Bayes, Oracle Bayes, and Empirical Bayes

Bradley Efron

Stanford University

Empirical Bayes Inference

- Robbins (1951) *Compound Decision Procedures*
- Robbins (1956) *Empirical Bayes*
- **Question** How does *empirical Bayes* relate to Bayesian and frequentist inference?
- Intermediate framework: Oracle Bayes

A Familiar Empirical Bayes Setup

- Unknown prior $g(\theta)$ produces unseen parameters

$$\boxed{g(\theta) \longrightarrow \theta_1, \theta_2, \ldots, \theta_N}$$

- Each θ_i independently gives observation $x_i \sim \mathcal{N}(\theta_i, 1)$,

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$$

- Robbins Use all of $\boldsymbol{x}$ to estimate each θ_i

Oracle Bayes

Jiang and Zhang (2009)

- Oracle tells us the order statistic of the θ's,

$$\boldsymbol{\theta}_{\text{ord}} = \left(\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(N)}\right)$$

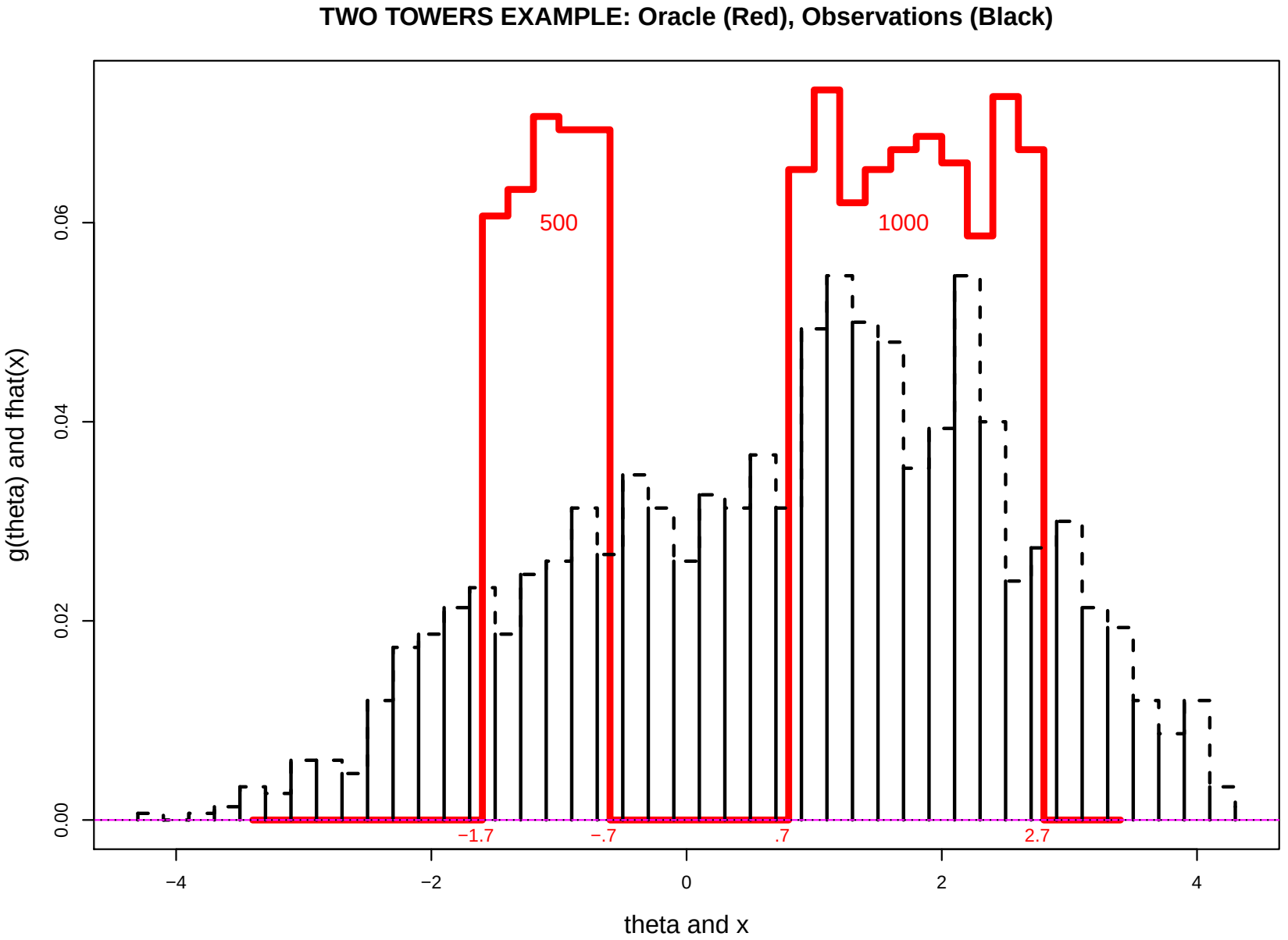
  (but not which x_i goes with which θ_i)

- We want to estimate $\hat{\theta}_i$ that minimizes

  Expected Average Squared Error

$$\boxed{\text{EASE} = E\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2\right\}}$$

- *Example:* "Two Towers"


TWO TOWERS EXAMPLE: Oracle (Red), Observations (Black)
g(theta) and fhat(x)
0.00
0.02
0.04
0.06
500
1000
−1.7
−.7
.7
2.7
−4
−2
0
2
4
theta and x

Using the Oracle

- $\bar{g}(\theta)$ empirical density of the θ's

  (probability $1/N$ on $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(N)}$)

- *Bayes estimate* (ϕ the standard normal density)

$$\hat{\theta}_{\bar{g}}(x) = \sum_1^N \theta_{(i)}\phi(x - \theta_{(i)}) \Big/ \sum_1^N \phi(x - \theta_{(i)})$$

- Minimizes EASE among rules $\hat{\theta}(x) = t(x)$
- Two Towers $\quad \text{EASE}_{\bar{g}} = 0.563$
- MLE estimates $\hat{\theta}_i = x_i$ has $\text{EASE}_{\text{MLE}} = 1.0$

Empirical Bayes

- No Oracle Use $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ to form estimate $\hat{\theta}_{\boldsymbol{x}}(x)$ of Oracle rule $\hat{\theta}_{\bar{g}}(x)$
- *Empirical Bayes regret*

$$\boxed{\text{EASE}_{\boldsymbol{x}} - \text{EASE}_{\bar{g}} = \text{EBregret}}$$

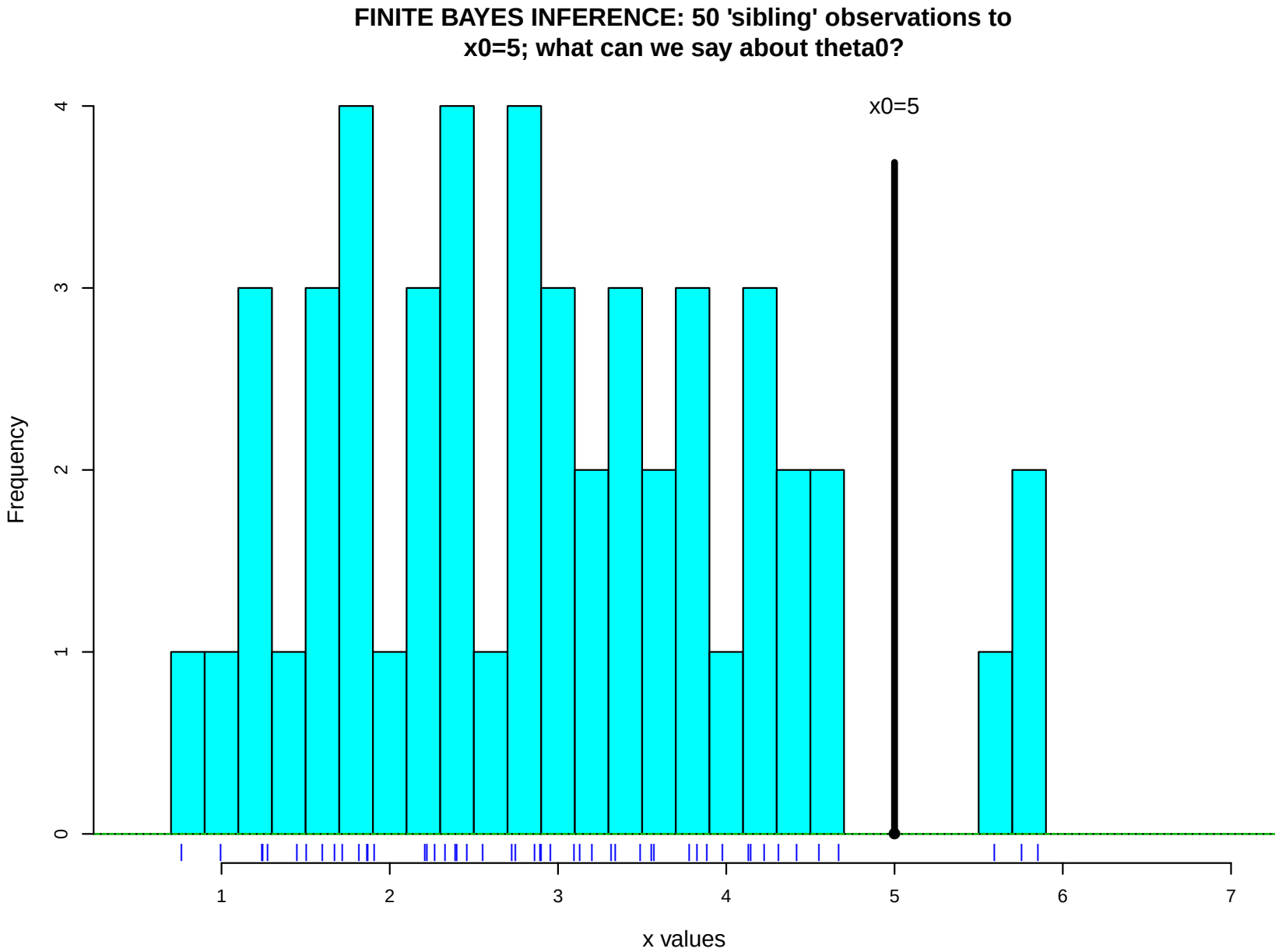
- Two Towers $\text{EBregret} = 0.008$
- $\text{EASE}_{\boldsymbol{x}} = 0.563 + 0.008 = 0.571$

