

# Empirical Likelihood

Art B. Owen

Department of Statistics  
Stanford University

## Thanks to

- University of Ottawa
- Fields Institute
- Mayer Alvo
- Jon Rao

## This talk

- based on the book “Empirical Likelihood” (2001)\*
- starts with central topics, spirals out, ends with challenges

\* Chapman & Hall/CRC holds copyright on figures here. They appear with permission. Book available at [www.amazon.com](http://www.amazon.com) or [www.crcpress.com](http://www.crcpress.com). Book website at [www-stat.stanford.edu/~owen/empirical](http://www-stat.stanford.edu/~owen/empirical)

## Empirical likelihood provides:

- **likelihood** methods for inference, especially
  - tests, and
  - confidence regions,
- **without** assuming a parametric model for data
- **competitive** power even when parametric model holds

# Parametric likelihoods

Data have *known* distribution  $f_\theta$  with *unknown parameter*  $\theta$

$$\Pr(X_1 = x_1, \dots, X_n = x_n) = f(x_1, \dots, x_n; \theta)$$

$$\Pr(x_1 \leq X_1 \leq x_1 + \Delta, \dots, x_n \leq X_n \leq x_n + \Delta) \propto f(x_1, \dots, x_n; \theta)$$

$f(\dots; \cdot)$  known,  $\theta \in \Theta \subseteq \mathbb{R}^p$  unknown

## Likelihood function

$$L(\theta) = L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta)$$

“Chance, under  $\theta$ , of getting the data we did get”

## Likelihood examples

$$X_i \sim \text{Poi}(\theta), \quad \theta \geq 0$$

$$L(\theta) = \prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{x_i!}$$

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad x_i \text{ fixed}$$

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}$$

# Likelihood inference

## Maximum likelihood estimate

$$\hat{\theta} = \arg \max_{\theta} L(\theta; x_1, \dots, x_n)$$

## Likelihood ratio inferences

$$-2 \log(L(\theta_0)/L(\hat{\theta})) \rightarrow \chi_{(q)}^2 \quad \text{Wilks}$$

Typically . . . Neyman-Pearson, Cramer-Rao, . . .

1.  $\hat{\theta}$  asymptotically normal
2.  $\hat{\theta}$  asymptotically efficient
3. Likelihood ratio tests powerful
4. Likelihood ratio confidence regions small

## Other likelihood advantages

- can model data distortion: bias, censoring, truncation
- can combine data from different sources
- can factor in prior information
- obey range constraints: MLE of correlation in  $[-1, 1]$
- transformation invariance
- data determined shape for  $\{\theta \mid L(\theta) \geq rL(\hat{\theta})\}$
- incorporates nuisance parameters

# Unfortunately

We might not know a correct  $f(\dots; \theta)$

No reason to expect that new data belong to one of our favorite families

Wrong models sometimes work (e.g. Normal mean via CLT) and sometimes fail (e.g. Normal variance)

## Also,

Usually easy to compute  $L(\theta)$ , but . . .

Sometimes hard to find  $\hat{\theta}$

Sometimes hard to compute  $\max_{\theta_2} L((\theta_1, \theta_2))$  (Profile likelihood)



# Nonparametric methods

Assume only  $X_i \sim F$  where

- $F$  is continuous, or,
- $F$  is symmetric, or,
- $F$  has a monotone density, or,
- $\dots$  other believable, but big, family

Nonparametric usually means infinite dimensional parameter

Sometimes lose power (e.g. sign test), sometimes not

# Nonparametric maximum likelihood

For  $X_i$  IID from  $F$ , 
$$L(F) = \prod_{i=1}^n F(\{x_i\})$$

The NPMLE is 
$$\hat{F} = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$$

where  $\delta_x$  is a point mass at  $x$

Kiefer and Wolfowitz, 1956

# Proof

Distinct values  $z_j$  appear  $n_j$  times in sample,  $j = 1, \dots, m$

Let  $F(\{z_j\}) = p_j \geq 0$  and  $\hat{F}(\{z_j\}) = \hat{p}_j = n_j/n$  with some  $p_j \neq \hat{p}_j$

$$\begin{aligned}\log\left(\frac{L(F)}{L(\hat{F})}\right) &= \sum_{j=1}^m n_j \log\left(\frac{p_j}{\hat{p}_j}\right) \\ &= n \sum_{j=1}^m \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) \\ &< n \sum_{j=1}^m \hat{p}_j \left(\frac{p_j}{\hat{p}_j} - 1\right) \\ &= 0. \quad \square\end{aligned}$$

## Other NPMLEs

|              |                                     |
|--------------|-------------------------------------|
| Kaplan-Meier | Right censored survival times       |
| Lynden-Bell  | Left truncated star brightness      |
| Hartley-Rao  | Sample survey data                  |
| Grenander    | Monotone density for actuarial data |

## Nonparametric likelihood ratios

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

Confidence region:  $\{T(F) \mid R(F) \geq r\}$

Profile likelihood:  $\mathcal{R}(\theta) = \sup\{R(F) \mid T(F) = \theta\}$

Confidence region:  $\{\theta \mid \mathcal{R}(\theta) \geq r\}$

In parametric setting,  $-2 \log(r) = \chi_{(q)}^{2,1-\alpha}$

## Suppose there are no ties

Let  $w_i = F(\{x_i\})$   $w_i \geq 0$   $\sum_{i=1}^n w_i \leq 1$

$$L(F) = \prod_{i=1}^n w_i \quad L(\hat{F}) = \prod_{i=1}^n 1/n \quad R(F) = \prod_{i=1}^n n w_i$$

$$\mathcal{R}(\theta) = \sup \left\{ \prod_{i=1}^n n w_i \mid T(F) = \theta \right\}$$

If there are ties . . .

$$L(F) \rightarrow L(F) \times \prod_j n_j^{n_j} \quad \text{and,} \quad L(\hat{F}) \rightarrow L(\hat{F}) \times \prod_j n_j^{n_j}$$

*R and  $\mathcal{R}$  unchanged*

## For the mean

$$T(F) = \int x dF(x), x \in \mathbb{R}^d$$

$$T(\hat{F}) = \frac{1}{n} \sum_{i=1}^n x_i$$

We get  $\{T(F) \mid R(F) \geq \epsilon\} = \mathbb{R}^d, \quad \forall r < 1$

$$\text{Let } F_{\epsilon,x} = (1 - \epsilon)\hat{F} + \epsilon\delta_x$$

For any  $r < 1$ ,

$$R(F_{\epsilon,x}) = \frac{L((1-\epsilon)\hat{F} + \epsilon\delta_x)}{L(\hat{F})} \geq (1 - \epsilon)^n \geq r \text{ for small enough } \epsilon$$

Then let  $\delta_x$  range over  $\mathbb{R}^d$

## Fix for the mean

Restrict to  $F(\{x_1, \dots, x_n\}) = 1$  i.e.  $\sum_{i=1}^n w_i = 1$

Confidence region is

$$C_{r,n} = \left\{ \sum_{i=1}^n w_i x_i \mid w_i \geq 0, \sum_{i=1}^n w_i = 1, \prod_{i=1}^n n w_i > r \right\}$$

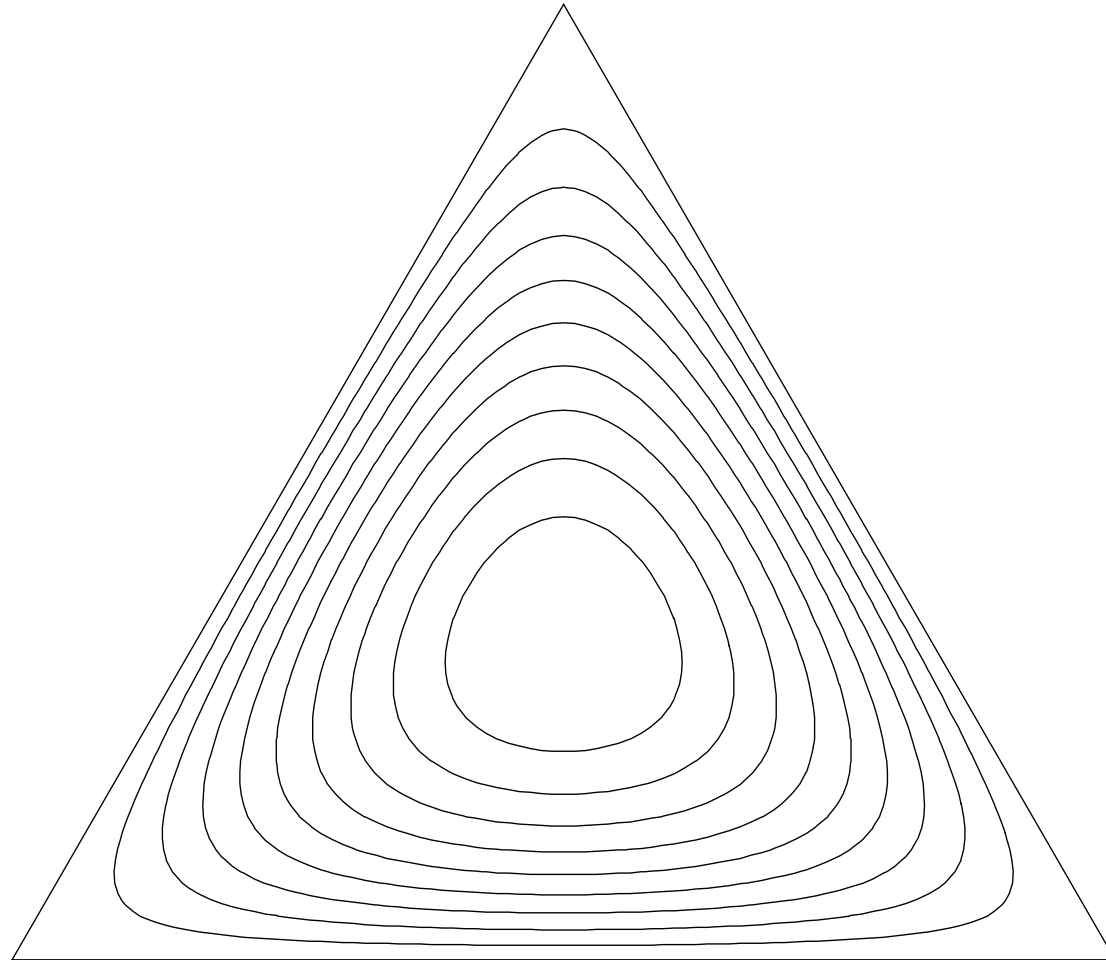
Profile likelihood

$$\mathcal{R}(\mu) = \sup \left\{ \prod_{i=1}^n n w_i \mid w_i > 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i x_i = \mu \right\}$$

We have a multinomial on the  $n$  data points, hence  $n - 1$  parameters



# Multinomial likelihood for $n = 3$



MLE at center     $LR = i/10, i = 0, \dots, 9$

# Empirical likelihood theorem

Suppose that  $X_i \sim F_0$  are IID in  $\mathbb{R}^d$

$$\mu_0 = \int x dF_0(x)$$

$$V_0 = \int (x - \mu_0)(x - \mu_0)^T dF_0(x) \text{ finite}$$

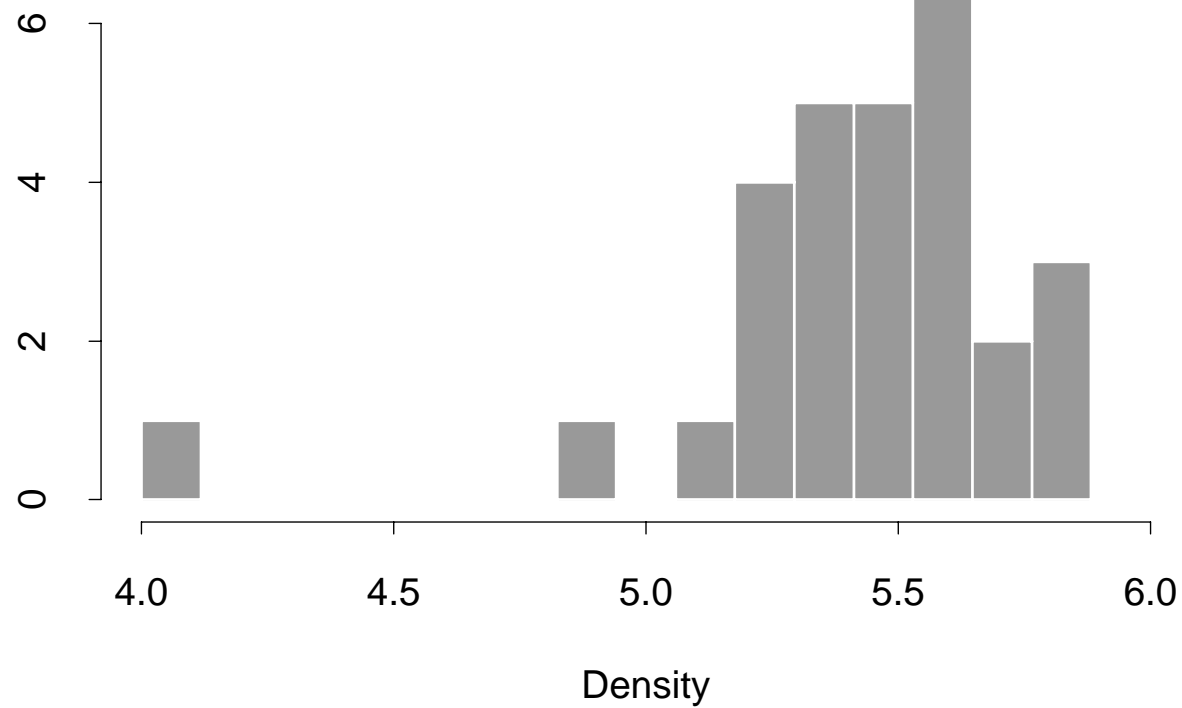
$$\text{rank}(V_0) = q > 0$$

Then as  $n \rightarrow \infty$

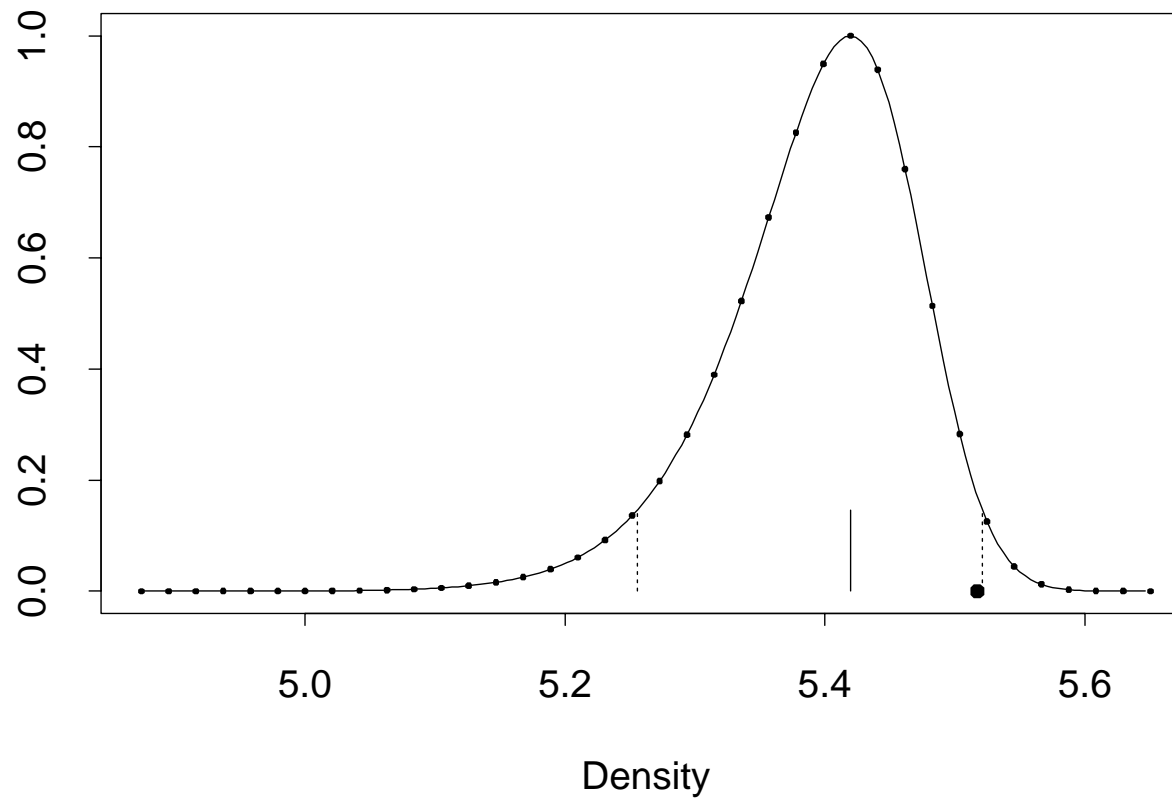
$$-2 \log \mathcal{R}(\mu_0) \rightarrow \chi_{(q)}^2$$

