

Likelihood inference in the presence of nuisance parameters

Nancy Reid, University of Toronto

www.utstat.utoronto.ca/reid/research

1. Notation, Fisher information, orthogonal parameters
2. Likelihood inference with no nuisance parameters; first and third order
3. Profile log-likelihood
4. Adjustments to profile log-likelihood
5. Third order p -values
6. Model classes

1. Notation...

- model $Y \sim f(y; \theta), \theta \in R^d$ $\theta = (\psi, \lambda)$
- likelihood $L(\theta) = L(\theta; y) = f(y; \theta), \ell(\theta)$
- i.i.d. sampling $\underline{y} = (y_1, \dots, y_n)$ $L(\theta; \underline{y}) =$
- m.l.e. $\sup_{\theta} L(\theta) = L(\hat{\theta})$
- observed information $j(\hat{\theta}) = -\ell''(\hat{\theta})$
- expected information $i(\theta) = E\{-\ell''(\theta)\}$
- partitioned information $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}$
- partitioned inverse $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix}$

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}$$

ψ is orthogonal to λ if $i_{\psi\lambda}(\theta) = 0$

implies in particular that $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent

Example: ratio of Poisson means

$y_1 \sim Po(\lambda), y_2 \sim Po(\psi\lambda)$:

$$L(\psi, \lambda; y_1, y_2) = e^{-2\lambda - \psi y_2} \lambda^{y_1 + y_2}$$

in fact $= L_1(\psi; y_2) L_2(\lambda; y_+)$, stronger than orthogonality

Example: exponential regression

y_i follows an exponential distribution

$$E(y) = \lambda \exp(-\psi x_i); \sum x_i = 0$$

$$\ell(\psi, \lambda; \underline{y}) = -n \log \lambda + \lambda \sum y_i \exp(-\psi x_i)$$

Likelihood inference with no nuisance parameters

- Plot the likelihood
- $\hat{\theta}$ is asymptotically normal, mean θ
variance $i^{-1}(\theta)$
- $r(\theta) = \pm[2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2}$

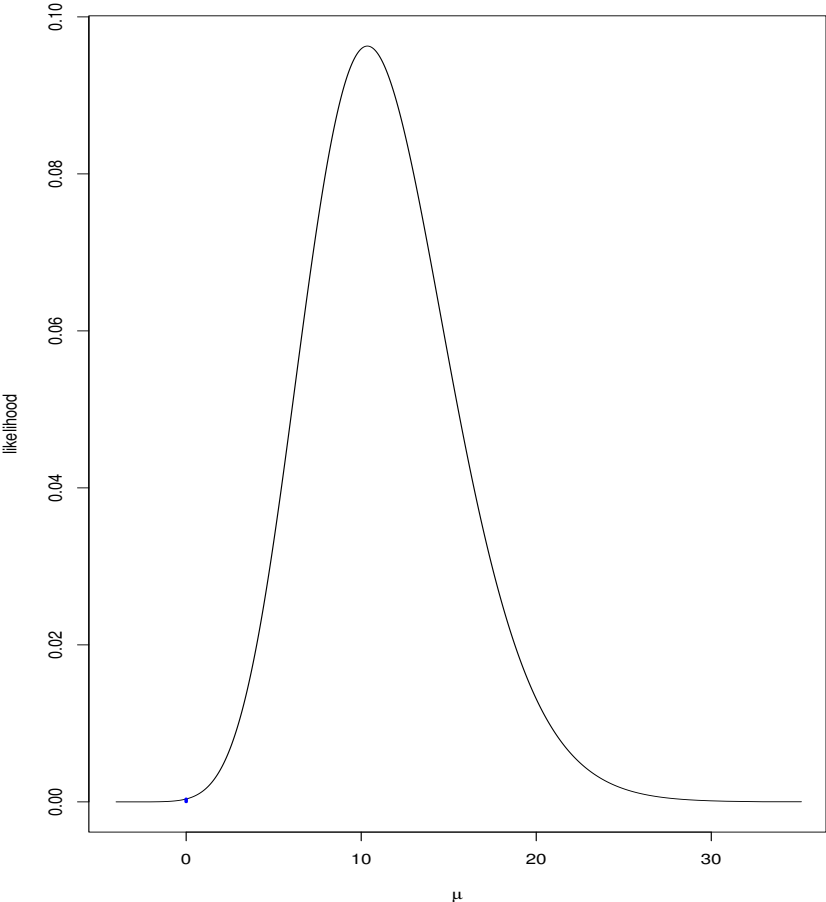
is asymptotically $N(0, 1)$ (better)

