Confidence Intervals for Nonparametric Empirical Bayes Analysis

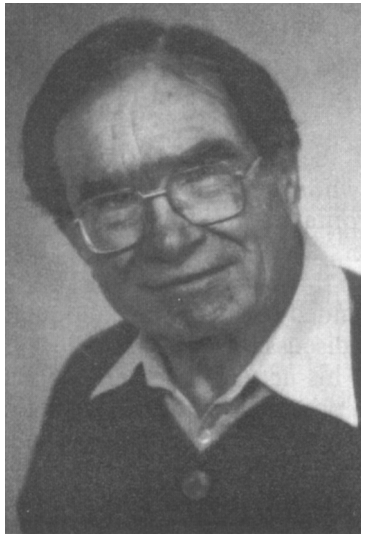
Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Στατιστικής

Οκτώβριος 7, 2022

Nikos Ignatiadis
Statistics Department, Columbia University

What is empirical Bayes?

- Regularization/Shrinkage estimation
- Deconvolution
- Inverse problem
- Hierarchical Bayes
- A Mixed Model with (potentially) some nonparametric components.
- Compound decision theory



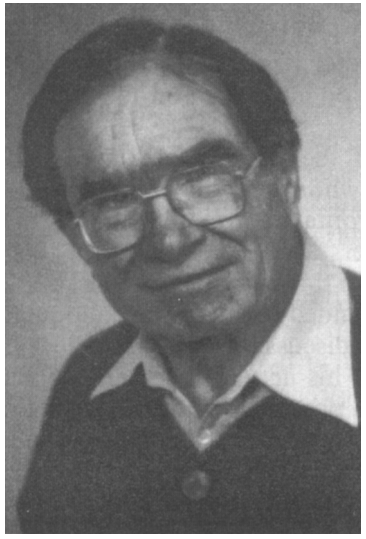
Experience rating (Bichsel, 1964)

Erfahrungs-Tarifierung
in der Motorfahrzeughaftpflicht-Versicherung

Von Fritz Bichsel, Muri bei Bern

$Z_i(1961)$: number of claims of i-th policy holder in 1961

z	$\#\{Z_i(1961) = z\}$
0	103704
1	14075
2	1766
3	255
4	45
≥ 5	8



Experience rating (Bichsel, 1964)

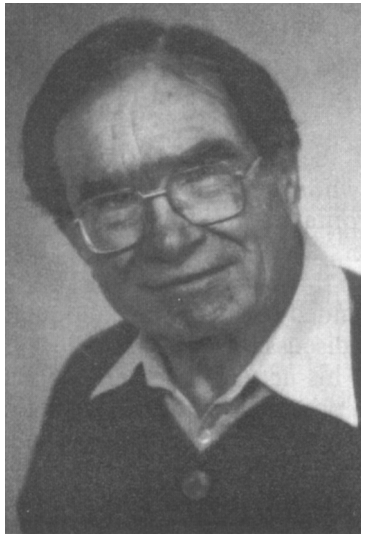
Erfahrungs-Tarifierung
in der Motorfahrzeughaftpflicht-Versicherung

Von Fritz Bichsel, Muri bei Bern

$Z_i(1961)$: number of claims of i-th policy holder in 1961

z	$\#\{Z_i(1961) = z\}$
0	103704
1	14075
2	1766
3	255
4	45
≥ 5	8

Goal: Assign premium for the next year as a function of $Z_i(1961)$.



Experience rating (Bichsel, 1964)

Erfahrungs-Tarifierung
in der Motorfahrzeughaftpflicht-Versicherung

Von Fritz Bichsel, Muri bei Bern

$Z_i(1961)$: number of claims of i-th policy holder in 1961

z	$\#\{Z_i(1961) = z\}$
0	103704
1	14075
2	1766
3	255
4	45
≥ 5	8

Goal: Assign premium for the next year as a function of $Z_i(1961)$.

Idea: Assess expected number of claims in next year,

$$\mathbb{E}[Z_i(1962) \mid Z_i(1961) = z]$$

Experience rating (Bichsel, 1964)

i: policy holder, $Z_i(t)$: number of claims in year t .

$$Z_i(1961), Z_i(1962) \overset{iid}{\sim} \text{Poisson}(\mu_i)$$

Experience rating (Bichsel, 1964)

i: policy holder, $Z_i(t)$: number of claims in year t .

$$Z_i(1961), Z_i(1962) \overset{iid}{\sim} \text{Poisson}(\mu_i)$$

$$\mu_i \sim G$$

"Structural function representing heterogeneity of the portfolio"

Experience rating (Bichsel, 1964)

i: policy holder, $Z_i(t)$: number of claims in year t.

$$Z_i(1961), Z_i(1962) \overset{iid}{\sim} \text{Poisson}(\mu_i)$$

$$\mu_i \sim G$$

"Structural function representing heterogeneity of the portfolio"

Best guess for the number of claims in 1962 for i is:

$$\begin{aligned}
&\mathbb{E}[Z_i(1962) \mid Z_i(1961) = z] \\
&= \mathbb{E}[\mu_i \mid Z_i(1961) = z] \\
&= \theta_G(z)
\end{aligned}$$

Experience rating (Bichsel, 1964)

i: policy holder, $Z_i(t)$: number of claims in year t.

$$Z_i(1961), Z_i(1962) \overset{iid}{\sim} \text{Poisson}(\mu_i)$$

$$\mu_i \sim G$$

"Structural function representing heterogeneity of the portfolio"

Best guess for the number of claims in 1962 for i is:

$$\begin{aligned}
&\mathbb{E}[Z_i(1962) \mid Z_i(1961) = z] \\
&= \mathbb{E}[\mu_i \mid Z_i(1961) = z] \\
&= \theta_G(z)
\end{aligned}$$

Bichsel did not know G… But had data.

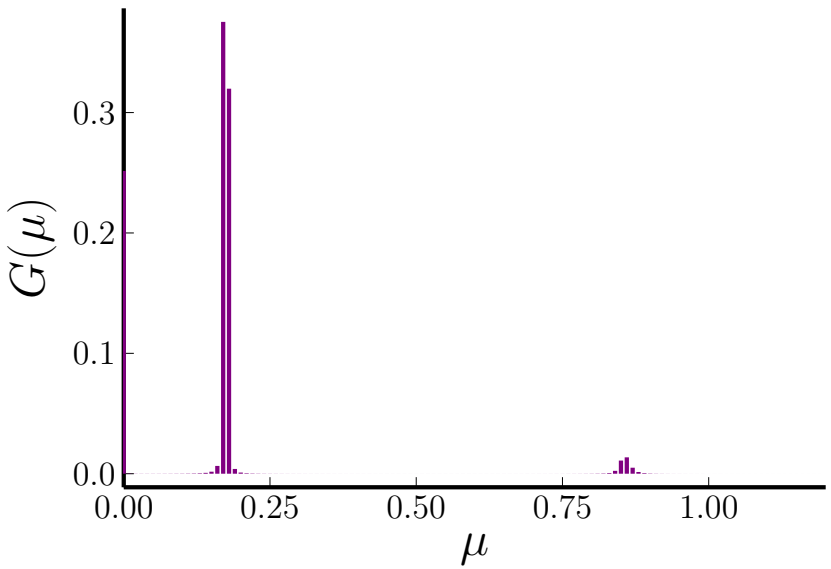
Nonparametric Maximum Likelihood (NPMLE)

i: policy holder

$Z_i = Z_i(1961)$: number of claims in 1961

$$Z_i \sim \text{Poisson}(\mu_i)$$
$$\mu_i \sim G$$

$$G \in \mathcal{G} = \{\text{all distributions supported on } [0, 5]\}$$



z	$\#\{Z_i = z\}$	$\widehat{\mathbb{E}}[\mu_i \mid Z_i = z]$
0	103704	0.14
1	14075	0.25
2	1766	0.44
3	255	0.69
4	45	0.82
≥ 5	8	

Kiefer-Wolfowitz (1956), Simar (1976), Laird (1978), Lindsay (1983),
Walhin and Paris (1999), Koenker and Mizera (2014), Koenker and Gu (2017)

What about confidence intervals?

Estimated posterior mean: $\widehat{\theta}_G(3) = \mathbb{E}_{\widehat{G}}[\mu \mid Z = 3] = 0.69$

What about confidence intervals?

Estimated posterior mean: $\widehat{\theta}_G(3) = \mathbb{E}_{\widehat{G}}[\mu \mid Z = 3] = 0.69$

SCANDINAVIAN ACTUARIAL JOURNAL, 2018
VOL. 2018, NO. 2, 129–144
https://doi.org/10.1080/03461238.2017.1307267

Taylor & Francis
Taylor & Francis Group

Check for updates

Confidence intervals of the premiums of optimal bonus malus systems

Dimitris Karlis[a], George Tzougas[b] and Nicholas Frangos[a]

[a]Department of Statistics, Athens University of Economics and Business, Athens, Greece; [b]Department of Statistics, London School of Economics, London, UK

$$\theta_G(3) = \mathbb{E}_G[\mu \mid Z = 3] \in [0.55,\ 0.80]\,.$$

Confidence intervals for $\theta_G(z) = \mathbb{E}_G[\mu \mid Z = z]$

1 Let $\widehat{G}$ be the NPMLE of G based on $Z_1, \ldots, Z_n$.
2 **for** $b = 1$ **to** B **do**
3 | Draw $\mu_i^b \sim \widehat{G}$, $Z_i^b \sim p(\cdot \mid \mu_i^b)$ for $i = 1, \ldots, n$ (iid).
4 | Let $\widehat{G}^b$ be the NPMLE of G based on $Z_1^b, \ldots, Z_n^b$.
5 | Let $\hat{\theta}^b(z) = \theta_{\widehat{G}^b}(z)$.
6 **end**
7 Form a percentile bootstrap confidence interval $[\widehat{\theta}_\alpha^-(z), \widehat{\theta}_\alpha^+(z)]$ of $\theta_G(z)$ based on $\hat{\theta}^b(z)$, $b = 1, \ldots, B$.

Confidence intervals for $\theta_G(z) = \mathbb{E}_G[\mu \mid Z = z]$

1 Let $\widehat{G}$ be the NPMLE of G based on $Z_1, \ldots, Z_n$.
2 **for** $b = 1$ **to** B **do**
3 | Draw $\mu_i^b \sim \widehat{G}$, $Z_i^b \sim p(\cdot \mid \mu_i^b)$ for $i = 1, \ldots, n$ (iid).
4 | Let $\widehat{G}^b$ be the NPMLE of G based on $Z_1^b, \ldots, Z_n^b$.
5 | Let $\hat{\theta}^b(z) = \theta_{\widehat{G}^b}(z)$.
6 **end**
7 Form a percentile bootstrap confidence interval $[\widehat{\theta}_\alpha^-(z), \widehat{\theta}_\alpha^+(z)]$ of $\theta_G(z)$ based on $\hat{\theta}^b(z)$, $b = 1, \ldots, B$.

Theorem [Karlis, Tzougas, and Frangos (2018)]: Assume independence and that $Z_i \mid \mu_i \sim \text{Poisson}(\mu_i)$, $\mu_i \sim G$, with

$$G \in \mathcal{G} = \left\{ G \text{ supported on compact interval } [0,M], \; \mathbb{P}_G[\mu \in (0,\varepsilon)] \text{ is sufficiently large for all } \varepsilon > 0 \right\}$$

Then:

$$\liminf_{n\to\infty} \left\{ \mathbb{P}_G \left[\theta_G(z) \in [\widehat{\theta}_\alpha^-(z), \widehat{\theta}_\alpha^+(z)] \right] \right\} \geq 1 - \alpha \,.$$

Confidence Intervals for Nonparametric Empirical Bayes Analysis (with Rejoinder)

N.I., and Stefan Wager, JASA T&M (2022)

Empirical Bayes (EB)

Robbins (1956), Efron (2010)

We have noisy data Z_i on n **related units** with latent parameter μ_i.
Three main ingredients to an EB analysis:

1) Known noise model: $Z_i \mid \mu_i \sim p(\cdot \mid \mu_i)$

$p(\cdot \mid \mu_i)$ is a density w.r.t. a measure λ, e.g., $Z_i \mid \mu_i \sim \text{Poisson}(\mu_i)$.

2) Class of priors: $\mu_i \sim G, \quad G \in \mathcal{G}$

G is unknown.

3) Empirical Bayes estimand: $\theta_G(z) = \mathbb{E}_G[h(\mu_i) \mid Z_i = z]$

for a known function $h(\cdot)$,
e.g., for $h(\mu) = \mu$, $\theta_G(z)$ is the posterior mean given $Z_i = z$

Empirical Bayes (EB)

Robbins (1956), Efron (2010)

We have noisy data Z_i on n **related units** with latent parameter μ_i.
Three main ingredients to an EB analysis:

1) Known noise model: $Z_i \mid \mu_i \sim p(\cdot \mid \mu_i)$

$p(\cdot \mid \mu_i)$ is a density w.r.t. a measure λ, e.g., $Z_i \mid \mu_i \sim \text{Poisson}(\mu_i)$.

2) Class of priors: $\mu_i \sim G, \quad G \in \mathcal{G}$

G is unknown.

