Approximate inference for vector parameters

Nancy Reid

March 3, 2010

with Don Fraser, Anthony Davison, Nicola Sartori

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Parametric models and likelihood

- ▶ model $f(y; \theta)$, $\qquad\qquad$ $\theta \in \mathbb{R}^d$

- ▶ data $y = (y_1, \ldots, y_n)$ $\qquad$ independent observations

- ▶ log-likelihood function $\qquad$ $\ell(\theta; y) = \log f(y; \theta)$

- ▶ parameter of interest $\qquad$ $\theta = (\psi, \lambda), \quad \psi \in \mathbb{R}^{d_0}$

- ▶ likelihood inference $\qquad$ $w(\psi) = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\}$

- ▶ standardized m.l.e. $\qquad$ $q(\psi) = (\hat{\psi} - \psi)^T (\hat{\jmath}^{\psi\psi})^{-1} (\hat{\psi} - \psi)$

- ▶ stand'd score $\qquad\qquad$ $t(\psi) = \ell'_p(\psi)^T \hat{\jmath}^{\psi\psi} \ell'_p(\psi)$

- ▶ $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$

log−likelihood function
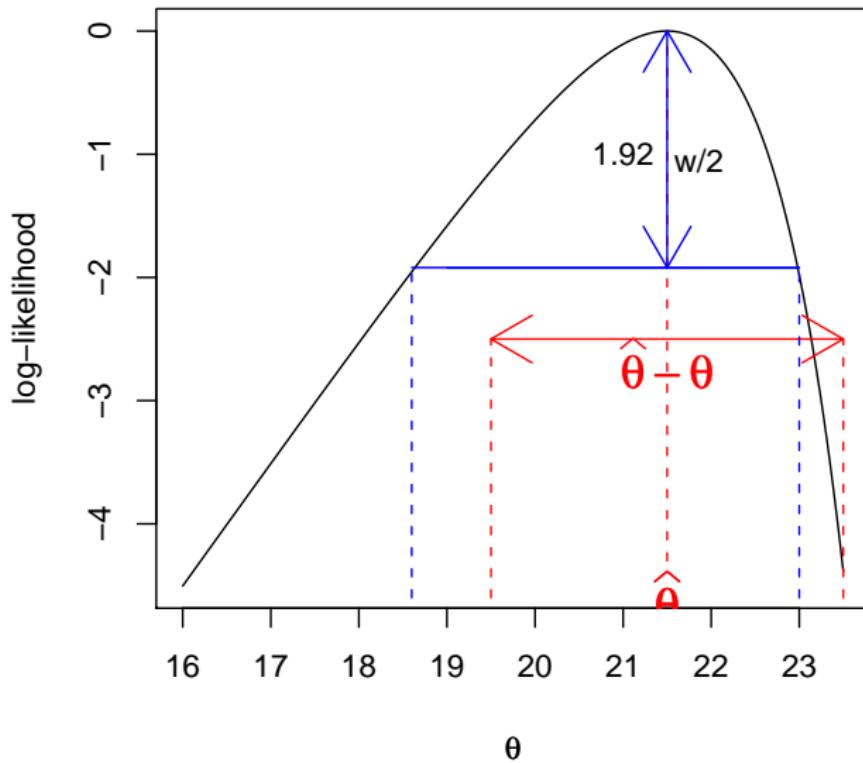
Likelihood statistics as pivots

Scalar parameter of interest

score statistic $\qquad t(\psi) = \ell_{\mathrm{p}}'(\psi)\{j_{\mathrm{p}}(\hat{\psi})\}^{-1/2}$

standardized m.l.e. $\qquad q(\psi) = (\hat{\psi} - \psi)\{j_{\mathrm{p}}(\hat{\psi})\}^{1/2}$

likelihood root $\qquad r(\psi) = \pm\sqrt{[2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\}]}$

First order p-values $\quad p(\psi) \doteq \Phi\{r(\psi)\}$
$$\doteq \Phi\{q(\psi)\}$$
$$\doteq \Phi\{t(\psi)\}$$