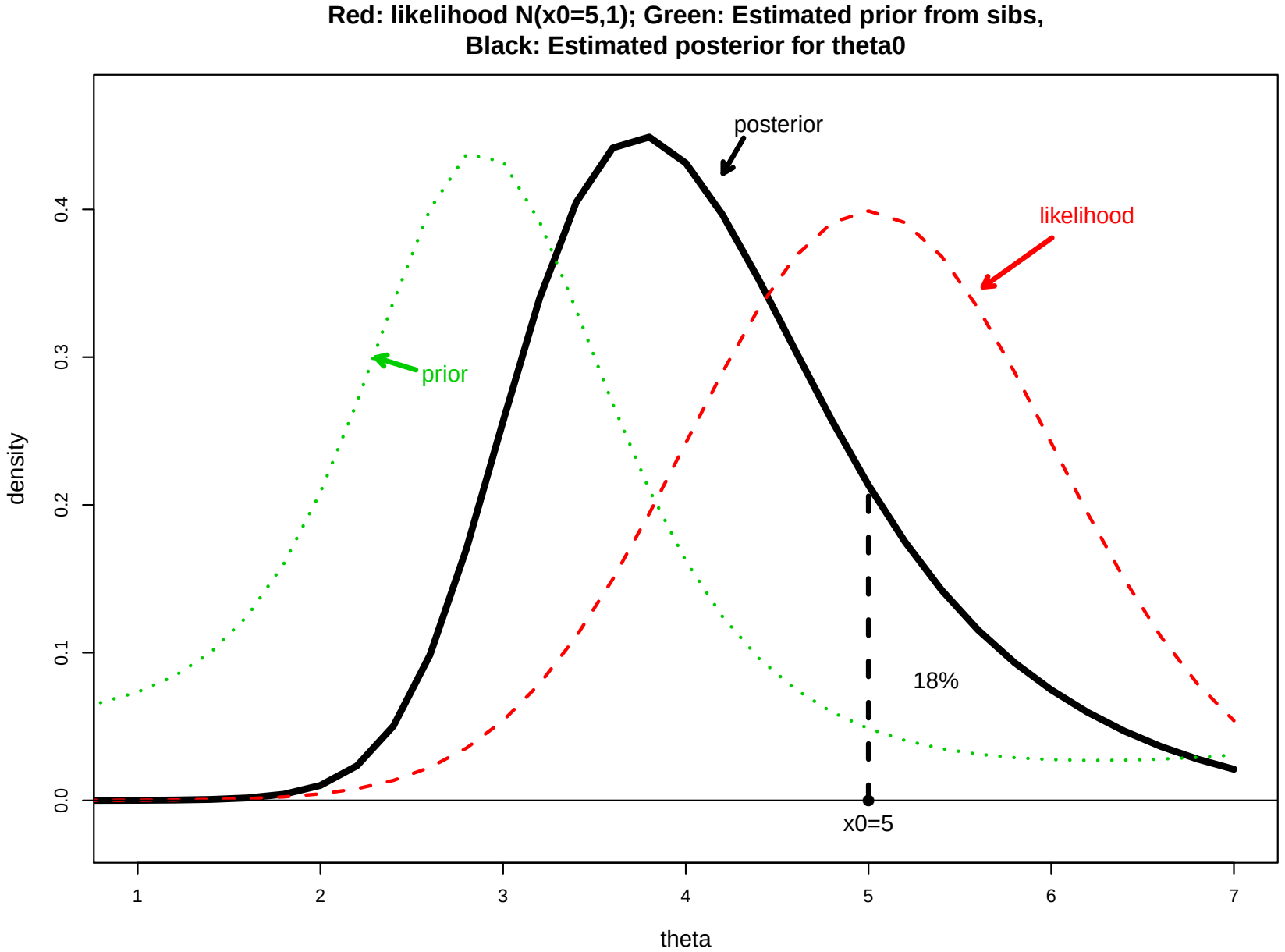
The Frequentist Face of Empirical Bayes

- Oracle application is entirely frequentist:
  - Minimizing EASE is a frequentist criterion
  - Assumption $\theta_i \sim g(\theta)$ is irrelevant!
- **Question** Is the 40% EASE reduction meaningful?
- Answer *Yes* for EASE, *No* for some other applications

Finite Bayes Inference

- $\theta_i \overset{\text{ind}}{\sim} g(\theta)$ and $x_i \mid \theta_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1)$ for $i = 0, 1, 2, \ldots, N$
- We *want* posterior inference for θ_0 given x_0 and the "sibling" observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$
- If $N \to \infty$ then finite Bayes $\to$ Bayes
- *Next example:* $x_0 = 5, N = 50$


FINITE BAYES INFERENCE: 50 'sibling' observations to
x0=5; what can we say about theta0?
x0=5
Frequency
0
1
2
3
4
1
2
3
4
5
6
7
x values


Red: likelihood N(x0=5,1); Green: Estimated prior from sibs,
Black: Estimated posterior for theta0
posterior
likelihood
prior
18%
x0=5
density
theta
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7

Standard Bayes (Just one θ and x)

- Prior $\theta \sim g(\theta)$
- Observe $x \sim p_\theta(x)$ $[\mathcal{N}(\theta, 1)]$
- Marginal density $f(x) = \int_{\mathcal{X}} g(\theta) p_\theta(x)\, d\theta$
- *Bayesian inference* $x \sim f(x)$ and

$$\theta \mid x \sim \left[e_g(x), v_g(x)\right]$$

conditional expectation ↗ ↖ conditional variance

- For squared error: $e_g(x)$ is Bayes estimate "$\hat{\theta}_g(x)$"
- *Bayes risk* $\mathcal{R}_g = E\left\{\left(\theta - e_g(x)\right)^2\right\} = \int_{\mathcal{X}} v_g(x) f(x)\, dx$
- $\mathcal{R}_{\bar{g}} = \text{EASE}_{\bar{g}}$ $(= 0.563)$

Tweedie's Formula (Efron, 2011)

- If $x \mid \theta \sim \mathcal{N}(\theta, 1)$:

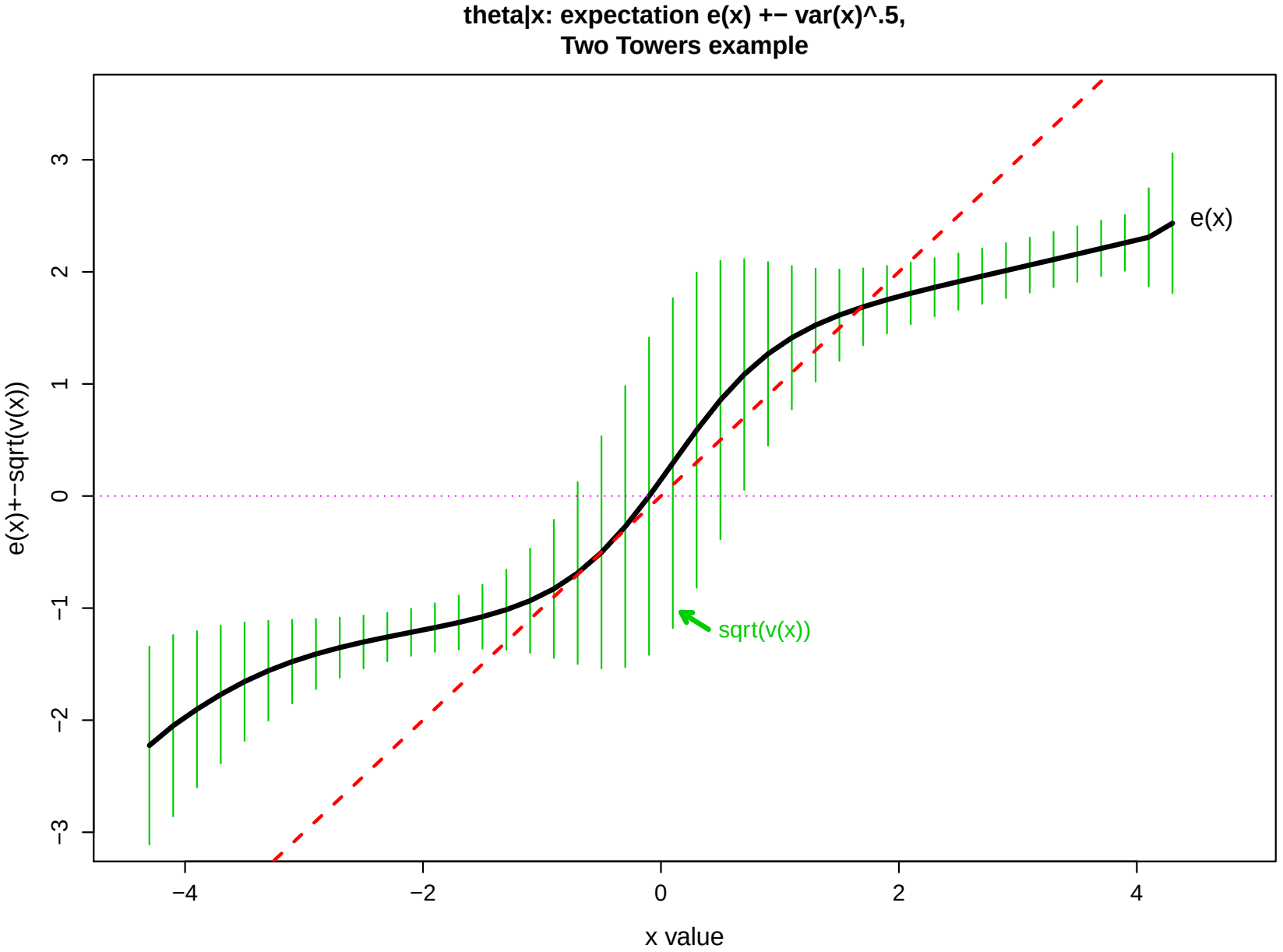
$$e_g(x) = x + l'(x) \qquad [l(x) = \log f(x)]$$

$$v_g(x) = 1 + l''(x)$$

- So $\hat{\theta}_g(x) = x + l'(x)$ and

$$\boxed{\mathcal{R}_g = \int_{\mathcal{X}} [1 + l''(x)]\, f(x) = 1 - \int_{\mathcal{X}} [l'(x)]^2\, f(x)}$$

- *Next:* $g(\theta) = \bar{g}(\theta)$, Two Towers


theta|x: expectation e(x) +− var(x)^.5,
Two Towers example
e(x)+−sqrt(v(x))
3
2
1
0
−1
−2
−3
−4
−2
0
2
4
x value
e(x)
sqrt(v(x))

Empirical Bayes Risk and Regret

- Idea Estimate "$\hat{f}(x)$" from data $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$
- $\hat{e}(x) = x + \hat{l}'(x)$ and $\hat{v}(x) = 1 + \hat{l}''(x)$ $\quad \left[\hat{l}(x) = \log \hat{f}(x)\right]$
- *Empirical Bayes risk* $\mathcal{R}(g, \hat{e}) = E\left\{(\hat{e}(x) - \theta)^2\right\}$

Lemma

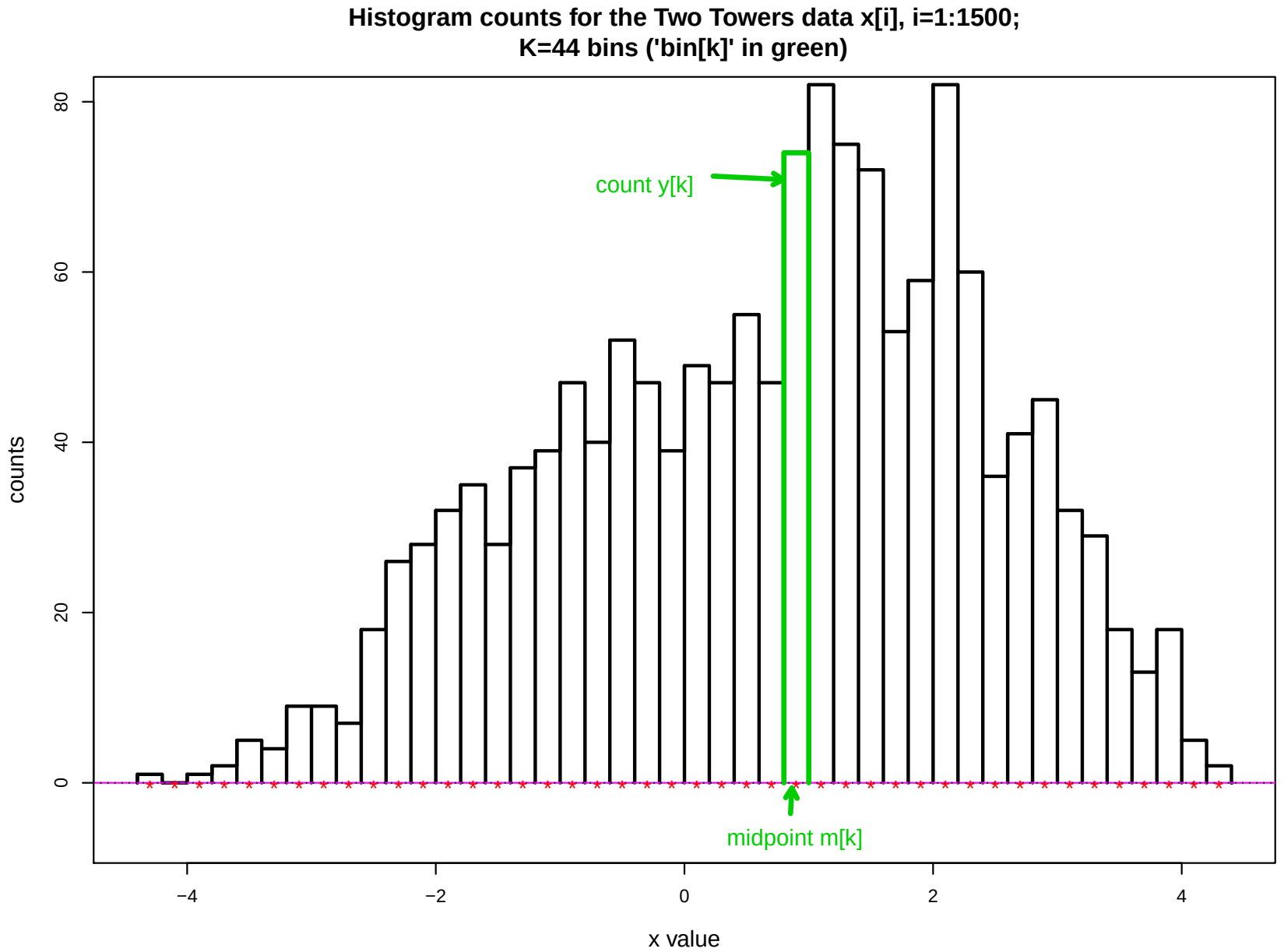
$$\boxed{\mathcal{R}(g, \hat{e}) = \mathcal{R}_g + \int_{\mathcal{X}} \left[\hat{e}(x) - e_g(x)\right]^2 f(x)\, dx}$$