3) Empirical Bayes estimand: $\theta_G(z) = \mathbb{E}_G[h(\mu_i) \mid Z_i = z]$

for a known function $h(\cdot)$,
e.g., for $h(\mu) = \mu$, $\theta_G(z)$ is the posterior mean given $Z_i = z$

Predominant approach: Point estimate $\widehat{\theta_G}(z)$ of $\theta_G(z)$.

EB confidence intervals: statistical task

1) Known noise model: $Z_i \mid \mu_i \sim p(\cdot \mid \mu_i)$

2) Class of priors: $\mu_i \sim G, \quad G \in \mathcal{G}$ $\qquad i = 1,\ldots,n$

G is unknown

3) Empirical Bayes estimand: $\theta_G(z) = \mathbb{E}_G[h(\mu_i) \mid Z_i = z]$

CI: $[\widehat{\theta}^-_\alpha(z), \widehat{\theta}^+_\alpha(z)]$ with pointwise frequentist coverage:

$$\liminf_{n\to\infty} \left\{ \mathbb{P}_G \left[\theta_G(z) \in [\widehat{\theta}^-_\alpha(z), \widehat{\theta}^+_\alpha(z)] \right] \right\} \geq 1 - \alpha$$

and also with simultaneous coverage.

EB confidence intervals: statistical task

1) Known noise model: $Z_i \mid \mu_i \sim p(\cdot \mid \mu_i)$

2) Class of priors: $\mu_i \sim G, \quad G \in \mathcal{G}$ $\qquad i = 1,\ldots,n$

G is unknown $\mathcal{G}$ **is a pre-specified convex class of priors.**

3) Empirical Bayes estimand: $\theta_G(z) = \mathbb{E}_G[h(\mu_i) \mid Z_i = z]$

CI: $[\widehat{\theta}^-_\alpha(z), \widehat{\theta}^+_\alpha(z)]$ with pointwise frequentist coverage:

$$\liminf_{n\to\infty} \left\{ \mathbb{P}_G \left[\theta_G(z) \in [\widehat{\theta}^-_\alpha(z), \widehat{\theta}^+_\alpha(z)] \right] \right\} \geq 1-\alpha \quad \textbf{for all } G \in \mathcal{G}$$

and also with simultaneous coverage.

RESEARCH ARTICLE

Statistics
in Medicine

The statistical properties of RCTs and a proposal for shrinkage

Erik van Zwet[1] | **Simon Schwab**[2,3] | **Stephen Senn**[4]

Z-Score from an RCT: $$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

RESEARCH ARTICLE

Statistics in Medicine

The statistical properties of RCTs and a proposal for shrinkage

Erik van Zwet[1] | **Simon Schwab**[2,3] | **Stephen Senn**[4]

Z-Score from an RCT: $Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$

For inference: $\mu_i \sim G$

RESEARCH ARTICLE

Statistics in Medicine

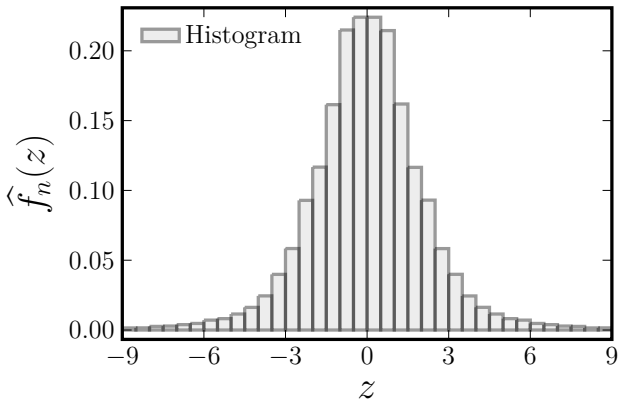
The statistical properties of RCTs and a proposal for shrinkage

Erik van Zwet[1] | **Simon Schwab**[2,3] | **Stephen Senn**[4]

Z-Score from an RCT: $Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$

For inference: $\mu_i \sim G$

$n = 23551$ Z-scores from RCTs in healthcare (Cochrane Database of Systematic Reviews).



RESEARCH ARTICLE

Statistics in Medicine

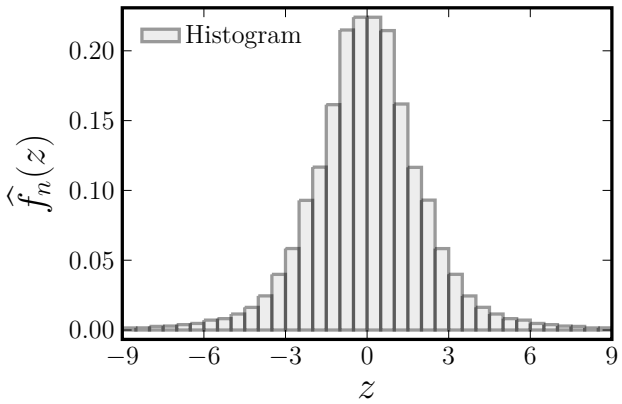
The statistical properties of RCTs and a proposal for shrinkage

Erik van Zwet[1] | **Simon Schwab**[2,3] | **Stephen Senn**[4]

Z-Score from an RCT: $Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$

For inference: $\mu_i \sim G$

$n = 23551$ Z-scores from RCTs in healthcare (Cochrane Database of Systematic Reviews).



For an RCT with Z-score $Z_i = -1.81$

Posterior mean:

$\mathbb{E}_{\widehat{G}}[\mu \mid Z = -1.81] = -1.12$

Local false sign rate:

$\mathbb{P}_{\widehat{G}}[\mu \geq 0 \mid Z = -1.81] = 0.091$

RESEARCH ARTICLE

Statistics in Medicine

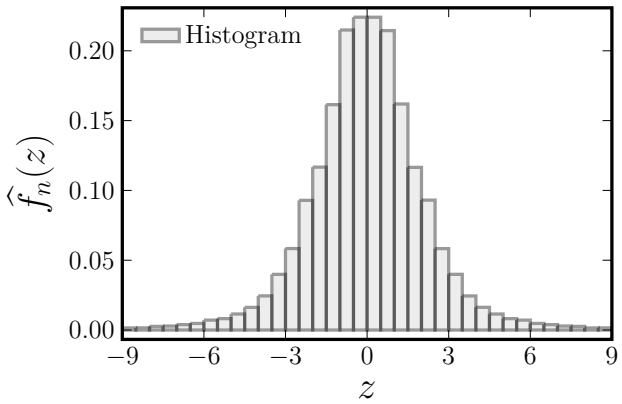
The statistical properties of RCTs and a proposal for shrinkage

Erik van Zwet[1] | **Simon Schwab**[2,3] | **Stephen Senn**[4]

Z-Score from an RCT: $Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$

For inference: $\mu_i \sim G$

$n = 23551$ Z-scores from RCTs in healthcare (Cochrane Database of Systematic Reviews).



For an RCT with Z-score $Z_i = -1.81$

Posterior mean:

$\mathbb{E}_{\widehat{G}}[\mu \mid Z = -1.81] = -1.12$

Local false sign rate:

$\mathbb{P}_{\widehat{G}}[\mu \geq 0 \mid Z = -1.81] = 0.091$

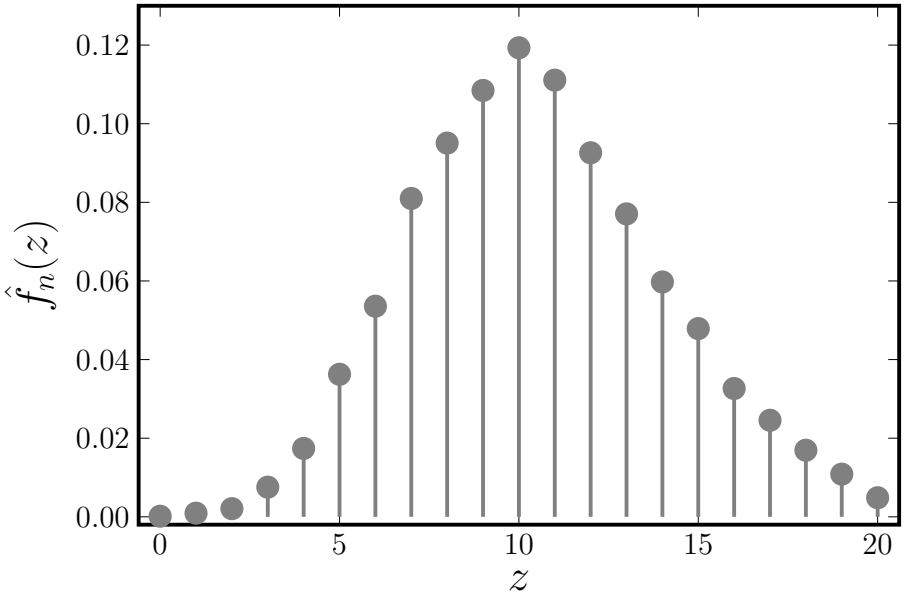
What about confidence intervals?

Educational Testing Service

Lord (1969), Lord and Cressie (1975), Lord and Stocking (1976)

$n = 12990$ students

Z_i: Score on multiple choice test with 20 questions (5 choices per question)

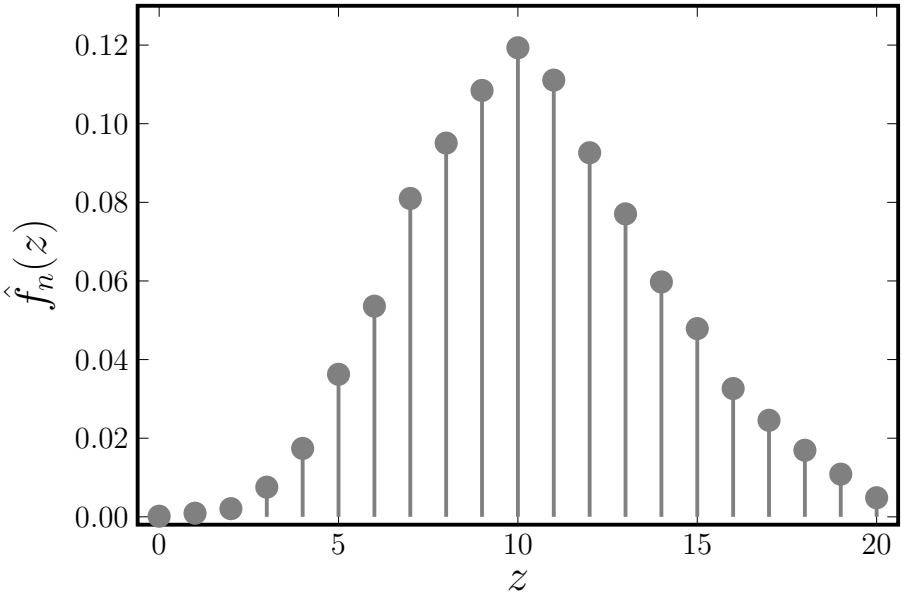


Educational Testing Service

Lord (1969), Lord and Cressie (1975), Lord and Stocking (1976)

$n = 12990$ students

Z_i: Score on multiple choice test with 20 questions (5 choices per question)



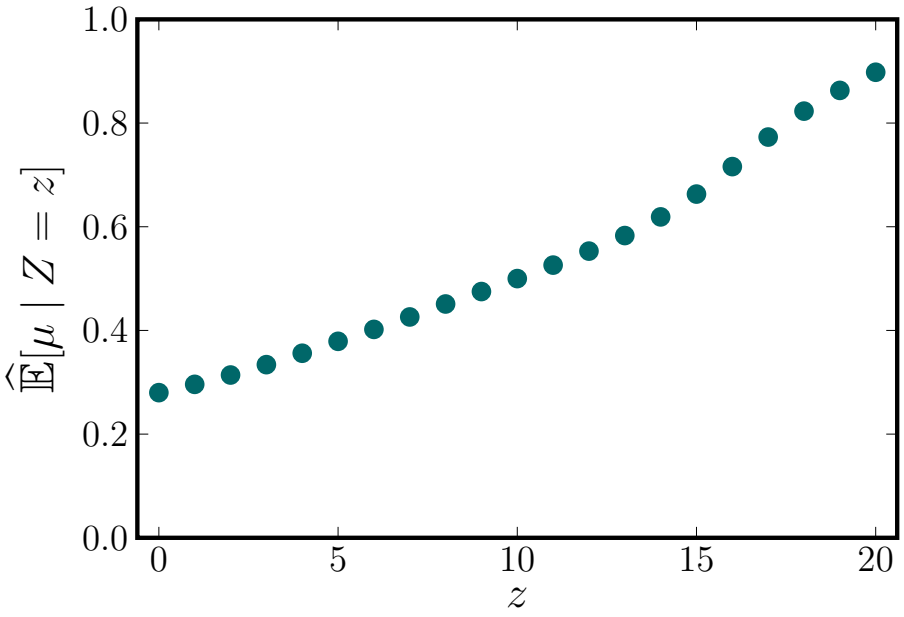
Empirical Bayes model:

$$\mu_i \sim G$$

$$Z_i \mid \mu_i \sim \text{Binomial}(20, \mu_i)$$

Posterior mean:

$$\theta_G(z) = \mathbb{E}_G[\mu_i \mid Z_i = z]$$

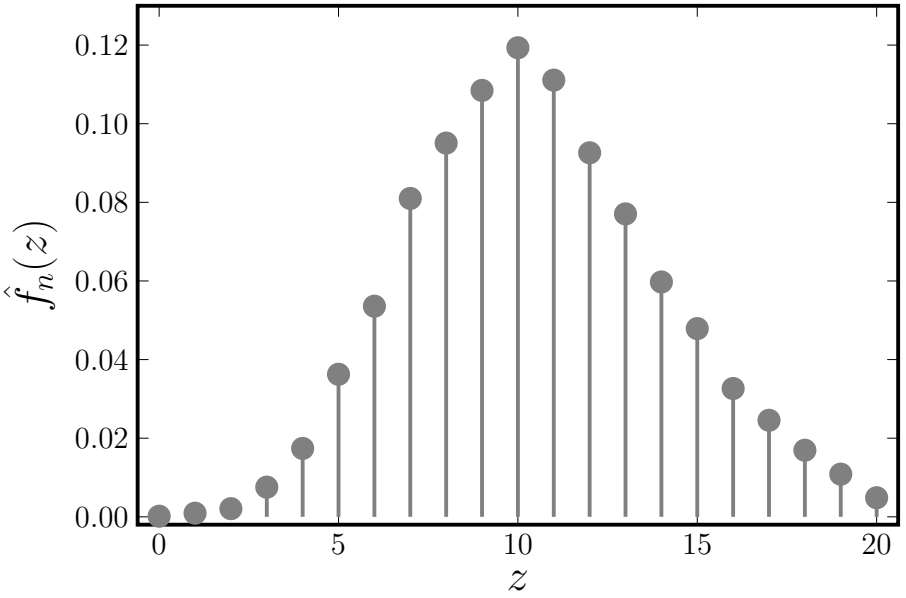


Educational Testing Service

Lord (1969), Lord and Cressie (1975), Lord and Stocking (1976)