Third order p-values

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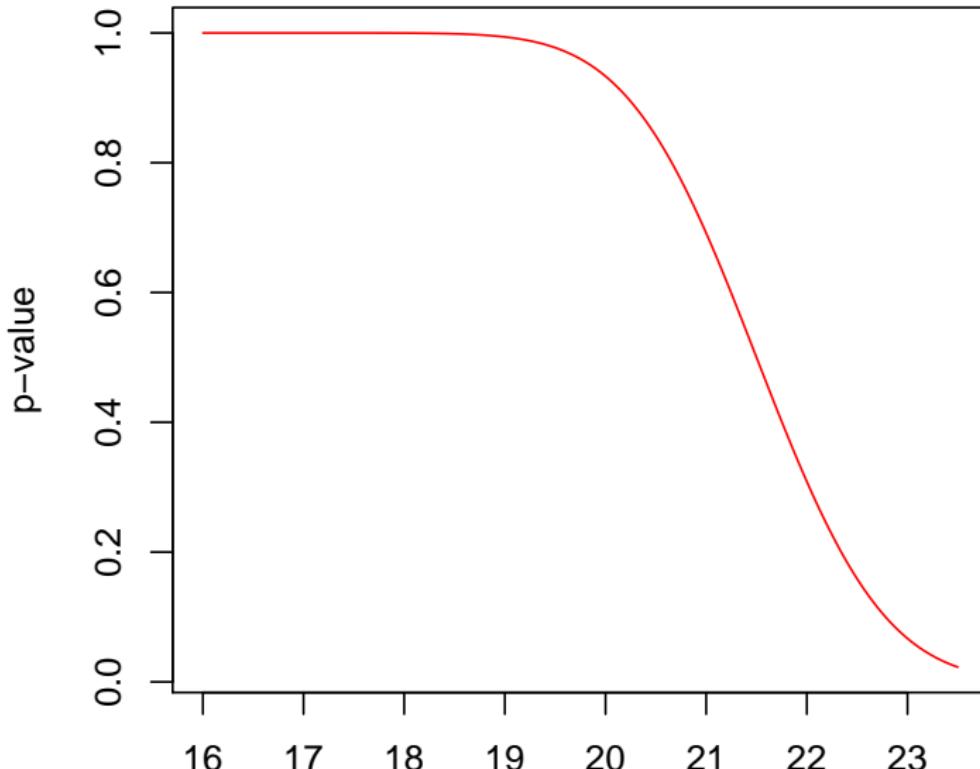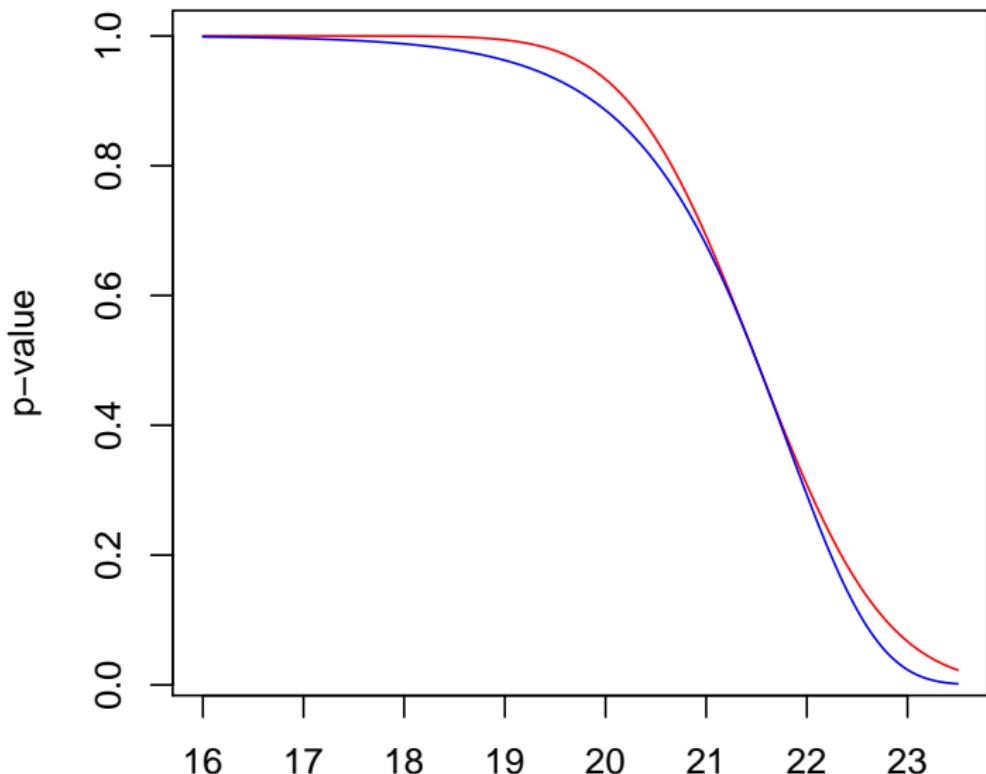
$$r^*(\psi) = r + \frac{1}{r(\psi)}\log\frac{Q(\psi)}{r(\psi)}$$