same as parametric limit

# Cavendish's measurements of Earth's density



# Profile empirical likelihood

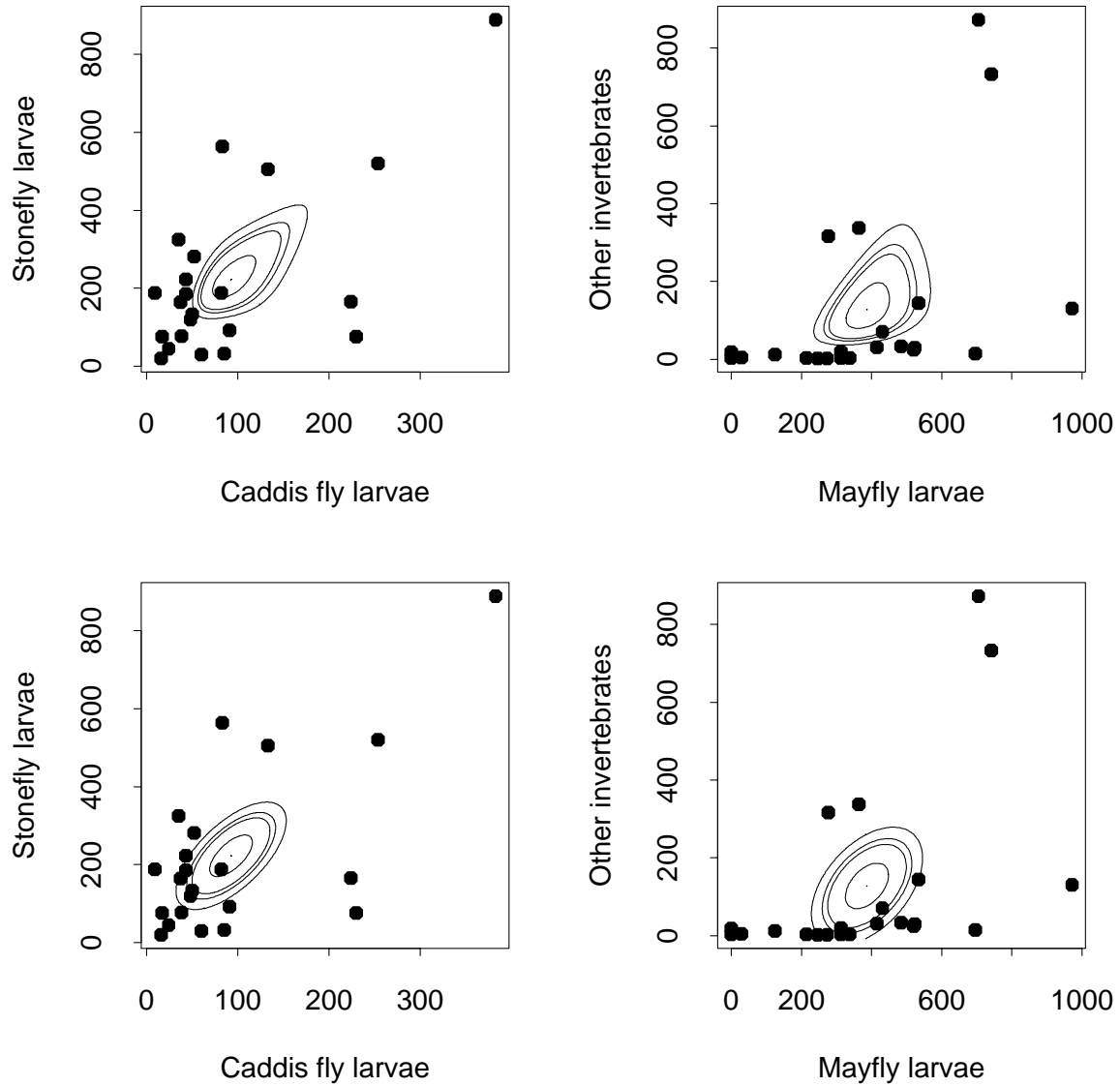


# Dipper, *Cinclus cinclus*

Possibly copyrighted picture of a dipper bird went here

Eats larvae of Mayflies, Stoneflies, Caddis flies, other

# Dipper diet means



# Convex Hull

$$\mathcal{H} = \mathcal{H}(x_1, \dots, x_n) = \left\{ \sum_{i=1}^n w_i x_i \mid w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}$$

$$\mu \notin \mathcal{H} \implies \log \mathcal{R}(\mu) = -\infty$$

If  $\mu \in \mathcal{H}$  we get  $\mathcal{R}(\mu)$  by Lagrange multipliers

# Lagrange multipliers

$$G = \sum_{i=1}^n \log(nw_i) - n\lambda' \left( \sum_{i=1}^n w_i(x_i - \mu) \right) + \gamma \left( \sum_{i=1}^n w_i - 1 \right)$$

$$\frac{\partial}{\partial w_i} G = \frac{1}{w_i} - n\lambda'(x_i - \mu) + \gamma = 0$$

$$\sum_i w_i \frac{\partial}{\partial w_i} G = n + \gamma = 0 \quad \implies \quad \gamma = -n$$

Solving,

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

Where  $\lambda = \lambda(\mu)$  solves

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



# Convex duality

$$\mathbb{L}(\lambda) \equiv - \sum_{i=1}^n \log(1 + \lambda'(x_i - \mu)) = \log R(F)$$

$$\frac{\partial \mathbb{L}}{\partial \lambda} = - \sum_{i=1}^n \frac{x_i - \mu}{1 + \lambda'(x_i - \mu)}$$

Maximize  $\log R$  or minimize  $\mathbb{L}$

$$\frac{\partial^2 \mathbb{L}}{\partial \lambda \partial \lambda'} = \sum_{i=1}^n \frac{(x_i - \mu)(x_i - \mu)'}{(1 + \lambda'(x_i - \mu))^2}$$

$\mathbb{L}$  is convex and  $d$  dimensional  $\implies$  easy optimization

## Sketch of ELT proof

WLOG  $q = d$ , and anticipate a small  $\lambda$

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \mu}{1 + (x_i - \mu)' \lambda} \quad 1/(1 + \epsilon) = 1 - \epsilon + \epsilon^2 - \epsilon^3 \dots$$

$$\doteq \frac{1}{n} \sum_{i=1}^n (x_i - \mu) - (x_i - \mu)(x_i - \mu)' \lambda, \quad \text{so,}$$

$$\lambda \doteq S^{-1}(\bar{x} - \mu), \quad \text{where,}$$

$$S = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)'$$

Left out: how  $E(\|X\|^2) < \infty$  implies small  $\lambda(\mu_0)$

## Sketch continued

$$\begin{aligned}
 -2 \log \prod_{i=1}^n n w_i &= -2 \log \prod_{i=1}^n \frac{1}{1 + \lambda'(x_i - \mu)} \\
 &= 2 \sum_{i=1}^n \log(1 + \lambda'(x_i - \mu)) \quad \log(1 + \epsilon) = \epsilon - (1/2)\epsilon^2 + \dots \\
 &\doteq 2 \sum_{i=1}^n \left( \lambda'(x_i - \mu) - \frac{1}{2} \lambda'(x_i - \mu)(x_i - \mu)' \lambda \right) \\
 &= n \left( 2\lambda'(\bar{x} - \mu) - \lambda' S \lambda \right) \\
 &= n \left( 2(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu) - (\bar{x} - \mu)' S^{-1} S S^{-1} (\bar{x} - \mu) \right) \\
 &= n (\bar{x} - \mu)' S^{-1} (\bar{x} - \mu) \\
 &\rightarrow \chi_{(d)}^2
 \end{aligned}$$

# Typical coverage errors

1.  $\Pr(\mu_0 \in C_{r,n}) = 1 - \alpha + O\left(\frac{1}{n}\right)$  as  $n \rightarrow \infty$
2. One-sided errors of  $O\left(\frac{1}{\sqrt{n}}\right)$  cancel
3. Bartlett correction **DiCiccio, Hall, Romano**
  - (a) replace  $\chi^{2,1-\alpha}$  by  $\left(1 + \frac{a}{n}\right)\chi^{2,1-\alpha}$  for carefully chosen  $a$
  - (b) get coverage errors  $O\left(\frac{1}{n^2}\right)$
  - (c)  $a$  does not depend on  $\alpha$
  - (d) data based  $\hat{a}$  gets same rate

same as for parametric likelihoods

# Calibrating empirical likelihood

|                           |                               |
|---------------------------|-------------------------------|
| Plain $\chi^{2,1-\alpha}$ | undercovers                   |
| $F_{d,n-d}^{1-\alpha}$    | is a bit better               |
| Bartlett correction       | asymptotics slow to take hold |
| Bootstrap                 | seems to work best            |

## Bootstrap calibration

### Recipe

Sample  $X_i^*$  IID  $\widehat{F}$

Get  $-2 \log \mathcal{R}(\bar{x}; x_1^*, \dots, x_n^*)$

Repeat  $B = 1000$  times

Use  $1 - \alpha$  sample quantile

### Results

Regions get empirical likelihood shape and bootstrap size

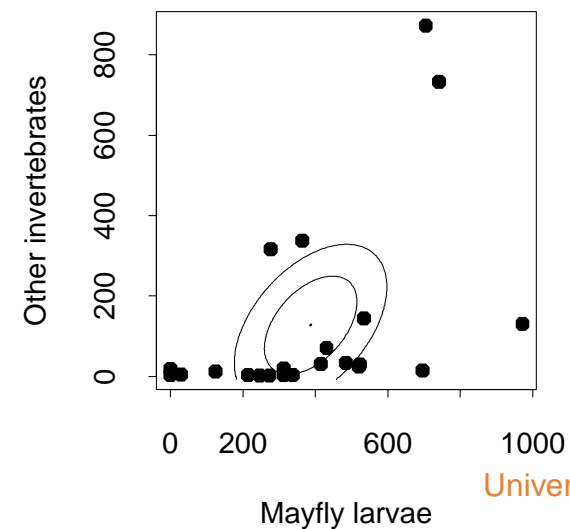
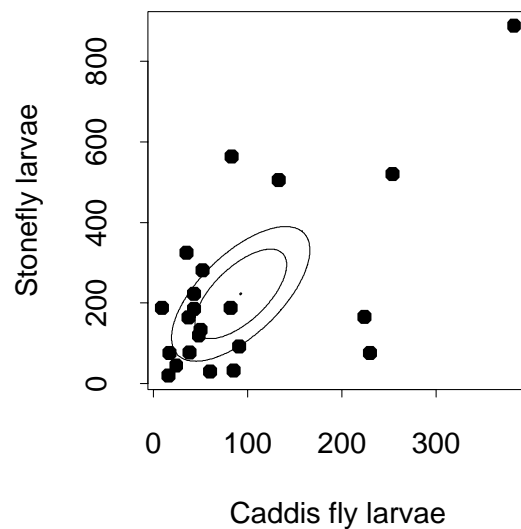
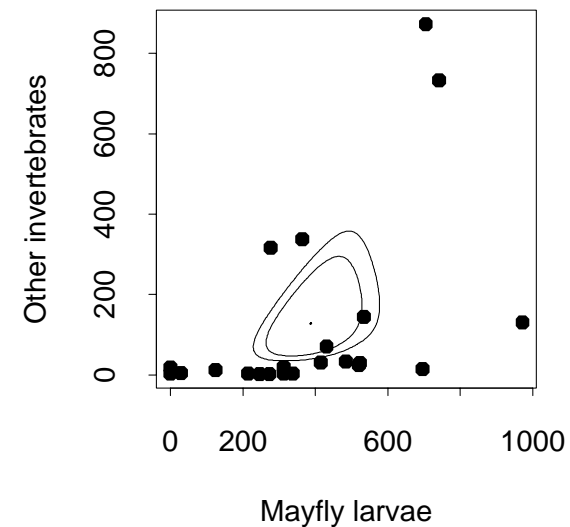
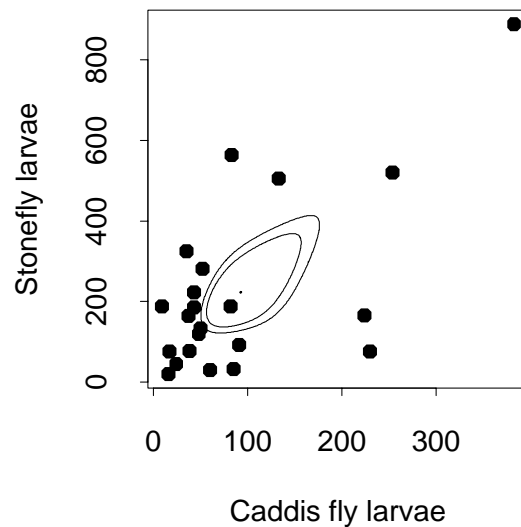
Coverage error  $O(n^{-2})$