$n = 12990$ students

Z_i: Score on multiple choice test with 20 questions (5 choices per question)



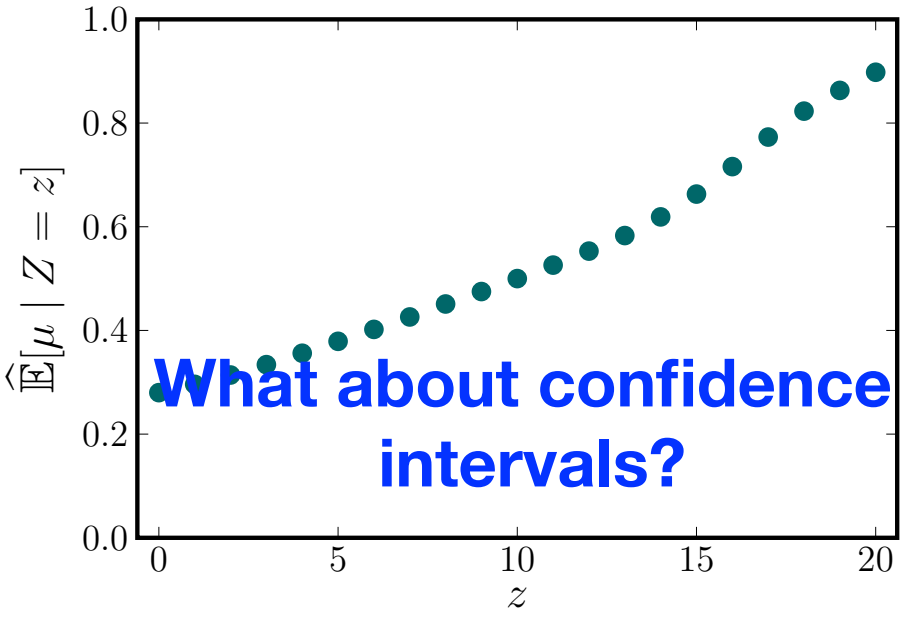
Empirical Bayes model:

$$\mu_i \sim G$$

$$Z_i \mid \mu_i \sim \text{Binomial}(20, \mu_i)$$

Posterior mean:

$$\theta_G(z) = \mathbb{E}_G[\mu_i \mid Z_i = z]$$



Related work: EB confidence intervals

Efron (2014, 2016, 2019), Narasimhan and Efron (2020)
Use a flexible parametric model for G, and estimate variance of point estimates $\widehat{\theta}_G(z)$.

Related work: EB confidence intervals

Efron (2014, 2016, 2019), Narasimhan and Efron (2020)
Use a flexible parametric model for G, and estimate variance of point estimates $\widehat{\theta}_G(z)$.

Nonparametrics:

Poisson model, Posterior Mean
Robbins (1980, *Proceedings of the National Academy of Sciences*)
Karlis, Tzougas and Frangos (2018, *Scandinavian Actuarial Journal*)

Binomial model, Posterior Mean
Lord and Cressie (1975, *Sankhyā Series B*)
Lord and Stocking (1976, *Psychometrika*)

Our work
A unified inference framework that works in the general nonparametric situation described.

Nonparametric EB estimation

Poisson model, Posterior Mean: ✓✓

It is possible to estimate $\theta_G(z)$ at the parametric rate $1/\sqrt{n}$

[Robbins (1956), Lambert and Tierney (1984)]

Gaussian model, Posterior Mean: ✓

It is possible to estimate $\theta_G(z)$ at the quasi-parametric rate $\log(n)^{3/4}/\sqrt{n}$ [Matias and Taupin (2004)]

Gaussian model, Local False Sign Rate: ✗

For Sobolev G, minimax rates for estimating $\theta_G(z)$ are polynomial in $1/\log(n)$ [Butucea and Comte (2009), Pensky (2017)]

Binomial model, Posterior Mean: ✗✗

If we impose no restrictions on G, $\theta_G(z)$ is only partially identified [Robbins (1956)]

Why does the difficulty of estimation vary?

Hierarchical model: $\mu_i \sim G, \quad Z_i \mid \mu_i \sim p(\cdot \mid \mu_i)$

Marginally: $Z_i \sim f_G, \quad f_G(z) = \int p(z \mid \mu) dG(\mu)$

Why does the difficulty of estimation vary?

Hierarchical model: $\mu_i \sim G, \quad Z_i \mid \mu_i \sim p(\cdot \mid \mu_i)$

Marginally: $Z_i \sim f_G, \quad f_G(z) = \int p(z \mid \mu) dG(\mu)$

$$f_{G_1} \approx f_{G_2} \not\Rightarrow G_1 \approx G_2$$

Why does the difficulty of estimation vary?

Hierarchical model: $\mu_i \sim G, \quad Z_i \mid \mu_i \sim p(\cdot \mid \mu_i)$

Marginally: $Z_i \sim f_G, \quad f_G(z) = \int p(z \mid \mu) dG(\mu)$

$$f_{G_1} \approx f_{G_2} \not\Rightarrow G_1 \approx G_2$$

$$f_{G_1} \approx f_{G_2} \overset{?}{\Rightarrow} \theta_{G_1}(z) \approx \theta_{G_2}(z)$$

Why is a unified inference approach difficult?

1) Distributional theory for e.g., NPMLE is notoriously difficult. Some results for discrete distributions by Lambert and Tierney (1984), Böhning and Patilea (2005), Karlis and Patilea (2008).

Why is a unified inference approach difficult?

1) Distributional theory for e.g., NPMLE is notoriously difficult. Some results for discrete distributions by Lambert and Tierney (1984), Böhning and Patilea (2005), Karlis and Patilea (2008).

2) Suppose we had asymptotic normality: $\widehat{\theta_G}(z) \pm 1.96 \cdot \widehat{\mathsf{se}}$

Why is a unified inference approach difficult?

1) Distributional theory for e.g., NPMLE is notoriously difficult. Some results for discrete distributions by Lambert and Tierney (1984), Böhning and Patilea (2005), Karlis and Patilea (2008).

2) Suppose we had asymptotic normality: $\widehat{\theta}_G(z) \pm 1.96 \cdot \widehat{\text{se}}$

Assumes that: $\text{Bias}^2 \ll \widehat{\text{se}}^2$

Not true under partial identification or
when minimax estimation rates are slow!

"when the minimax convergence rate is slower than any algebraic rate, the optimal linear estimator must have maximum squared bias completely dominating the variance" [Cai and Low (2003)]

Instead: **Bias-aware inference**

[Armstrong and Kolesár (2018), Imbens and Manski (2004), Imbens and Wager (2018)]

F-Localization

F-Localization

$\mathscr{F}_n(\alpha)$: a set of distributions, such that

$$\liminf_{n\to\infty} \left\{ \mathbb{P}_G[F_G \in \mathscr{F}_n(\alpha)] - (1-\alpha) \right\} \geq 0.$$

F-Localization

$\mathscr{F}_n(\alpha)$: a set of distributions, such that

$$\liminf_{n\to\infty} \left\{ \mathbb{P}_G[F_G \in \mathscr{F}_n(\alpha)] - (1-\alpha) \right\} \geq 0.$$

Set of all priors consistent with the F-Localization

$$\mathscr{G}(\mathscr{F}_n(\alpha)) = \{G \in \mathscr{G} : F_G \in \mathscr{F}_n(\alpha)\}$$

Confidence intervals for empirical Bayes estimand

$$\hat{\theta}^+_\alpha(z) = \sup\{\theta_G(z) : G \in \mathscr{G}(\mathscr{F}_n(\alpha))\},\ \mathscr{I}_\alpha(z) = \left[\hat{\theta}^-_\alpha(z), \hat{\theta}^+_\alpha(z)\right]$$

F-Localization

$\mathscr{F}_n(\alpha)$: a set of distributions, such that

$$\liminf_{n\to\infty} \left\{ \mathbb{P}_G[F_G \in \mathscr{F}_n(\alpha)] - (1-\alpha) \right\} \geq 0.$$

Set of all priors consistent with the F-Localization

$$\mathscr{G}(\mathscr{F}_n(\alpha)) = \{G \in \mathscr{G} : F_G \in \mathscr{F}_n(\alpha)\}$$

Confidence intervals for empirical Bayes estimand

$$\hat{\theta}^+_\alpha(z) = \sup\{\theta_G(z) : G \in \mathscr{G}(\mathscr{F}_n(\alpha))\},\ \mathscr{I}_\alpha(z) = \left[\hat{\theta}^-_\alpha(z), \hat{\theta}^+_\alpha(z)\right]$$

Proposition (I., Wager)**:** $\liminf_{n\to\infty} \left\{ \mathbb{P}_G \left[\theta_G(z) \in [\hat{\theta}^-_\alpha(z), \hat{\theta}^+_\alpha(z)]\right] \right\} \geq 1-\alpha$

Robbins (1956), Anderson (1964), Deely and Kruse (1968), Romano and Wolf (2000), Stark (1992), Donoho and Reeves (2013), Kuusela and Stark (2017), Greenshtein and Itskov (2018), Brennan et al. (2020), ….

Dvoretzky–Kiefer–Wolfowitz

$$\mathscr{F}_n(\alpha) = \left\{ F : \sup_{t \in \mathbb{R}} |F(t) - \widehat{F_n}(t)| \leq \sqrt{\log(2/\alpha)/(2n)} \right\}$$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Z_i \leq t)$$

Theorem [Massart (1990)]: $\mathbb{P}_G[F_G \in \mathscr{F}_n(\alpha)] \geq 1 - \alpha$

Dvoretzky–Kiefer–Wolfowitz

$$\mathcal{F}_n(\alpha) = \left\{ F : \sup_{t \in \mathbb{R}} |F(t) - \widehat{F_n}(t)| \leq \sqrt{\log(2/\alpha)/(2n)} \right\}$$

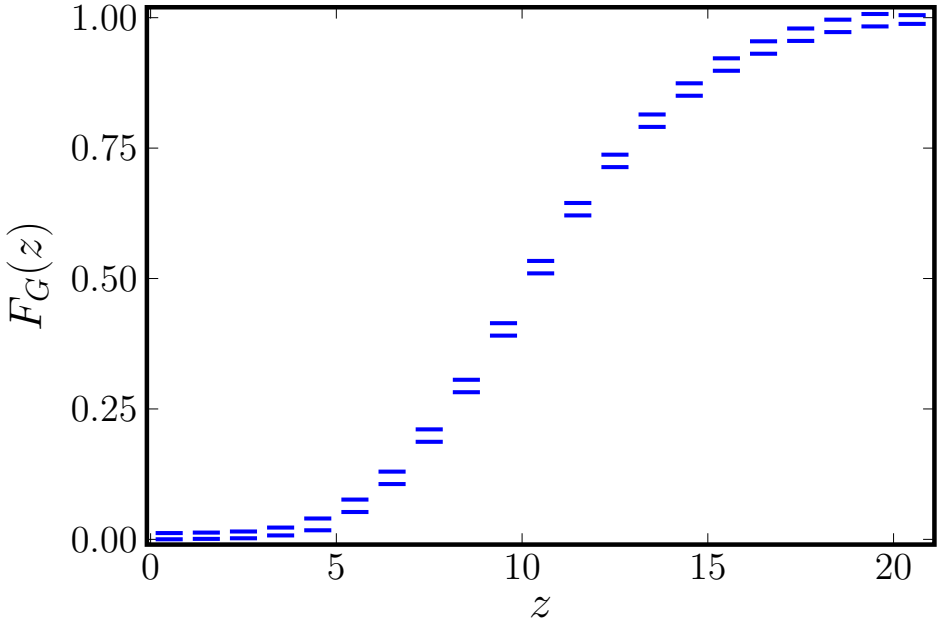
$$\widehat{F_n}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Z_i \leq t)$$