$$Q = q(\psi) \text{ or } t(\psi) \text{ or } \dots$$

Pvalue functions

Pvalue functions

Pvalue functions

Pvalue functions

p-value

Pvalue functions

p−value

lik. root r
mle. q
score s
rstar
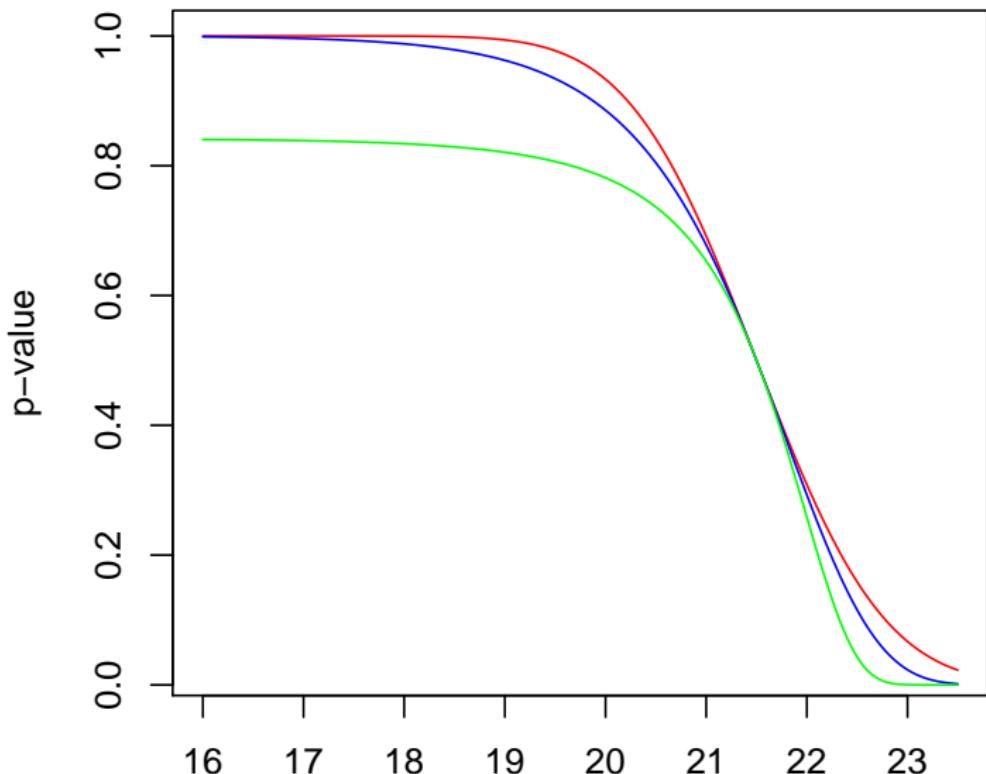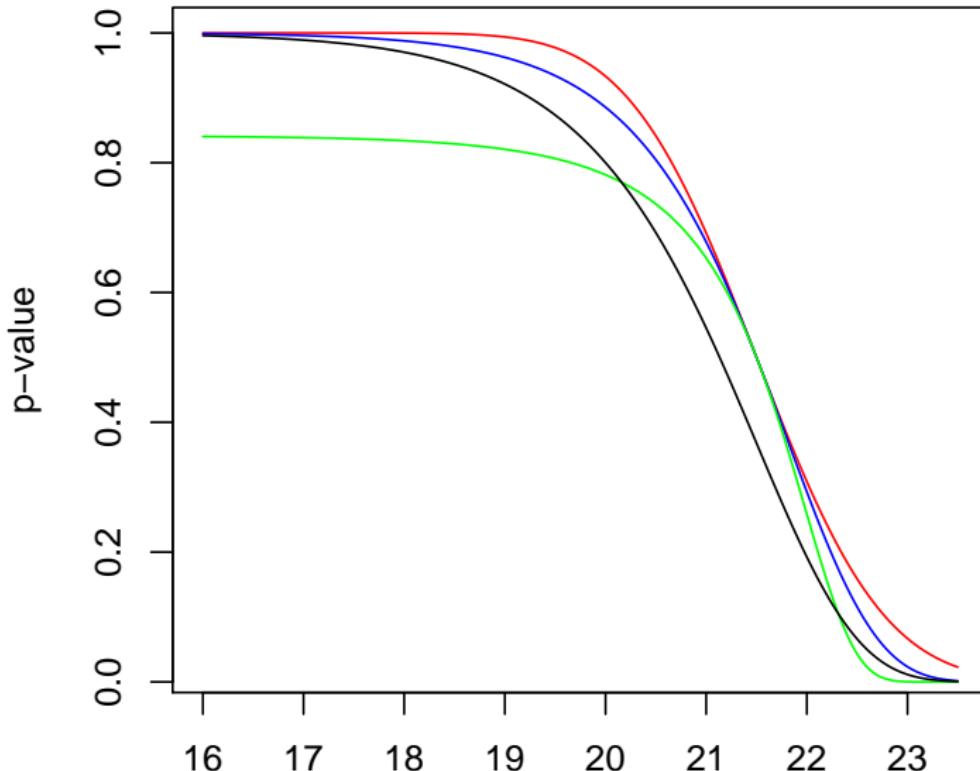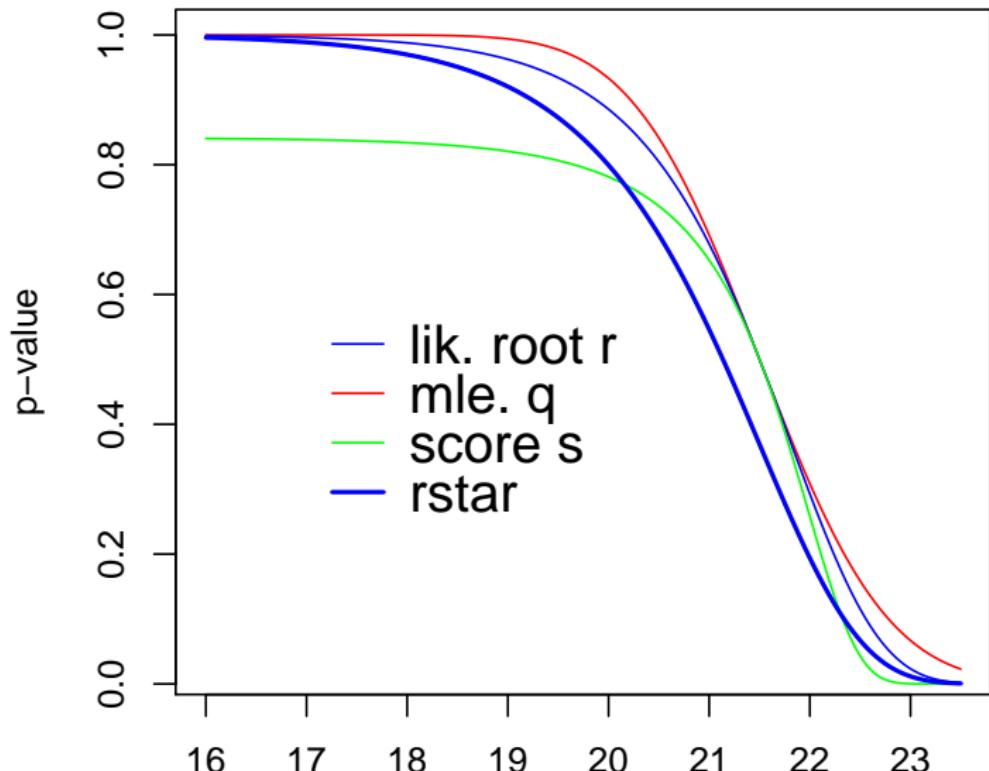
... likelihood statistics as pivots
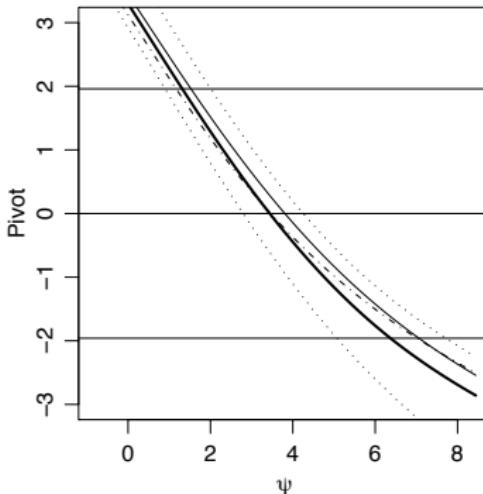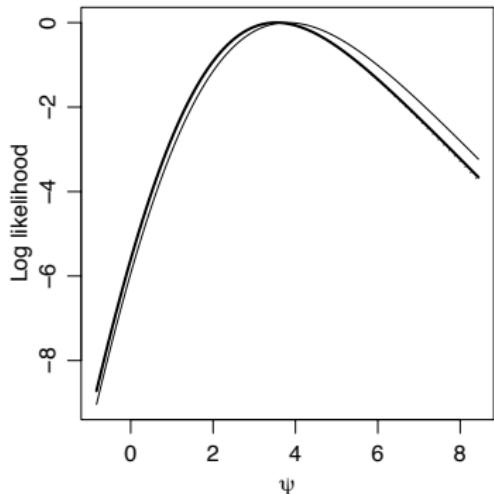
	M	S	
M	1	18	19
F	5	2	7
	6	20	26