- $r^*(\theta) = r(\theta) + \frac{1}{r(\theta)} \log \left\{ \frac{q(\theta)}{r(\theta)} \right\}$

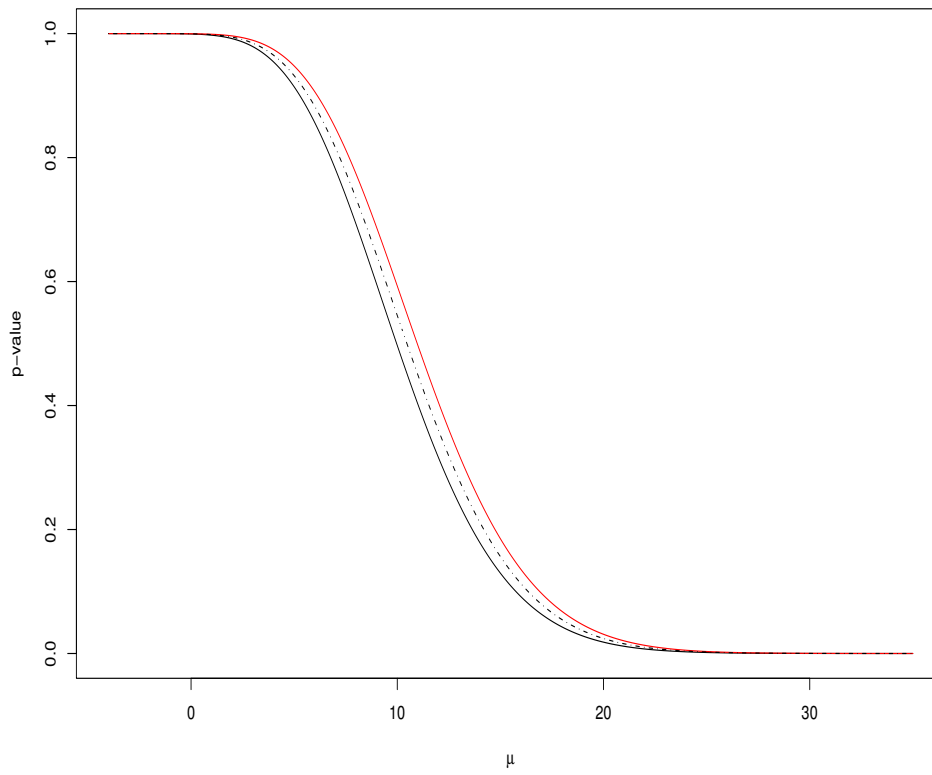
is asymptotically $N(0, 1)$ (even better)

Example: $Y \sim Po(\theta), \theta > b, b$ known;

$b = 6.7, y = 17$:

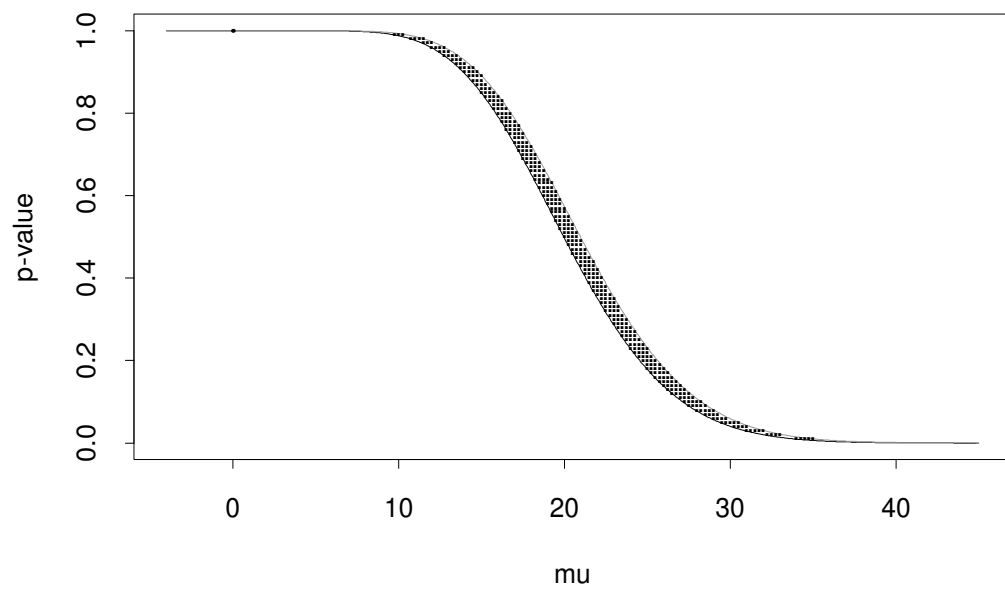
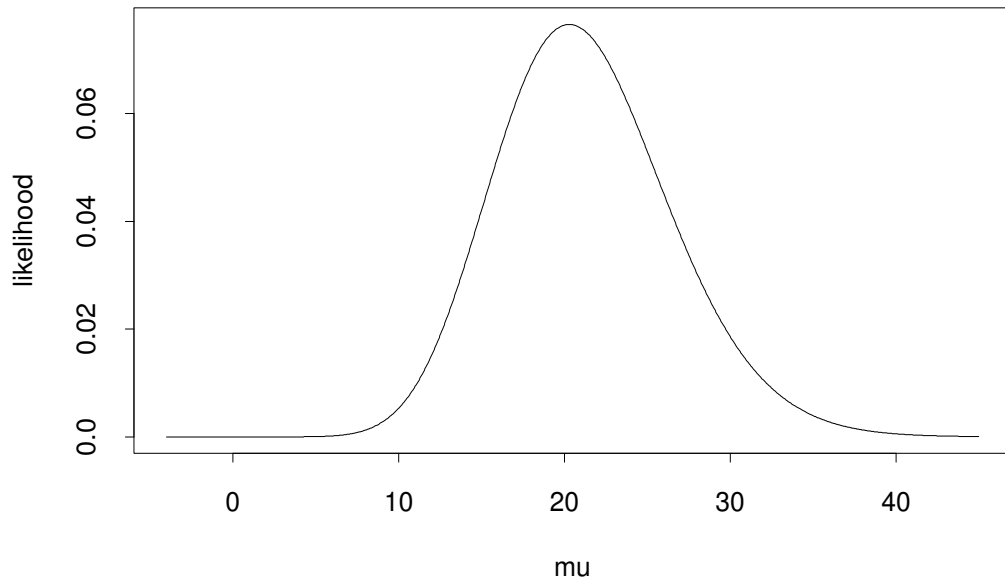


Fraser, Reid, Wong 2003



p-values:

upper	0.0005993
lower	0.0002170
mid	0.0004081
r^*	0.0003779
r	0.0004416
$\hat{\theta}$	0.0062427



Nuisance parameters: profile likelihood

$$\theta = (\psi, \lambda)$$

restricted m.l.e. $\hat{\lambda}_\psi: \sup_\lambda L(\psi, \lambda)$

$$L_p(\psi) = L(\psi, \hat{\lambda}_\psi) \text{ (concentrated likelihood)}$$

for λ of fixed dimension, i.i.d. sampling y :

