Accurate approximation for inference on vector parameters

Nancy Reid

January 15, 2010

with Don Fraser, Anthony Davison, Nicola Sartori
Approximate exponential families

- model \( f(y; \theta) \), \( \theta \in \mathbb{R}^d \)
- data \( y = (y_1, \ldots, y_n) \) independent observations
- log-likelihood function \( \ell(\theta; y) = \log f(y; \theta) \)

linear exponential family:

\[
f(y; \theta) = \exp\{\varphi(\theta)'s(y) - c(\theta) - d(y)\}
\]

- canonical parameter obtained as

\[
\frac{\partial \ell(\theta; y)}{\partial s(y)} = \varphi(\theta)
\]
- up to affine transformations
- Example: \( \mathcal{N}(\mu, \sigma^2) \):

\[
\ell(\theta; y) = \frac{\mu}{\sigma^2} \sum y_i - \frac{1}{2\sigma^2} \sum y_i^2 - \frac{n\mu^2}{2\sigma^2} - n \log \sigma
\]
Tangent exponential model

- In general, can find an approximate exponential model:

\[
f_{TEM}(s; \theta)ds = \exp\{\varphi(\theta)'s + \ell(\theta)\}h(s)ds \tag{1}
\]

- \(s\) is a score variable on \(\mathbb{R}^d\):

\[s(y) = -\ell_\varphi(\hat{\theta}^0; y)\]

- \(\ell(\theta) = \ell(\theta; y^0)\) is the observed log-likelihood function

- \(\varphi(\theta) = \varphi(\theta; y^0)\) is the canonical parameter \(\in \mathbb{R}^d\) to be determined

- Has the same observed log likelihood function as the original model

- Has same first derivative on the sample space, at \(y^0\), as the original model

- (1) approximates \(f(y \mid a; \theta)\) to \(O(n^{-1})\)
Likelihood Inference

- assume $\theta = (\psi, \lambda)$, $\psi$ scalar parameter of interest
- likelihood ratio statistic $w(\psi) = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\}$ approximately $\chi^2_1$
- or a directional departure $r(\psi) = \pm \sqrt{w(\psi)}$
- approximately $N(0, 1)$ $O(n^{-1/2})$
- can do much better:

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

- $r^* \sim N(0, 1)$ with relative error $O(n^{-3/2})$ under $f(y; \theta)$
- $r$ and $q$ functions only of $\{\ell(\theta; y^0), \varphi(\theta, y^0)\}$
- and their derivatives; also need $\hat{\theta}$ and $\hat{\theta}_\psi$, full and constrained mle, as with $w$
log-likelihood function

The graph shows the log-likelihood function with values ranging from -4 to 0 on the y-axis and from 16 to 23 on the x-axis. The peak of the function is labeled with a value of 1.92 w/2, and there are annotations indicating points at \( \theta \) and \( \theta' \).
Inference with TEM

\[ f_{TEM}(s; \theta) = \exp\{\varphi(\theta)'s + \ell(\theta)\}h(s) \]

\[ \varphi(\theta) = \varphi(\theta; y^0), \quad \ell(\theta) = \ell(\theta; y^0) \]

- why \( y^0 \)?
- \( p \)-value: probability of data as or more extreme than that observed
- can be plotted as a function of the parameter
- provides tests of particular values, and confidence bounds or intervals
Example: $2 \times 2$ table

$$\begin{array}{cc}
M & S \\
M & 1 & 18 \\
F & 5 & 2 \\
\end{array}$$

$$\psi = \text{log-odds ratio}$$

BDR, 2007, Fig.3.4
Details

- \{\ell(\theta), \varphi(\theta)\} \rightarrow \text{TEM} \rightarrow p\text{-value}
- using \( r^* = r^*(\psi) = r + \frac{1}{r} \log\left(\frac{q}{r}\right) \sim N(0, 1) \)
- \( r(\psi) = \pm \sqrt{2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}} \) likelihood root
- \( q(\psi) = \frac{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)|}{|\varphi_\theta(\hat{\theta})|} \frac{|j(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}} \)
- observed information \( j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta' \)
- nuisance parameter integrated out via Laplace
Canonical parameter $\varphi(\theta)$

