Empirical Likelihood:

Some History and New Directions

Art B. Owen

Stanford University

Gottfried Noether (1915–1991)

- Pioneer and advocate for nonparametrics
- Mathematical family: father Fritz, aunt Emmy, grandfather Max

Survey of early nonparametrics

Gottfried E. Noether (1984)

Nonparametrics: The Early Years-Impressions and Recollections

The American Statistician 38(3), pp173–178

Nugget: Wilcoxon wanted to avoid the computational cost of t-tests

Other contributions

Key results underlying linear rank statistics

Understanding of Pitman tests and efficiency

Advocate of nonparametrics for its simplicity of explanation

Parametric likelihood

Let $X_i \overset{\text{iid}}{\sim} f(x;\theta) \quad \theta \in \mathbb{R}^p$

Observe $X_i = x_i, i = 1, \ldots, n.$

Likelihood is

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

Uses

1) MLE: $\hat{\theta} = \arg\max_\theta L(\theta)$
2) Wilks: reject $H_0 : \theta = \theta_0$ if $-2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right) \geqslant \chi^{2,1-\alpha}_{(p)}$
3) Confidence set: unrejected θ_0s

However

Why do the data have to follow one of **our** distributions?

Nonparametric MLE

Let $X_i \overset{\text{iid}}{\sim} F$ and

$$L(F) = \prod_{i=1}^{n} F(\{x_i\}) \qquad \text{prob of exactly those } x_i$$

By Kiefer & Wolfowitz (1956):

$$\widehat{F} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad \text{empirical CDF}$$

Other NPMLEs

- Kaplan & Meier (1958) censored data
- Hartley & Rao (1968) survey data (Jon Rao)
- Grenander (1956) monotone density for actuarial data
- Linden-Bell (1971) truncated data in astronomy

Likelihood ratios

Inference	Parametric	Nonparametric
Point	MLE $\hat{\theta}$	NPMLE $\widehat{F}$
Interval	$-2\log\left(\frac{L(\theta)}{L(\hat{\theta})}\right) \to \chi^2_{(p)}$	???

Empty corner

Empirical likelihood goes there.

Test based on $R(F) = L(F)/L(\widehat{F})$.

Original motivation

Which way should a bootstrap confidence interval skew?

Inspiration: Thomas & Grunkemeier (1975) for survival analysis.

Nonparametric likelihood ratio

Likelihood $$L(F) = \prod_{i=1}^{n} F(\{x_i\})$$

Likelihood ratio $$R(F) = L(F)/L(\widehat{F})$$

Confidence region for $T(F)$

E.g., T is mean or median or regression coefficient $\cdots$

Consider

$$\{T(F) \mid R(F) \geqslant r\}$$

Parametric: $-2\log(r) = \chi^{2,1-\alpha}_{(p)}$

Profile likelihood

$$\mathcal{R}(\theta) \equiv \sup\{R(F) \mid T(F) = \theta\}$$

$$\text{Region} = \{\theta \mid \mathcal{R}(\theta) \geqslant r\}$$

A big multinomial

1) Time today is short

2) Story kind of long

See the EL monograph

Upshot

- EL ratio usually reduces to a multinomial on $x_1, \ldots, x_n$
- Even when there are ties!

Parameter count

- n observations x_i, and
- n parameters $w_i = F(\{x_i\})$

Could be trouble: Neyman-Scott (1948)