Theorem [Massart (1990)]: $\mathbb{P}_G[F_G \in \mathcal{F}_n(\alpha)] \geq 1 - \alpha$

$n = 12990$ students

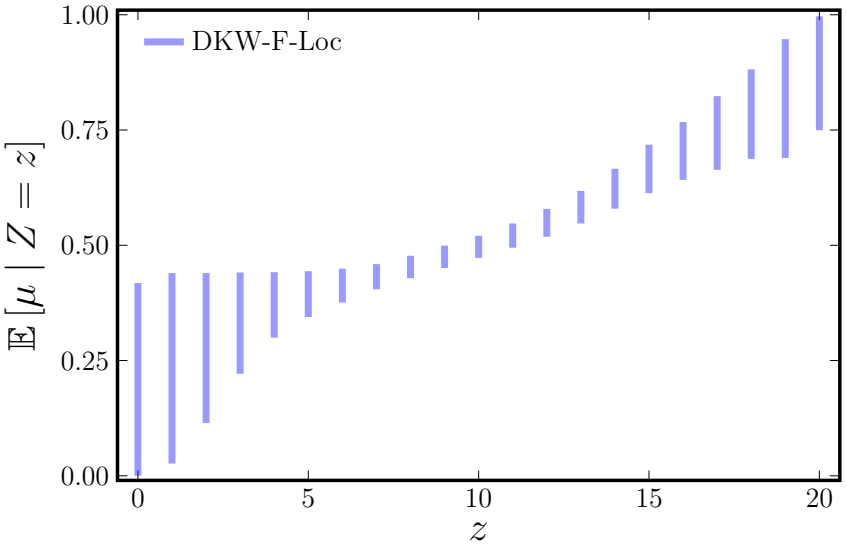
Z_i: Score on multiple choice test with 20 questions (5 choices per question)

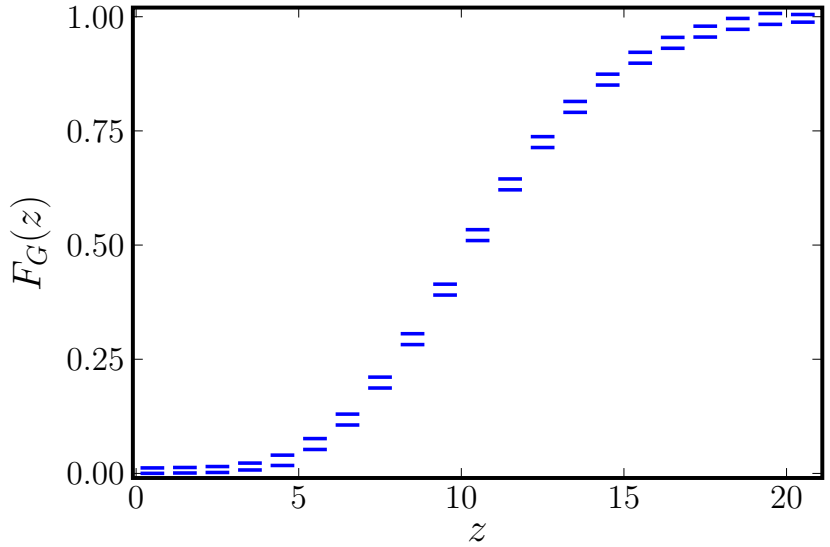
$Z_i \mid \mu_i \sim \text{Binomial}(20, \mu_i)$



Example (ETS)

$\theta_G(z) = \mathbb{E}_G[\mu_i \mid Z_i = z]$, $\mathcal{G} = \{$all distributions supported on $[0,1]\}$

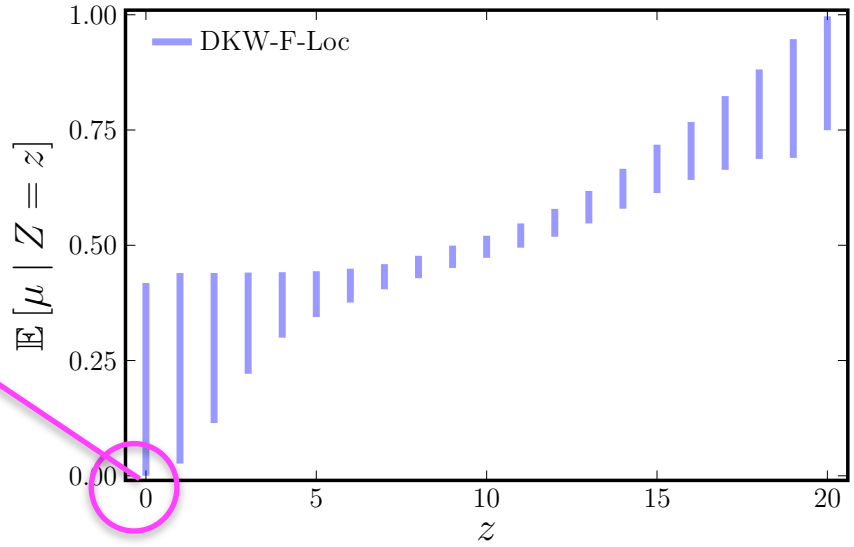


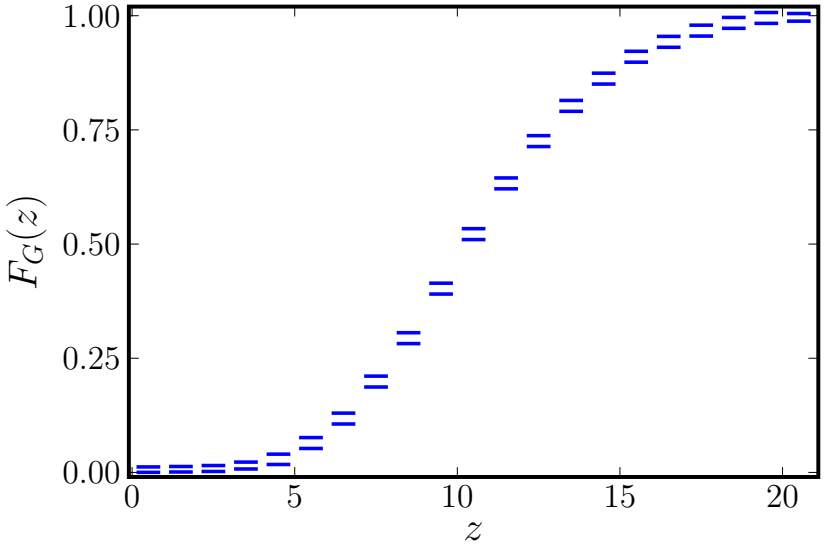


Example (ETS)

$\theta_G(z) = \mathbb{E}_G[\mu_i \mid Z_i = z]$, $\mathcal{G} = \{$all distributions supported on $[0,1]\}$

$$\widehat{\theta}^{-}_{\alpha}(0) = 2 \cdot 10^{-4}$$





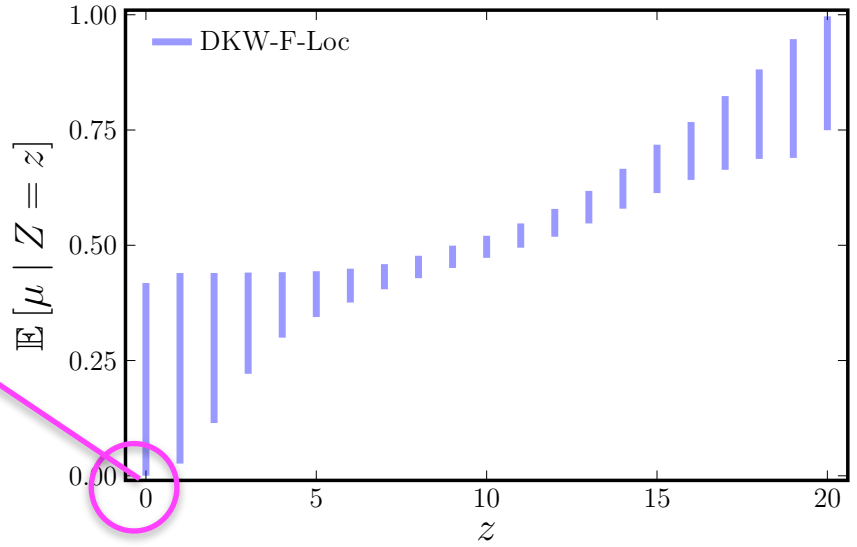
Example (ETS)

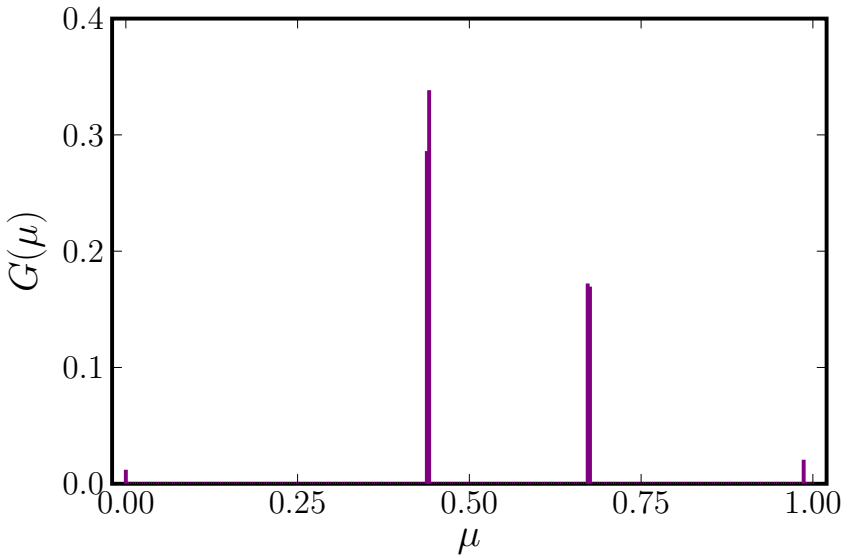
$\theta_G(z) = \mathbb{E}_G[\mu_i \mid Z_i = z]$, $\mathcal{G} = \{$all distributions supported on [0,1]$\}$

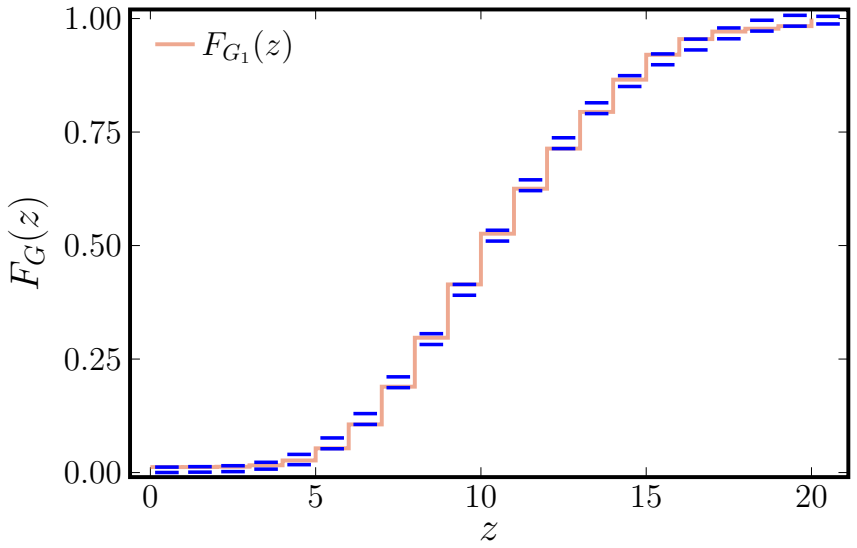
$$\widehat{\theta}_{\alpha}^{-}(0) = 2 \cdot 10^{-4}$$

Linear programming!