$\psi = $ log-odds ratio

BDR, 2007, Fig.3.4

Exponential family models

- linear exponential family:

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

- canonical parameter obtained as

$$\frac{\partial \ell(\theta; y)}{\partial s(y)} = \varphi(\theta)$$

- Example $N(\mu, \sigma^2)$:

$$\ell(\theta; y) = \frac{\mu}{\sigma^2} \sum y_i - \frac{1}{2\sigma^2} \sum y_i^2 - \frac{n\mu^2}{2\sigma^2} - n \log \sigma$$

- Example $Bin(n, p)$:

$$\ell(\theta; y) = \log \frac{p}{1-p} y + n \log(1-p)$$

Tangent exponential model

- Every model $f(y; \theta)$ on $\mathbb{R}^n$ can be approximated by an exponential family model:

$$f_{TEM}(s; \theta)ds = \exp\{\varphi(\theta)'s + \ell(\theta)\}h(s)ds \qquad (1)$$

- 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 described

- 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 by definition of φ
- (1) approximates $f(y \mid a; \theta)$ to $O(n^{-1})$

Inference with TEM

- $f_{TEM}(s; \theta) = \exp\{\varphi(\theta)'s + \ell(\theta)\}h(s)$
- $\varphi(\theta) = \varphi(\theta; y^0), \qquad \ell(\theta) = \ell(\theta; y^0)$

- likelihood pivot
  $$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)}\log\{\frac{Q(\psi)}{r(\psi)}\} \dot\sim N(0,1) \qquad O(n^{-3/2})$$

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

- $Q(\psi) = \dfrac{|\varphi(\hat\theta) - \varphi(\hat\theta_\psi) \quad \varphi_\lambda(\hat\theta_\psi)|}{|\varphi_\theta(\hat\theta)|} \dfrac{|j(\hat\theta)|^{1/2}}{|j_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}}$

- observed information $j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta'$

- nuisance parameter has been integrated out
  using Laplace approximation

Canonical parameter $\varphi(\theta)$

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

$$V = \left.\frac{dy}{d\theta}\right|_{y=y^0, \theta=\hat\theta^0} \qquad y = (y_1, \ldots, y_n)$$

- ??
- $z_i = z_i(y_i; \theta)$ with a fixed distribution, e.g. $(y_i - \mu)/\sigma$
- $V = -\left.\left(\dfrac{\partial z}{\partial y}\right)^{-1}\dfrac{\partial z}{\partial \theta}\right|_{y=y^0, \theta=\hat\theta^0}$ $n \times p$
-

$$\varphi(\theta) = \varphi(\theta; y^0) = \left.\frac{\partial \ell(\theta; y)}{\partial V}\right|_{y=y^0} = \sum_{i=1}^{n} \frac{\partial \ell(\theta; y^0)}{\partial y_i} V_i$$

Example: regression

- Model: $y_i = x_i'\beta + \sigma\epsilon_i$
- Canonical parameter: $\varphi(\theta) = \sum_{i=1}^{n} \ell_{;y_i}(\theta; y^0) V_i$

- $V_i = [x_i' \quad (y_i^0 - x_i'\hat{\beta})/\hat{\sigma}]$

- $\varphi(\theta; y^0) = \sum_{i=1}^{n} \dfrac{1}{\sigma}\dot{g}(\dfrac{y_i^0 - x_i'\beta}{\sigma})[x_i' \quad \hat{\epsilon}_i]$

	Normal errors		t_4 errors	
	Est (SE)	z	Est (SE)	z
Constant	-13.26 (3.140)	-4.22	-11.86 (3.70)	-3.21
date	0.212 (0.043)	4.91	0.196 (0.049)	4.02
log(cap)	0.723 (0.119)	6.09	0.682 (0.129)	5.31
NE	0.249 (0.074)	3.36	0.239 (0.080)	2.97
CT	0.140 (0.060)	2.32	0.143 (0.063)	2.26
log(N)	-0.088 (0.042)	-2.11	-0.072 (0.048)	-1.51
PT	-0.226 (0.114)	-1.99	-0.265 (0.110)	-2.42

Discrete responses y

- φ is a data-derivative of log-likelihood, $\ell_{;V}(\theta; y^0)$
- doesn't work if if the data is discrete
- $y \longrightarrow s$ score variable
- $\dfrac{dy}{d\theta} \longrightarrow \dfrac{dE(s;\theta)}{d\theta}$ DFR, 2006
-
$$s_i = s_i(y_i) = \left.\frac{\partial\ell(\theta; y_i)}{\partial\theta}\right|_{\theta=\hat\theta^0}$$

-
$$V_i = \left.\frac{\partial}{\partial\theta}E(s_i;\theta)\right|_{\theta=\hat\theta_0}$$

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

Models and inference   Tangent exponential model   **Canonical parameter**   Vector parameter   Conclusion
ooooooooo   ooo   ooooooo   ooooooooo
                                                                                ooooooooo

Example: Poisson counts

Likelihood for discrete data   9

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.