- if $f(y; \theta)$ is an exponential family, $\varphi$ is sitting in the model
- if not
- if $y$ is continuous, define

$$V = \frac{dy}{d\theta} \bigg|_{y=y_0, \theta=\hat{\theta}}$$

$$y = (y_1, \ldots, y_n)$$

- $??$
- $z_i = z_i(y_i; \theta)$ with a fixed distribution, e.g. $(y_i - \mu)/\sigma$

$$V = -\left(\frac{\partial z}{\partial y}\right)^{-1} \frac{\partial z}{\partial \theta} \bigg|_{y=y_0, \theta=\hat{\theta}}$$

$$n \times p$$

$$\varphi(\theta) = \varphi(\theta; y^0) = \frac{\partial \ell(\theta; y)}{\partial V} \bigg|_{y=y^0} = \sum_{i=1}^{n} \frac{\partial \ell(\theta; y^0)}{\partial y_i} V_i$$
Example: regression

- Model: \( y_i = x_i' \beta + \sigma \epsilon_i \)
- Canonical parameter: \( \varphi(\theta) = \sum_{i=1}^{n} \ell_i y_i(\theta; y^0) V_i \)
- \( V_i = [x_i' (y_i^0 - x_i' \hat{\beta})/\hat{\sigma}] \)
- \( \varphi(\theta; y) = \sum_{i=1}^{n} \frac{1}{\sigma} g'(\frac{y_i^0 - x_i' \beta}{\sigma})[x_i' \hat{\epsilon}_i] \)

<table>
<thead>
<tr>
<th>Normal</th>
<th>t_4 errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{Est (SE)}</td>
<td>\text{z}</td>
</tr>
<tr>
<td>Constant</td>
<td>(-13.26 (3.140))</td>
</tr>
<tr>
<td>date</td>
<td>0.212 (0.043)</td>
</tr>
<tr>
<td>log(cap)</td>
<td>0.723 (0.119)</td>
</tr>
<tr>
<td>NE</td>
<td>0.249 (0.074)</td>
</tr>
<tr>
<td>CT</td>
<td>0.140 (0.060)</td>
</tr>
<tr>
<td>log(N)</td>
<td>(-0.088 (0.042))</td>
</tr>
<tr>
<td>PT</td>
<td>(-0.226 (0.114))</td>
</tr>
</tbody>
</table>
... canonical parameter $\varphi$

- a sample space derivative of log-likelihood $\ell; \mathcal{V}(\theta; y^0)$
- if the sample space is discrete
- $\begin{align*}
y \rightarrow s & \quad \text{score variable} \\
\frac{dy}{d\theta} & \rightarrow \frac{dE(s; \theta)}{d\theta} \quad \text{DFR, 2006}
\end{align*}$

\[
s_i = s_i(y_i) = \left. \frac{\partial \ell(\theta; y_i)}{\partial \theta} \right|_{\theta = \hat{\theta}^0}
\]

\[
V_i = \left. \frac{\partial}{\partial \theta} E(s_i; \theta) \right|_{\theta = \hat{\theta}^0}
\]

\[
\varphi(\theta) = \sum_{i=1}^{n} \left. \frac{\partial \ell(\theta; y^0)}{\partial s_i} \right|_{\theta = \hat{\theta}^0} V_i
\]
relative error $O(n^{-1})$

DFR, 2006
Example: Poisson counts

Likelihood for discrete data

Table 1. Lung cancer deaths in British male physicians (Frome, 1983). The table gives man-years at risk/number of cases of lung cancer, $T/y$, cross-classified by years of smoking, taken to be age minus 20 years, and number of cigarettes smoked per day.

<table>
<thead>
<tr>
<th>Years of smoking $t$</th>
<th>Daily cigarette consumption $x$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonsmokers</td>
</tr>
<tr>
<td></td>
<td>1–9</td>
</tr>
<tr>
<td>15–19</td>
<td>10366/1</td>
</tr>
<tr>
<td>20–24</td>
<td>8162</td>
</tr>
<tr>
<td>25–29</td>
<td>5969</td>
</tr>
<tr>
<td>30–34</td>
<td>4496</td>
</tr>
<tr>
<td>35–39</td>
<td>3512</td>
</tr>
<tr>
<td>40–44</td>
<td>2201</td>
</tr>
<tr>
<td>45–49</td>
<td>1421</td>
</tr>
<tr>
<td>50–54</td>
<td>1121</td>
</tr>
<tr>
<td>55–59</td>
<td>826/2</td>
</tr>
</tbody>
</table>

$$E_\theta(Y) = T\lambda(x, t) = \exp(\theta_1)t^{\theta_2}\{1 + \exp(\theta_3)x^{\theta_4}\}$$

