.0274	0.9509	0.0217

simulations based on data from Davison & Hinkley (1997);

$\psi \simeq 0.21$; 10,000 simulations

B1: flat prior on β

B2: matching prior for ψ (guarantees frequentist coverage)

Well-calibrated priors

- ▶ can we find priors that are guaranteed to be well-calibrated, at least approximately
- ▶ what structure do such priors need to have
- ▶ can we do this for all component parameters simultaneously
- ▶ matching priors using Edgeworth expansions (Welch & Peers, 1963; Peers, 1965; Tibshirani (1989), ...)
- ▶ $\pi(\theta)d\theta \propto i_{\psi\psi}^{1/2}(\theta)g(\lambda)$: “expected matching”
Staicu/R 2008
- ▶ $r_F^* = r_B^* \implies \pi(\hat{\theta})/\pi(\hat{\theta}_\psi) \propto \dots$: “strong matching”
FR 2002

Well-calibrated priors

- ▶ can we find priors that are guaranteed to be well-calibrated, at least approximately
- ▶ what structure do such priors need to have
- ▶ can we do this for all component parameters simultaneously
- ▶ matching priors using Edgeworth expansions (Welch & Peers, 1963; Peers, 1965; Tibshirani (1989), ...)
- ▶ $\pi(\theta)d\theta \propto i_{\psi\psi}^{1/2}(\theta)g(\lambda)$: “expected matching”
Staicu/R 2008
- ▶ $r_F^* = r_B^* \implies \pi(\hat{\theta})/\pi(\hat{\theta}_\psi) \propto \dots$: “strong matching”
FR 2002



Approximate location models

- ▶ Location model: $Y \sim f(y - \theta) \implies \pi(\theta)d\theta \propto d\theta$
- ▶ Location model: $\theta \rightarrow \theta + d\theta, \quad y \rightarrow y + d\theta$
- ▶ $F(y; \theta)$ unchanged, i.e. $dF(y; \theta) = 0$

- ▶ General model $Y \sim f(y; \theta)$ **require** $dF(y^0; \theta) = 0$
- ▶ $F_y(y^0; \theta)dy + F_{;\theta}(y^0; \theta)d\theta = 0$ (scalar or vector θ)
- ▶

$$dy = -\frac{F_{;\theta}(y^0; \theta)}{F_y(y^0; \theta)}d\theta = V(\theta)d\theta$$

- ▶ $y_1, \dots, y_n : V(\theta) = \begin{bmatrix} V_1(\theta) \\ \vdots \\ V_n(\theta) \end{bmatrix}$ $n \times p$ matrix

Default prior



$$dy = V(\theta)d\theta$$



$$\ell_{\theta}(\hat{\theta}; y) = 0 \implies \ell_{\theta\theta}(\hat{\theta}; y)d\hat{\theta} + \ell_{\theta;y}(\hat{\theta}; y)dy = 0$$



$$d\hat{\theta} = \hat{j}^{-1}Hdy$$

- ▶ proposed default prior

$$\pi(\theta)d\theta \propto |\hat{j}^{-1}HV(\theta)|d\theta$$



Example $y \sim N(X\beta, \sigma^2)$

▶ $\theta = (\beta, \sigma)$: length $p = q + 1$

▶ $V(\hat{\theta}) = \begin{pmatrix} X & \frac{y^0 - X\hat{\beta}}{\hat{\sigma}} \end{pmatrix}$

▶ *design* *residuals*

▶ $V(\theta) = \begin{pmatrix} X & \frac{y^0 - X\beta}{\sigma} \end{pmatrix}$

▶ $\pi(\theta)d\theta \propto d\beta d\sigma/\sigma$

▶ $y \sim N(X\beta, \Sigma)$



Targetted priors

- ▶ If parameter of interest is **curved**, then prior needs to be targetted on the parameter of interest
- ▶ marginalization paradox (Dawid, Stone and Zidek)
- ▶ proposal: $d\hat{\theta} = \hat{j}^{-1}HV(\theta)d\theta = W(\theta)d\theta$, say
- ▶ $W(\theta) = (W_\psi(\theta), W_\lambda(\theta))$
- ▶ if ψ is fixed then $\pi(\lambda | \psi)d\lambda \propto |W_\lambda(\theta)|d\lambda$
- ▶ composite prior, targetted on ψ



$$\pi_\psi(\psi, \lambda)d\psi d\lambda \propto |W_\psi(\psi, \hat{\lambda}_\psi)| |W_\lambda(\psi, \lambda)|d\psi d\lambda$$