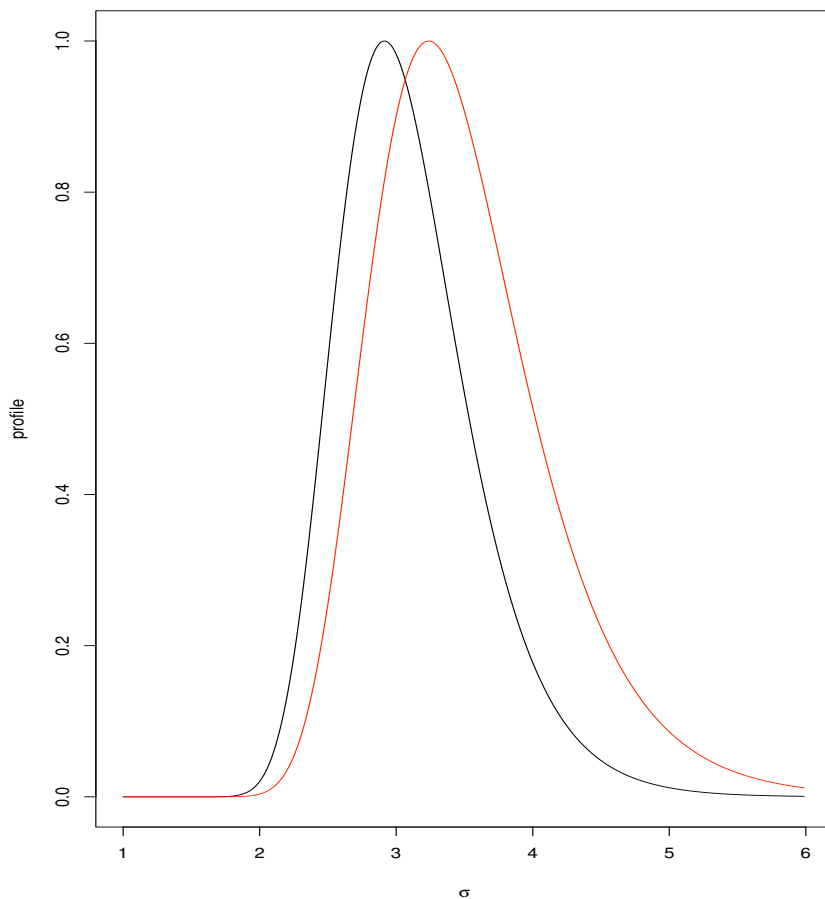
- $\sup_\psi L_p(\psi) = L(\hat{\psi}, \hat{\lambda})$
- $r_p(\psi) = \pm [2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2} \xrightarrow{d} N(0, 1)$
- $\hat{\psi}$ asymptotically normal with mean ψ and variance consistently estimated by $\{-\ell_p''(\hat{\psi})\}^{-1} = j^{\psi\psi}(\hat{\psi}, \hat{\lambda})$

But, profile likelihood can be too concentrated, and maximized at 'wrong' point:

Example: linear regression

$$y_i = \underline{x}'_i \beta + \epsilon_i, \quad \underline{x}_i = (x_{i1}, \dots, x_{ip})$$

$$\epsilon_i \sim N(0, \psi) \quad \hat{\psi} = \frac{1}{n} \sum (y_i - \underline{x}'_i \hat{\beta})^2$$



Adjustments to profile log-likelihood

If ψ is orthogonal to λ :

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

$$j(\theta) = -\ell''(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$$

ℓ_p is $O_p(n)$, $\log |j|$ is $O_p(1)$

Example: product of exponential means

$$y_{1i} \sim \text{Exp}(\psi \lambda_i) \quad y_{2i} \sim \text{Exp}(\psi / \lambda_i), i = 1, \dots, n$$

$$\hat{\psi} \longrightarrow \frac{\pi}{4} \psi \quad \hat{\psi}_a \longrightarrow \frac{\pi}{3} \psi$$

not invariant to (one-one) reparametrizations of λ ; better to use

$$\ell_a(\psi) = \ell_p(\psi) + \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| + B(\psi)$$