[Charnes-Cooper (1962)]





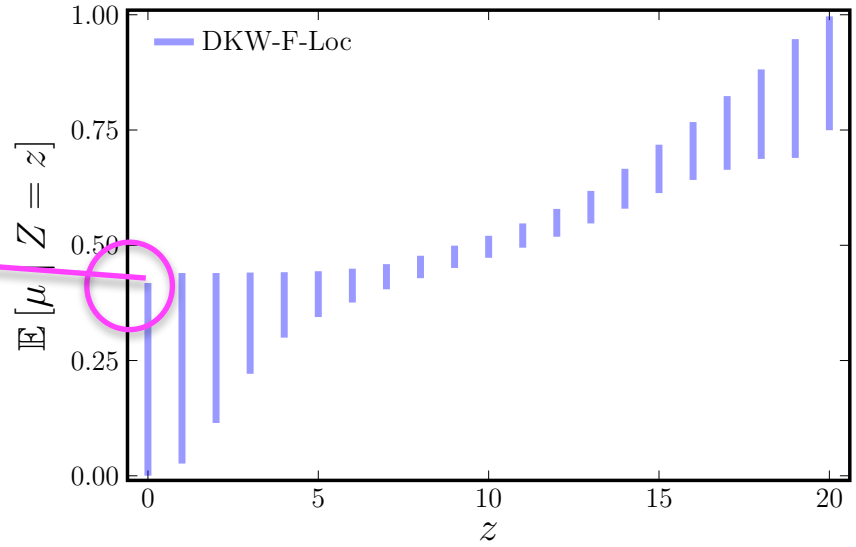


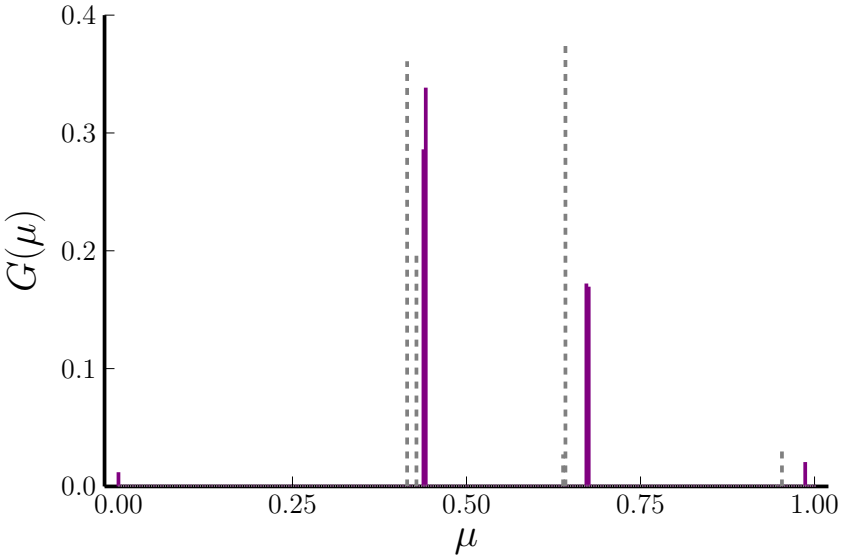
Example (ETS)

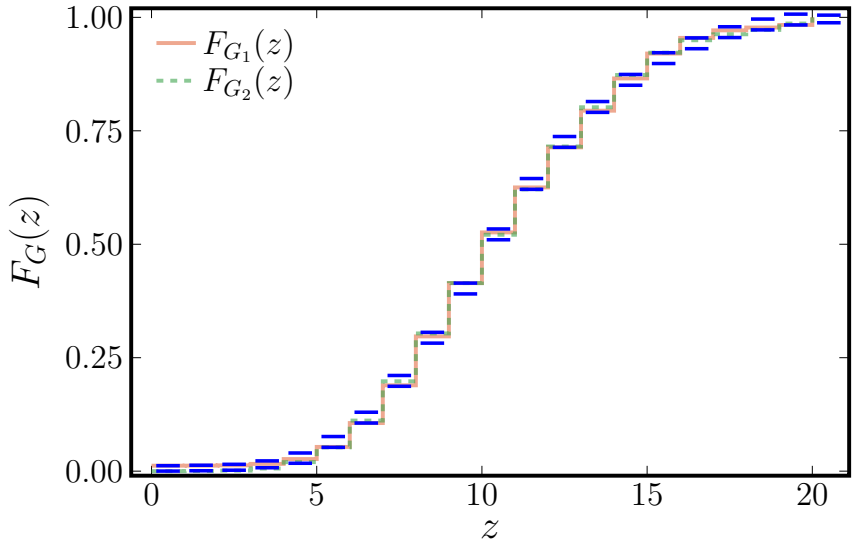
$\theta_G(z) = \mathbb{E}_G[\mu_i \mid Z_i = z]$, $\mathcal{G} = \{$all distributions supported on [0,1]$\}$

$$\widehat{\theta}^{-}_{\alpha}(0) = 2 \cdot 10^{-4}$$

$$\widehat{\theta}^{+}_{\alpha}(0) = 0.42$$





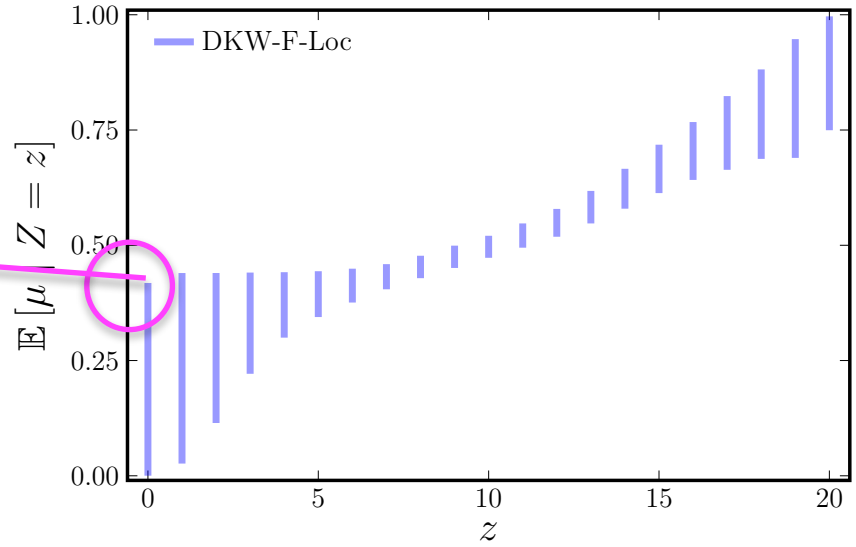


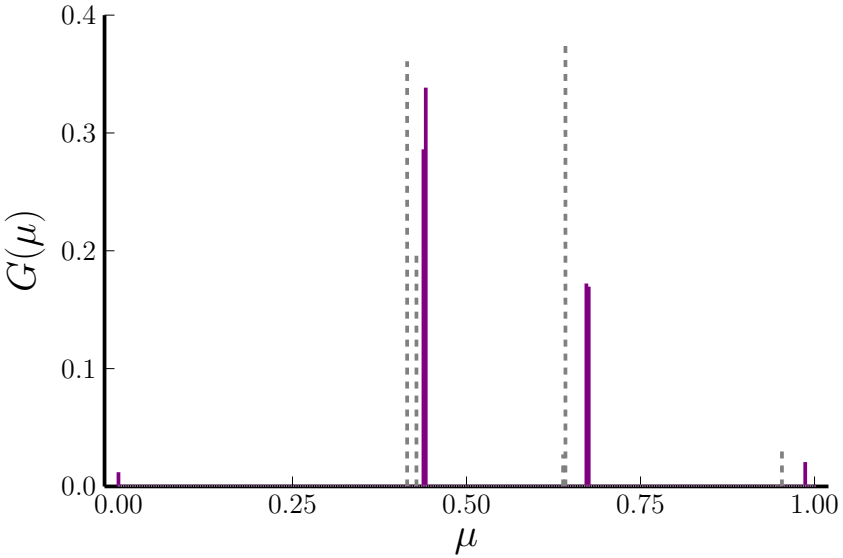
Example (ETS)

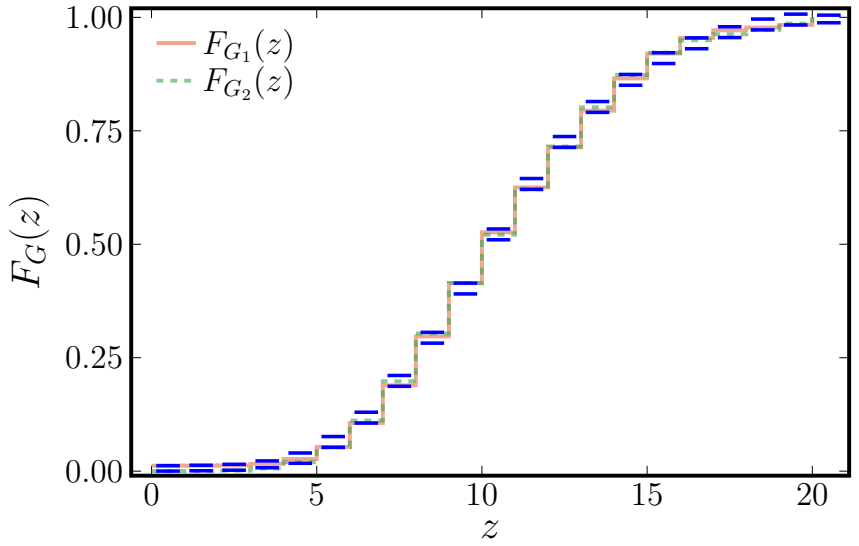
$\theta_G(z) = \mathbb{E}_G[\mu_i \mid Z_i = z]$, $\mathcal{G} = \{$all distributions supported on [0,1]$\}$

$$\widehat{\theta}^{-}_{\alpha}(0) = 2 \cdot 10^{-4}$$

$$\widehat{\theta}^{+}_{\alpha}(0) = 0.42$$





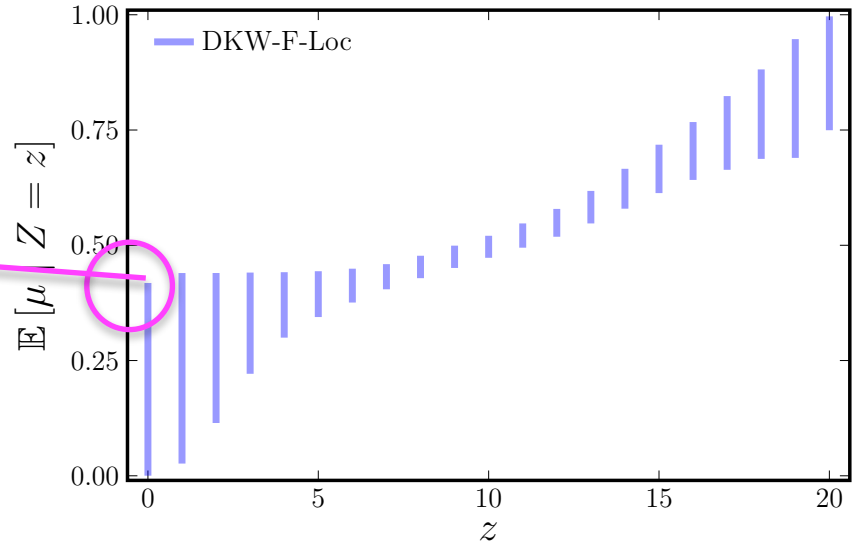


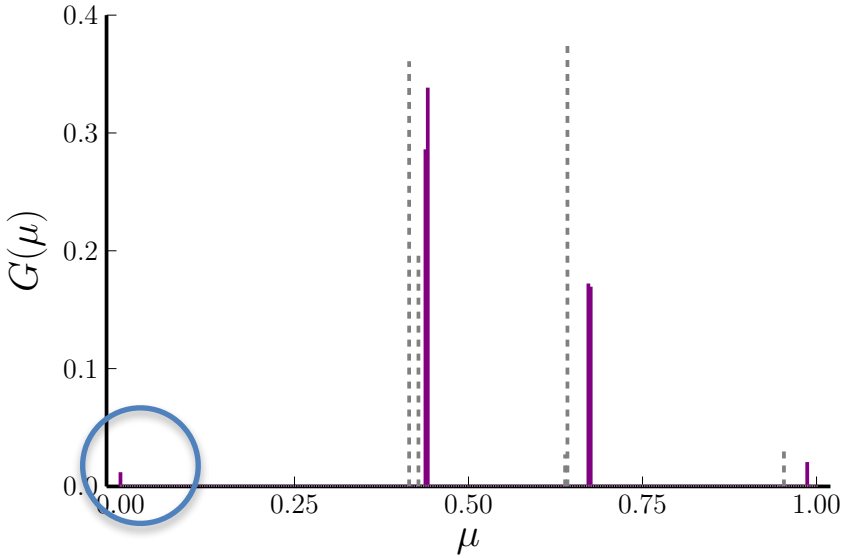
Example (ETS)

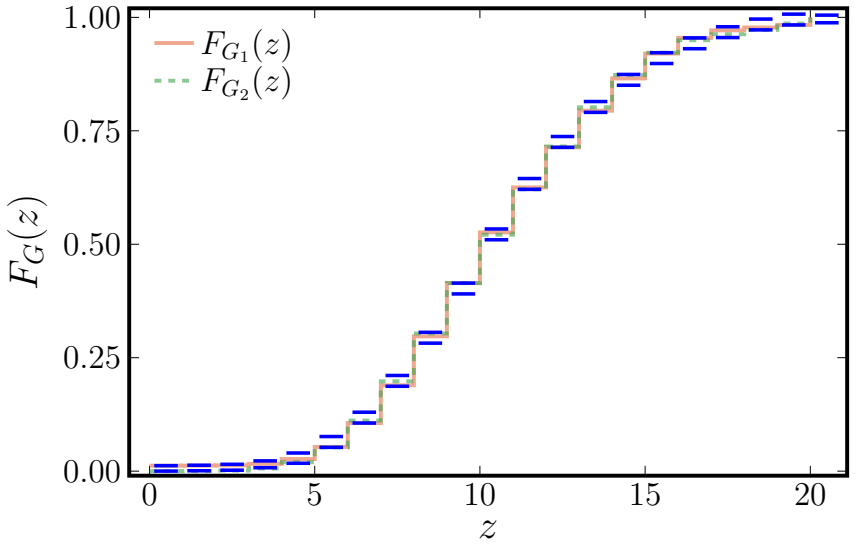
$\theta_G(z) = \mathbb{E}_G[\mu_i \mid Z_i = z]$, $\mathcal{G} = \{$all distributions supported on [0,1]$\}$

$$\widehat{\theta}^{-}_{\alpha}(0) = 2 \cdot 10^{-4}$$

$$\widehat{\theta}^{+}_{\alpha}(0) = 0.42$$







Choice of F-localization

DKW F-Localization: Universal and finite-sample

χ^2-F-Localization: When $Z_i \mid \mu_i \sim \text{Binomial}(N, \mu_i)$,

$$\mathscr{F}_n(\alpha) = \left\{ F \text{ with pmf } f \text{ on } \{0,\ldots,N\} \ : \ \sum_{z=0}^{N} \frac{(n\hat{f}_n(z) - nf(z))^2}{nf(z)} \le \chi^2_{N,1-\alpha} \right\}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Z_i = z)$$