Years of smoking t	Daily cigarette consumption x						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

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

**T** yrs. at risk   **x** # cigarettes   **t** Years smoking θ_4 parameter of interest

... Poisson regression

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

- linear increase in death rate with 'dose' $\longrightarrow H_0 : \theta_4 = 1$

      signed root
of log-likelihood ratio statistic $r = 1.506$ $p = 0.066$

      higher order
      approximation $r^* = 1.491$ $p = 0.068$

Models and inference   Tangent exponential model   **Canonical parameter**   Vector parameter   Conclusion
○○○○○○○○○   ○○○   ○○○○○●○   ○○○○○○○○○
                                                      ○○○○○○○

Relative error $O(n^{-1})$
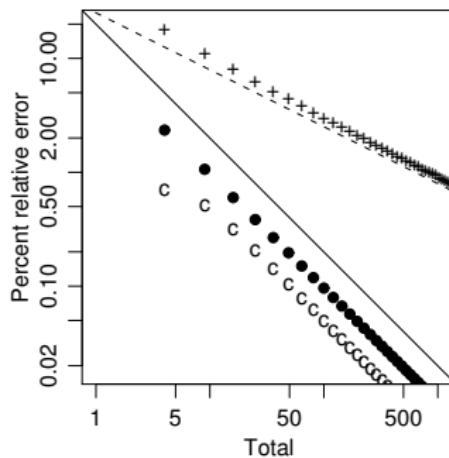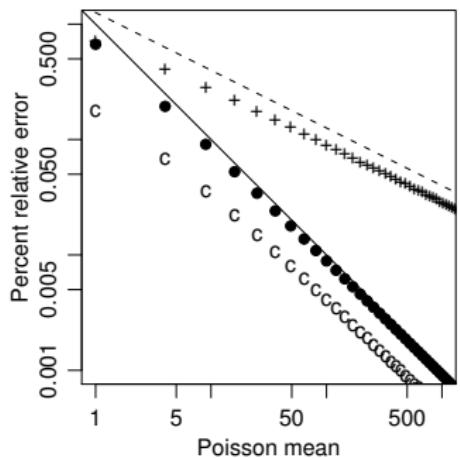
DFR, 2006

|	scalar ψ	vector ψ
continuous response	$r^* : O(n^{-3/2})$	directional: $O(n^{-1})$
discrete response	$r^* : O(n^{-1})$	directional: $O(n^{-1})$

Relative error $O(n^{-1})$

DFR, 2006

|	scalar ψ	vector ψ
continuous response	r^* : $O(n^{-3/2})$	directional: $O(n^{-1})$
discrete response	r^* : $O(n^{-1})$	directional: $O(n^{-1})$

Vector parameter of interest

- $\theta = (\psi, \lambda), \quad \psi \in \mathbb{R}^{d_0} \quad H_0 : \psi = \psi_0$
- likelihood ratio test:

$$W(\psi_0) = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \hat{\lambda}_{\psi_0})\} \;\dot\sim\; \chi^2_{d_0}$$

- Bartlett correction:

$$\widetilde{W}(\psi_0) = \frac{W(\psi_0)}{1 + B(\psi_0)/n} \;\dot\sim\; \chi^2_{d_0}\{1 + O(n^{-2})\}$$
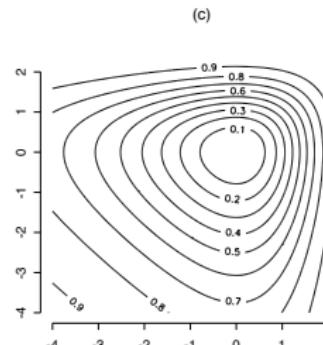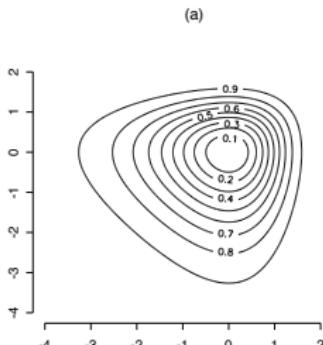
Directional tests

- given a vector of 'departures' from ψ_0
  e.g. $(\hat{\psi}_1 - \psi_{01}, \ldots, \hat{\psi}_{d_0} - \psi_{0d_0})$
- compute a directional departure based on the magnitude of the vector, conditional on its length
- preferred departure measure based on score function
- Propose: directed departure on profile sample space $\mathcal{S}_{\psi_0}$
- all sample points that give the same estimate for the nuisance parameter $\hat{\lambda}_\psi$
-

$$\mathcal{S}_\psi = \{s : \ell_\lambda(\hat{\theta}^0_\psi; s) = 0\} \qquad = \{s : \hat{\varphi}_\psi = \hat{\varphi}^0_\psi\}$$

a surface of dimension d_0, passing through data point y^0

... directional tests

- ▶ Directed departure on $\mathcal{S}_{\psi_0}$ $\qquad \lambda = \hat{\lambda}_{\psi_0}$
- ▶ observed value $s^0 = 0$, corresponding to m.l.e. $\hat{\theta}^0 = \hat{\varphi}^0$
- ▶ expected value s_0 under H_0, corresponding to constrained m.l.e. $\hat{\varphi} = \hat{\varphi}^0_{\psi_0}$ $\qquad s_0 = -\ell_\varphi(\hat{\theta}_{\psi_0})$