with $B(\psi) = O_p(1)$

this can make ℓ_a invariant and remove the need for orthogonal parametrization

$$B(\psi) = -\frac{1}{2} \log |\varphi'_\lambda(\psi, \hat{\lambda}_\psi) j_{\varphi\varphi}(\hat{\psi}, \hat{\lambda}) \varphi'_\lambda(\psi, \hat{\lambda}_\psi)|$$

with $\varphi = \varphi(\theta) = \ell_{;V}(\theta; y^0)$

comes from an approximating location model: Fraser 2003 Biometrika

p-values from profile likelihood

First order: $p(\psi) \doteq \Phi(r_p)$

$$r_p(\psi) = \pm[2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2} \sim N(0, 1)$$

Third order: $p(\psi) \doteq \Phi(r^*)$

$$r^*(\psi) = r_p(\psi) + \frac{1}{r_p} \log \frac{r_p}{Q}$$

$$Q = (\hat{\nu} - \hat{\nu}_\psi) \hat{\sigma}_\nu^{-1/2}$$

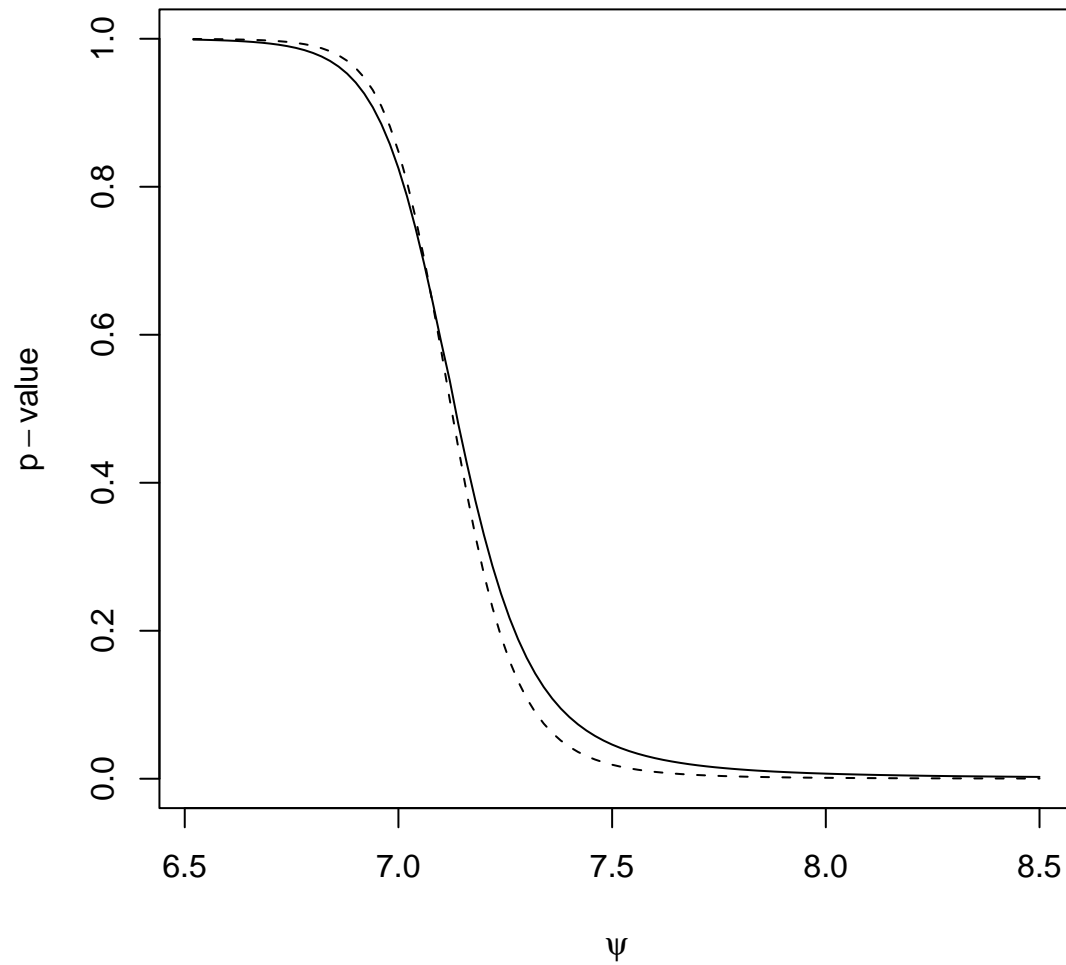
$$\begin{aligned} \nu(\theta) &= e_\psi^T \varphi(\theta) , \\ e_\psi &= \psi_{\varphi'}(\hat{\theta}_\psi) / |\psi_{\varphi'}(\hat{\theta}_\psi)| , \\ \hat{\sigma}_\nu^2 &= |j_{(\lambda\lambda)}(\hat{\theta}_\psi)| / |j_{(\theta\theta)}(\hat{\theta})| , \\ |j_{(\theta\theta)}(\hat{\theta})| &= |j_{\theta\theta}(\hat{\theta})| |\varphi_{\theta'}(\hat{\theta})|^{-2} , \\ |j_{(\lambda\lambda)}(\hat{\theta}_\psi)| &= |j_{\lambda\lambda}(\hat{\theta}_\psi)| |\varphi_{\lambda'}(\hat{\theta}_\psi)|^{-2} . \end{aligned}$$