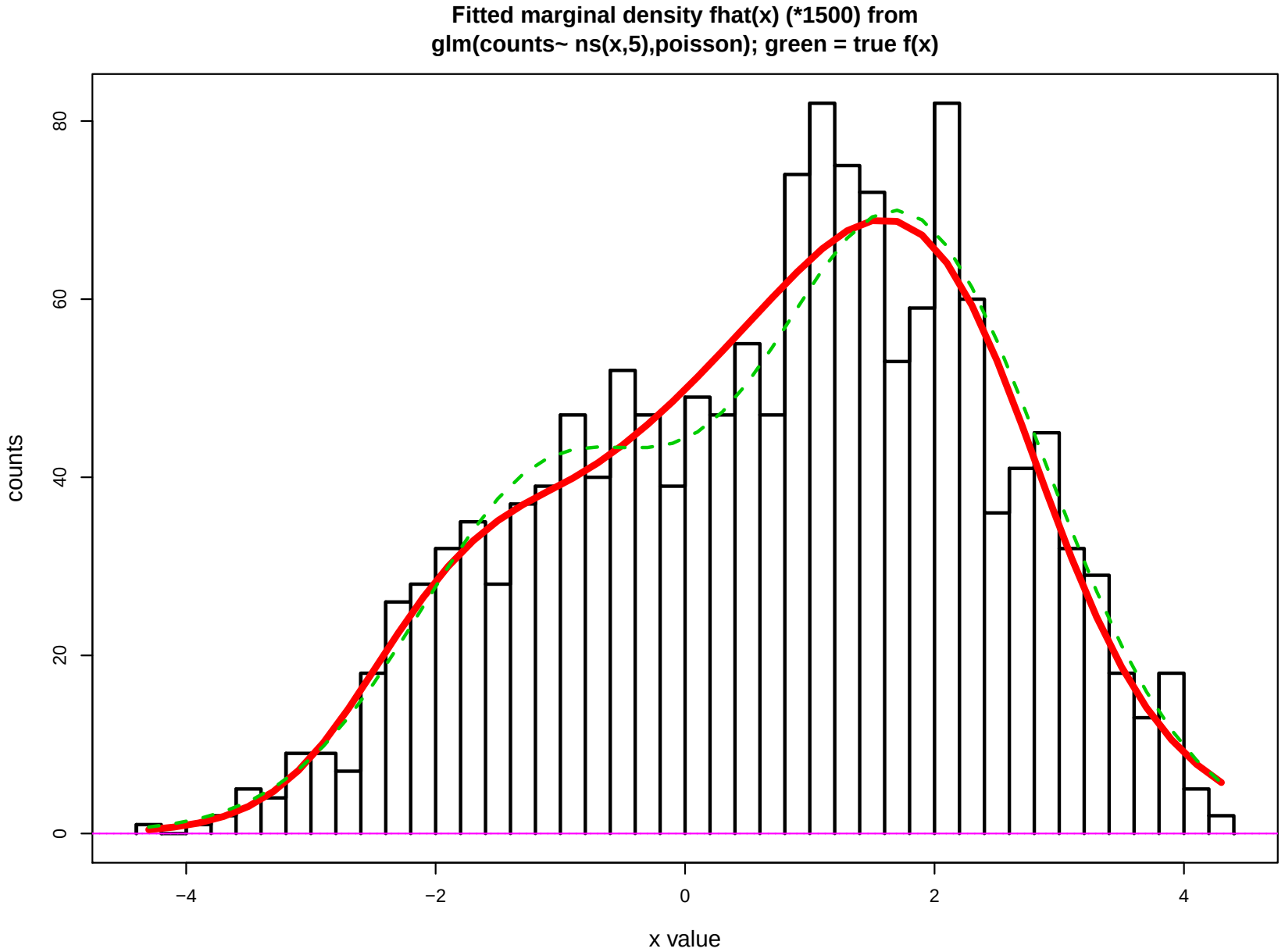
Bayes risk ↗ ↖ *EBregret*

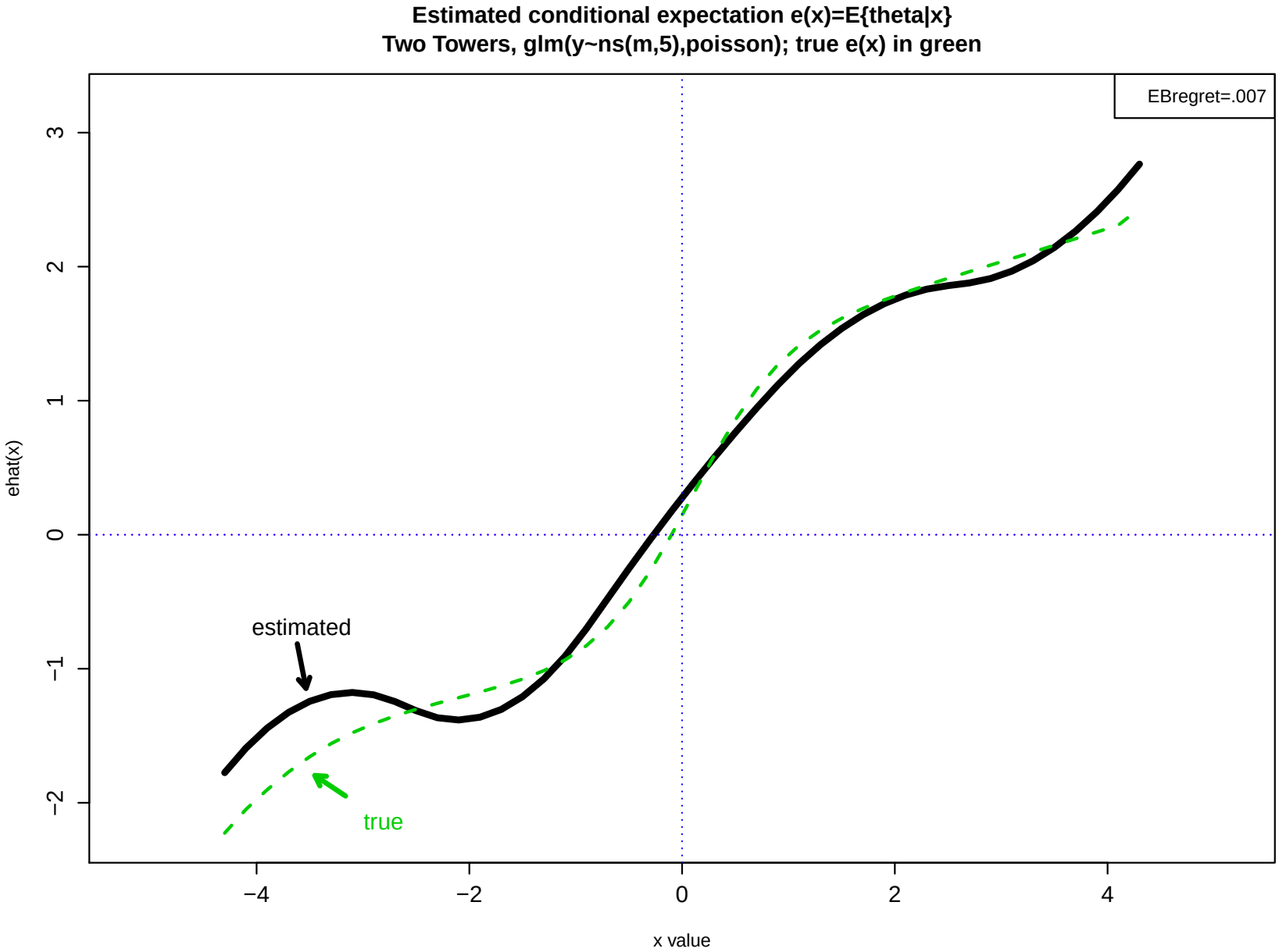
- Only need f and its estimate $\hat{f}$

Estimating $f(x)$: "f-Modeling"

- Bin data $x_1, x_2, \ldots, x_N$
- y_k = number of x's in bin k
- m_k = midpoint of bin k
- $\hat{f} = \texttt{glm(y~ns(m,df),poisson)\$est/N}$
- Two Towers $K = 44$ bins ($\texttt{df} = 5$)


Histogram counts for the Two Towers data x[i], i=1:1500;
K=44 bins ('bin[k]' in green)
counts
0
20
40
60
80
count y[k]
midpoint m[k]
−4
−2
0
2
4
x value


Fitted marginal density fhat(x) (*1500) from
glm(counts~ ns(x,5),poisson); green = true f(x)
counts
0
20
40
60
80
−4
−2
0
2
4
x value


Estimated conditional expectation e(x)=E{theta|x}
Two Towers, glm(y~ns(m,5),poisson); true e(x) in green
EBregret=.007
ehat(x)
3
2
1
0
−1
−2
estimated
true
−4
−2
0
2
4
x value