Same error rate as bootstrapping the bootstrap

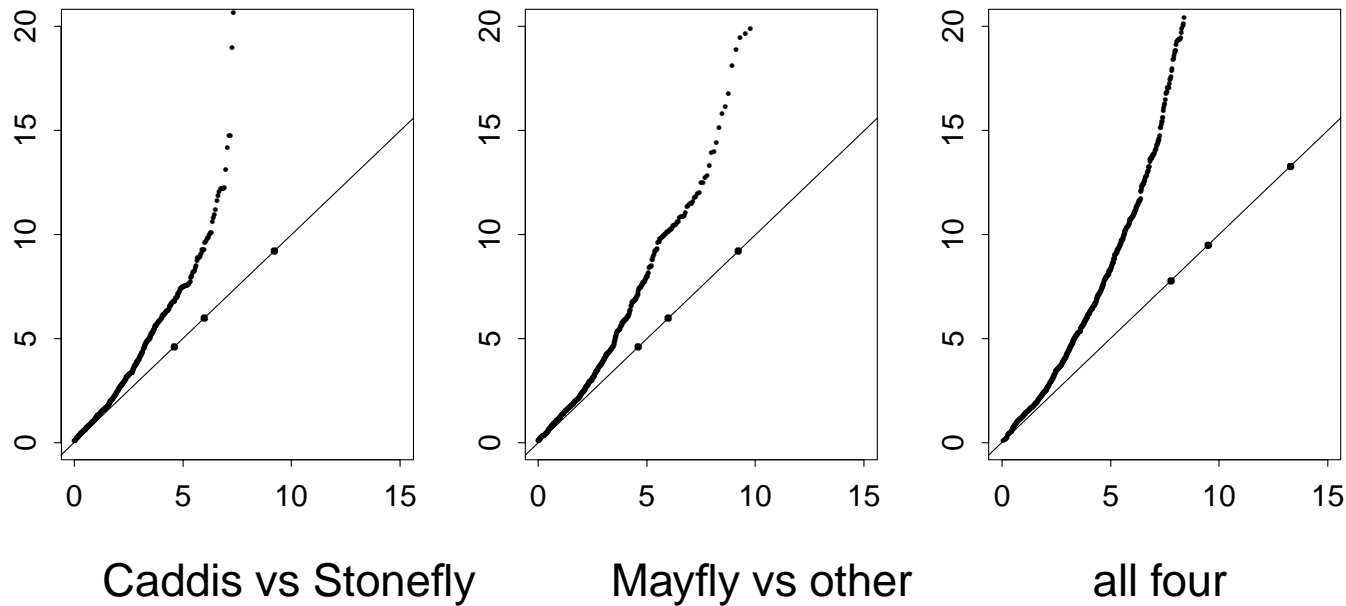
Sets in faster than Bartlett correction

Need further adjustments for one-sided inference

# Bootstrap (and $\chi^2$ ) calibrated Dipper regions



# Resampled $-2 \log \mathcal{R}(\mu)$ values vs $\chi^2$





## Smooth functions of means

$$\sigma = \sqrt{E(X^2) - E(X)^2}$$

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2} \sqrt{E(Y^2) - E(Y)^2}}$$

$$\theta = h(E(U, V, \dots, Z))$$

Generally

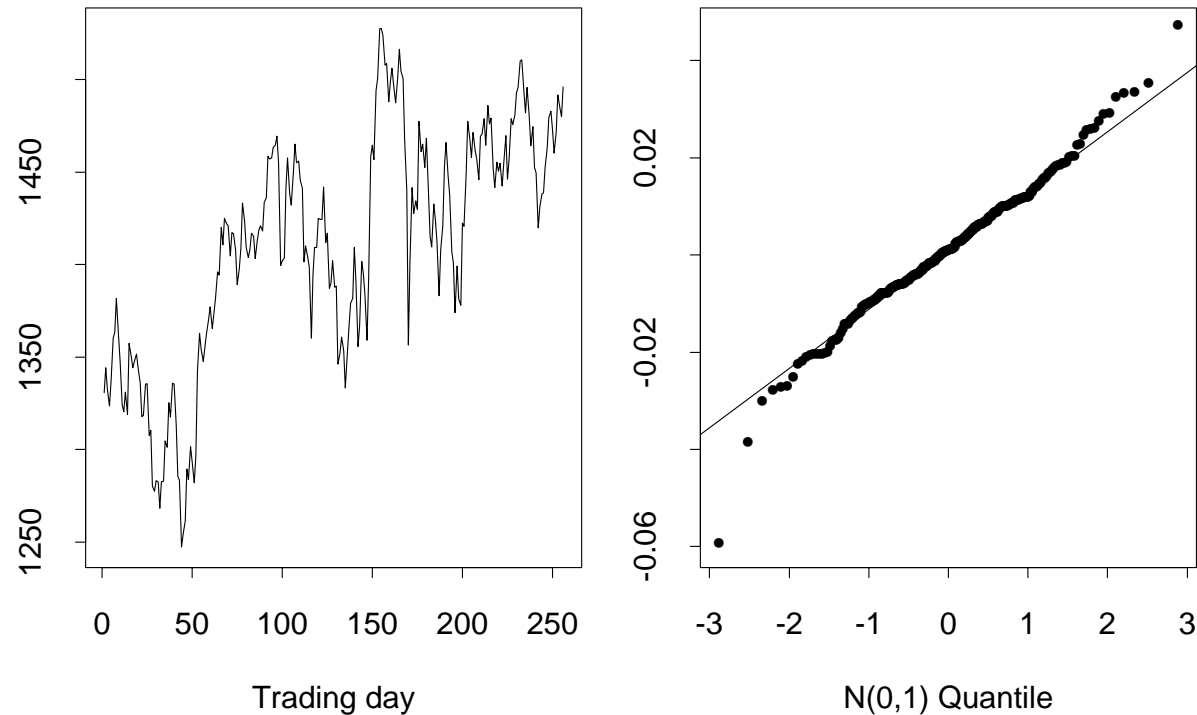
$$X = (U, V, \dots, Z)$$

$$\theta = E(h(X))$$

$$\hat{\theta} = h(\bar{x}) \doteq h(E(X)) + (\bar{x} - E(X))' \frac{\partial}{\partial x} h(E(X))$$

$h$  nearly linear near  $E(X) \implies \theta$  nearly a mean

# S&P 500 returns

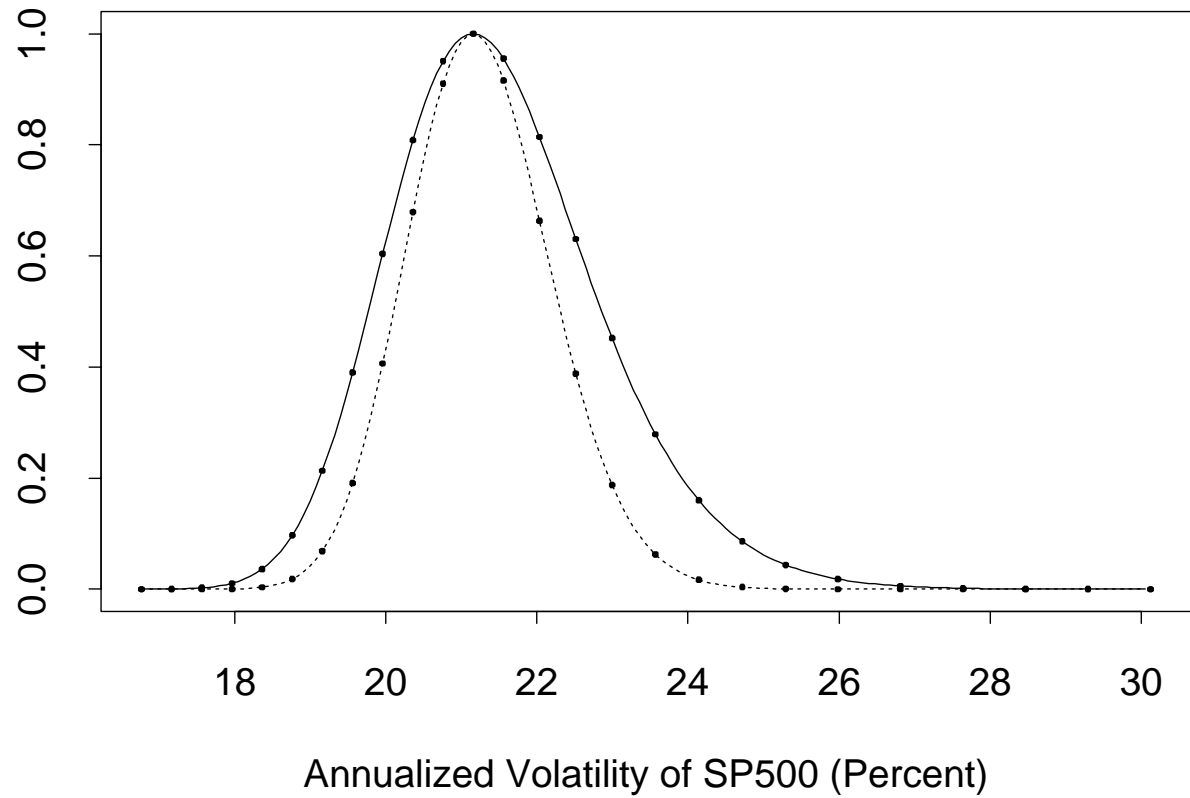


$$\text{Return} = \log(x_{i+1}/x_i)$$

Nearly  $N(0, \sigma^2)$  but heavy tails

Volatility  $\sigma$  is Standard deviation of returns

# S&P 500 returns



Solid = Empirical likelihood

Dashed = Normal likelihood

# Estimating equations

More powerful and general than smooth functions

Define  $\theta$  via  $E(m(X, \theta)) = 0$

Define  $\hat{\theta}$  via  $\frac{1}{n} \sum_{i=1}^n m(x_i, \hat{\theta}) = 0$

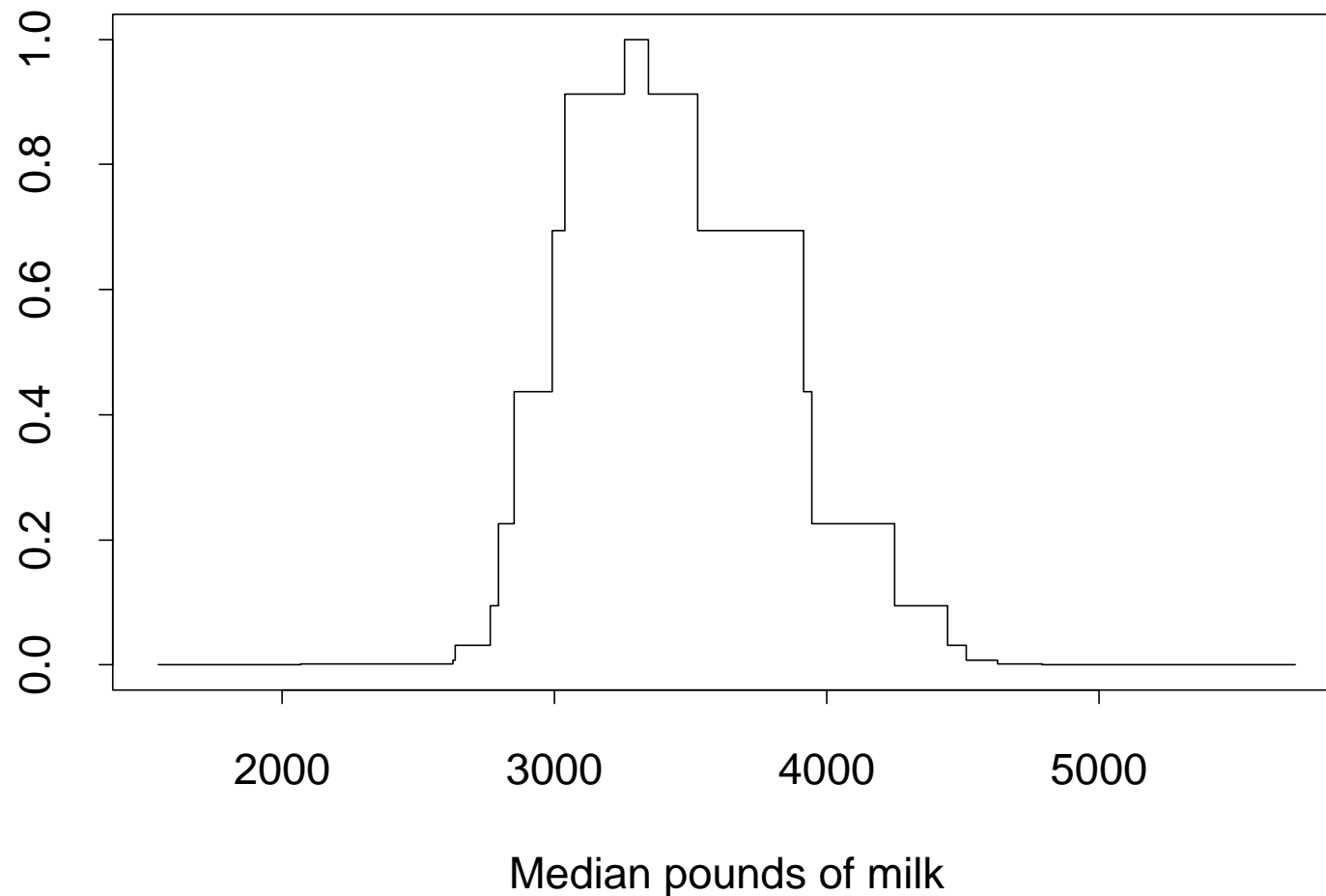
Usually  $\dim(m) = \dim(\theta)$

Basic examples:  $\dim(m) = \dim(\theta) = 1$

| $m(X, \theta)$                                   | Statistic              |
|--|------------------------|
| $X - \theta$                                     | Mean                   |
| $1_{X \in A} - \theta$                           | Probability of set $A$ |
| $1_{X \leq \theta} - \frac{1}{2}$                | Median                 |
| $\frac{\partial}{\partial x} \log(f(X; \theta))$ | MLE under $f$          |

$$-2 \log \mathcal{R}(\theta_0) \rightarrow \chi_{\text{Rank}(\text{Var}(m(X, \theta_0)))}^2$$

# Empirical likelihood for a median



LR is constant between observations

# Est. eq. with nuisance parameters

For  $\theta = (\rho)$  and  $\nu = (\mu_x, \mu_y, \sigma_x, \sigma_y)$

$$E(m(X, \theta, \nu)) = 0 = \frac{1}{n} \sum_{i=1}^n m(X_i, \hat{\theta}, \hat{\nu})$$

## Correlation example

$$0 = E(X - \mu_x)$$

$$0 = E(Y - \mu_y)$$

$$0 = E((X - \mu_x)^2 - \sigma_x^2)$$

$$0 = E((Y - \mu_y)^2 - \sigma_y^2)$$

$$0 = E((X - \mu_x)(Y - \mu_y) - \rho\sigma_x\sigma_y)$$

Profile empirical likelihood  $\mathcal{R}(\theta) = \sup_{\nu} \mathcal{R}(\theta, \nu)$

Typically  $-2 \log \mathcal{R}(\theta_0) \rightarrow \chi_{\dim(\theta)}^2$

# Huber's robust estimation

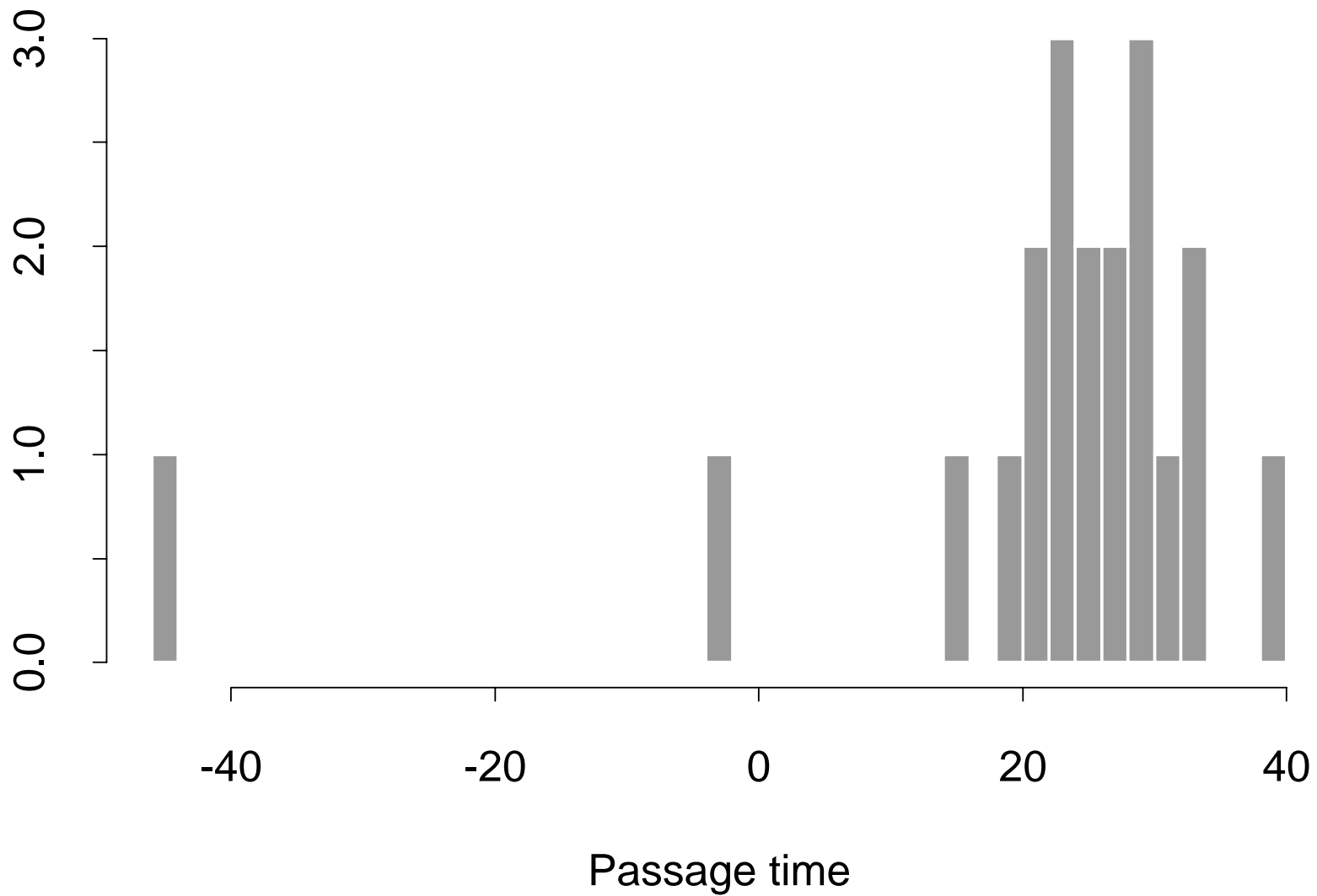
$$0 = \frac{1}{n} \sum_{i=1}^n \psi\left(\frac{x_i - \mu}{\sigma}\right) 0 = \frac{1}{n} \sum_{i=1}^n \left[ \psi\left(\frac{x_i - \mu}{\sigma}\right)^2 - 1 \right]$$