$T$ yrs. at risk  $x$ # cigarettes  $t$ Years smoking  $\theta_4$ parameter of interest
... Poisson regression

$\left. \begin{array}{l}
\quad E_{\theta}(Y) = T \lambda(x, t) = \exp(\theta_1) t^{\theta_2} \{1 + \exp(\theta_3) x^{\theta_4}\} \\
\quad \text{linear increase in death rate with ‘dose’} \rightarrow H_0 : \theta_4 = 1 \\
\quad \text{signed root} \\
\quad \text{of log-likelihood ratio statistic} \quad r = 1.506 \quad p = 0.066 \\
\quad \text{higher order approximation} \quad r^* = 1.491 \quad p = 0.068
\end{array} \right.$
Vector parameter of interest

- \( \theta = (\psi, \lambda), \quad \psi \in \mathbb{R}^{d_0} \quad H_0 : \psi = \psi_0 \)
- usual:
  \[
  W(\psi_0) = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \hat{\lambda}_{\psi_0})\} \sim \chi^2_{d_0}
  \]
- proposed: profile sample space \( S_\psi \),
  all sample points that give the same estimate for the nuisance parameter \( \hat{\lambda}_\psi \)

\[
S_\psi = \{ s : \hat{\varphi}_\psi = \hat{\varphi}^0_\psi \} = \{ s : \ell_\lambda(\hat{\theta}^0_\psi; s) = 0 \}
\]
a surface of dimension \( d_0 \), passing through data point \( y^0 \)
- Directed departure on \( S_\psi \)
- observed value \( s^0 = 0 \)
- expected value \( s_H \) under \( H_0 \)
  \[ s_H = -\ell_\varphi(\hat{\theta}^0_\psi) \]
- Distribution of magnitude of \( |s - s_H| \)
- given the direction \( (s - s_H)/|s - s_H| \)
- Skovgaard, 1988; Fraser & Massam, 1988; Cheah, Fraser, Reid, 1994
Directional $p$-value

- line $s(t)$ from hypothesis, $s_H$, to data, $s^0 : s_H + t(s^0 - s_H)$
- $f(s; \psi_0)$ used to compute the probability at and beyond the observed $s^0$ ($t \geq 1$), conditional on being on the line $s(t)$.
- along the line $s(t)$ we have
  \[ f(s; \psi_0) ds = f\{s(t); \psi_0\} dt = f\{s_H + t(s^0 - s_H); \psi_0\} dt. \]
- directional $p$-value:
  \[ p(\psi_0) = \frac{\int_1^{+\infty} t^{d_0 - 1} f\{s(t); \psi_0\} dt}{\int_0^{+\infty} t^{d_0 - 1} f\{s(t); \psi_0\} dt} \]
- one-dimensional integrals computed numerically
Log-likelihood along the line $s(t)$

$t=0$

$t=0.5$

$t=1$

$t=1.33$

$t=1.67$

$t=2$
Score variable?

- exponential family model

\[ f(y; \theta) = \exp\{\varphi(\theta)'s(y) - c(\theta) - d(y)\} \]

- \( f(s; \theta) \) available from saddlepoint approximation

- tangent exponential family model

\[ f_{TEM}(s; \theta) = \exp\{\varphi(\theta; y^0)'s + \ell(\theta; y^0)\} h(s) \]

- saddlepoint approximation

\[ f(s; \psi) \doteq \frac{e^{c/n}}{(2\pi)^d} \exp\{((\psi - \hat{\psi})'s + \ell(\hat{\theta}) - \ell(\hat{\theta}_\psi))|\hat{j}_{\varphi\varphi}|^{-1/2} \]

- on line \( s(t) = s_H + t(s^0 - s_H) \)
Directional $p$-value

- The directional $p$-value is equal to 0.050

- First order $\chi^2$ approximation
  - $W(\psi_0) = 0.047$

- Skovgaard (2001 SJS) modified version
  - $W^*(\psi) = 0.048$

- Simulated conditional
  - 0.051
Testing independence in $2 \times 3$ contingency table

contingency table on activity amongst psychiatric patients (Everitt, 1992 CH)