Data-based Formula for EBregret

- $\underset{44\times 6}{\boldsymbol{M}} = \texttt{(ns(m,5), 1)}$, kth row $\boldsymbol{M}_k$
- $\underset{44\times 6}{\dot{\boldsymbol{M}}}$, kth row $\dfrac{d\boldsymbol{M}_k}{dm_k}$
- $\hat{\boldsymbol{f}}$ = binned estimate of $f(x)$ from `glm(y~M,poisson)`

Theorem

$$E\Big\{\text{EBregret}\Big\} \doteq \frac{1}{N}\,\text{trace}\left\{\left[\boldsymbol{M}'\,\text{diag}\left(\hat{\boldsymbol{f}}\right)\boldsymbol{M}\right]^{-1}\left[\dot{\boldsymbol{M}}'\,\text{diag}\left(\hat{\boldsymbol{f}}\right)\dot{\boldsymbol{M}}\right]\right\}$$

- Decreases as $1/N$
- Doesn't account for bias of $\hat{e}$

The Poisson Case

- $\theta \sim g(\cdot)$ and $x \mid \theta \sim \text{Poi}(\theta)$ $\quad \left[p_\theta(x) = e^{-\theta}\theta^x/x!\right]$
- $f(x) = \int g(\theta)p_\theta(x)\,d\theta$ and $\theta \mid x \sim \left[e_g(x), v_g(x)\right]$
- Robbins $\quad \boxed{e_g(x) = \dfrac{(x+1)f(x+1)}{f(x)}}$

  and $\quad v_g(x) = e_g(x)\left[e_g(x+1) - e_g(x)\right]$
- *Bayes estimate* $\hat{\theta}_g(x) = e_g(x)$ $\quad$ with $\mathcal{R}_g = \sum_{\mathcal{X}} f(x)v_g(x)$

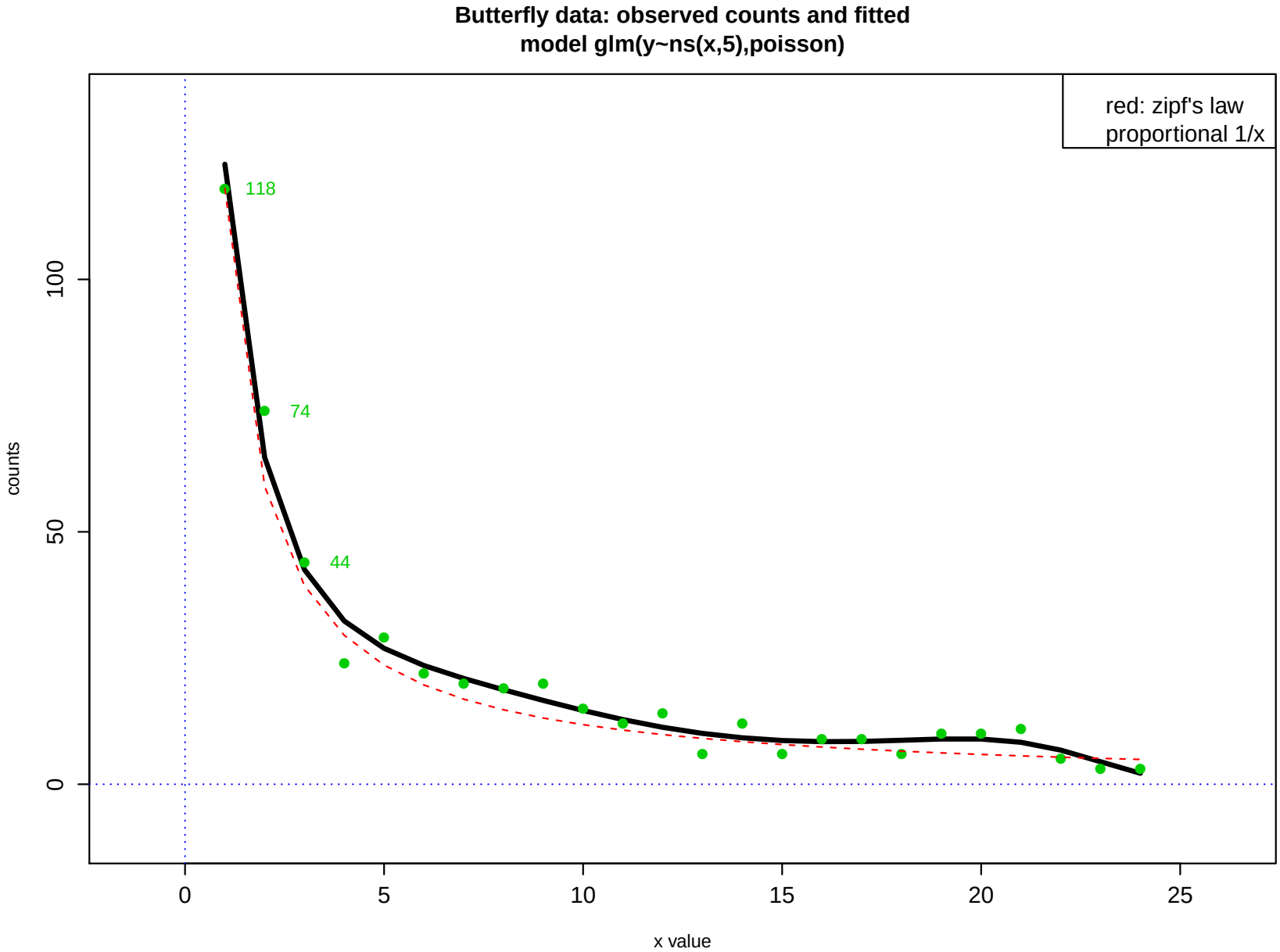
$$\mathcal{R}\left[g, e(x)\right] = \mathcal{R}_g + \sum_{\mathcal{X}} f(x)\left[e(x) - e_g(x)\right]^2$$

Butterfly Data

Corbet in Malaysia, 1941–42

- $y_1 = 118$ species trapped just one time each
- $y_2 = 74$ species trapped twice each. . .

x	1	2	3	4	$\cdots$	24
y	118	74	44	24	$\cdots$	3


Butterfly data: observed counts and fitted
model glm(y~ns(x,5),poisson)
red: zipf's law
proportional 1/x
118
74
44
counts
0
50
100
0
5
10
15
20
25
x value

Zipf's Law and Robbins' Formula

- If $f(x) = c/x$ then Robbins' formula gives

$$e_g(x) = \frac{(x+1)f(x+1)}{f(x)} = x$$

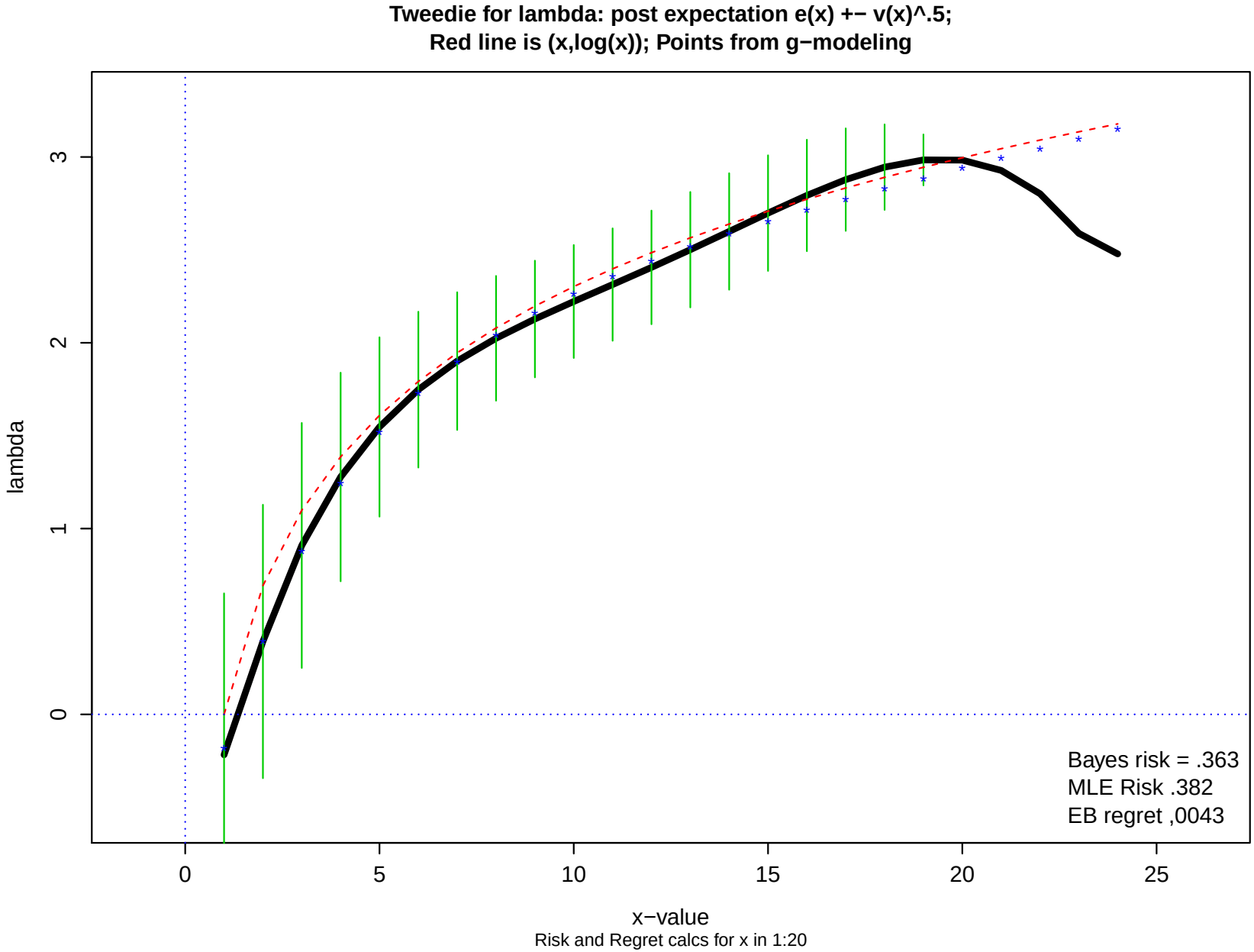
- So the MLE $\hat{\theta}_i = x_i$ is Bayes!
- Butterfly estimate $\mathcal{R}_g = 5.24$ and $\mathcal{R}(g, x) = 5.39$