Like mean for small obs, median for outliers

$$\psi(z) = \begin{cases} z, & |z| \leq 1.35 \\ 1.35 \operatorname{sign}(z), & |z| \geq 1.35. \end{cases}$$

$$\mathcal{R}(\mu) = \max_{\sigma} \max \left\{ \prod_{i=1}^n n w_i \mid 0 \leq w_i, \sum_i w_i = 1, \sum_i w_i \psi\left(\frac{x_i - \mu}{\sigma}\right) = 0, \right. \\ \left. \sum_i w_i \left[ \psi\left(\frac{x_i - \mu}{\sigma}\right)^2 - 1 \right] = 0 \right\}$$

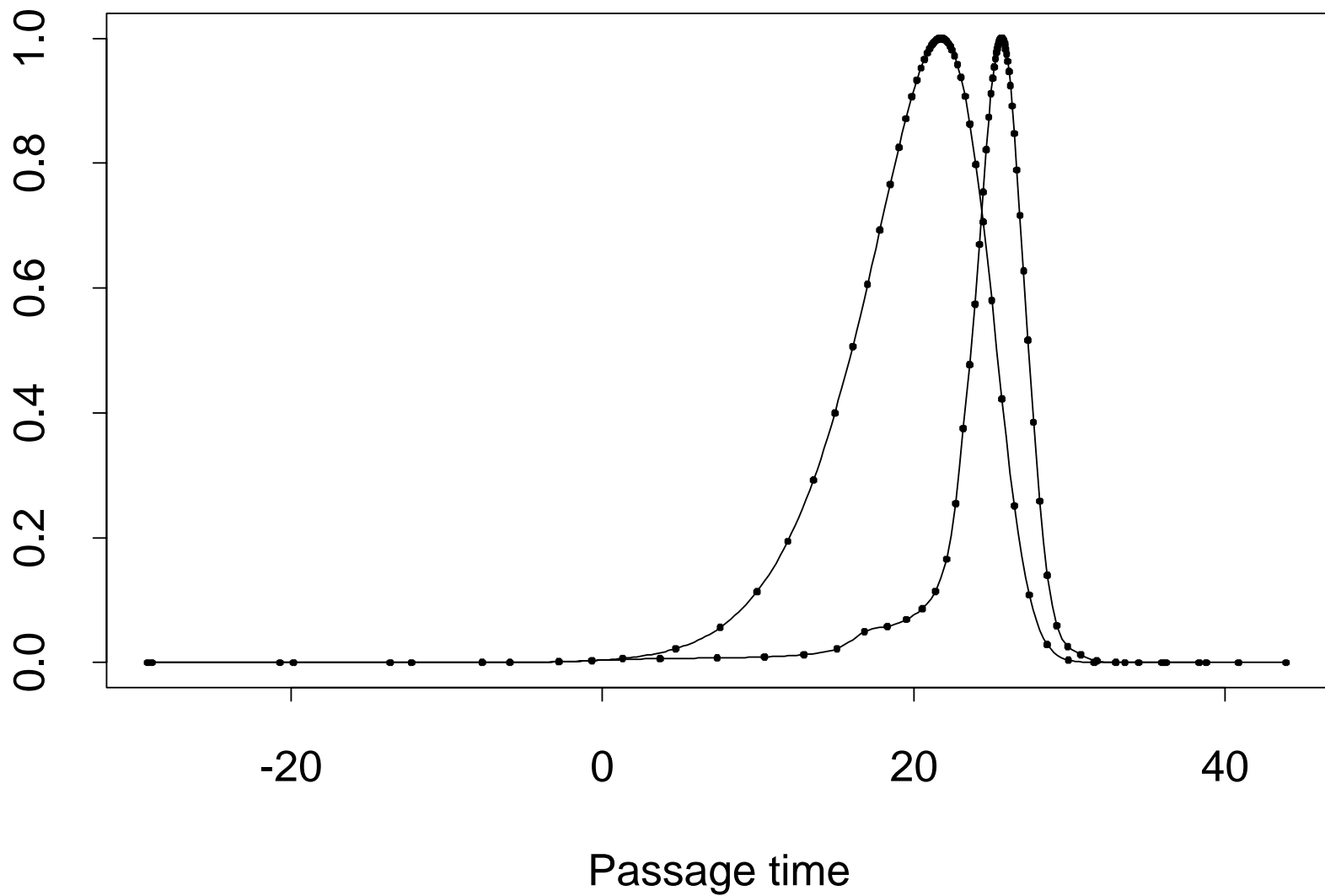
# Newcomb's passage times of light



From Stigler



# EL for mean and Huber's location



# Maximum empirical likelihood estimates

|               |      |                            |
|---------------|------|----------------------------|
| Hartley & Rao | 1968 | means & finite populations |
| Owen          | 1991 | means IID                  |
| Qin & Lawless | 1993 | estimating eqns IID        |

## Simple MELEs

Observe  $(X_i, Y_i)$  pairs with mean  $(\mu_x, \mu_y)$  and  $\mu_x = \mu_{x0}$  *known*

Let  $w_i$  maximize  $\prod_{i=1}^n n w_i$  st:

$w_i \geq 0$  and  $\sum_{i=1}^n w_i = 1$  and  $\sum_{i=1}^n w_i x_i = \mu_x$

$$\text{MELE } \tilde{\mu}_y = \sum_{i=1}^n w_i y_i \doteq \bar{Y} - \Sigma_{yx} \Sigma_{xx}^{-1} (\bar{X} - \mu_{x0})$$

# Conditional empirical likelihood

$\mu_x = \mu_{x0}$  known

$$\mathcal{R}_{X,Y}(\mu_x, \mu_y) = \max \left\{ \prod_{i=1}^n n w_i \mid w_i \geq 0, \sum_i w_i x_i = \mu_x, \sum_i w_i y_i = \mu_y \right\}$$

$$\mathcal{R}_X(\mu_x) = \max \left\{ \prod_{i=1}^n n w_i \mid w_i \geq 0, \sum_i w_i x_i = \mu_x \right\}$$

$$\mathcal{R}_{Y|X}(\mu_y \mid \mu_x) = \frac{\mathcal{R}_{X,Y}(\mu_x, \mu_y)}{\mathcal{R}_X(\mu_x)}$$

$$-2 \log \mathcal{R}_{Y|X}(\mu_y \mid \mu_{x0}) \rightarrow \chi_{\dim(Y)}^2$$

$$-2 \log \mathcal{R}_Y \doteq n(\mu_{y0} - \bar{y})' \Sigma_{yy}^{-1} (\mu_{y0} - \tilde{\mu}_y)$$

$$-2 \log \mathcal{R}_{Y|X} \doteq n(\mu_{y0} - \tilde{\mu}_y)' \Sigma_{y|x}^{-1} (\mu_{y0} - \tilde{\mu}_y)$$

$$\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \leq \Sigma_{yy}$$

## Side or auxiliary information

---

| Known parameter   | Estimating equation     |
|-------------------|-------------------------|
| mean              | $X - \mu_x$             |
| $\alpha$ quantile | $1_{X \leq Q} - \alpha$ |
| $P(A   B)$        | $(1_A - \rho)1_B$       |
| $E(X   B)$        | $(X - \mu)1_B$          |

---

# Overdetermined equations

$$E(m(X, \theta)) = 0, \quad \dim(m) > \dim(\theta)$$

Approaches:

1. Drop  $\dim(m) - \dim(\theta)$  equations
2. Replace  $m(X, \theta)$  by  $m(X, \theta)A(\theta)$  where  
 $A$  a  $\dim(m) \times \dim(\theta)$  matrix      (IE pick  $\dim(\theta)$  linear comb. of  $m$ )
3. GMM: estimate the optimal  $A$
4. MELE:  $\tilde{\theta} = \arg \max_{\theta} \max_{w_i} \prod_i n w_i \quad \text{st} \quad \sum_{i=1}^n w_i m(x_i, \theta) = 0$

MELE has same asymptotic variance as using optimal  $A(\theta)$

Bias scales more favorably with dimensions for MELE than for  $\hat{A}$  methods

## Qin and Lawless result

$$\dim(m) = p + q \geq p = \dim(\theta) \quad \text{MELE } \tilde{\theta}$$

$$-2 \log(\mathcal{R}(\theta_0)/\mathcal{R}(\tilde{\theta})) \rightarrow \chi_{(p)}^2 \quad \text{conf regions for } \theta_0$$

$$-2 \log \mathcal{R}(\tilde{\theta}) \rightarrow \chi_{(q)}^2 \quad \text{goodness of fit tests when } q > 0$$

Requires considerable smoothness

What happens for  $\text{IQR} = Q^{0.75} - Q^{0.25}$  ?

$$0 = E(1_{X \leq Q^{.75}} - 0.75) = E(1_{X \leq Q^{.25}} - 0.25)$$

$$0 = E(1_{X \leq Q^{.25} + \text{IQR}} - 0.75) = E(1_{X \leq Q^{.25}} - 0.25)$$

Need to max over  $Q^{.25}$

# Euclidean log likelihood

Replace  $-\sum_{i=1}^n \log(nw_i)$  by

$$\ell_E = -\frac{1}{2} \sum_{i=1}^n (nw_i - 1)^2$$

Reduces to Hotelling's  $T^2$  for the mean **Owen**

Reduces to Huber-White covariance for regression

Reduces to continuous updating GMM **Kitamura**

Quadratic approx to EL, like Wald test is to parametric likelihood

**Allows  $w_i < 0$ , and so**

1. confidence regions for means can get out of the convex hull
2. confidence regions no longer obey range restrictions

# Exponential empirical likelihood

Replace  $-\sum_{i=1}^n \log(nw_i)$  by

$$\text{KL} = \sum_{i=1}^n w_i \log(nw_i)$$

relates to entropy and exponential tilting

## Hellinger distance

$$\sum_{i=1}^n (w_i^{1/2} - n^{-1/2})^2$$



# Renyi, Cressie-Read

$$\frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^n ((nw_i)^{-\lambda} - 1)$$

---

| $\lambda$        | Method                           |
|------------------|----------------------------------|
| $-2$             | Euclidean log likelihood         |
| $\rightarrow -1$ | Exponential empirical likelihood |
| $-1/2$           | Freeman-Tukey                    |
| $\rightarrow 0$  | Empirical likelihood             |
| $1$              | Pearson's                        |

---

## Alternate artificial likelihoods

All Renyi Cressie-Read families have  $\chi^2$  calibrations. [Baggerly](#)

Only EL is Bartlett correctable [Baggerly](#)

$-2 \sum_{i=1}^n \widetilde{\log}(nw_i)$  Bartlett correctable if

$$\widetilde{\log}(1+z) = z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \frac{1}{4}z^4 + o(z^4), \quad \text{as } z \rightarrow 0$$

[Corcoran](#)

# Regression

$$E(Y | X = x) \doteq \beta_0 + \beta_1 x$$

## Models (Freedman)

Correlation  $(X_i, Y_i) \sim F_{XY}$  IID

Regression  $x_i$  fixed,  $Y_i \sim F_{Y|X=(1,x_i)}$  indep

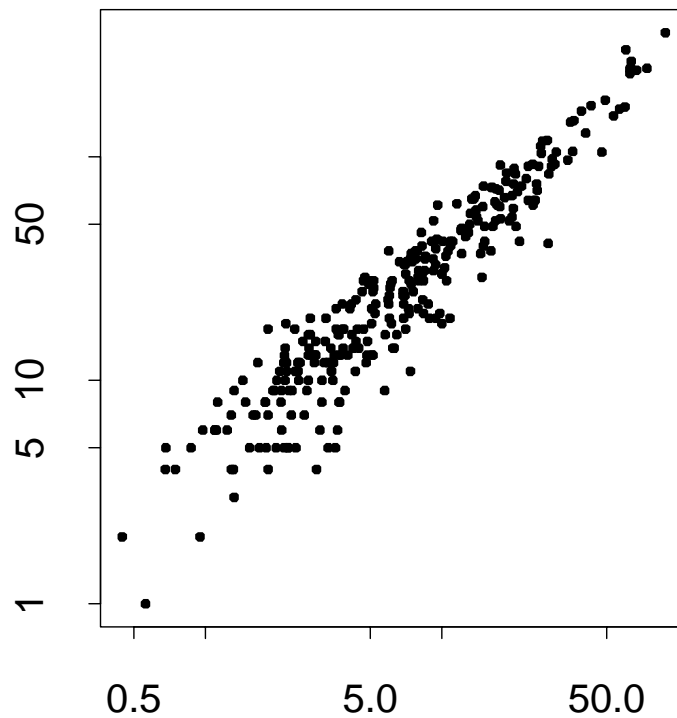
## Correlation model

$$\beta = E(X'X)^{-1}E(X'Y)$$

$$\hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^n X_i' X_i \right)^{-1} \frac{1}{n} \sum_{i=1}^n X_i' Y_i$$

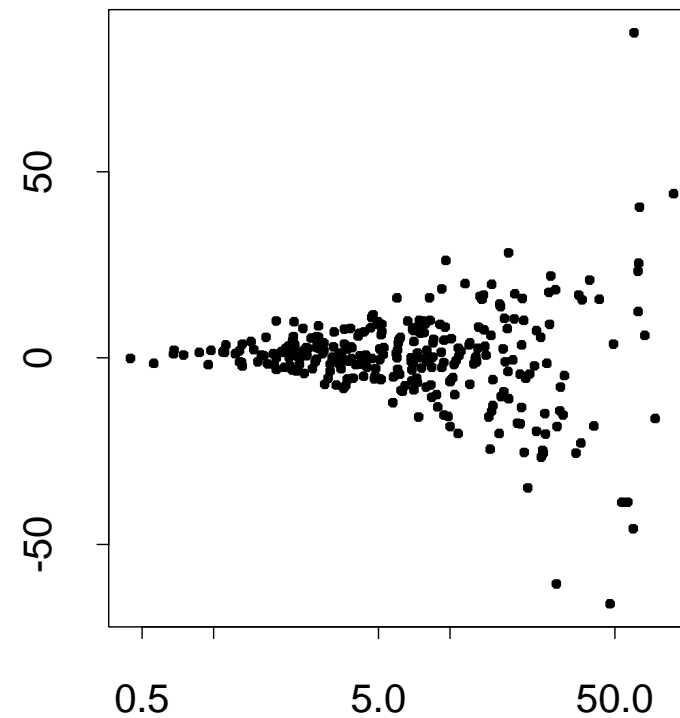
$\beta$  and  $\hat{\beta}$  well defined even for lack of fit

# Cancer deaths vs population, by county



Population (1000s)