<table>
<thead>
<tr>
<th></th>
<th>Affective disorders</th>
<th>Schizophrenics</th>
<th>Neurotics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retarded</td>
<td>12</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Not retarded</td>
<td>18</td>
<td>17</td>
<td>25</td>
</tr>
</tbody>
</table>

model: log-linear $y \sim \text{Poisson}, \quad \log\{E(y)\} = X\beta$

$H_0$: independence

nuisance parameter $\lambda \in \mathbb{R}^4$

full model has an additional $(\psi_1, \psi_2)$: interaction between the variables

$H_0: \psi = \psi_0 = (0, 0)$. 
... $2 \times 3$ contingency table

- expected frequencies under the null hypothesis $t = 0$

<table>
<thead>
<tr>
<th></th>
<th>Affective disorders</th>
<th>Schizophrenics</th>
<th>Neurotics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retarded</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Not retarded</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

- need to stop at $t = t_{\text{max}} = 2$.
- the expected frequencies corresponding to $t_{\text{max}} = 2$

<table>
<thead>
<tr>
<th></th>
<th>Affective disorders</th>
<th>Schizophrenics</th>
<th>Neurotics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retarded</td>
<td>14</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Not retarded</td>
<td>16</td>
<td>14</td>
<td>30</td>
</tr>
</tbody>
</table>

- All tables along the line $s(t)$ have the same margins.
Another $2 \times 3$ table

Consider the following data on party identification by race
Agresti, 2002 Wiley

<table>
<thead>
<tr>
<th></th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>103</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>White</td>
<td>341</td>
<td>105</td>
<td>405</td>
</tr>
</tbody>
</table>

$H_0$: independence nested in saturated model

First order likelihood ratio $p$-value: $2.43 \times 10^{-20}$

Directional $p$-value: $3.14 \times 10^{-20}$.
... party identification example

- **Expected** \((t = 0)\)

<table>
<thead>
<tr>
<th></th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>58.44</td>
<td>15.80</td>
<td>54.76</td>
</tr>
<tr>
<td>White</td>
<td>385.56</td>
<td>104.20</td>
<td>361.24</td>
</tr>
</tbody>
</table>

- **Observed** \((t = 1)\)

<table>
<thead>
<tr>
<th></th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>103</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>White</td>
<td>341</td>
<td>105</td>
<td>405</td>
</tr>
</tbody>
</table>

- **Boundary** \((t_{\text{max}} = 1.251)\)

<table>
<thead>
<tr>
<th></th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>114.20</td>
<td>14.80</td>
<td>0.00</td>
</tr>
<tr>
<td>White</td>
<td>329.80</td>
<td>105.20</td>
<td>416.00</td>
</tr>
</tbody>
</table>
Log-linear models for contingency tables

- easier; already an exponential family model
- \( y = (y_1, \ldots, y_C) \), \( X \) a \( C \times d \) design matrix
- \( E(y) = \exp(X\beta) \)

\[ \ell(\beta) = \beta'X'y - 1'\exp(X\beta) \]

- \( \varphi(\beta) = \beta \)
- \( \beta = (\lambda, \psi) \), and design matrix partitioned as \( X = (X_1 \ X_2) \)

\[ \ell_\beta(\beta) = \begin{bmatrix} \ell_\lambda(\lambda, \psi) \\ \ell_\psi(\lambda, \psi) \end{bmatrix} = \begin{bmatrix} X_1'(y - e^{X\hat{\beta}}) \\ X_2'(y - e^{X\hat{\beta}}) \end{bmatrix} \cdot \]

- constrained maximum likelihood estimate:
  \( \ell_\lambda(\hat{\beta}_\psi) = X_1'(y - e^{X\hat{\beta}_\psi}) = 0 \)

- tangent exponential model = double saddlepoint distribution of \( X_2'y \), given \( X_1'y \)
... details on log-linear models

- observed data point $s^0 = 0$
- expected value when $\psi = \psi_0$

$$s_H = -\ell_\beta(\hat{\beta}_{\psi_0}) = \begin{bmatrix} 0 \\ -X_2^T(y - e^{X\hat{\beta}_{\psi_0}}) \end{bmatrix}.$$