- ▶ Example: linear regression, leads back to $d\beta d\sigma/\sigma$ when targetting on subvector of β

... well-calibrated priors

- ▶ can we find priors that are guaranteed to be well-calibrated, at least approximately
- ▶ what structure do such priors need to have
- ▶ can we do this for all component parameters simultaneously
- ▶ matching priors using Edgeworth expansions (Welch & Peers, 1963; Peers, 1965; Tibshirani (1989), ...)
- ▶ $\pi(\theta)d\theta \propto i_{\psi\psi}^{1/2}(\theta)g(\lambda)$: “expected matching”
Staicu/R 2008
- ▶ $r_F^* = r_B^* \implies \pi(\hat{\theta})/\pi(\hat{\theta}_\psi) \propto \dots$: “strong matching”
FR 2002

Strong matching

▶ $s(\psi) = \Phi(r + \frac{1}{r} \log \frac{q_B}{r})$: Bayesian survivor value

▶ $p(\psi) = \Phi(r + \frac{1}{r} \log \frac{q_F}{r})$: Frequentist p -value

$$\text{▶ } q_B = -\ell'_p(\psi) j_p(\hat{\psi})^{-1/2} \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)}$$

$$\text{▶ } q_F = \frac{|\ell_{;\nu}(\hat{\theta}) - \ell_{;\nu}(\theta) \quad \ell_{\lambda;\nu}(\hat{\theta}_\psi)|}{|\ell_{\theta;\nu}(\hat{\theta})|} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}$$

$$\text{▶ } q_B = q_F \Leftrightarrow \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} = \dots$$

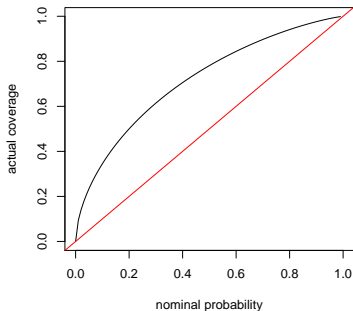
▶ default prior along the curve $\theta = \hat{\theta}_\psi$ F&R, 2002

▶ extend to full parameter space using location model approximation



Location based priors not calibrated

- ▶ Example $y_i \sim N(\mu_i, 1)$, $\psi = \|\mu\|$, default prior $\propto d\mu$
- ▶ Posterior probability limits for ψ **do not** have frequentist coverage





... not calibrated

- ▶ Cox and Hinkley, 1974, after Mitchell, 1969:

$$E(y | x) = \alpha + \beta(1 - x)^\rho, \quad x \in (0, 1)$$
- ▶ Bayesian probit regression (Jones, 2008; Siddhartha and Chib, 1984): $p(y = 1) = \Phi(\alpha + \beta x_i)$: flat priors on α, β
- ▶ hierarchical Poisson models (Gelman et al, 2007):

$$E(y_{ij}) = c_0 x_{ij} \exp(\mu + \alpha_i + \beta_j + \gamma_{ij})$$
- ▶ “non-informative uniform priors on $\mu, \underline{\alpha}, \sigma_\beta, \sigma_\gamma$ ”



Conclusions

- ▶ calibrated priors are data dependent
- ▶ focus motivated by asymptotic theory for likelihood inference
- ▶ **reference priors** use a different asymptotic approach based on K-L divergence from prior to posterior
- ▶ also need to target on parameter of interest
- ▶ also complicated to calculate
- ▶ **marginalization to curved parameters using flat priors may lead to poorly calibrated inferences**
- ▶ checking sensitivity to prior specification can be done simply using asymptotic approximation $\Phi(r_B^*)$

Some references

- ▶ Fraser, D.A.S. and Reid, N. (2002). Strong matching of frequentist and Bayesian inference. *J. Statist. Plan. Infer.* **103**, 263–285.
- ▶ Staicu, A.M. and Reid, N. (2008). Uniqueness of matching priors. *Canad. J. Statist.*, to appear.
- ▶ Datta, G.S. and Ghosh, M. (1995). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.
- ▶ Dawid, A.P., Stone, M., and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189–233.
- ▶ Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion).
- ▶ Reid, N. and Sun, Y. (2008). Assessing sensitivity to priors using higher order approximations. *Commun. Statist. Th. Methods*, to appear.