Nearly linear regression



Population (1000s)

nonconstant residual variance

Royall via Rice

# Estimating equations for regression

$$E(X'(Y - X'\beta)) = 0, \quad \frac{1}{n} \sum_{i=1}^n X'_i(Y_i - X'_i\hat{\beta}) = 0$$

$$\mathcal{R}(\beta) = \max \left\{ \prod_{i=1}^n nw_i \mid \sum_{i=1}^n w_i Z_i(\beta) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}$$

$$Z_i(\beta) = X'_i(Y_i - X'_i\beta)$$

$$\text{need } E(\|Z\|^2) \leq E\left(\|X\|^2(Y - X'\beta)^2\right) < \infty$$

Don't need:

normality, constant variance, exact linearity

## For cancer data

$P_i$  = population of  $i$ 'th county in 1000s

$C_i$  = cancer deaths of  $i$ 'th county in 20 years

$$C_i \doteq \beta_0 + \beta_1 P_i$$

$$\hat{\beta}_1 = 3.58 \quad \implies 3.58/20 = 0.18 \text{ deaths per thousand per year}$$

$$\hat{\beta}_0 = -0.53 \quad \text{near zero, as we'd expect}$$

# Regression through the origin

$$C_i \doteq \beta_1 P_i$$

Residuals should have mean zero and be orthogonal to  $P_i$

We want two equations in one unknown  $\beta_1$

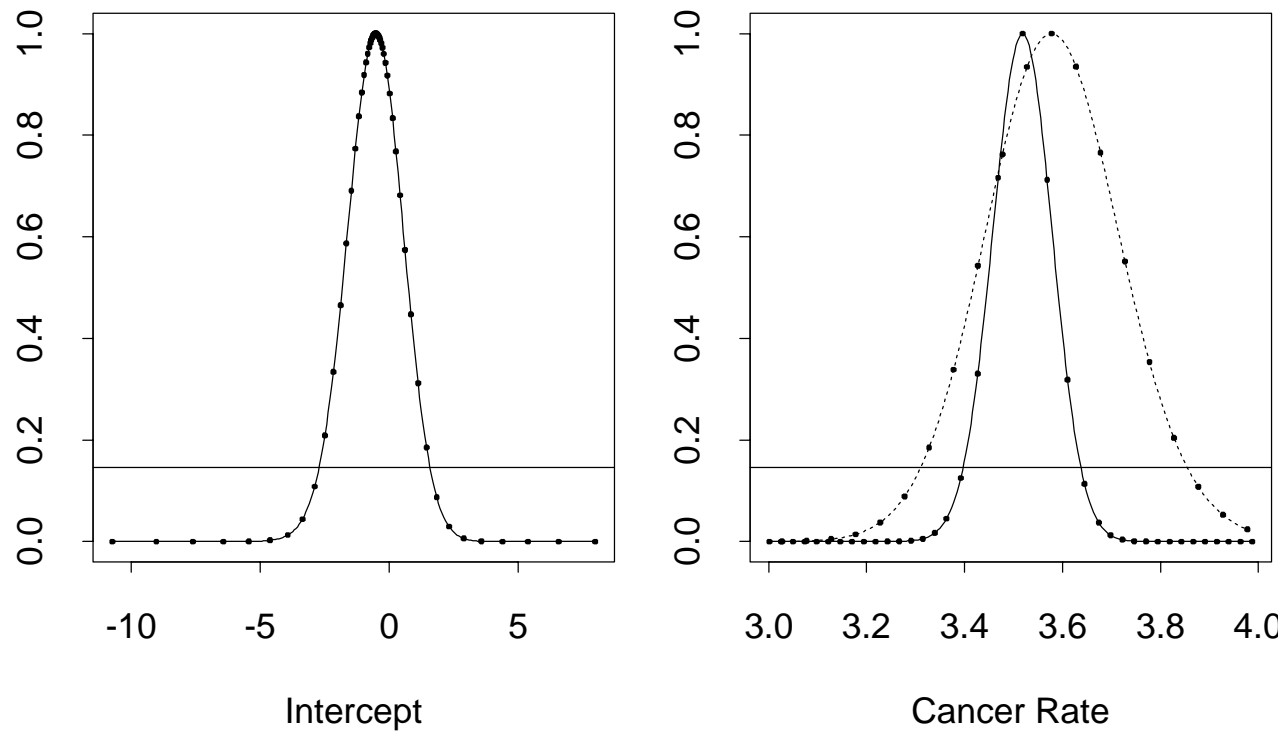
Equivalently, side information  $\beta_0 = 0$

Least squares regression through origin does not solve both equations

$$\text{MELE } \tilde{\beta}_1 = \arg \max_{\beta_1} \mathcal{R}(\beta_1)$$

$$\mathcal{R}(\beta_1) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i (C_i - P_i \beta_1) = 0, \right. \\ \left. \sum_{i=1}^n w_i P_i (C_i - P_i \beta_1) = 0, \sum_{i=1}^n w_i = 1, w_i \geq 0 \right\}$$

# Regression parameters



Intercept nearly 0, MELE smaller than MLE

CI based on conditional empirical likelihood

Constraint narrows CI for slope by over half



# Fixed predictor regression model

$$E(Y_i) = \mu_i \doteq \beta_0 + \beta_1 x_i \text{ fixed, and } V(Y_i) = \sigma_i^2$$

With lack of fit  $\mu_i \neq \beta_0 + \beta_1 x_i$

No good definition of 'true'  $\beta$  given L.O.F.

$$Z_i = x_i(Y_i - x_i'\beta) \text{ have}$$

1. mean  $E(Z_i) = x_i(\mu_i - x_i'\beta)$       0 may be the common value
2. variance  $V(Z_i) = x_i x_i' \sigma_i^2$       non-constant, even if  $\sigma_i^2$  constant

# Triangular array ELT

$$\begin{array}{cccc}
 Z_{11} & & & \\
 Z_{12} & Z_{22} & & \\
 Z_{13} & Z_{23} & Z_{33} & \\
 \vdots & \vdots & \vdots & \ddots \\
 Z_{1n} & Z_{2n} & Z_{3n} & \cdots & Z_{nn} \\
 \vdots & \vdots & \vdots & & \ddots
 \end{array}$$

Row  $n$  has indep  $Z_{1n}, \dots, Z_{nn}$ , common mean 0 not ident distributed

Different rows have different distns

Still get  $-\log \mathcal{R}(\text{Common mean} = 0) \rightarrow \chi_{\dim(Z)}^2$  under mild conditions

Applies for fixed  $x$  regression:  $Z_{in} = x_i(Y_i - x_i'\beta)$

# Variance modelling

Working model  $Y \sim N(x'\beta, e^{2z'\gamma})$

$$0 = \frac{1}{n} \sum_{i=1}^n x_i (y_i - x_i'\beta) e^{-2z_i'\gamma} \quad (\text{weight} \propto 1/\text{var})$$

$$0 = \frac{1}{n} \sum_{i=1}^n z_i \left( 1 - \exp(-2z_i'\gamma) (y_i - x_i'\beta)^2 \right)$$

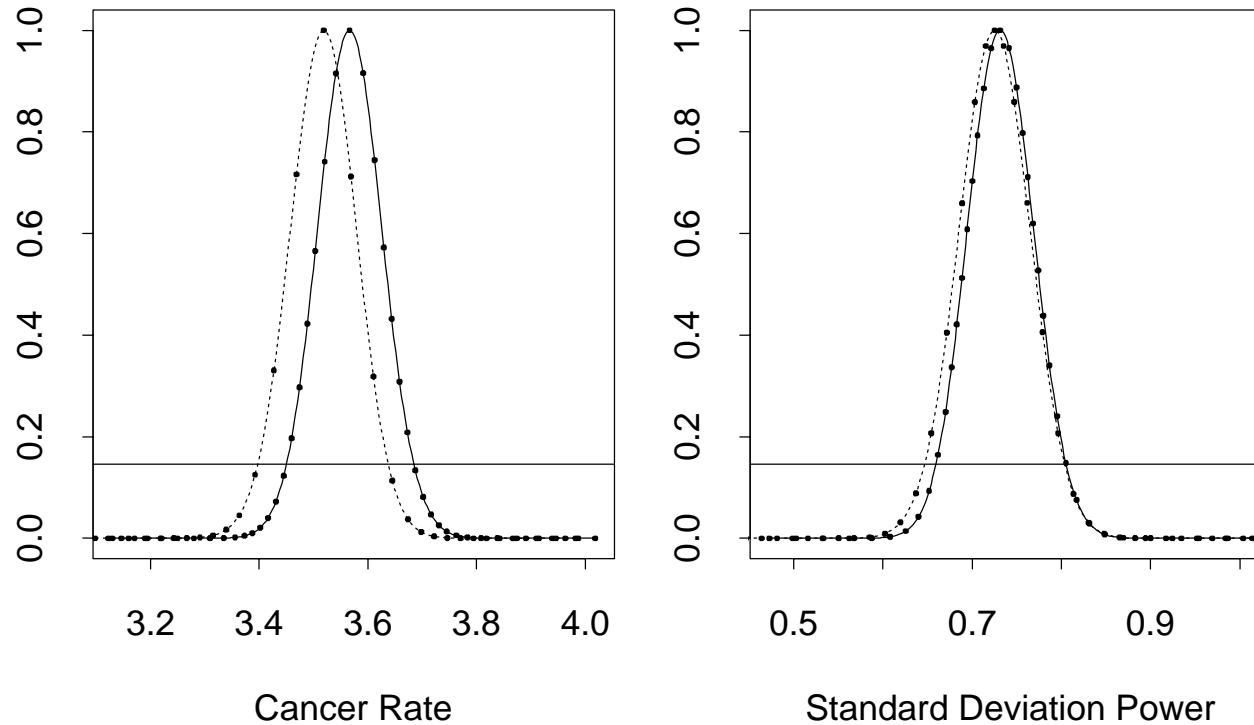
For cancer data

$$x_i = (1, P_i) \quad z_i = (1, \log(P_i))$$

$$E(Y_i) = \beta_0 + \beta_1 P_i \quad \sqrt{V(Y_i)} = \exp(\gamma_0 + \gamma_1 \log(P_i)) = e^{\gamma_0} P_i^{\gamma_1}$$

and  $\beta_0 = 0$

# Heteroscedastic model

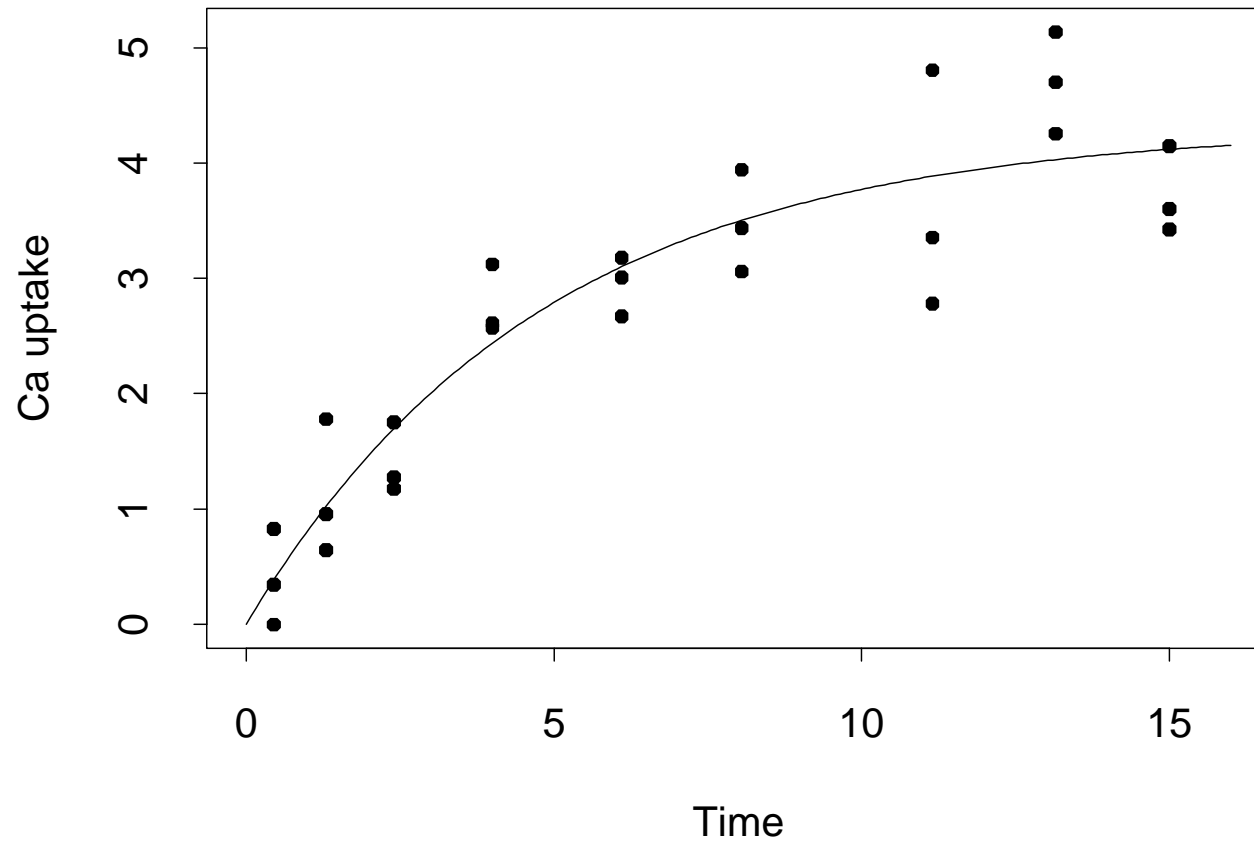


Left: solid curve accounts for nonconstant variance

Right: solid curve forces  $\beta_0 = 0$ , and,

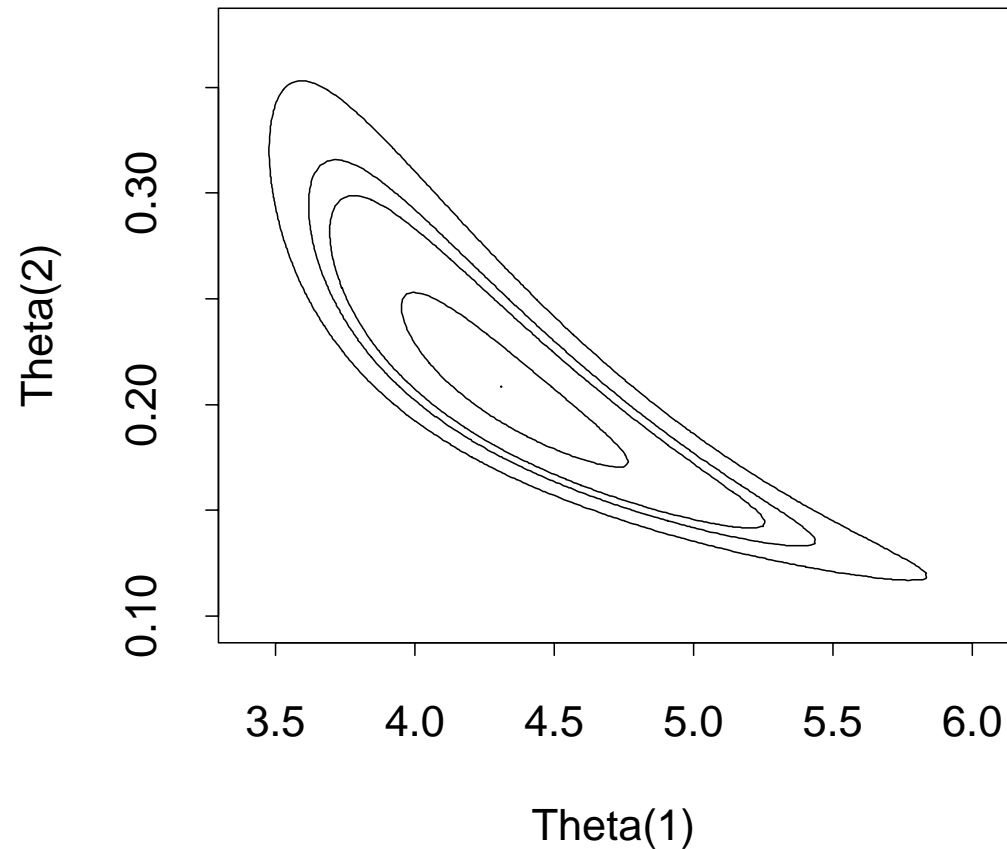
rules out  $\gamma_1 = 1/2$  (Poisson) and  $\gamma_1 = 1$  (Gamma)