The multinomial

$$\delta_x = \text{ point mass at } x$$

$$F \equiv \sum_{i=1}^{n} w_i \delta_{x_i}$$

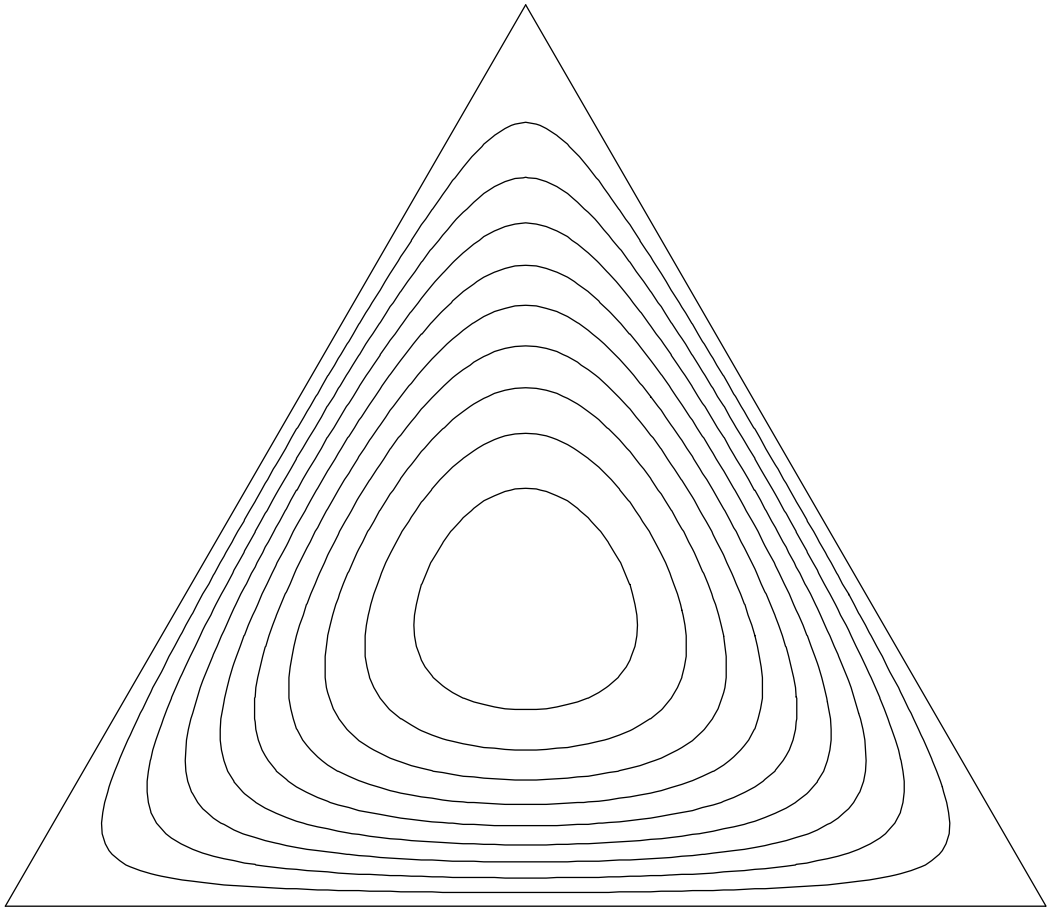
$$\sum_{i=1}^{n} w_i = 1$$

Likelihood and likelihood ratio

$$L(F) = \prod_{i=1}^{n} w_i$$

$$R(F) = \frac{L(F)}{L(\widehat{F})} = \prod_{i=1}^{n} (n w_i)$$

Likelihood ratio for $n = 3$



Contours of $\prod_i nw_i$ Levels $0.0, 0.1, 0.2, \cdots, 0.9$

EL for the mean

$$x_i \in \mathbb{R}^d \qquad T(F) = \sum_{i=1}^{n} w_i x_i$$

Profile EL

$$\log(\mathcal{R}(\mu)) \equiv \max_{w} \left\{ \sum_{i=1}^{n} \log(n w_i) \;\middle|\; \sum_{i=1}^{n} w_i (x_i - \mu) = 0, \quad w_i \geqslant 0, \quad \sum_{i=1}^{n} w_i = 1 \right\}$$

- Easy convex optimization in $\mathbb{R}^d$ (duality)
- Iteratively reweighted least squares