Tweedie Poisson Estimation

- $\theta \sim g(\cdot)$ and $x \mid \theta \sim \text{Poi}(\theta)$
- $f(x) = \int g(\theta)\left(e^{-\theta}\theta^x/x!\right)\, d\theta$
- $\lambda = \log\theta$ is "natural parameter"

$$E\{\lambda \mid x\} = e_g(x) = \text{lgamma}(x+1)' + l'(x)$$
$$\text{Var}\{\lambda \mid x\} = v_g(x) = \text{lgamma}(x+1)'' + l''(x)$$

- Same for truncated Poisson!
- *Next:* Butterfly data, $\hat{f}$ from `ns(df=5)`


Tweedie for lambda: post expectation e(x) +− v(x)^.5;
Red line is (x,log(x)); Points from g−modeling
lambda
0
1
2
3
0
5
10
15
20
25
x−value
Risk and Regret calcs for x in 1:20
Bayes risk = .363
MLE Risk .382
EB regret ,0043

Empirical Bayes: g-Modeling

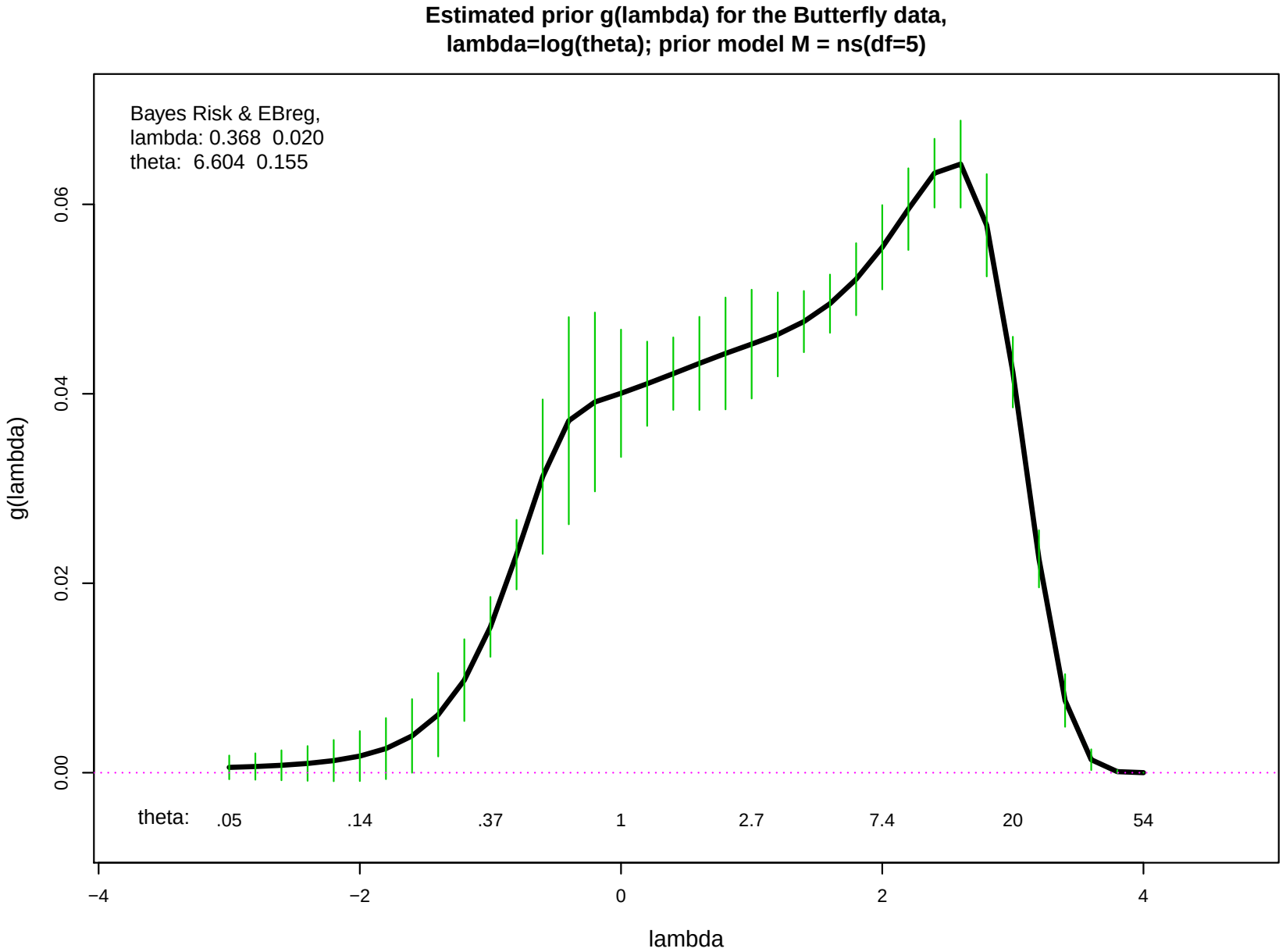
- *f-modeling:* Never need to estimate prior $g(\theta)$
- But only for criteria that depend just on $f(x)$:
  - EASE
  - Tweedie
  - Robbins
  - false discovery rates...
- *g-modeling* (Efron, 2016): Direct parametric modeling of prior $g(\theta)$

Parametric Models for $g(\theta)$

- *Exponential family* $\log g(\theta) = m(\theta)'\beta$ ("hidden GLM")
- β a p-dimensional parameter; $m(\theta)$ say $\texttt{ns}(\theta, 5)$
- MLE $\hat{\beta}$ found by nonlinear maximization:

$$\boxed{\beta \to g_\beta(\theta) \to f_\beta(x) \to y \sim \text{Mult}(N, f_\beta)}$$

- **Trouble** $y \sim \text{Mult}(N, f_\beta)$ not exponential family
- Advantage: Estimate $g(\theta \mid x)$, $\Pr\{\theta > 2 \mid x\}$, etc.