- ▶ Distribution of magnitude of $|s - s_0|$; given the direction $(s - s_0)/|s - s_0|$
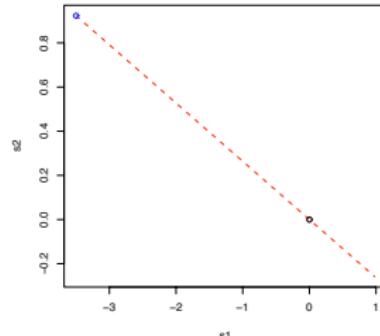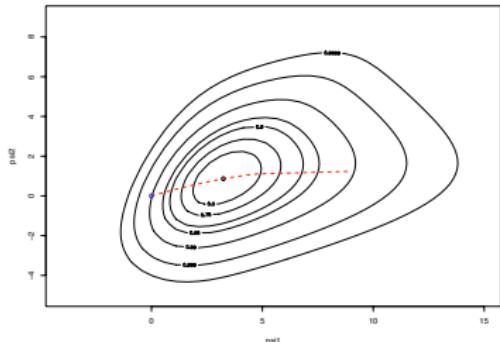
.... line through s_0, s^0

Directional *p*-value

- line $s(t)$ from hypothesis, s_0, to data, $s^0 : s_0 + t(s^0 - s_0)$
- $f(s; \psi_0)$: distribution of s under H_0. Observed value $s^0 = 0$ ($t = 1$)
- Expected value s_0 ($t = 0$)
- along the line $s(t)$ we have

  $f(s; \psi_0)ds = f\{s(t); \psi_0\}dt = f\{s_0 + t(s^0 - s_0); \psi_0\}dt$.

- directional *p*-value:

$$p(\psi_0) \quad = \quad \frac{\int_1^\infty t^{d_0-1} f\{s(t); \psi_0\}dt}{\int_0^\infty t^{d_0-1} f\{s(t); \psi_0\}dt}$$

- one-dimensional integrals computed numerically

Example: 2×3 contingency table

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

	Affective disorders	Schizophrenics	Neurotics
Retarded	12	13	5
Not retarded	18	17	25

► model: log-linear $y \sim$ Poisson, $\quad \log\{E(y)\} = X\varphi \in \mathbb{R}^6$
► nuisance parameter $\lambda \in \mathbb{R}^4$
► parameter of interest (ψ_1, ψ_2): interaction
► $H_0 : \psi = \psi_0 = (0, 0)$ $\qquad\qquad$ independence

... 2×3 contingency table

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

	Affective disorders	Schizophrenics	Neurotics
Retarded	10	10	10
Not retarded	20	20	20

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

	Affective disorders	Schizophrenics	Neurotics
Retarded	14	16	0
Not retarded	16	14	30

▶ All tables along the line $s(t)$ have the same margins.

Log-likelihood along the line $s(t)$
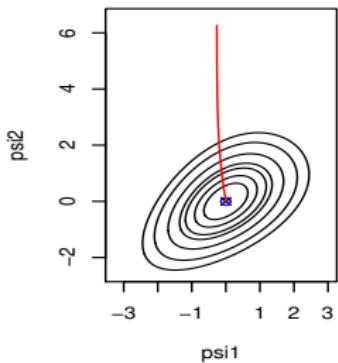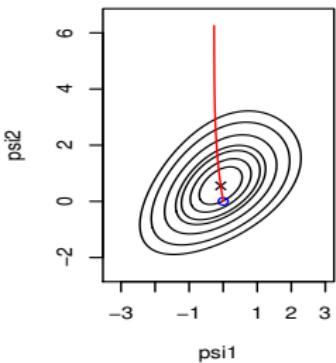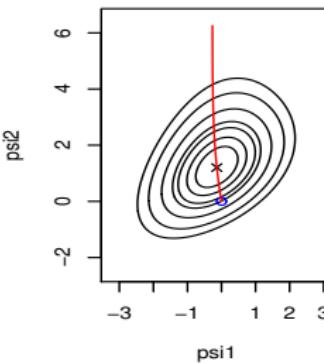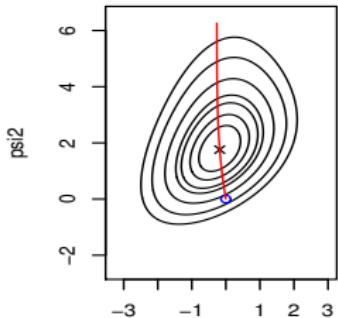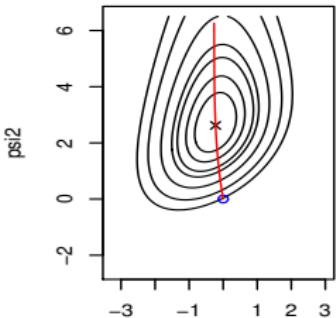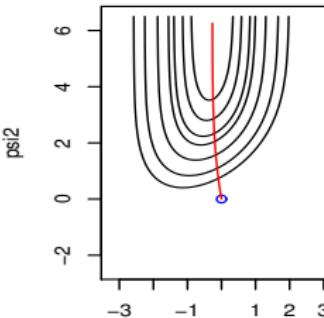
Directional *p*-value