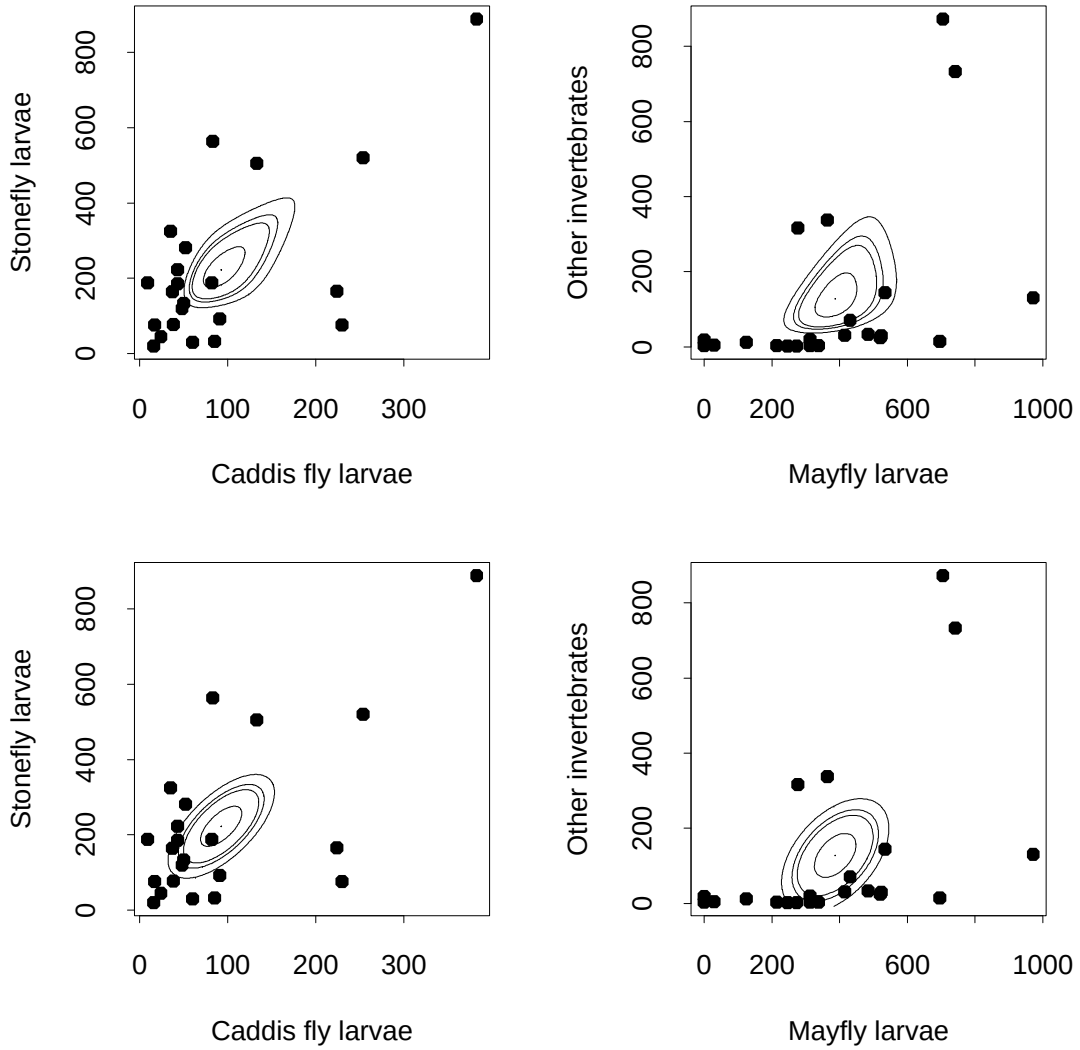
Wilks for EL

$$-2 \log(\mathcal{R}(\mu_0)) \to \chi^2_{(d)} \qquad \text{O (1987,1990)}$$

A dipper

From Wikipedia “White-throated dipper”, credit: Mark Medcalf

Dipper diet means



Top row shows EL; bottom Hotelling's T^2 ellipses

Data from Iles (1993) 22 sites in Wales.

Hall (1990) quantifies correctness of shape.

Convex duality

We maximize $\sum_{i=1}^n \log(nw_i)$ subject to

- $w_i \geqslant 0$
- $\sum_{i=1}^n w_i = 1$
- $\sum_{i=1}^n w_i(x_i - \mu) = 0$

μ in convex hull of $x_1, \ldots, x_n$

Lagrange multipliers

Optimal w_i are

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda^\mathsf{T}(x_i - \mu)}$$

where $\lambda = \lambda(\mu) \in \mathbb{R}^d$ uniquely satisfies

$$\frac{1}{n} \sum_{i=1}^n \frac{x_i - \mu}{1 + \lambda^\mathsf{T}(x_i - \mu)} = 0$$