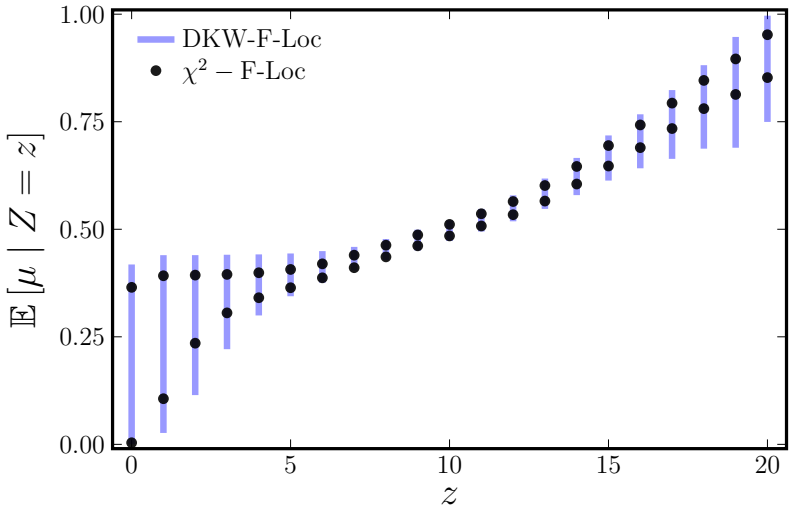
Choice of F-localization

DKW F-Localization: Universal and finite-sample

χ^2-F-Localization: When $Z_i \mid \mu_i \sim \text{Binomial}(N, \mu_i)$,

$$\mathscr{F}_n(\alpha) = \left\{ F \text{ with pmf } f \text{ on } \{0,\ldots,N\} \;:\; \sum_{z=0}^{N} \frac{(n\hat{f}_n(z) - nf(z))^2}{nf(z)} \leq \chi^2_{N,1-\alpha} \right\}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Z_i = z)$$



Choice of F-localization

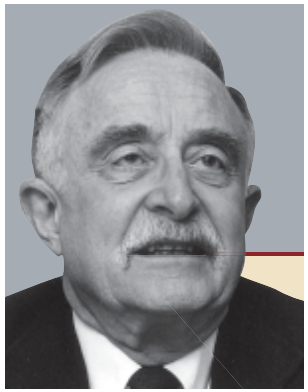
DKW F-Localization: Universal and finite-sample

χ^2-F-Localization: When $Z_i \mid \mu_i \sim \text{Binomial}(N, \mu_i)$,

$$\mathscr{F}_n(\alpha) = \left\{ F \text{ with pmf } f \text{ on } \{0,\ldots,N\} \;:\; \sum_{z=0}^{N} \frac{(n\hat{f}_n(z) - nf(z))^2}{nf(z)} \leq \chi^2_{N,1-\alpha} \right\}$$

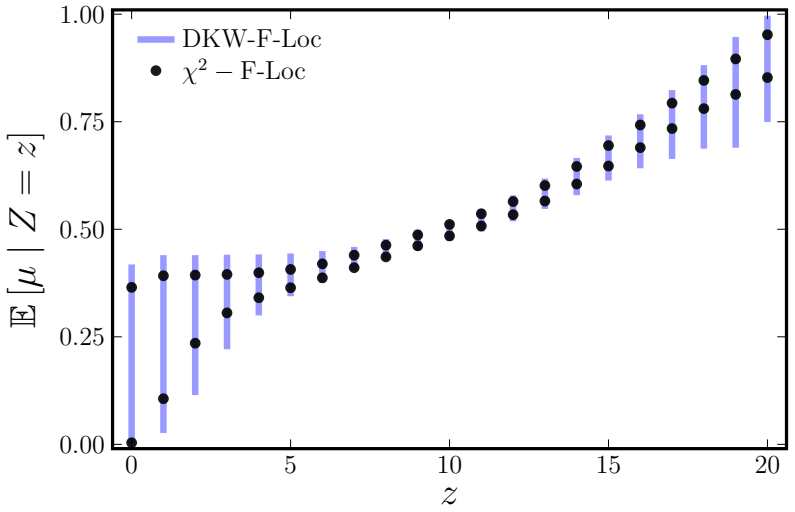
$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Z_i = z)$$







Lord and Cressie (1975), Lord and Stocking (1976)



Beyond F-localization



Easy to construct

“Common sense approach”

Simultaneous coverage



What about power? Power strong only if F-localization gives tight characterization of uncertainty. Not true in general, because of simultaneous coverage.

Conservative!

Choice of F-Localization matters.

AMARI

(Affine Minimax Anderson Rubin Inference)

AMARI

(Affine Minimax Anderson Rubin Inference)

$$H_0 : \theta_G(z) = c$$

AMARI

(Affine Minimax Anderson Rubin Inference)

$$H_0 : \theta_G(z) = c$$

$$\theta_G(z) = \mathbb{E}_G[h(\mu_i) \mid Z_i = z] = \frac{\int h(\mu) p(z \mid \mu)\ dG(\mu)}{\int p(z \mid \mu)\ dG(\mu)} = \frac{a_G(z)}{f_G(z)}$$

$$H_0 : \theta_G(z) = c \iff H_0 : a_G(z) - c \cdot f_G(z) = 0$$

AMARI

(Affine Minimax Anderson Rubin Inference)

$$H_0 : \theta_G(z) = c$$

$$\theta_G(z) = \mathbb{E}_G[h(\mu_i) \mid Z_i = z] = \frac{\int h(\mu) p(z \mid \mu)\, dG(\mu)}{\int p(z \mid \mu)\, dG(\mu)} = \frac{a_G(z)}{f_G(z)}$$

$$H_0 : \theta_G(z) = c \iff H_0 : a_G(z) - c \cdot f_G(z) = 0$$

$L(G)$ Linear in G

Jiaying Gu

Noack and Rothe (2019)

Anderson and Rubin (1949)

Fieller (1954)

Upshot:

Can focus on inference for linear functionals!

AMARI

(Affine Minimax Anderson Rubin Inference)

Seek to conduct inference for the linear functional

$$L(G)$$

Ansatz: Consider affine estimators of the form

$$\widehat{L} = \widehat{L}(G) = \frac{1}{n}\sum_{i=1}^{n} Q_n(Z_i)\,.$$

Why?

- Convenient computationally.
- Class of estimators that includes kernel density estimators and minimax optimal estimators of linear functionals in the Gaussian deconvolution problem [Butucea and Comte (2009), Pensky (2017)]

AMARI

(Affine Minimax Anderson Rubin Inference)

$$\hat{L} = \frac{1}{n}\sum_{i=1}^{n} Q_n(Z_i)\,. \quad \text{How to choose } Q = Q_n?$$

$$\min_{Q:\mathbb{R}\to\mathbb{R}} \left\{ \max_{G\in\mathcal{G}_n} \left\{ \text{Bias}_G[\hat{L}]^2 \right\} + \widehat{\text{Var}}\,[\hat{L}] \right\}$$

AMARI

(Affine Minimax Anderson Rubin Inference)

$$\hat{L} = \frac{1}{n}\sum_{i=1}^{n} Q_n(Z_i)\,. \quad \text{How to choose } Q = Q_n?$$

$$\min_{Q:\mathbb{R}\to\mathbb{R}} \left\{ \max_{G\in\mathcal{G}_n} \left\{ \text{Bias}_G[\hat{L}]^2 \right\} + \widehat{\text{Var}}\,[\hat{L}] \right\}$$

Bias:
$$\max_{G\in\mathcal{G}_n} \left\{ \text{Bias}_G[\hat{L}]^2 \right\} = \max_{G\in\mathcal{G}_n} \left\{ \left(\mathbb{E}_G[Q(Z_i)] - L(G) \right)^2 \right\}$$

$$\mathcal{G}_n = \{G \in \mathcal{G} : F_G \in \mathcal{F}_n(\alpha_n)\},\ \ \alpha_n \to 0$$

AMARI

(Affine Minimax Anderson Rubin Inference)

$$\hat{L} = \frac{1}{n}\sum_{i=1}^{n} Q_n(Z_i)\,. \quad \text{How to choose } Q = Q_n?$$

$$\min_{Q:\mathbb{R}\to\mathbb{R}} \left\{ \max_{G\in\mathcal{G}_n} \left\{ \text{Bias}_G[\hat{L}]^2 \right\} + \widehat{\text{Var}}\,[\hat{L}] \right\}$$

Bias:
$$\max_{G\in\mathcal{G}_n} \left\{ \text{Bias}_G[\hat{L}]^2 \right\} = \max_{G\in\mathcal{G}_n} \left\{ \left(\mathbb{E}_G[Q(Z_i)] - L(G) \right)^2 \right\}$$

$$\mathcal{G}_n = \{G \in \mathcal{G} : F_G \in \mathcal{F}_n(\alpha_n)\},\ \ \alpha_n \to 0$$

Variance: We use a pilot estimator $\bar{f}_n(z)$ of the marginal density.

$$\widehat{\text{Var}}\,[\hat{L}] = \frac{1}{n}\left\{ \int Q^2(z)\bar{f}_n(z)d\lambda(z) - \left(\int Q(z)\bar{f}_n(z)d\lambda(z) \right)^2 \right\}$$

Minimax optimization

$$\min_{Q:\mathbb{R}\to\mathbb{R}} \left\{ \max_{G\in\mathcal{G}_n} \left\{ \text{Bias}_G[\widehat{L}]^2 \right\} + \widehat{\text{Var}}[\widehat{L}] \right\}$$

The above optimization problem can be solved by building upon:

Donoho (1994), Armstrong and Kolesár (2018, 2020, 2021, …), Sacks and Ylvisacker (1978), Ibragimov and Hasminskii (1984), Donoho and Liu (1989, 1991), Low (1995), Zhao (1997), Cai and Low (2003, 2004), …

Bias-aware inference

Armstrong and Kolesár (2018), Imbens and Manski (2004), Imbens and Wager (2018)

Estimate $L(G)$ by $\widehat{L} = \sum_{i=1}^{n} Q_n(Z_i) \,/\, n$

$$\widehat{V} = \widehat{\mathrm{Var}}\left(Q_n(Z_i)\right) / n \qquad \widehat{B} = \sup_{G \in \mathcal{G}_n} \left| \mathrm{Bias}_G[\widehat{L}] \right|$$

$$\widehat{L} \pm t_\alpha(\widehat{B}, \widehat{V}) \qquad t_\alpha(B, V) = \inf\{t : \mathbb{P}[\,|b + V^{1/2}W| > t] < \alpha \text{ for all } |b| \le B\}$$
$$W \sim \mathcal{N}(0,1)$$

Theorem (I., Wager), Informal

Suppose we choose $Q_n(\,\cdot\,)$ as piecewise constant outside $[-M, M]$. If $\mathcal{F}_n(\alpha_n)$, $\bar{f}_n(\,\cdot\,)$ are constructed by sample splitting, then in the Binomial, Gaussian and Poisson empirical Bayes models, our intervals have asymptotic coverage $\geq 1 - \alpha$.