# Nonlinear regression



$$y \doteq f(x, \theta) \equiv \theta_1 (1 - \exp(-\theta_2 x))$$

# Nonlinear regression regions



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

Don't need: normality or constant variance

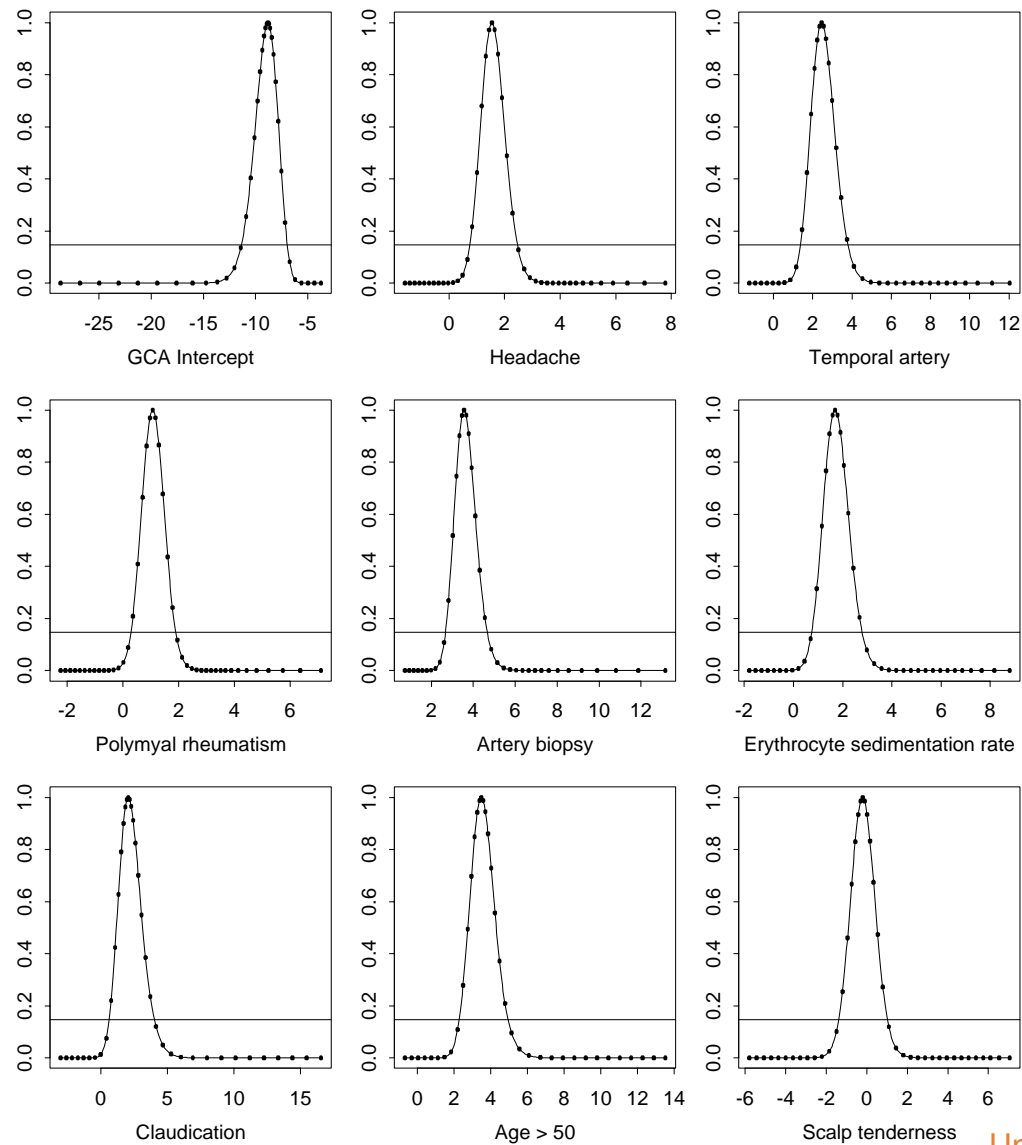
# Logistic regression

- Giant cell arteritis is a type of vasculitis (inflammation of blood or lymph vessels)
- Not all vasculitis is GCA
- Try to predict GCA from 8 binary predictors

$$\Pr(GCA) \doteq \tau(X'\beta) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_8 X_8)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_8 X_8)}$$

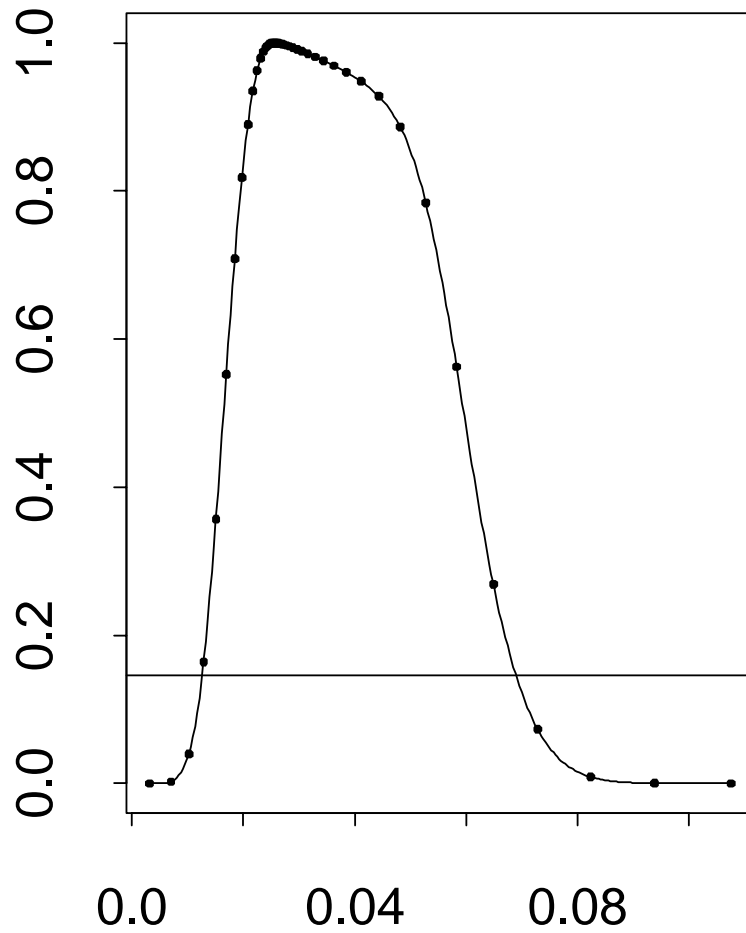
Likelihood estimating equations reduce to:  $Z_i(\beta) = X_i(Y_i - \tau(X'\beta))$

# Logistic regression coefficients

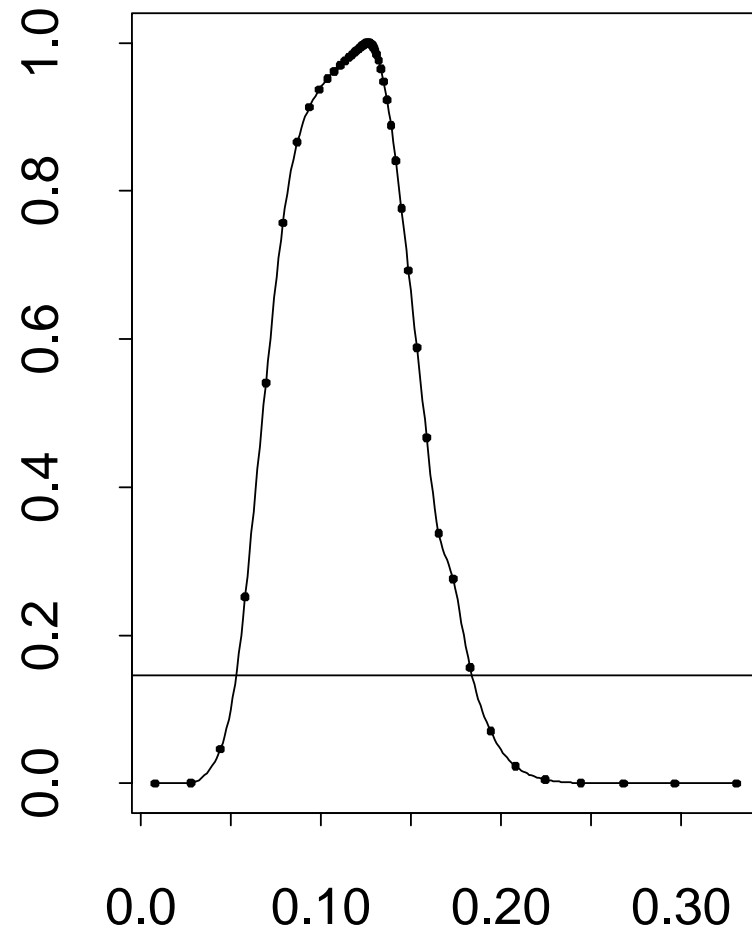




# Prediction accuracy



Smoothed  $P(\text{Err}|Y=0)$



Smoothed  $P(\text{Err}|Y=1)$

# Biased sampling

## Examples

1. Sample children, then record family sizes.
2. Draw blue line over cotton, sample fibers that are partly blue.
3. When  $Y = y$  it is recorded as  $X$  with prob.  $u(y)$ , lost with prob.  $1 - u(y)$ .

$Y \sim F$ ,    observe  $X \sim G$ ,    but we really want  $F$

$$G(A) = \frac{\int_A u(y) dF(y)}{\int u(y) dF(y)}$$

$$L(F) = \prod_{i=1}^n G(\{x_i\}) = \prod_{i=1}^n \frac{F(\{x_i\}) u(x_i)}{\int u(x) dF(x)}$$

# NPMLE

$$\widehat{G}(\{x_i\}) = \frac{1}{n} \quad (\text{for simplicity, suppose no ties})$$

$$\widehat{G}(\{x_i\}) \propto \widehat{F}(\{x_i\}) \times u(x_i)$$

$$\widehat{F}(\{x_i\}) = \frac{u_i^{-1}}{\sum_{j=1}^n u_j^{-1}}$$

## For the mean

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i/u_i}{\sum_{i=1}^n 1/u_i}$$

Horvitz-Thompson estimator is NPMLE

$$\hat{\mu} = \left( \frac{1}{n} \sum_{i=1}^n x_i^{-1} \right)^{-1}$$

when  $u_i \propto x_i$ , so length bias  $\implies$  harmonic mean

## Biased sampling again

$$0 = \int m(x, \theta) dF(x) = \int \frac{m(x, \theta)}{u(x)} dG(x)$$

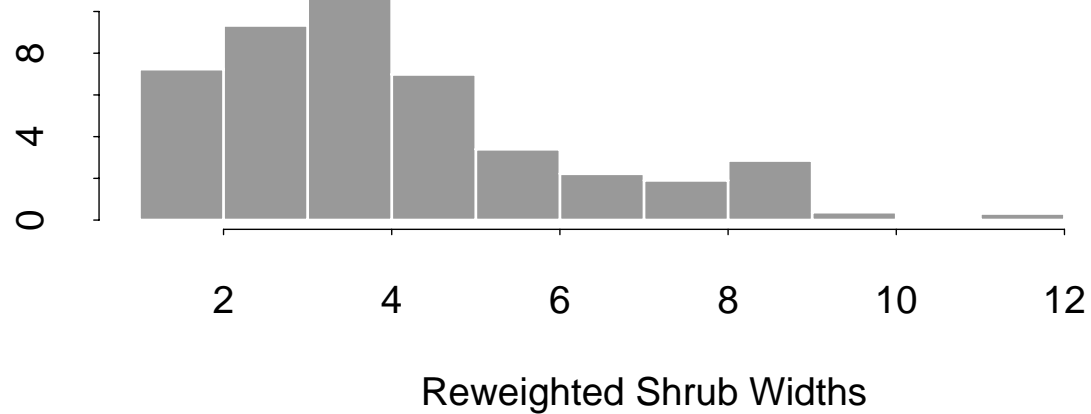
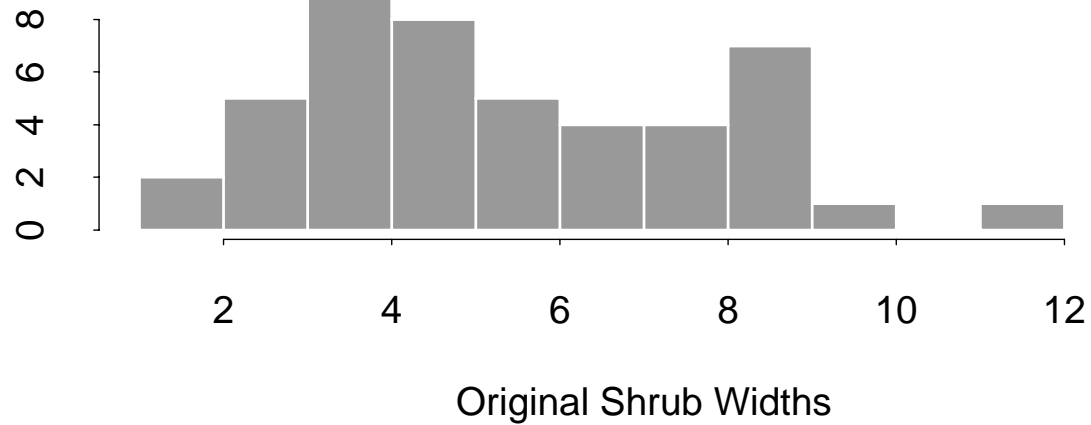
$$G(\{x_i\}) = w_i \implies F(\{x_i\}) = \frac{w_i/u_i}{\sum_{j=1}^n 1/u_j}$$

Very simple recipe

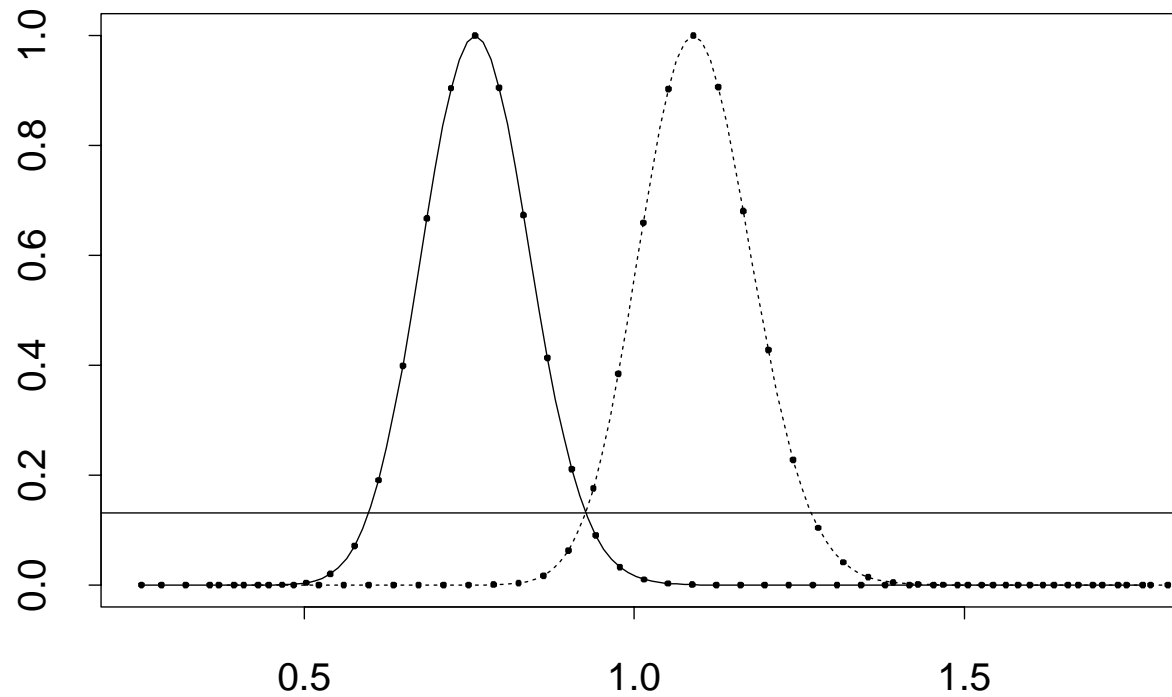
$$m(x, \theta) \longrightarrow \tilde{m}(x, \theta) \equiv \frac{m(x, \theta)}{u(x)}$$

$$\mathcal{R}(\theta) = \max \left\{ \prod_{i=1}^n n w_i \mid w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i \tilde{m}(x_i, \theta) = 0 \right\}$$