Estimating equations

θ solves $\mathbb{E}(m(\boldsymbol{x}, \theta)) = 0$

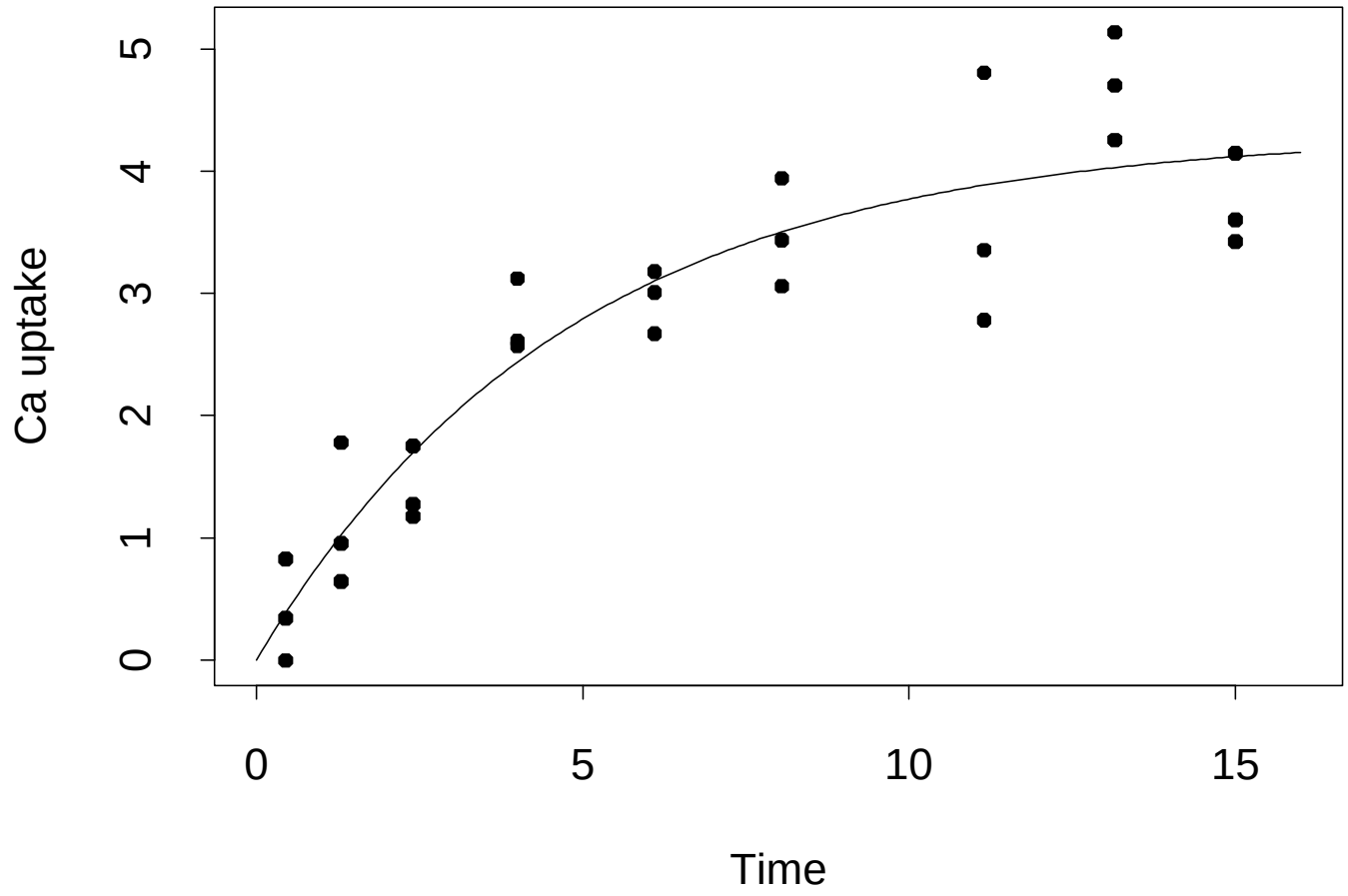
Examples

$m(\boldsymbol{x}, \theta)$	Estimand θ
$\boldsymbol{x} - \theta$	$\mathbb{E}(\boldsymbol{x})$
$1_{x<\theta} - 0.5$	median(x)
$\boldsymbol{x}(y - \boldsymbol{x}^{\mathsf{T}}\beta)$	regression
$\frac{\partial}{\partial\theta} \log f(\boldsymbol{x}, \theta)$	MLE