▶ The directional *p*-value is equal to 0.050
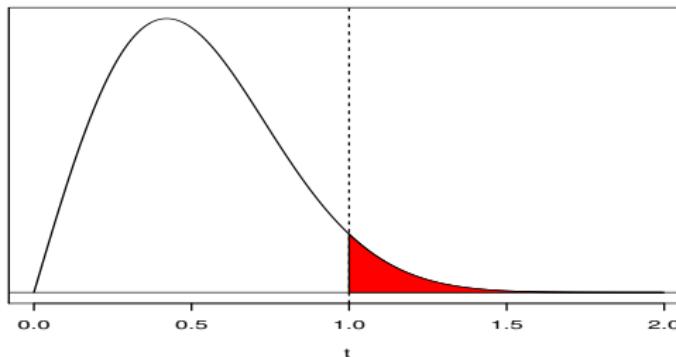
first order χ_2^2 approximation	$W(\psi_0)$	0.047
Skovgaard (2001 SJS) modified version	$W^*(\psi_0)$	0.048
simulated conditional		0.051

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

$$f(s_1, s_2; \psi) \doteq \frac{e^{c/n}}{(2\pi)^{d_0}} \exp[\{(\psi - \hat{\psi})'s + \ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]|\hat{\jmath}_{\varphi\varphi}|^{-1/2}$$

- on line $s(t) = s_0 + t(s^0 - s_0)$

Another 2×3 table

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

	Democrat	Independent	Republican
Black	103	15	11
White	341	105	405

- H_0: independence nested in saturated model
- first order likelihood ratio p-value: 2.43×10^{-20}
- directional p-value: 3.14×10^{-20}.

... party identification example

- Expected ($t = 0$)

	Democrat	Independent	Republican
Black	58.44	15.80	54.76
White	385.56	104.20	361.24

- Observed ($t = 1$)

	Democrat	Independent	Republican
Black	103	15	11
White	341	105	405

- Boundary ($t_{max} = 1.251$)

	Democrat	Independent	Republican
Black	114.20	14.80	0.00
White	329.80	105.20	416.00

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
- $E(y) = \exp(X\beta)$
-
$$\ell(\beta) = \beta' X' y - \mathbf{1}' \exp(X\beta)$$

- $\varphi(\beta) = \beta$
- $\beta = (\lambda, \psi)$, and design matrix partitioned as $X = (X_1 \quad X_2)$
-
$$\ell_\beta(\beta) = \begin{bmatrix} \ell_\lambda(\lambda, \psi) \\ \ell_\psi(\lambda, \psi) \end{bmatrix} = \begin{bmatrix} X_1'(y - e^{X\beta}) \\ X_2'(y - e^{X\beta}) \end{bmatrix}.$$

- 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 $X_1'y$

... log-linear models

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

$$s_0 = -\ell_\beta(\hat{\beta}_{\psi_0}) = \begin{bmatrix} \mathbf{0} \\ -X_2^T(y - e^{X\hat{\beta}_{\psi_0}}) \end{bmatrix}.$$

- ▶ directional test goes radially from s_0 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 $\texttt{glm}$ with numerical integration (univariate)

Example: suicide behavior data

		age	m	o	y
cause	sex				
drug	f		450	154	259
	m		399	93	398
gas	f		13	5	15
	m		82	6	121
gun	f		26	7	14
	m		168	33	155
hang	f		450	185	95
	m		797	316	455
jump	f		71	38	40
	m		51	26	55
other	f		60	10	38
	m		82	14	124

... suicide behavior data

- ▶ suicide classified by sex, age and cause Everitt, 1992 CH
- ▶ H_0: independence, nested in saturated model
- ▶ nuisance parameter λ: intercept, main effects, first order interaction parameters $\in \mathbb{R}^{26}$
- ▶ saturated model: add parameter of interest $\psi \in \mathbb{R}^{10}$; second order interaction between the three variables.
- ▶ The first order likelihood ratio *p*-value is 0.136.
- ▶ The directional *p*-value is 0.141.

Example: Infant survival data

	survival	gestation	smoking	age	Freq
1	No	$\leq$260	<5	<30	50
2	Yes	$\leq$260	<5	<30	315
3	No	>260	<5	<30	24
4	Yes	>260	<5	<30	4012
5	No	$\leq$260	>5	<30	9
6	Yes	$\leq$260	>5	<30	40
7	No	>260	>5	<30	6
8	Yes	>260	>5	<30	459
9	No	$\leq$260	<5	>30	41
10	Yes	$\leq$260	<5	>30	147
11	No	>260	<5	>30	14
12	Yes	>260	<5	>30	1594
13	No	$\leq$260	>5	>30	4
14	Yes	$\leq$260	>5	>30	11
15	No	>260	>5	>30	1
16	Yes	>260	>5	>30	124

...infant survival data

- ▶ four dichotomous variables: age of mother (A), length of gestation (G), infant survival (I) and number of cigarettes smoked per day during gestation (S) Agresti, 2002
- ▶ 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 ψ for: IS and GA
- ▶ first order likelihood ratio *p*-value = 0.052.
- ▶ directional *p*-value = 0.056.

Conclusion

|	scalar ψ	vector ψ
continuous response	$r^* : O(n^{-3/2})$	directional: $O(n^{-1})$
discrete response	$r^* : O(n^{-1})$	directional: $O(n^{-1})$

– tangent exponential model
– saddlepoint approximation
– easy, accurate

– extensions to nonlinear hypotheses, more complex models
for categorical data...