Estimated prior g(lambda) for the Butterfly data,
lambda=log(theta); prior model M = ns(df=5)
Bayes Risk & EBreg,
lambda: 0.368 0.020
theta: 6.604 0.155
g(lambda)
0.00
0.02
0.04
0.06
theta:
.05
.14
.37
1
2.7
7.4
20
54
−4
−2
0
2
4
lambda

Missing Species Problem

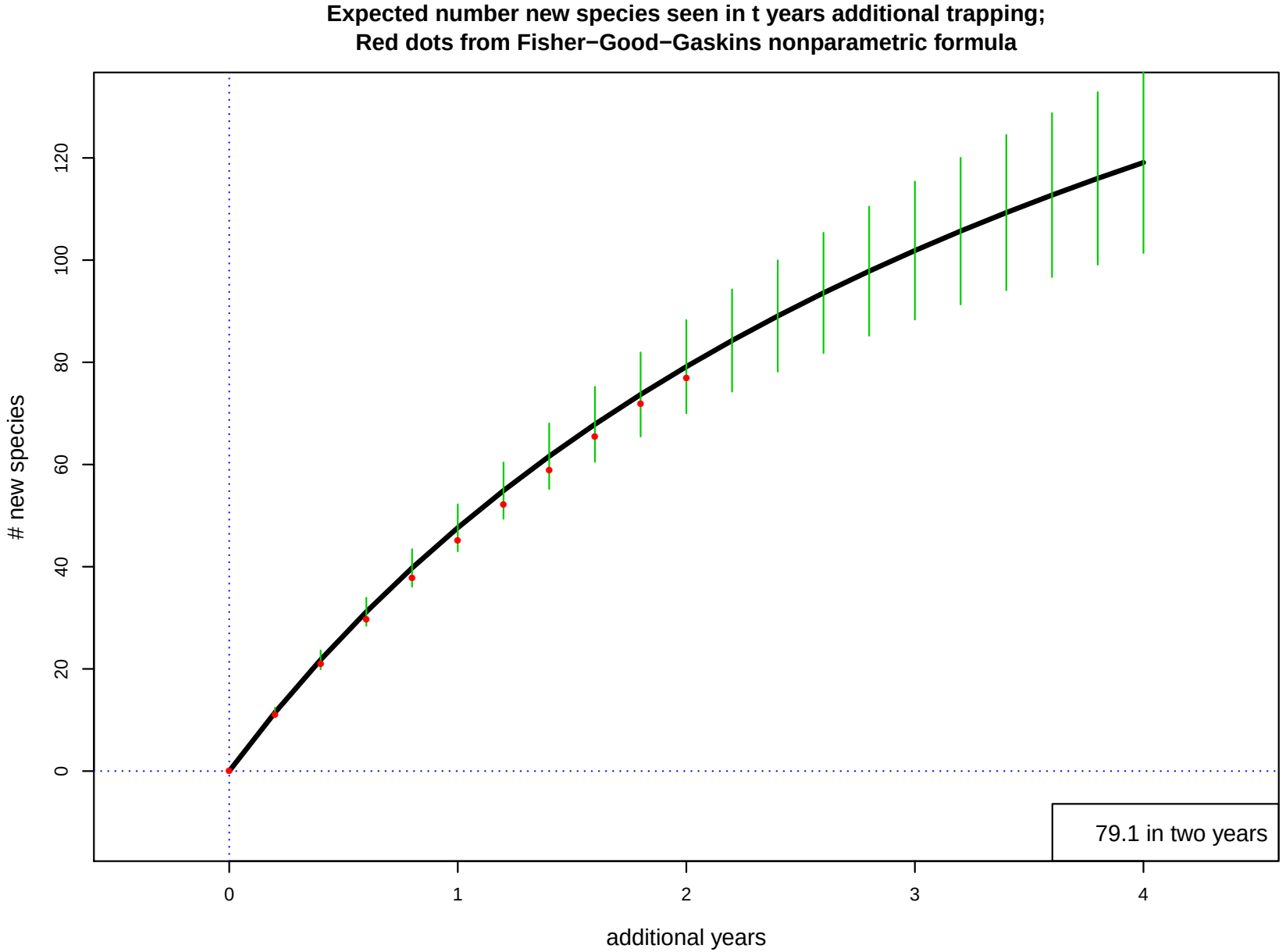
- Corbet How many new species in t more years of observation?
- Fisher, Good (Turing) $\longrightarrow$ Poisson process model

$$\boxed{\frac{E\{\#\text{new}\}}{\#\text{old}} = \int e^{-\theta} \frac{1 - e^{-\theta t/2}}{1 - e^{-\theta}} g(\theta)\, d\theta}$$

- *f-modeling:* Clever formula in terms of $\hat{f}$

  (Good and Toulmin, 1956)

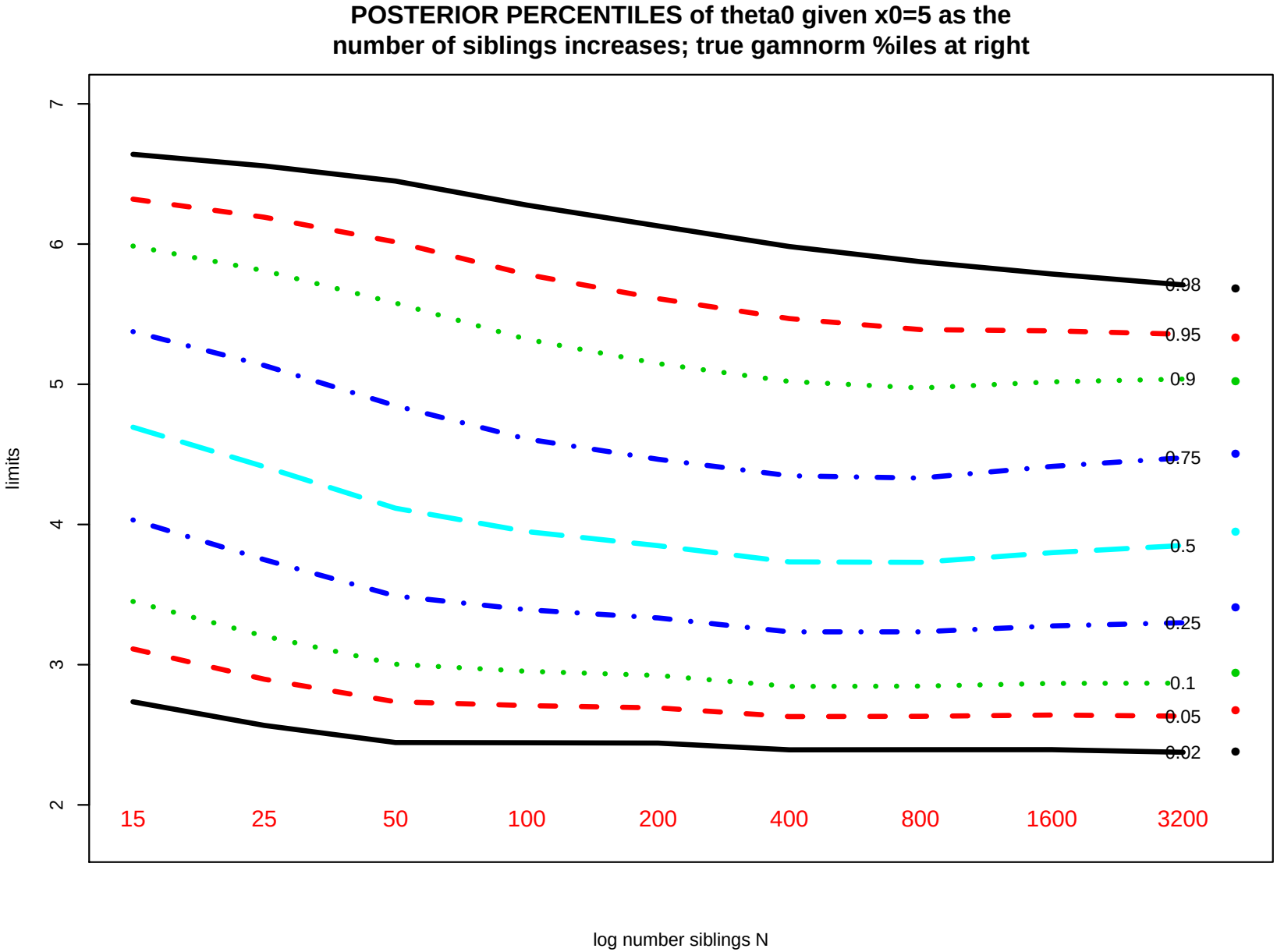
- *g-modeling:* Substitute $\hat{g}(\theta)$ for $g(\theta)$


Expected number new species seen in t years additional trapping;
Red dots from Fisher−Good−Gaskins nonparametric formula
new species
0
20
40
60
80
100
120
0
1
2
3
4
additional years
79.1 in two years