Godambe, Huber, White, Thompson $\cdots$

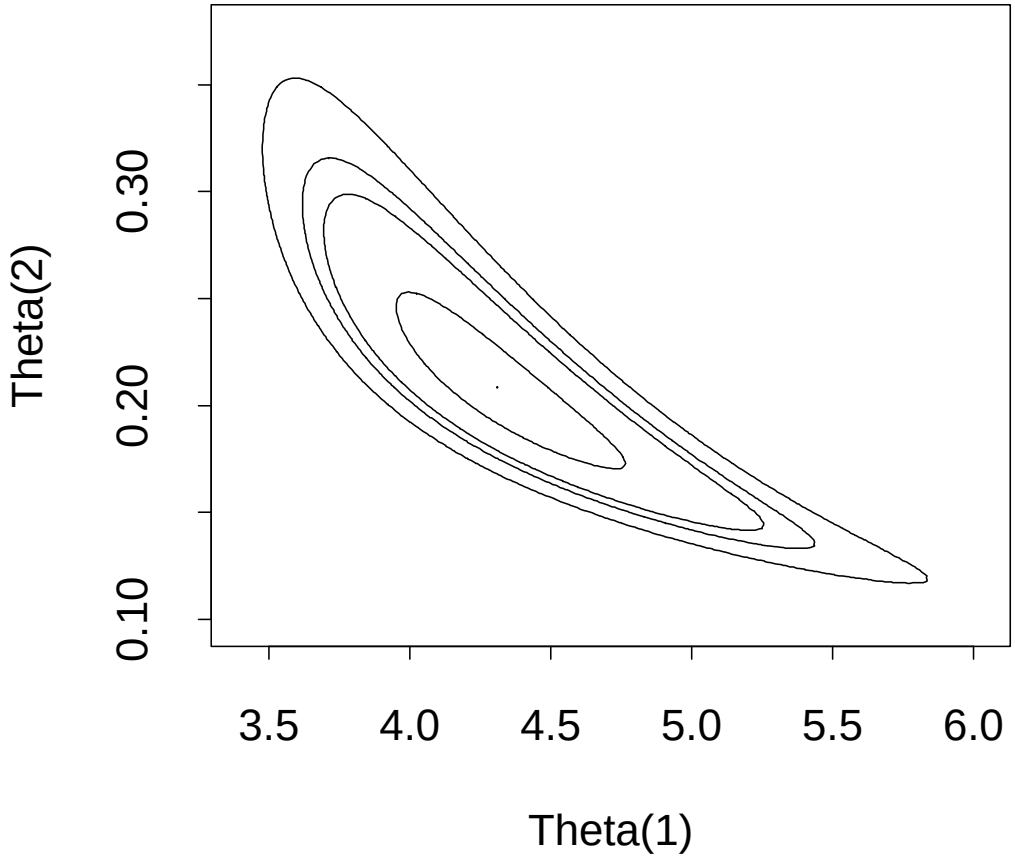
Nonlinear regression

Data from Bates & Watts (1988)



$$y_i = \underbrace{\theta_1(1 - \exp(-\theta_2 x_i))}_{f(x_i,\theta)} + \varepsilon_i$$

Nonlinear regression



$$0 = \sum_{i=1}^{n} w_i \Big(y_i - f(x_i, \theta) \Big) \frac{\partial}{\partial \theta} f(x_i, \theta)$$

Don’t need: normality or constant variance

Extensions

Apologies for many omissions

Maximum EL Estimate	J. Qin & Lawless (1994)
GLM	Kolaczyk (1994)
Kernel densities	Hall & O (1993)
Time series	Kitamura (1997)
” ”	Monte, Haerdle, J. Chen, Gao, Nordman, Lahiri
Survival analysis	M. Zhou (2015)
” ”	McKeague, Li, Zhou, Jing, G. Qin
Econometrics	Imbens, Newey, Smith, Kitamura, Guggenberger, Schennach
” ”	Mittlehammer, Judge, Miller (2000)
Survey sampling	J. Chen, Qin, Rao, Wu, Sitter
Case control	Qin, YH Chen, N Chatterjee, Carroll, Maas
Power	Kitamura, Mykland, Lazar
Theory	Hall, Romano, DiCiccio, McKeague, Hjort, Van Keilegom, . . .
Robust optimization	Duchi, Glynn, Namkoong, 2016++