Fraser, Reid, Wu (1999) *Biometrika*

Example: $\log y \sim N(\mu, \sigma^2)$:

inference for $\psi = \log(EY)$

solid: 3rd order dotted: 1st order



Example: comparing two binomials

Employment of men and women at the Space Telescope Science Institute, 1998–2002 (from *Science* magazine, Volume 299, page 993, 14 February 2003).

	Left	Stayed	Total
Men	1	18	19
Women	5	2	7
Total	6	20	26

$$Y_1 \sim \text{Bin}(19, p_1), \quad Y_2 \sim \text{Bin}(7, p_2)$$

$$\psi = \log \frac{p_1(1 - p_2)}{p_2(1 - p_1)}$$

p -value for testing $\psi = 0$ is

0.00203 using normal approx to maximum likelihood

0.00028 using normal approx to r_p (1st order)

0.00048 using normal approx to r^* (3rd order)

Some more technical points

In special model classes, it is possible to eliminate nuisance parameters by either *conditioning* or *marginalizing*. The conditional or marginal likelihood then gives essentially exact inference for the parameter of interest, if this likelihood can itself be computed exactly.

The main example is the canonical parameter of an exponential family:

$$f(\underline{y}; \psi, \lambda) = \exp\{\psi s + \lambda' t - c(\psi, \lambda) - d(y)\};$$

$$f(s | t; \psi) = \exp\{\psi s - C_t(\psi) - D_t(s)\}$$

$$\ell_{cond}(\psi) = \psi s - C_t(\psi)$$

The adjusted log-likelihood

$$\ell_a(\psi) = \ell_p(\psi) - (1/2) \log |j_{\lambda\lambda}| \text{ approximates } \ell_{cond}$$

The 3rd order p -value approximation is particularly simple:

$$r^* = r_a + \frac{1}{r_a} \log\left(\frac{Q}{r_a}\right)$$

$$r = r_a = \pm[2\{\ell_a(\hat{\psi}_a) - \ell_a(\psi)\}]^{1/2}$$

$$Q = (\hat{\psi}_a - \psi)\{j_a(\hat{\psi})\}^{1/2}$$

A similar discussion applies to the class of transformation models, using marginal approximations. Both class are reviewed in Reid 1996

The approximations given earlier reduce to these special cases.

Some References

Fraser, D.A.S., Reid, N., Wong, A. (2003). Inference for bounded parameters. xxx.lanl.gov/0303111.

Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327–339.

Fraser, D.A.S., Reid, N., Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 246–264.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.*, to appear.

Reid, N. (1992). Aspects of modified profile likelihood. in *Nonparametric Statistics and Related Topics*, A.K.Md.E. Saleh, ed. North-Holland, Amsterdam.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc. B*, **49**, 1–39.

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