A *g*-Modeling Example

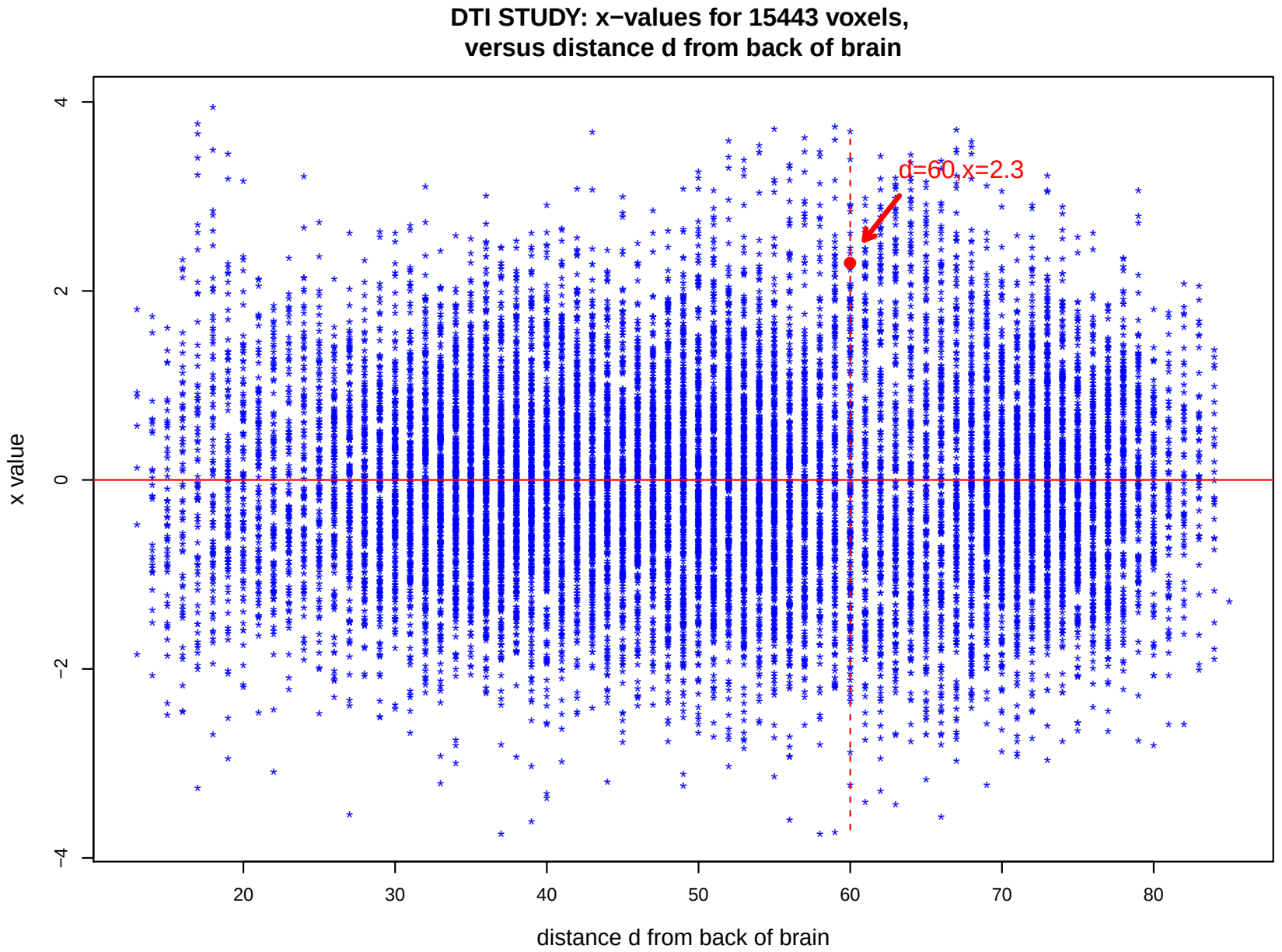
- "gamnorm" $\theta_i \sim \text{Gamma}_9\,/3,\ x_i \sim \mathcal{N}(\theta_i, 1), \quad i = 1, 2, \ldots, 3200$
- $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ for $N = 15, 25, 50, \ldots, 3200$
- $g(\theta) = \mathtt{ns}(\theta, 5)$
- *Next:* posterior quantiles of $\theta_0 \mid x_0 = 5$

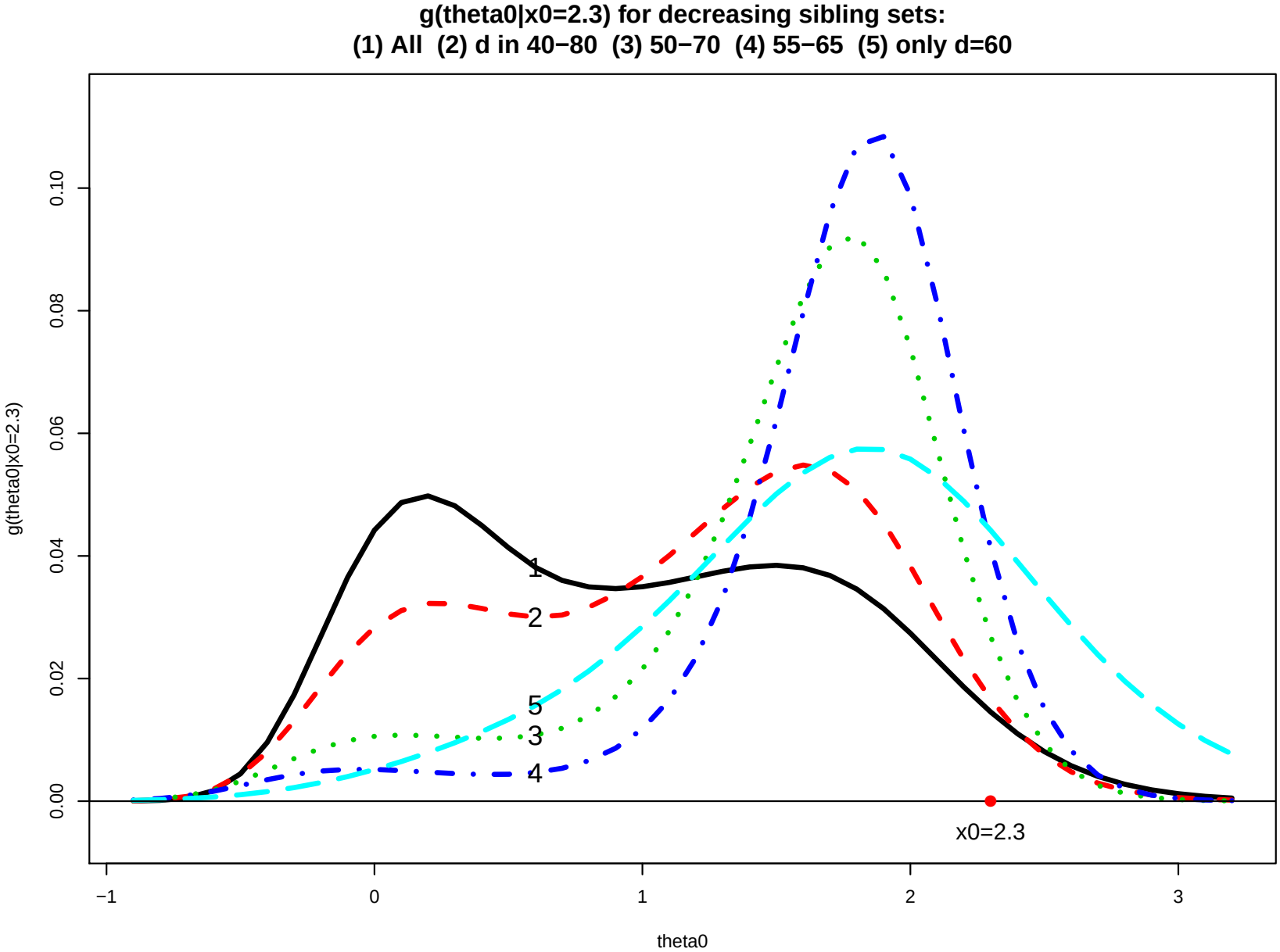
  (finite Bayes inference as N increases)


POSTERIOR PERCENTILES of theta0 given x0=5 as the
number of siblings increases; true gamnorm %iles at right
limits
2
3
4
5
6
7
0.98
0.95
0.9
0.75
0.5
0.25
0.1
0.05
0.02
15
25
50
100
200
400
800
1600
3200
log number siblings N

DTI Study

- 12 children, 6 dyslexic vs 6 normal controls
- $N = 15433$ brain voxels
- "x_i", x-value for ith voxel, $x_i \sim \mathcal{N}(\theta_i, 1)$
- $\theta_i =$ true effect size $\quad (= E\{x_i\})$


DTI STUDY: x−values for 15443 voxels,
versus distance d from back of brain
d=60, x=2.3
x value
distance d from back of brain
4
2
0
−2
−4
20
30
40
50
60
70
80


g(theta0|x0=2.3) for decreasing sibling sets:
(1) All (2) d in 40−80 (3) 50−70 (4) 55−65 (5) only d=60
g(theta0|x0=2.3)
0.00
0.02
0.04
0.06
0.08
0.10
−1
0
1
2
3
theta0
1
2
5
3
4
x0=2.3

References

Efron, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* 106: 1602–1614.

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* 103: 1–20.

Good, I. and Toulmin, G. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* 43: 45–63.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* 37: 1647–1684.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*. UC Press, 131–148.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I*. UC Press, 157–163.