Calibration

Exact nonparametric confidence intervals on $\mathbb{E}(X)$ do not exist

Bahadur & Savage (1956)

Better coverage

O (1992)	Bootstrap calibration
DiCiccio, Hall & Romano (1991)	Bartlett correction
J. Chen, A. Variyath & B. Abraham (2012)	Adding "undata"
Min Tsao (2013)	warping space
Sarah Emerson (2009)	thesis

Some of these methods escape the convex hull

Nuisance parameters

Computing

$\mathcal{R}(\theta_1, \ldots, \theta_p)$ easy for estimating equations

$\mathcal{R}(\theta_j) = \max_{\theta_j} \mathcal{R}(\theta_1, \ldots, \theta_p)$ can be hard to do

Probably best to Taylor approximate $\log(\mathcal{R}(\theta))$ around $\hat{\theta}$

Wish list

1) Log concave densities

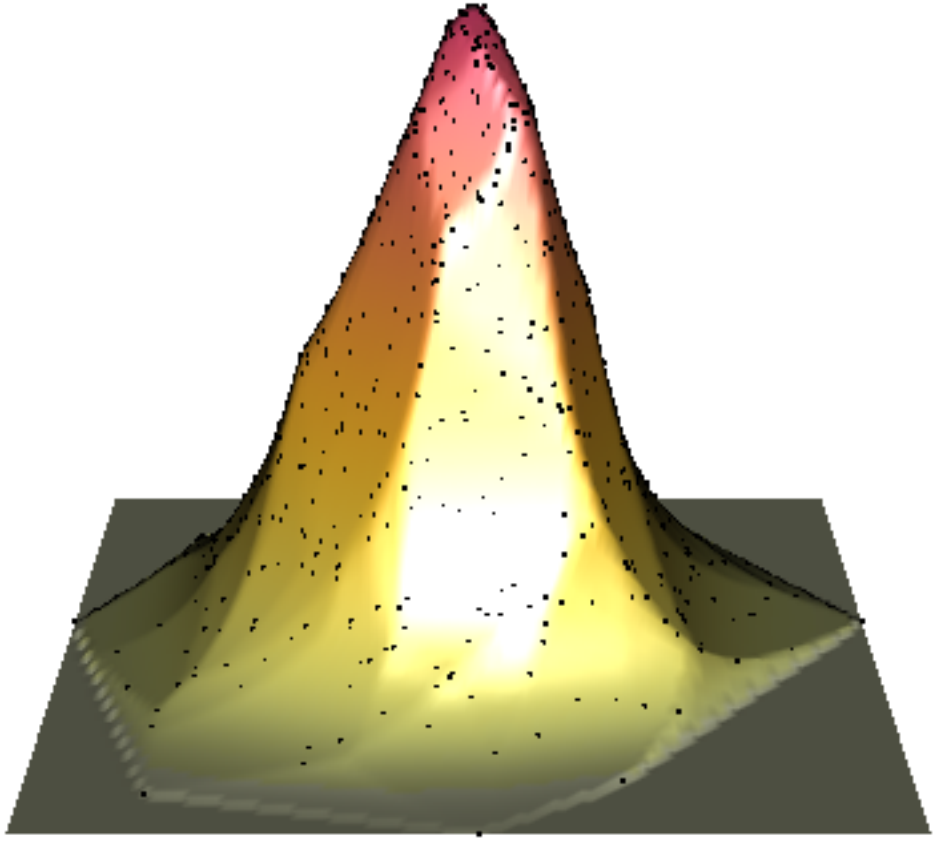
2) (More) random effects

3) (More) Bayesian empirical likelihood

Some fear

Maybe what I want has been done and I just didn't find it!

A log concave MLE



Cule, Samworth, Stewart (2010)

x continuous with $\log f(x)$ concave

Can we get nonparametric likelihood ratio regions for moments?

Random effects

EL: # param. $=$ # obs.

Nested effects

$$Y_{ij} = x_{ij}^{\mathsf{T}}\beta + a_i + \varepsilon_{ij}$$

Crossed effects

$$Y_{ij} = x_{ij}^{\mathsf{T}}\beta + a_i + b_j + \varepsilon_{ij}$$

Could this work?

$$\varepsilon_{ij} \overset{\text{iid}}{\sim} F \qquad a_i \overset{\text{iid}}{\sim} G \qquad b_j \overset{\text{iid}}{\sim} H$$

More parameters than data.