1) Sample / Parameter space notation
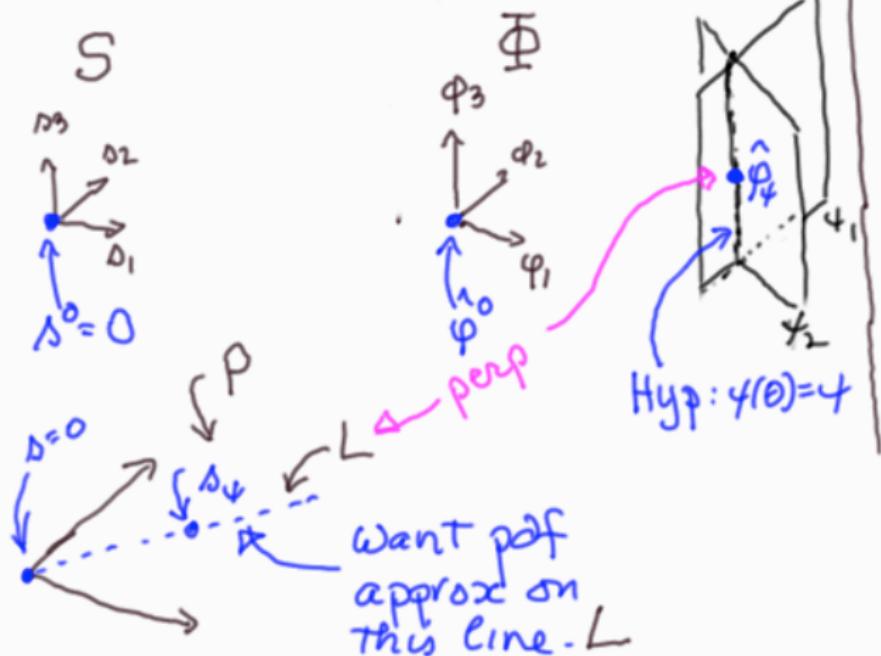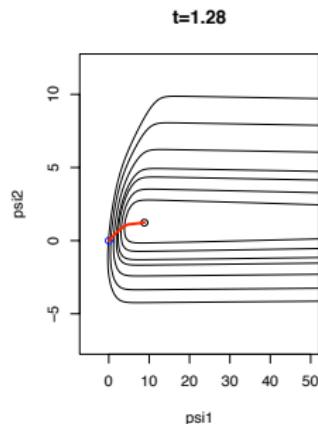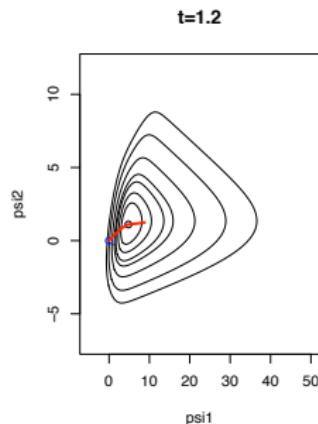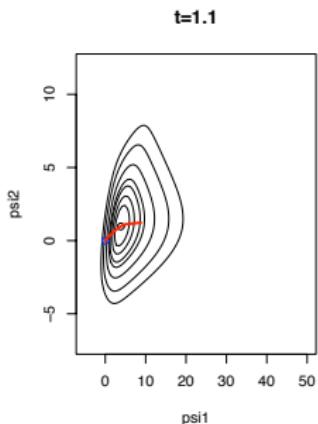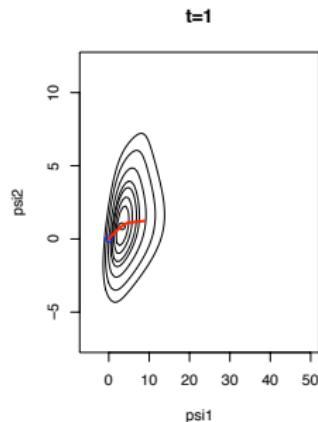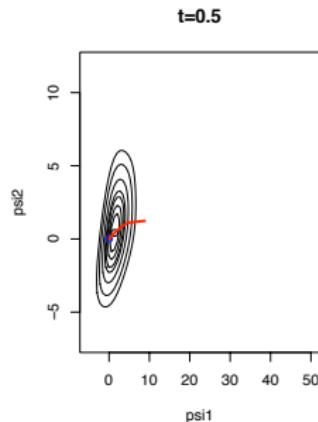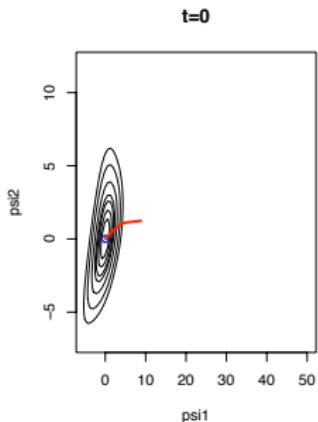
$$S \qquad \Phi$$

s_3, s_2, s_1

$s^0 = 0$

φ_3, φ_2, φ_1

φ_0
φ

perp

$\hat{\varphi}_\varphi$

φ_1

φ_2

Hyp: $\varphi(\theta) = \varphi$

P

L

$s = 0$

s_φ

Want pdf
approx on
this line. L

Plane on s-space
⊥ to φ-planes on φ
space : contains
$s = 0$ and s_φ

References

Fraser, D.A.S. and Massam, H. (1985). Conical tests. *Stat. Hefte*

Skovgaard, I.M. (1988). Saddlepoint expansions for directional test probabilities. *JRSS B*

Cheah, P.K., Fraser, D.A.S., Reid, N. (1994). Multiparameter testing in exponential models. *Biometrika*

Davison, A.C., Fraser, D.A.S., Reid, N. (2006). Improved likelihood inference for discrete data. *JRSS B*

Davison, A.C., Fraser, D.A.S., Reid, N., Sartori, N. (2009). On assessing vector valued parameters. in progress