## Transect sampling of shrubs (Muttlak & McDonald)



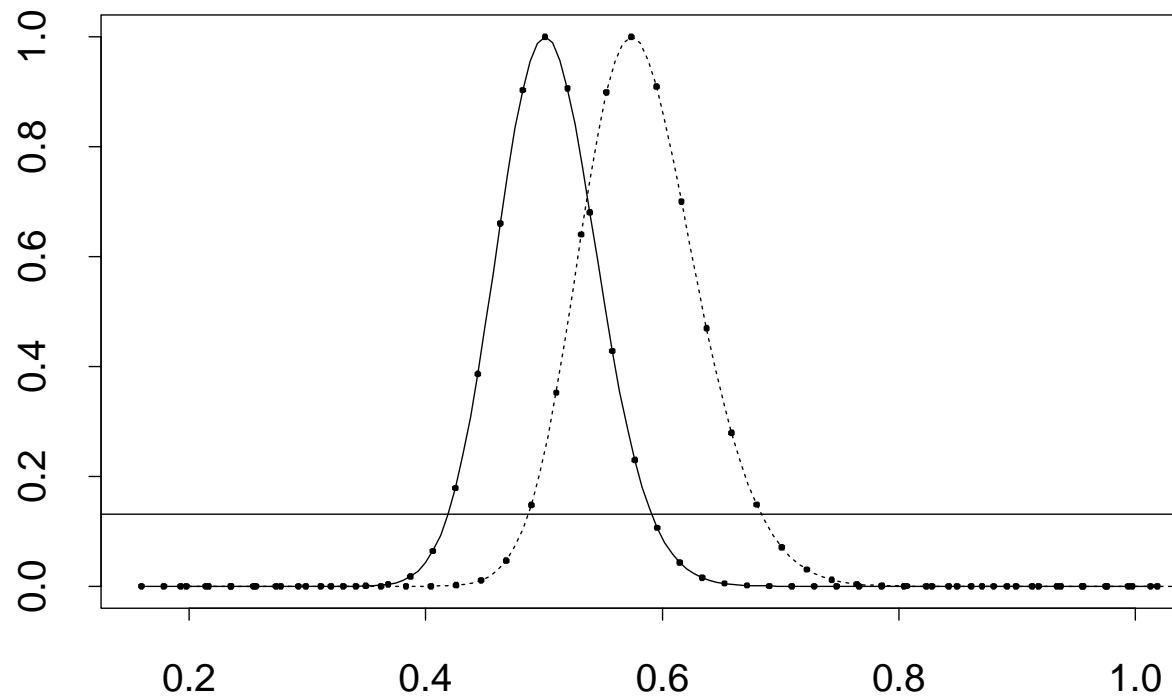
# Mean shrub width



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

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

# Standard dev. of shrub width



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

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

# Multiple biased samples

Population  $k$  sampled from  $F$  with bias  $u_k(\cdot)$ ,  $k = 1, \dots, s$

$$X_{ik} \sim G_k, \quad i = 1, \dots, n_k, \quad k = 1, \dots, s$$

$$G_k(A) = \frac{\int_A u_k(y) dF(y)}{\int u_k(y) dF(y)}, \quad k = 1, \dots, s$$

## Examples

1. clinical trials with varying enrolment criteria
2. mix of length biased and unbiased samples
3. telescopes with varying detection limits
4. sampling from different frames

NPMLs **Vardi** and ELTs **Qin** by multiplying likelihoods



# Truncation

Extreme sample bias with

$$u(x) = \begin{cases} 1, & x \in T \\ 0, & x \notin T \end{cases}$$

## Examples

1. Heights of military recruits, above a minimum
2. Swim times of olympic qualifiers, below a maximum
3. Star too dim to be seen

$$L(F) = \prod_{i=1}^n \frac{F(\{x_i\})}{\int_{T_i} dF(x)} = \prod_{i=1}^n \frac{F(\{x_i\})}{\sum_{j:x_j \in T_i} F(\{x_j\})}$$

# Censoring

Instead of exact value, only find that  $X_i \in C_i$

$C_i = \{x_i\}$  incorporates uncensored values

Famous example: right censoring of survival time

$$C_i = \begin{cases} \{X_i\}, & X_i \leq Y_i \\ (Y_i, \infty), & X_i > Y_i \end{cases}$$

## Censoring vs truncation

Censoring: Swim times over 3 minutes reported as  $(3, \infty)$

Truncation: Swim times over 3 minutes not reported at all

# Coarsening at random

Following truncation to set  $T_i$ ,

1. Set  $T_i$  partitioned into subsets  $C_{i,\omega}$ ,  $\omega \in \Omega_i$
2.  $X_i$  is drawn
3. We only learn which  $C_i$  contained  $X_i$

## Conditional likelihood for censoring

$$L(F) = \prod_{i=1}^n \frac{\int_{C_i} dF(x)}{\int_{T_i} dF(x)} = \prod_{i=1}^n \frac{\sum_{j:x_j \in C_i} F(\{x_j\})}{\sum_{j:x_j \in T_i} F(\{x_j\})}$$

conditional on the coarsening

# More examples

## Left truncation:

$x_i$  = brightness of star

$y_i$  = distance

$(x_i, y_i)$  observed  $\iff x_i \geq h(y_i)$

## Double censoring:

$x_i$  = age when child learns to read

$y_i$  = age when observation ends, right censoring

$z_i$  = age when observation begins, left censoring

Observe  $\{x_i\}$  or  $[0, z_i)$  or  $(y, \infty]$

## Left truncation and right censoring:

As above but only non-readers are observed

## Some NPMLEs

Kaplan-Meier for right censored data

$$\hat{F}((-\infty, t]) = 1 - \prod_{j|t_j \leq t} \frac{r_j - d_j}{r_j}$$

$r_j$  = Number alive at  $t_j$  -

$d_j$  = Number dying at  $t_j$

Lynden-Bell (conditional likelihood) for left truncated data

$$\hat{F}((-\infty, t]) = 1 - \prod_{i=1}^n \left( 1 - \frac{1_{x_i \leq t}}{\sum_{\ell=1}^n 1_{y_\ell < x_i \leq x_\ell}} \right)$$

Can have  $\hat{F}((-\infty, x_{(i)}]) = 1$  for some  $i < n$

## Some ELTs

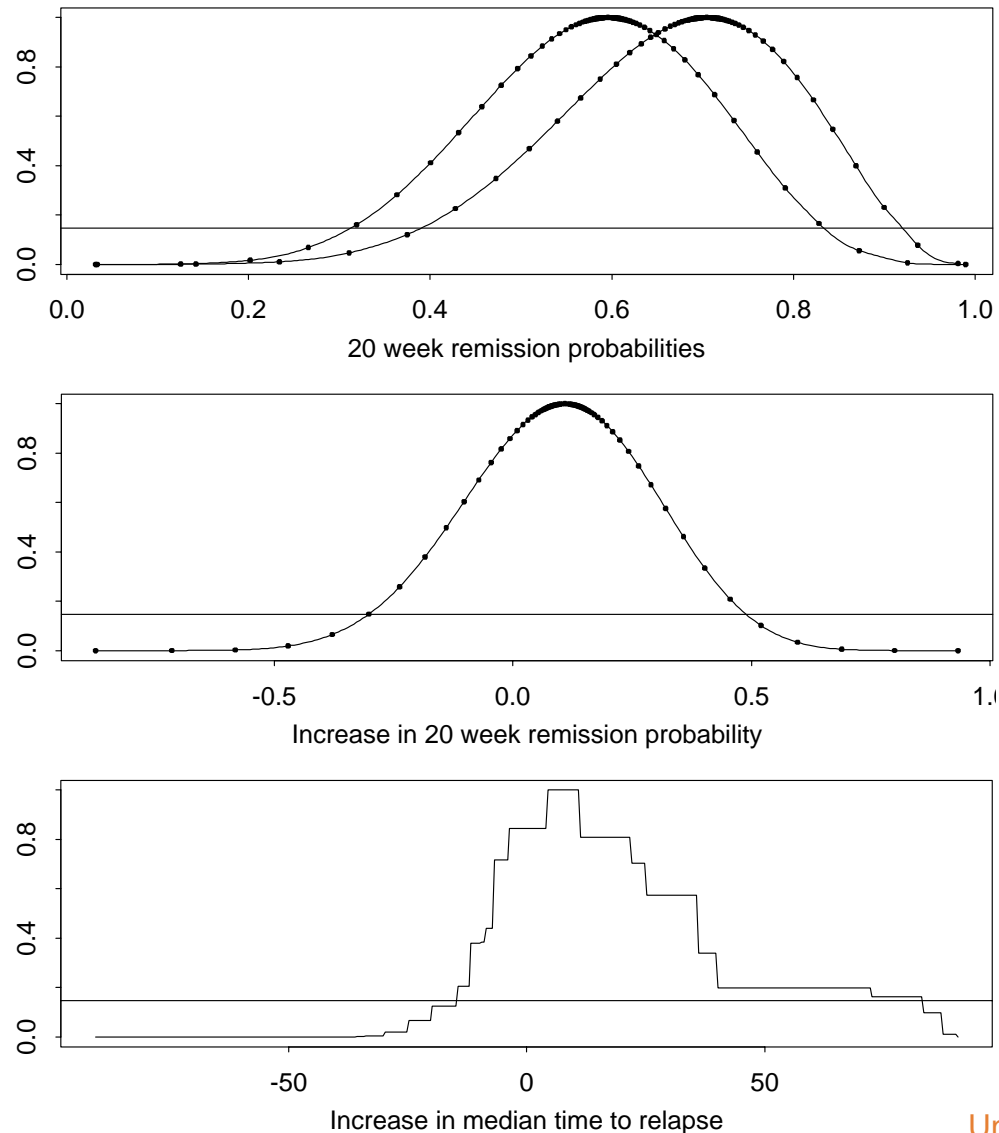
---

| Data type              | Statistic                 | Reference                        |
|------------------------|---------------------------|----------------------------------|
| Right censoring        | Survival prob             | Thomas & Grunkemeier, Li, Murphy |
| Left truncation        | Survival prob             | Li                               |
| Left trunc, right cens | Mean                      | Murphy & van der Vaart           |
| Right censoring        | proportional hazard param | Murphy & van der Vaart           |
| Right censoring        | integral vs cum hazard    | Pan & Zhou                       |

---

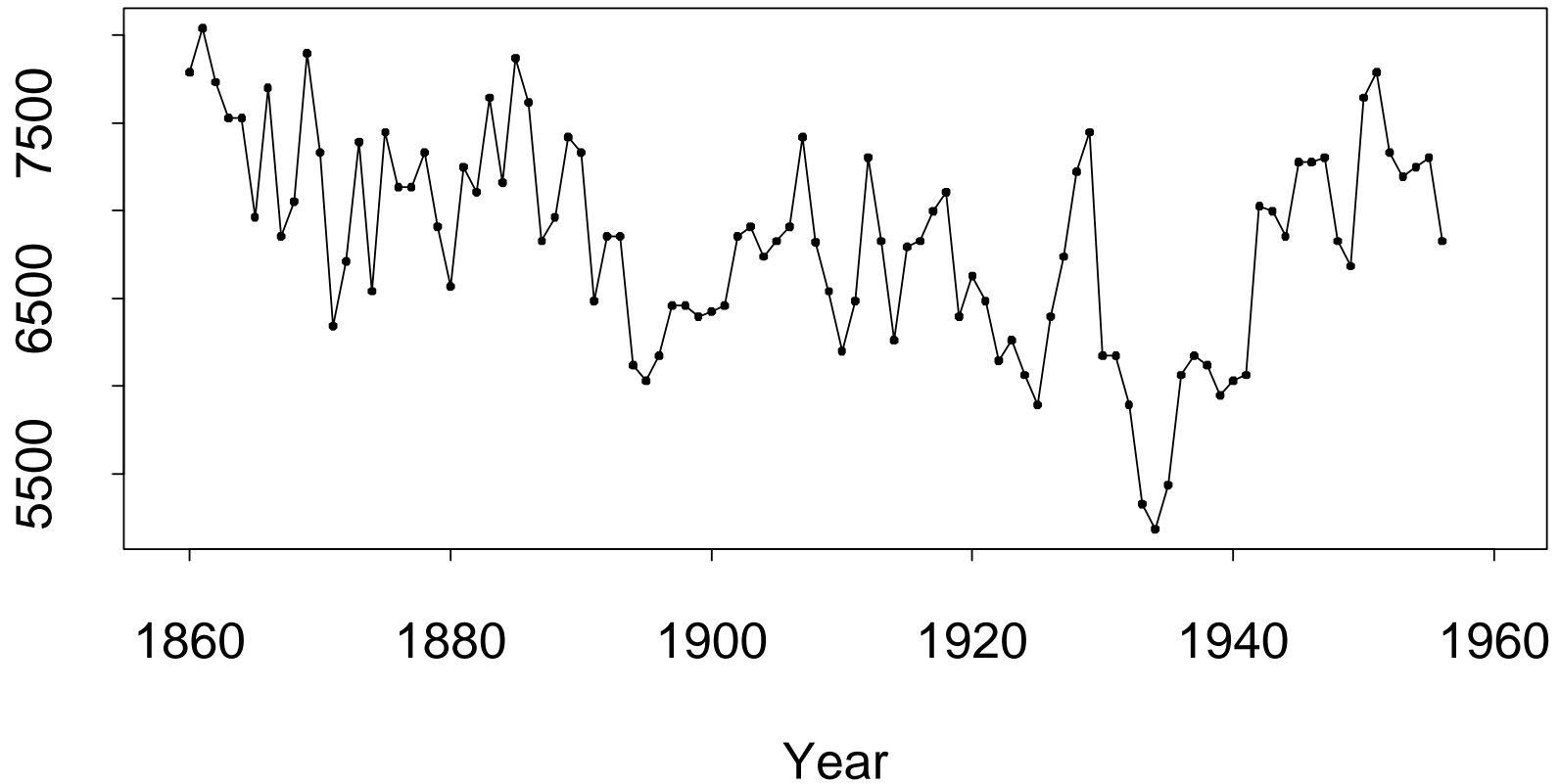
# Acute myelogenous leukemia (AML)

Embury et al. Weeks until relapse for 11 with maintenance chemotherapy and 12 non-maintained



# Time series

## St. Lawrence River flow



at Ogdensburg Yevjevich



# Reduce to independence

$$Y_i - \mu = \beta_1(Y_{i-1} - \mu) + \cdots + \beta_k(Y_{i-k} - \mu) + \epsilon_i$$

$$E(\epsilon_i) = 0$$

$$E(\epsilon_i^2) = \exp(2\tau)$$

$$E(\epsilon_i(Y_{i-j} - \mu)) = 0$$

| $j$ | $\hat{\beta}_j$ | $-2 \log \mathcal{R}(\beta_j = 0)$ |
|-----|-----------------|------------------------------------|
| 1   | 0.627           | 30.16                              |
| 2   | -0.093          | 0.48                               |
| 3   | 0.214           | 4.05                               |