One example is Sieve EL: Shen, Shi, Wong (1999)

Bayesian EL

The dream:

Science $\longrightarrow$ prior

Your data $\longrightarrow$ likelihood

The "other" nonparametric Bayes

Approach

Lazar (2003)

$$p(\theta \mid \boldsymbol{x}) \propto p(\theta)\mathcal{R}(\theta \,|\, \boldsymbol{x})$$

Posterior calibration via Bayesian CLT

Like Boos & Monahan (1992)

Also

Hjort (2017) Bernstein von-Mises for EL (at BNP 11 in Paris)

Bayes approaches

$$p(\theta \mid \boldsymbol{x}) \propto p(\theta)\mathcal{R}(\theta \mid \boldsymbol{x})$$

Approximate Bayesian Computation with EL

Mengersen, Pudlo & Robert (2013)

Hamiltonian MCMC

Chaudhuri, Mondal & Yin (2017)

Gradients of posterior get steep just where you need it

Bayesian exponentially tilted EL

Schennach (2005)

EL on a p-dimensional family formed by exponential tilting

for overdetermined estimating equation

Thanks

- Emiliana Noether (1917–2018) Professor of Italian History, U. Conn.
- Monica Noether V.P. Charles River Associates
- Elizabeth Henry & Naomi Friedman, ASA
- Ray Carroll, session chair
- Jiahua Chen & Noether committee
- Several NSF grants

Some extras

Here are a few extra EL ideas.

Least favorable families

There is a **worst** p-dimensional parametric family through F

DiCiccio & Romano (1990) show that EL uses that least favorable family (LFF), asymptotically.

The connection

$$
\begin{aligned}
p(\theta) &\quad \text{induces a prior on the LFF,} \quad \text{and} \\
\mathcal{R}(\theta \mid \boldsymbol{x}) &\doteq \text{likelihood on LFF} \\
\Longrightarrow \; p(\theta) \times \mathcal{R}(\theta \mid \boldsymbol{x}) &\doteq \text{posterior on LFF} \\
&= \text{posterior on } \theta
\end{aligned}
$$

Selection bias

Sample

- families proportionally to # kids
- cotton fibers proportional to length
- shrubs proportional to width