Confidence intervals in the RCT example

$n = 23{,}551$ RCTs

$$\mu_i \sim G,$$
$$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

$\mathcal{G} \triangleq$ Scale Mixture of centered Gaussians

$$= \left\{ G \text{ with Leb. density } g(\mu) = \int \frac{1}{\tau}\varphi\left(\frac{\mu}{\tau}\right) d\Pi(\tau), \ \Pi \text{ supported on } [0.1, 60] \right\}$$

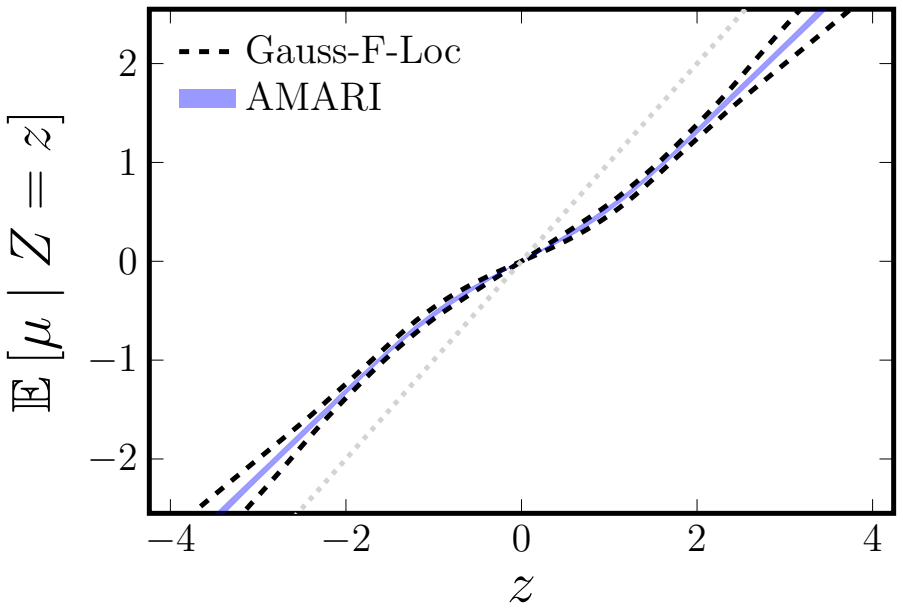
Confidence intervals in the RCT example

$n = 23{,}551$ RCTs

$$\mu_i \sim G,$$
$$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

$\mathcal{G} \triangleq$ Scale Mixture of centered Gaussians

$$= \left\{ G \text{ with Leb. density } g(\mu) = \int \frac{1}{\tau}\varphi\left(\frac{\mu}{\tau}\right) d\Pi(\tau), \ \ \Pi \text{ supported on } [0.1, 60] \right\}$$



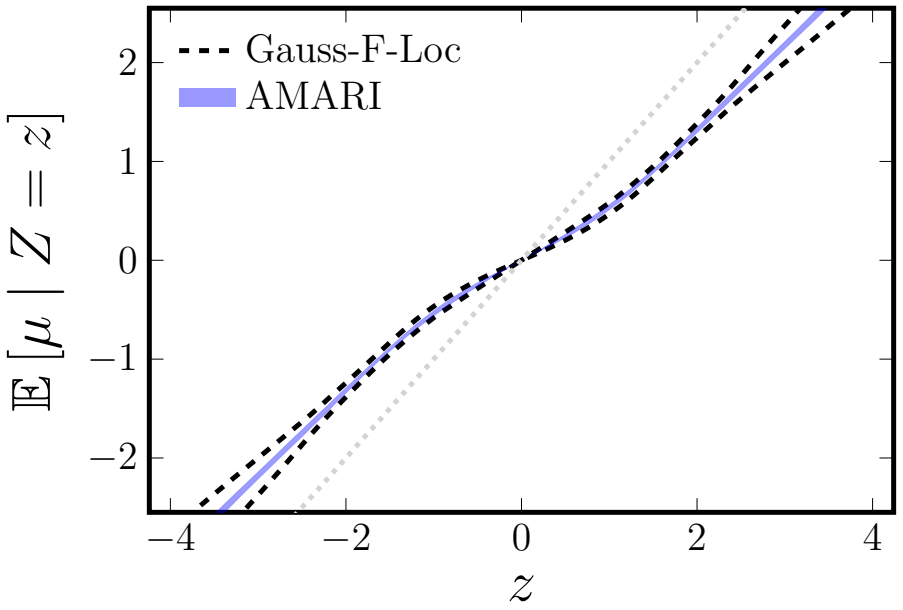
Confidence intervals in the RCT example

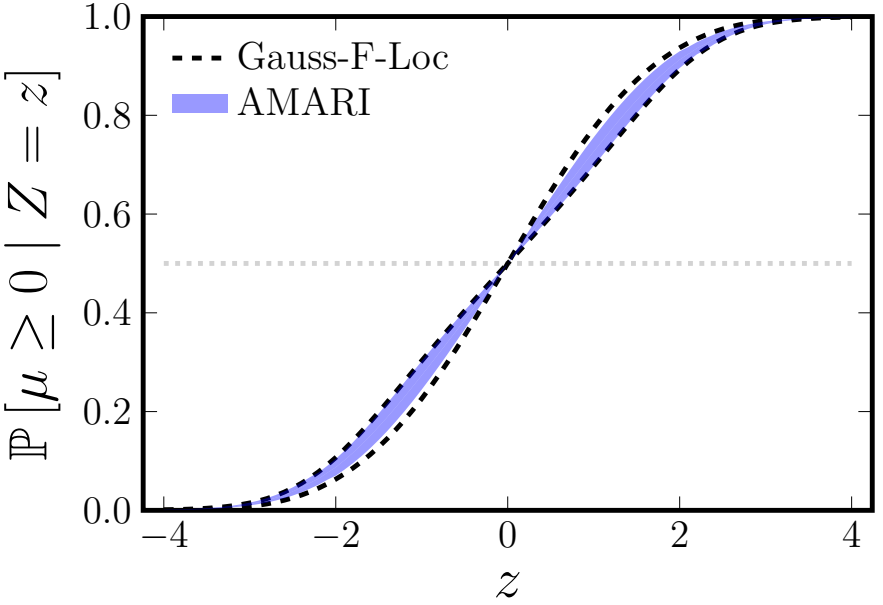
$n = 23{,}551$ RCTs

$$\mu_i \sim G,$$
$$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

$\mathcal{G} \triangleq$ Scale Mixture of centered Gaussians

$$= \left\{ G \text{ with Leb. density } g(\mu) = \int \frac{1}{\tau}\varphi\left(\frac{\mu}{\tau}\right) d\Pi(\tau), \ \ \Pi \text{ supported on } [0.1, 60] \right\}$$



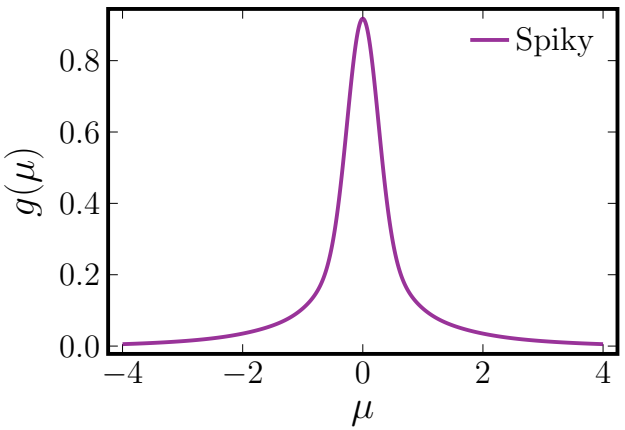


Simulation

$$\mu_i \sim G$$

$$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

$$n = 5000$$



$\mathcal{G}_{loc} \triangleq$ Location Mixture of Gaussians $\boxed{\tau = 0.25}$

$$= \left\{ G \text{ with Leb. density } g(\mu) = \int \frac{1}{\tau}\varphi\left(\frac{\mu - u}{\tau}\right) d\Pi(u), \ \ \Pi \text{ supported on } [-3,3] \right\}$$

[Magder and Zeger (1996), Cordy and Thomas (1997)]

$\mathcal{G}_{sc} \triangleq$ Scale Mixture of centered Gaussians

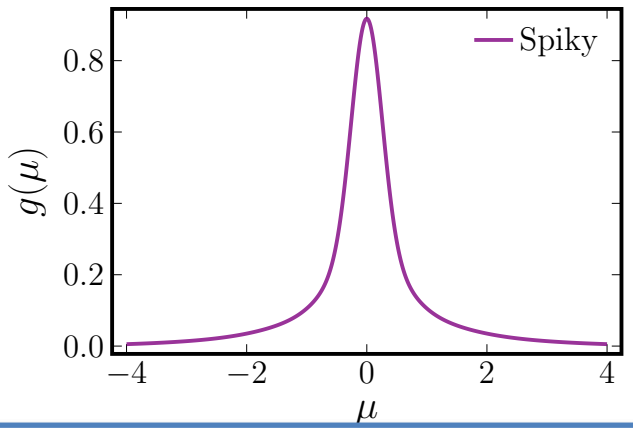
$$= \left\{ G \text{ with Leb. density } g(\mu) = \int \frac{1}{\tau}\varphi\left(\frac{\mu}{\tau}\right) d\Pi(\tau), \ \ \Pi \text{ supported on } [0.1, 15.6] \right\}$$

Simulation

$$\mu_i \sim G$$

$$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

$$n = 5000$$



$$\mathcal{G}_{loc} \triangleq$$

$$= \left\{ G \text{ w} \right\}$$

[M

$$\mathcal{G}_{sc} \triangleq \text{Sc}$$

$$= \left\{ G \text{ with} \right\}$$

Biostatistics (2017) **18**, 2, *pp.* 275–294

False discovery rates: a new deal

MATTHEW STEPHENS*

Department of Statistics and Department of Human Genetics, University of Chicago, 5801 S Ellis Ave, Chicago, IL 60637 USA

mstephens@uchicago.edu

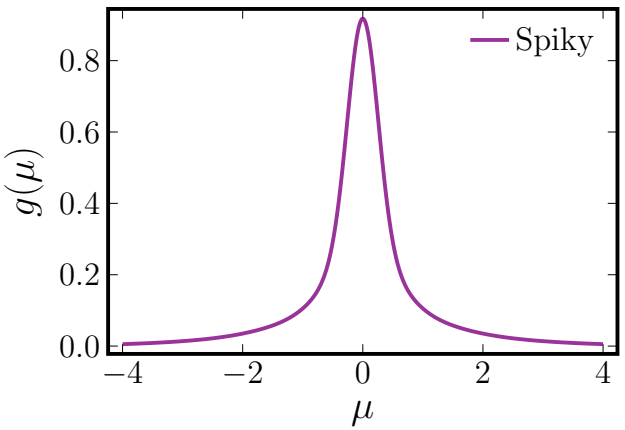
"enables more accurate inferences provided that it holds"

Simulation

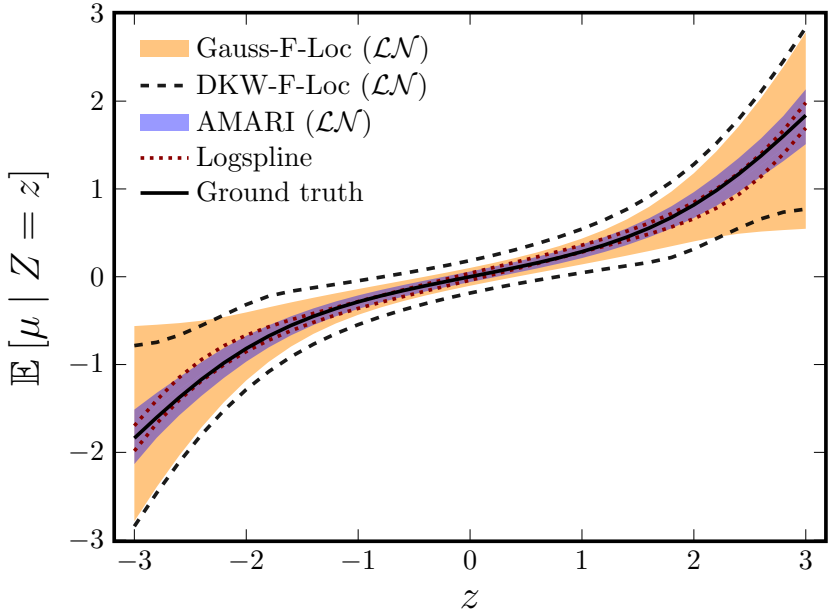
$$\mu_i \sim G$$

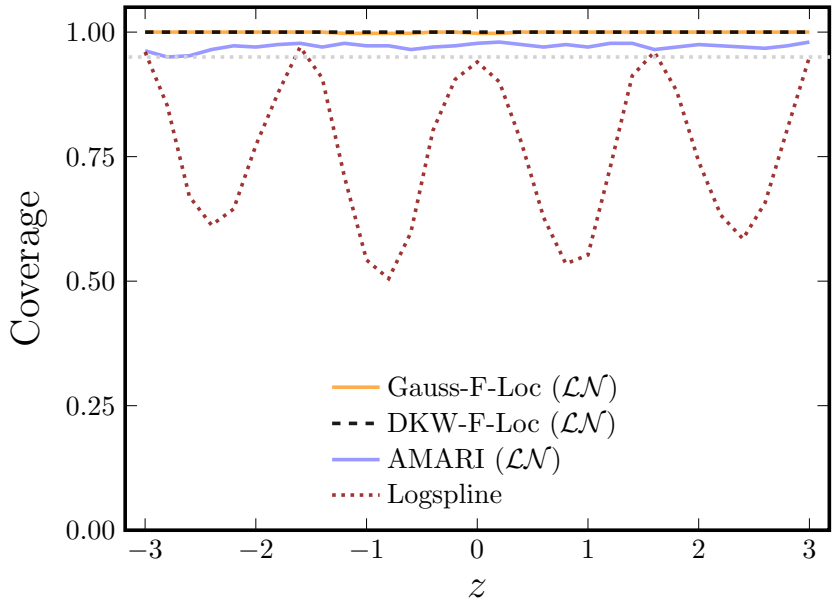
$$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