- directional test goes radially from $s_H$ towards the data point $s^0$ and beyond to the boundary in that direction.
- $f(s(t); \psi_0)$ along the line $s(t)$ computed using
  $$\ell(\psi; t) = \ell(\hat{\beta}_\psi) + \hat{\beta}_\psi^T s(t).$$
- Nicola: use glm with numerical integration (univariate)
## Infant survival data

<table>
<thead>
<tr>
<th>survival</th>
<th>gestation</th>
<th>smoking</th>
<th>age</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>&lt;=260</td>
<td>&lt;5</td>
<td>&lt;30</td>
<td>50</td>
</tr>
<tr>
<td>Yes</td>
<td>&lt;=260</td>
<td>&lt;5</td>
<td>&lt;30</td>
<td>315</td>
</tr>
<tr>
<td>No</td>
<td>&gt;260</td>
<td>&lt;5</td>
<td>&lt;30</td>
<td>24</td>
</tr>
<tr>
<td>Yes</td>
<td>&gt;260</td>
<td>&lt;5</td>
<td>&lt;30</td>
<td>4012</td>
</tr>
<tr>
<td>No</td>
<td>&lt;=260</td>
<td>&gt;5</td>
<td>&lt;30</td>
<td>9</td>
</tr>
<tr>
<td>Yes</td>
<td>&lt;=260</td>
<td>&gt;5</td>
<td>&lt;30</td>
<td>40</td>
</tr>
<tr>
<td>No</td>
<td>&gt;260</td>
<td>&gt;5</td>
<td>&lt;30</td>
<td>6</td>
</tr>
<tr>
<td>Yes</td>
<td>&gt;260</td>
<td>&gt;5</td>
<td>&lt;30</td>
<td>459</td>
</tr>
<tr>
<td>No</td>
<td>&lt;=260</td>
<td>&lt;5</td>
<td>&gt;30</td>
<td>41</td>
</tr>
<tr>
<td>Yes</td>
<td>&lt;=260</td>
<td>&lt;5</td>
<td>&gt;30</td>
<td>147</td>
</tr>
<tr>
<td>No</td>
<td>&gt;260</td>
<td>&lt;5</td>
<td>&gt;30</td>
<td>14</td>
</tr>
<tr>
<td>Yes</td>
<td>&gt;260</td>
<td>&lt;5</td>
<td>&gt;30</td>
<td>1594</td>
</tr>
<tr>
<td>No</td>
<td>&lt;=260</td>
<td>&gt;5</td>
<td>&gt;30</td>
<td>4</td>
</tr>
<tr>
<td>Yes</td>
<td>&lt;=260</td>
<td>&gt;5</td>
<td>&gt;30</td>
<td>11</td>
</tr>
<tr>
<td>No</td>
<td>&gt;260</td>
<td>&gt;5</td>
<td>&gt;30</td>
<td>1</td>
</tr>
<tr>
<td>Yes</td>
<td>&gt;260</td>
<td>&gt;5</td>
<td>&gt;30</td>
<td>124</td>
</tr>
</tbody>
</table>
...infant survival data

- Data (Agresti, 2002 Wiley) four dichotomous variables: age of mother (A), length of gestation (G), infant survival (I) and number of cigarettes smoked per day during gestation (S).
- response: length of gestation and infant survival
- null model: with all main effects and three first order interactions (IG, IA and SA) as the null model $\lambda \in \mathbb{R}^8$
- full model has two additional first order interaction parameters: IS and GA
- first order likelihood ratio $p$-value = 0.052.
- directional $p$-value = 0.056.
## Conclusion

<table>
<thead>
<tr>
<th></th>
<th>Scalar $\psi$</th>
<th>Vector $\psi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous response</td>
<td>$r^* : O(n^{-3/2})$</td>
<td>Directional: $O(n^{-1})$</td>
</tr>
<tr>
<td>Discrete response</td>
<td>$r^* : O(n^{-1})$</td>
<td>Directional: $O(n^{-1})$</td>
</tr>
</tbody>
</table>

- Tangent exponential model
- Saddlepoint approximation
- Easy, accurate

- Extensions to nonlinear hypotheses, more complex models for categorical data...