$x_i \sim F$ but seen with prob $\propto u(x)$

True x_i has PDF / CDF f, F,

Measured x_i have g, G

$$g(x) = u(x)f(x) \Big/ \int_0^\infty u(x)f(x)\,\mathrm{d}x$$

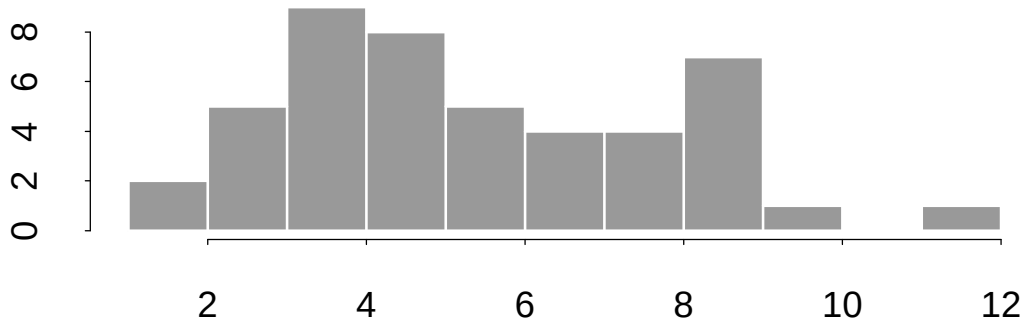
Adjust the estimating equations

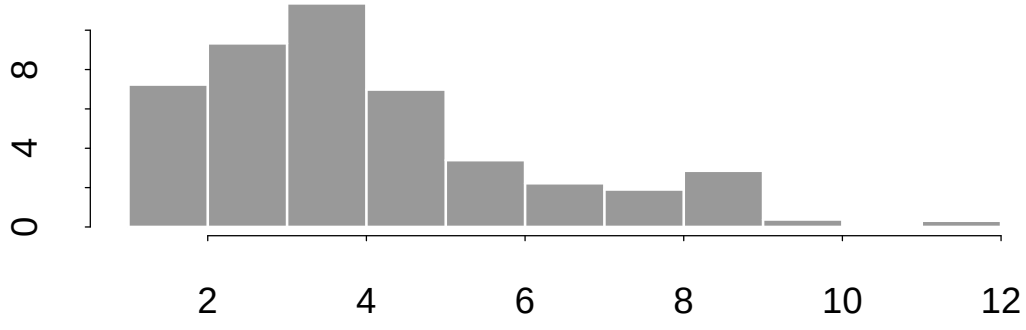
$$\sum_{i=1}^{n} w_i \frac{m(x_i, \theta)}{u(x_i)}$$

Horvitz-Thompson adjustment

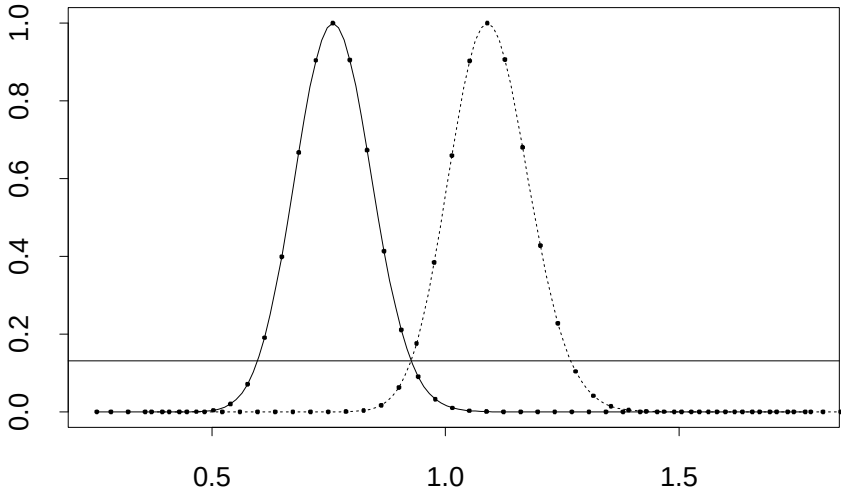
Transect sampling of shrubs

Muttlak & McDonald





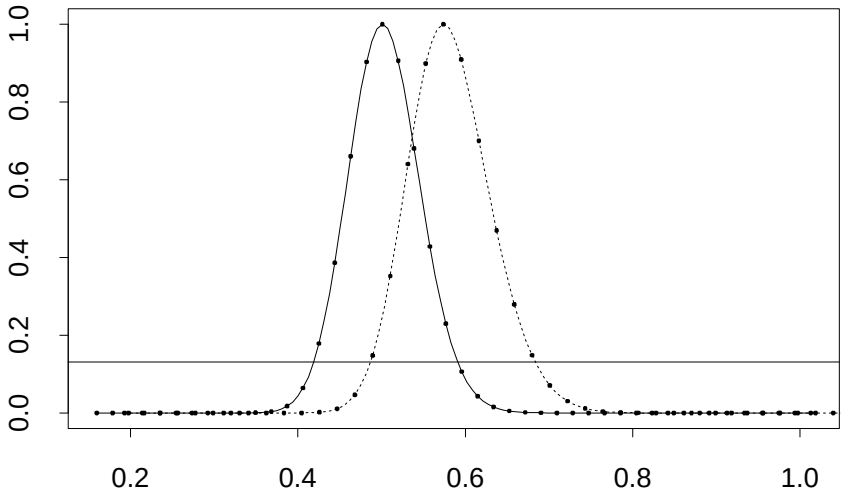
Mean shrub width



$$0 = \sum_{i=1}^{n} w_i \frac{x_i - \mu}{x_i}$$ Solid

$$0 = \sum_{i=1}^{n} w_i (x_i - \mu)$$ Dotted

Standard dev. of shrub width



$$0 = \sum_{i=1}^{n} w_i \frac{(x_i - \mu)^2 - \sigma^2}{x_i}$$ Solid

$$0 = \sum_{i=1}^{n} w_i ((x_i - \mu)^2 - \sigma^2)$$ Dotted