$$n = 5000$$



$\mathcal{G}_{loc} \triangleq$ Location Mixture of Gaussians



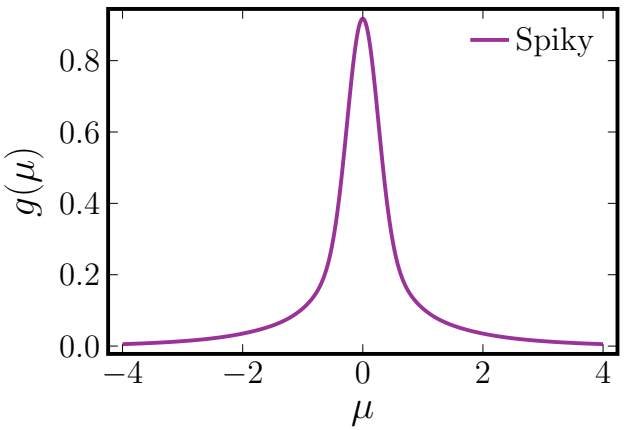


Simulation

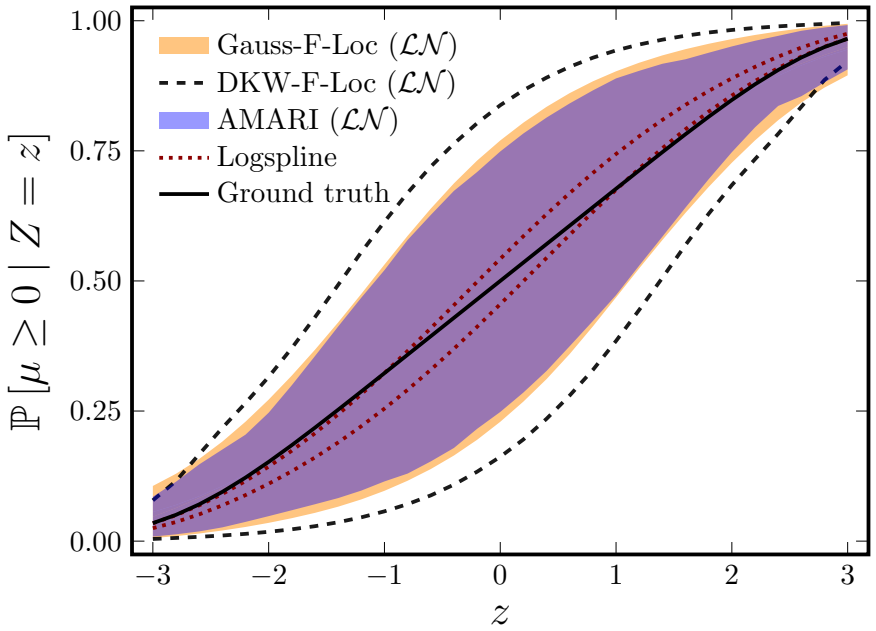
$$\mu_i \sim G$$

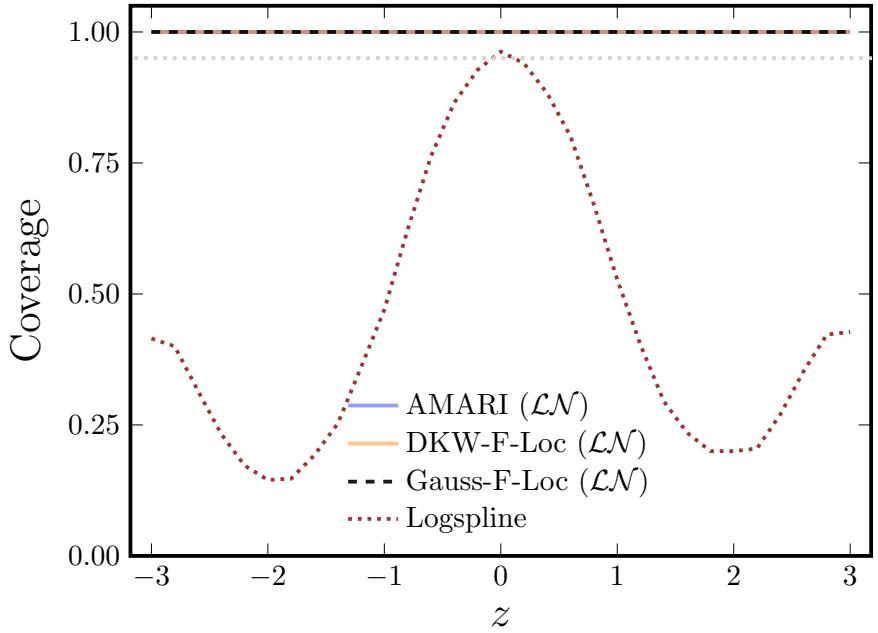
$$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

$$n = 5000$$



$\mathcal{G}_{loc} \triangleq$ Location Mixture of Gaussians



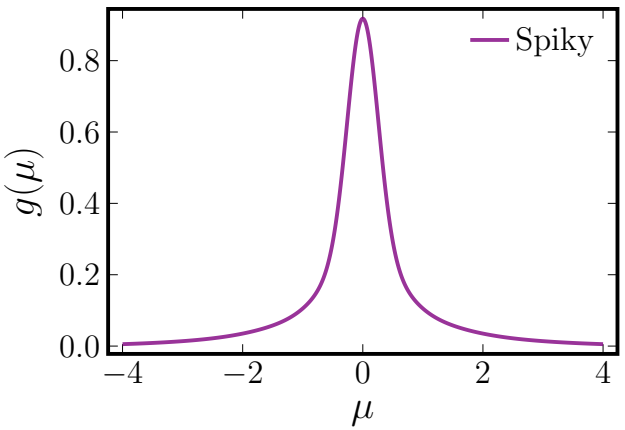


Simulation

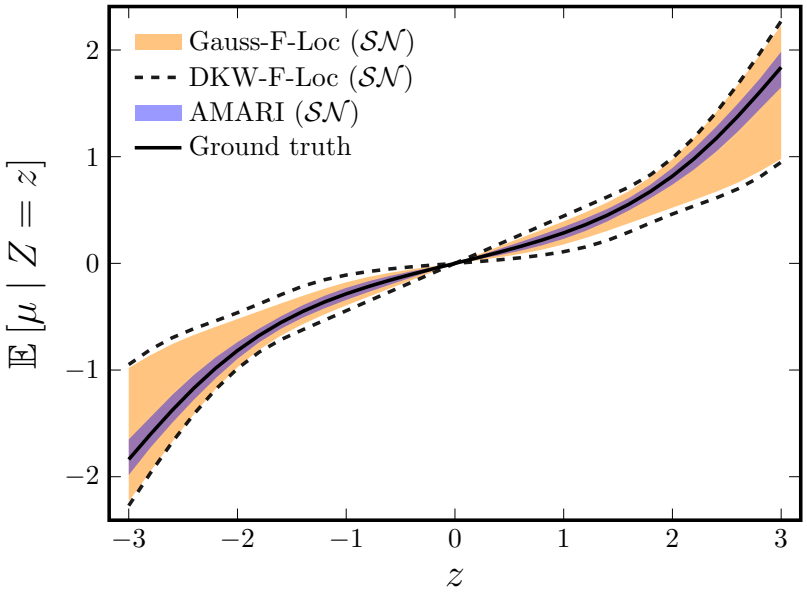
$$\mu_i \sim G$$

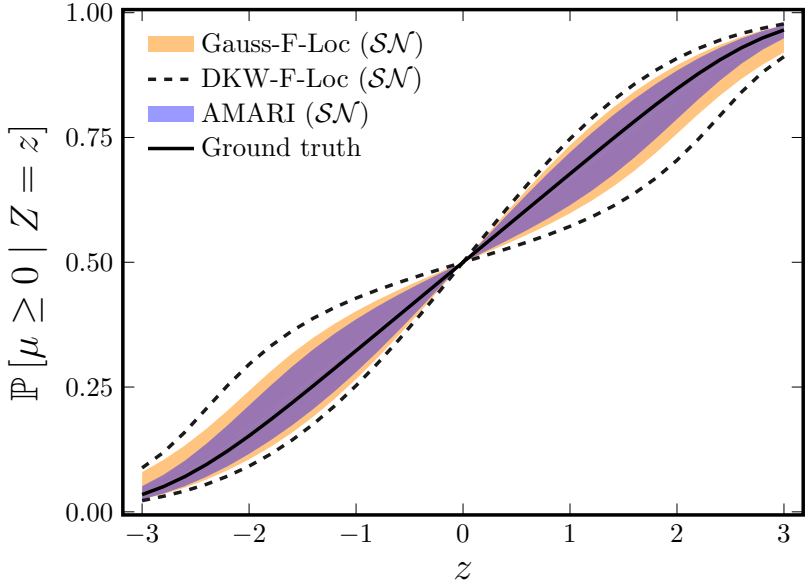
$$Z_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1)$$

$$n = 5000$$



$\mathcal{G}_{sc} \triangleq$ Scale Mixture of centered Gaussians





Conclusion

Even if n is large, there may be substantial uncertainty in the estimation of empirical Bayes quantities.

Here we describe two approaches to conduct inference for empirical Bayes estimands that can accompany empirical Bayes point estimation in practice.

Our approaches can also be used to assess the sensitivity to the choice of prior class $\mathscr{G}$.

Thank you for your attention!

AMARI

(Affine Minimax Anderson Rubin Inference)

Seek to conduct inference for the linear functional

$$L(G)$$

We consider **affine** estimators

Butucea and Comte (2009), Pensky (2017)

We choose the affine estimator by **minimax** optimization

Donoho (1994), Armstrong and Kolesár (2018, 2020, 2021, …), Sacks and Ylvisacker (1978), Ibragimov and Hasminskii (1984), Donoho and Liu (1989, 1991), Low (1995), Zhao (1997), Cai and Low (2003, 2004), …

We conduct bias-aware inference

Armstrong and Kolesár (2018), Imbens and Manski (2004), Imbens and Wager (2018)

Stein's hardest 1-dimensional subfamily

Suppose $G_1, G_2 \in \mathcal{G}_n$ solve the following optimization problem ($\delta > 0$):

$$\text{maximize} \quad L(G_1) - L(G_2) \qquad \text{s. t.} \quad \int \frac{(f_{G_1}(z) - f_{G_2}(z))^2}{\bar{f}_n(z)} \, d\lambda(z) \leq \frac{\delta^2}{n}$$

We call $\text{Conv}(G_1, G_2) = \left\{ \eta G_1 + (1-\eta) G_2 \ : \ \eta \in [0,1] \right\}$
a hardest 1-dimensional subfamily.

Stein's hardest 1-dimensional subfamily

Suppose $G_1, G_2 \in \mathcal{G}_n$ solve the following optimization problem ($\delta > 0$):

$$\text{maximize} \quad L(G_1) - L(G_2) \qquad \text{s. t.} \quad \int \frac{(f_{G_1}(z) - f_{G_2}(z))^2}{\bar{f}_n(z)} \, d\lambda(z) \leq \frac{\delta^2}{n}$$

We call $\text{Conv}(G_1, G_2) = \left\{ \eta G_1 + (1-\eta) G_2 \ : \ \eta \in [0,1] \right\}$
a hardest 1-dimensional subfamily.

$$\min_{Q:\mathbb{R}\to\mathbb{R}} \max_{G \in \mathcal{G}_n} \left\{ \text{Bias}_G[\widehat{L}]^2 \ + \ \Gamma_n \ \cdot \widehat{\text{Var}}[\widehat{L}] \right\}$$

Stein's hardest 1-dimensional subfamily

Suppose $G_1, G_2 \in \mathcal{G}_n$ solve the following optimization problem ($\delta > 0$):

$$\text{maximize} \quad L(G_1) - L(G_2) \qquad \text{s. t.} \quad \int \frac{(f_{G_1}(z) - f_{G_2}(z))^2}{\bar{f}_n(z)} \, d\lambda(z) \leq \frac{\delta^2}{n}$$

We call $\text{Conv}(G_1, G_2) = \left\{ \eta G_1 + (1-\eta) G_2 \ : \ \eta \in [0,1] \right\}$
a hardest 1-dimensional subfamily.

$$\min_{Q:\mathbb{R}\to\mathbb{R}} \max_{G \in \mathcal{G}_n} \left\{ \text{Bias}_G[\hat{L}]^2 \ + \ \Gamma_n(\delta) \cdot \widehat{\text{Var}}[\hat{L}] \right\}$$

$$G \in \text{Conv}(G_1, G_2)$$

The above optimization problem can be solved analytically!

Stein's hardest 1-dimensional subfamily

Suppose $G_1, G_2 \in \mathscr{G}_n$ solve the following optimization problem ($\delta > 0$):

$$\text{maximize} \quad L(G_1) - L(G_2) \qquad \text{s. t.} \quad \int \frac{(f_{G_1}(z) - f_{G_2}(z))^2}{\bar{f}_n(z)}\, d\lambda(z) \le \frac{\delta^2}{n}$$

We call $\text{Conv}(G_1, G_2) = \left\{\eta G_1 + (1-\eta) G_2 \ : \ \eta \in [0,1]\right\}$
a hardest 1-dimensional subfamily.

$$\min_{Q:\mathbb{R}\to\mathbb{R}} \max_{G \in \mathscr{G}_n} \left\{ \text{Bias}_G[\widehat{L}]^2 \ + \ \Gamma_n(\delta) \cdot \widehat{\text{Var}}[\widehat{L}] \right\}$$
$$G \in \text{Conv}(G_1, G_2)$$

The above optimization problem can be solved analytically!

Donoho (1994), Armstrong and Kolesár (2018, 2020, 2021, …),
Sacks and Ylvisacker (1978), Ibragimov and Hasminskii (1984), Donoho and Liu (1989, 1991), Low (1995), Zhao (1997), Cai and Low (2003, 2004), …