# Blocking of time series

Block  $i$  of observations, out of  $n = \lfloor (T - M)/L + 1 \rfloor$  blocks

$$B_i = (Y_{(i-1)L+1}, \dots, Y_{(i-1)L+M})$$

$M$  = length of blocks

$L$  = spacing of start points

Large  $M = L \implies$  block dependence small

Large  $M \implies$  block dependence predictable given  $L$

Blocked estimating equation, replace  $m$  by  $b$

$$b(B_i, \theta) = \frac{1}{M} \sum_{j=1}^M m(X_{(i-1)L+j}, \theta)$$

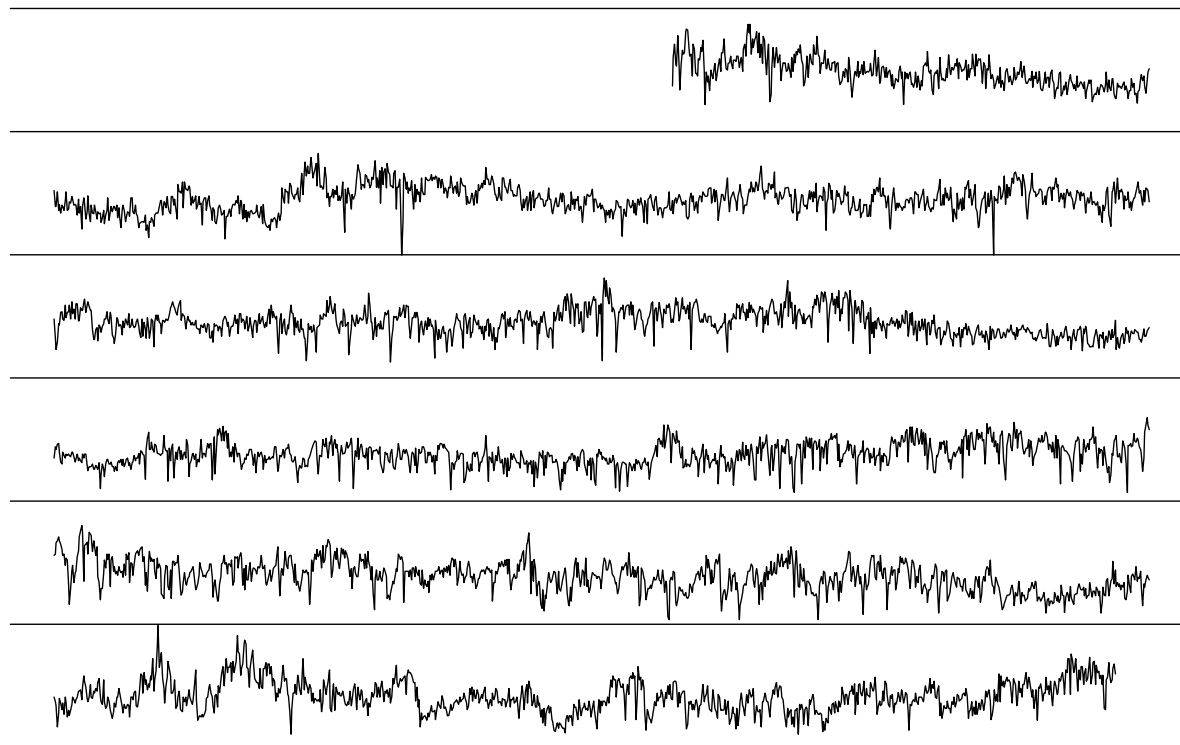
$$-2 \left( \frac{T}{nM} \right) \log \mathcal{R}(\theta_0) \rightarrow \chi^2 \quad \text{as } M \rightarrow \infty, MT^{-1/2} \rightarrow 0 \quad \text{Kitamura}$$

# Bristlecone pine

Possibly copyrighted picture of a bristlecone pine went here

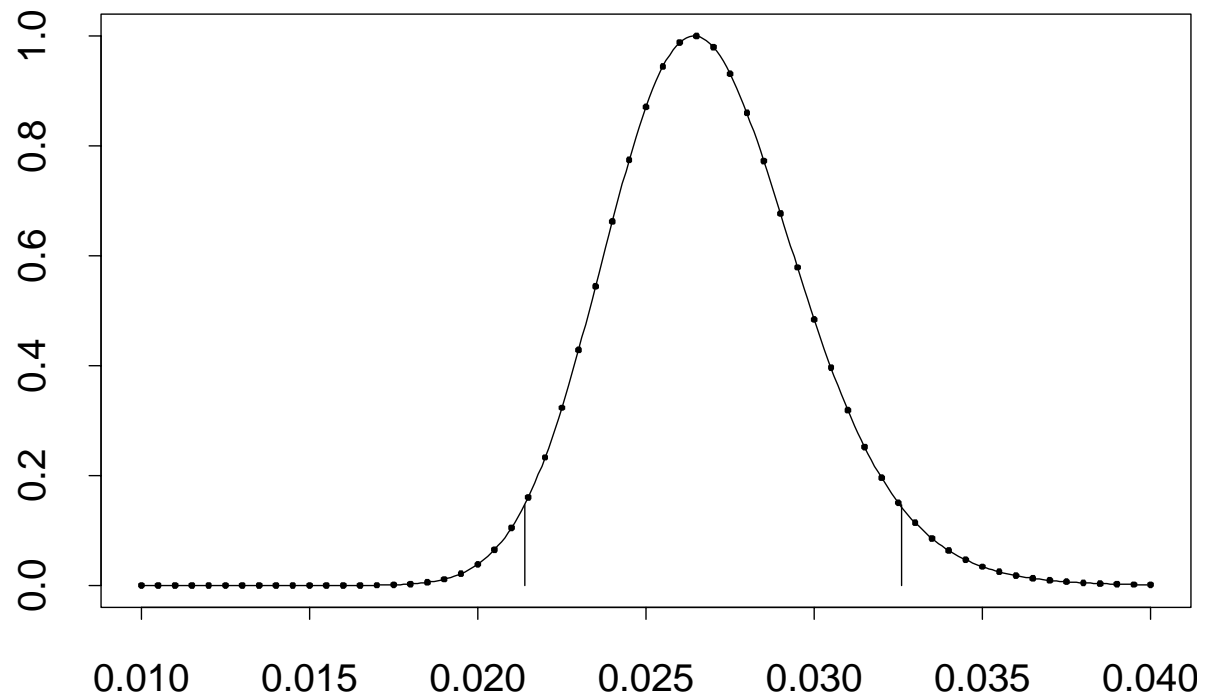
# 5405 years of Bristlecone pine tree ring widths

## Campito tree ring data



0 to 100 in 0.01 mm **Fritts et al.**

# Probability of sharp decrease



Sharp  $\equiv$  drop of over 0.2 mm from average of previous 10 years.

# MELs for finite population sampling

1. use side information
  - (a) population means, totals, sizes
  - (b) stratum means, totals, sizes
2. take unequal sampling probabilities
3. use non-negative observation weights

Hartley & Rao, Chen & Qin, Chen & Sitter

## More finite population results

---

|                  |   |                    |
|------------------|---|--------------------|
| ELTs             | $-2\left(1 - \frac{n}{N}\right)\mathcal{R}(\mu) \rightarrow \chi^2$ | Zhong & Rao        |
| EL variance ests | via pairwise inclusion probabilities                                | Sitter & Wu        |
| Multiple samples | varying distortions   | Zhong, Chen, & Rao |

---

## EL hybrids (mostly Jing Qin)

Part of the problem parametric

We want to use that knowledge

Rest of the problem non-parametric

# One parametric sample, one not

$Y$  well studied and has parametric distribution

$X$  new and/or does not follow parametric distribution

$$X_i \sim F, \quad i = 1, \dots, n$$

$$Y_j \sim G(y; \theta), \quad j = 1, \dots, m$$

$$0 = \int \int h(x, y, \phi) dF(x) dG(y; \theta)$$

e.g.  $\phi = E(Y) - E(X)$



## Multiply the likelihoods

$$L(F, \theta) = \prod_{i=1}^n F(\{x_i\}) \prod_{j=1}^m g(y_j; \theta)$$

$$R(F, \theta) = L(F, \theta) / L(\hat{F}, \hat{\theta})$$

$$\mathcal{R}(\phi) = \max_{F, \theta} R(F, \theta) \quad \text{such that}$$

$$0 = \sum_{i=1}^n w_i \int h(x_i, y, \phi) dG(y; \theta)$$

Qin gets an ELT

# Parametric model for data ranges

$$X \sim \begin{cases} f(x; \theta) & x \in P_0 \\ ??? & x \notin P_0 \end{cases}$$

## Examples

- Extreme values with exponential tails on  $P_0 = [T, \infty)$
- Normal data on  $P_0 = [-T, T]$  with outliers

$$L = \prod_{i=1}^n f(x_i; \theta)^{x_i \in P_0} w_i^{x_i \notin P_0}$$

Define  $\mathcal{R}$  using

$$1 = \int_{P_0} dF(x; \theta) + \sum_{i=1}^n w_i 1_{x \notin P_0}$$

Qin & Wong get an ELT for means

## More hybrids

Parametric

$$g(y | x; \theta)$$

$$x \sim f(x; \theta)$$

$$x \sim f(x; \theta)$$

Nonparametric

$$X \sim F$$

$$y | x \sim G_x$$

$$(y - \mu(x))/\sigma(x) \sim G$$

Few  $x$  vals

# Bayesian empirical likelihood (Lazar)

Prior  $\theta \sim \pi(\theta)$

$x \sim F$  nonparametric

Posterior  $\propto \pi(\theta)\mathcal{R}(\theta)$

Here we have informative prior nonparametric likelihood

Reverse of common practice

Posterior regions asymptotically properly calibrated

Justify via least favorable families

# Curve estimation problems

$$\hat{f}_h(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \quad \text{density}$$

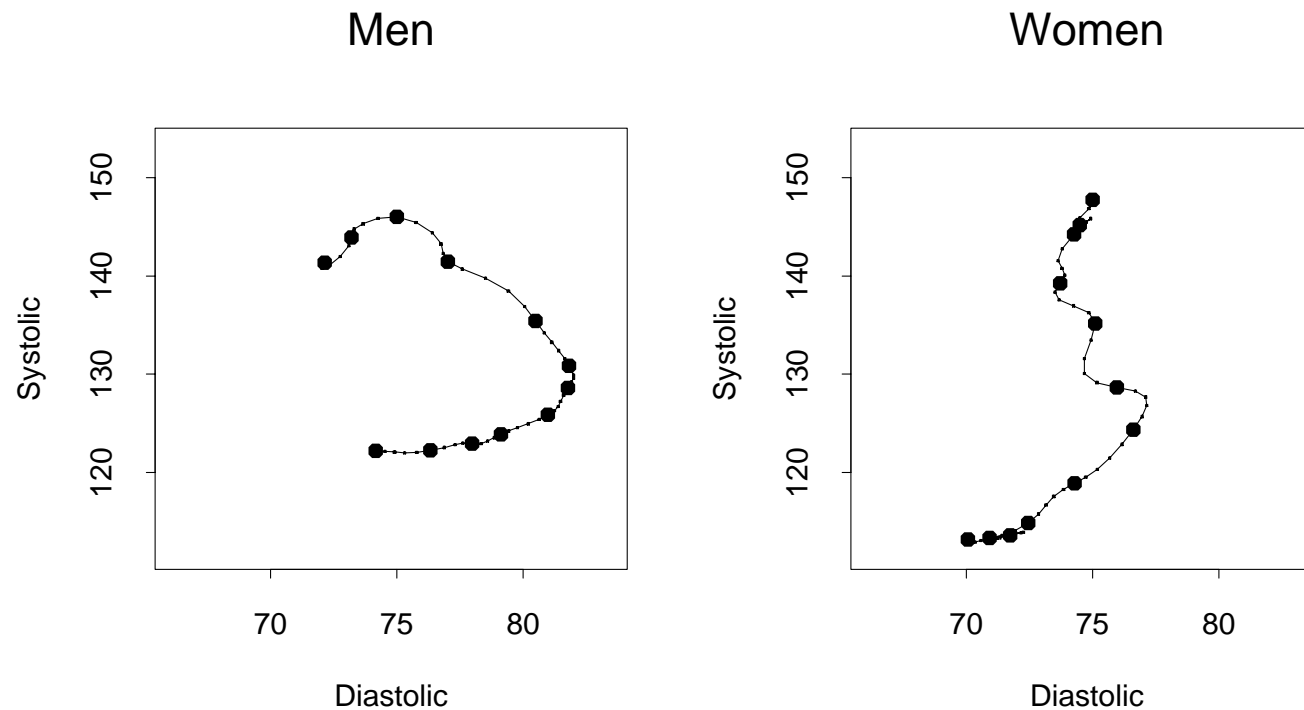
$$\hat{\mu}_h(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) Y_i \quad \text{regression}$$

Triangular array ELT applies      Bias adjustment issues

## Dimensions and geometry

| Dim(x)   | Dim(y)   | Estimate        | Region              |
|----------|----------|-----------------|---------------------|
| 1        | $\geq 2$ | space curve     | confidence tube     |
| $\geq 2$ | 1        | (hyper)-surface | confidence sandwich |

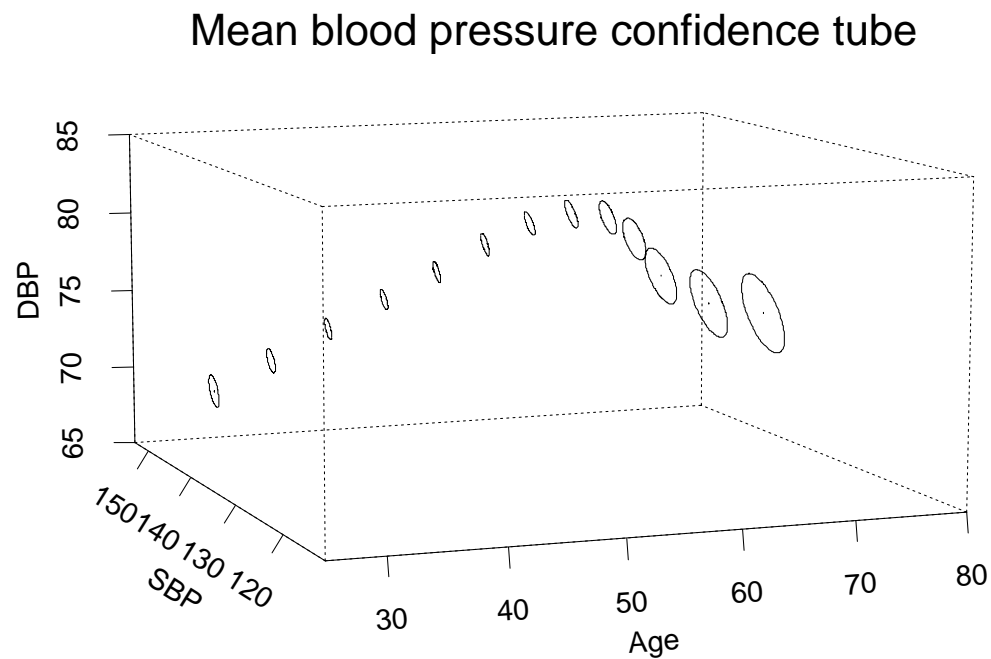
# Trajectories of mean blood pressure



dots at ages 25, 30, ..., 80

data from Jackson et al., courtesy of Yee

# Confidence tube for men's mean SBP, DBP



# Empirical likelihood vs bootstrap

1. EL gives shape of regions for  $d > 1$
2. EL Bartlett correctable, bootstrap not
3. EL can be faster, but,
4. EL optimization can be hard



# Why use anything else?

1. Computation is hard
2. Convex hull is binding

# Computation

$$\begin{aligned}\log \mathcal{R}(\theta) &= \max_{\nu} \log \mathcal{R}(\theta, \nu) \\ &= \max_{\nu} \min_{\lambda} \mathbb{L}(\theta, \nu, \lambda), \quad \text{where,} \\ \mathbb{L}(\theta, \nu, \lambda) &= - \sum_{i=1}^n \log(1 + \lambda' m(x_i, \theta, \nu))\end{aligned}$$

Inner and outer optimizations  $\ll n$  dimensional

Used NPSOL, expensive and not public domain (but it works)

# Convex hull

confidence regions nested inside convex hull of data

restrictive if  $d$  not small

not so bad for one and two dimensional subparameters

## possible remedies

1. Empirical likelihood  $t$  [Baggerley](#)
2. Hybrid with Euclidean likelihood