# FAB Inference

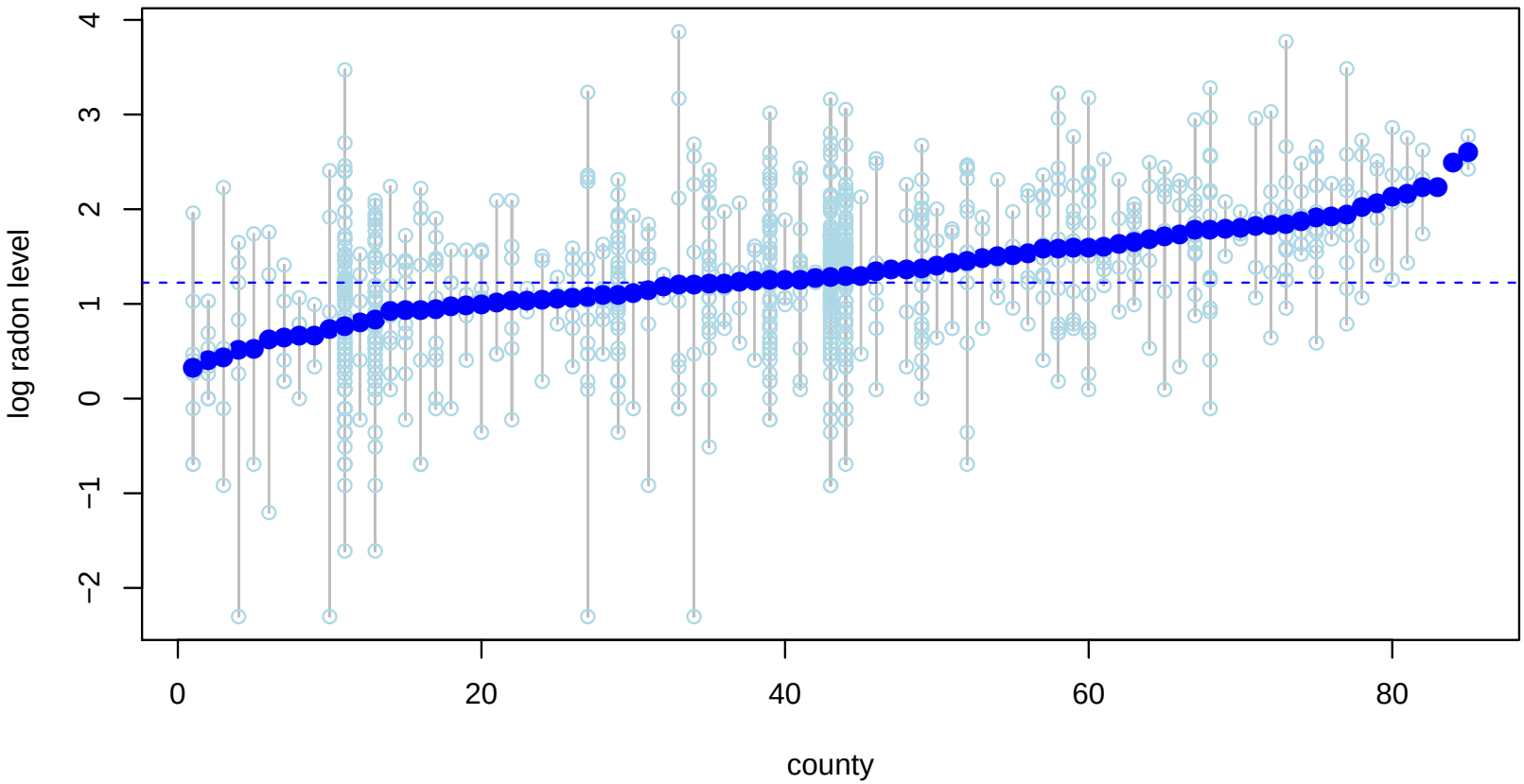
Peter Hoff
Duke University

In collaboration with Chaoyu Yu (Google) and Kyle Burris (Duke)

# Outline

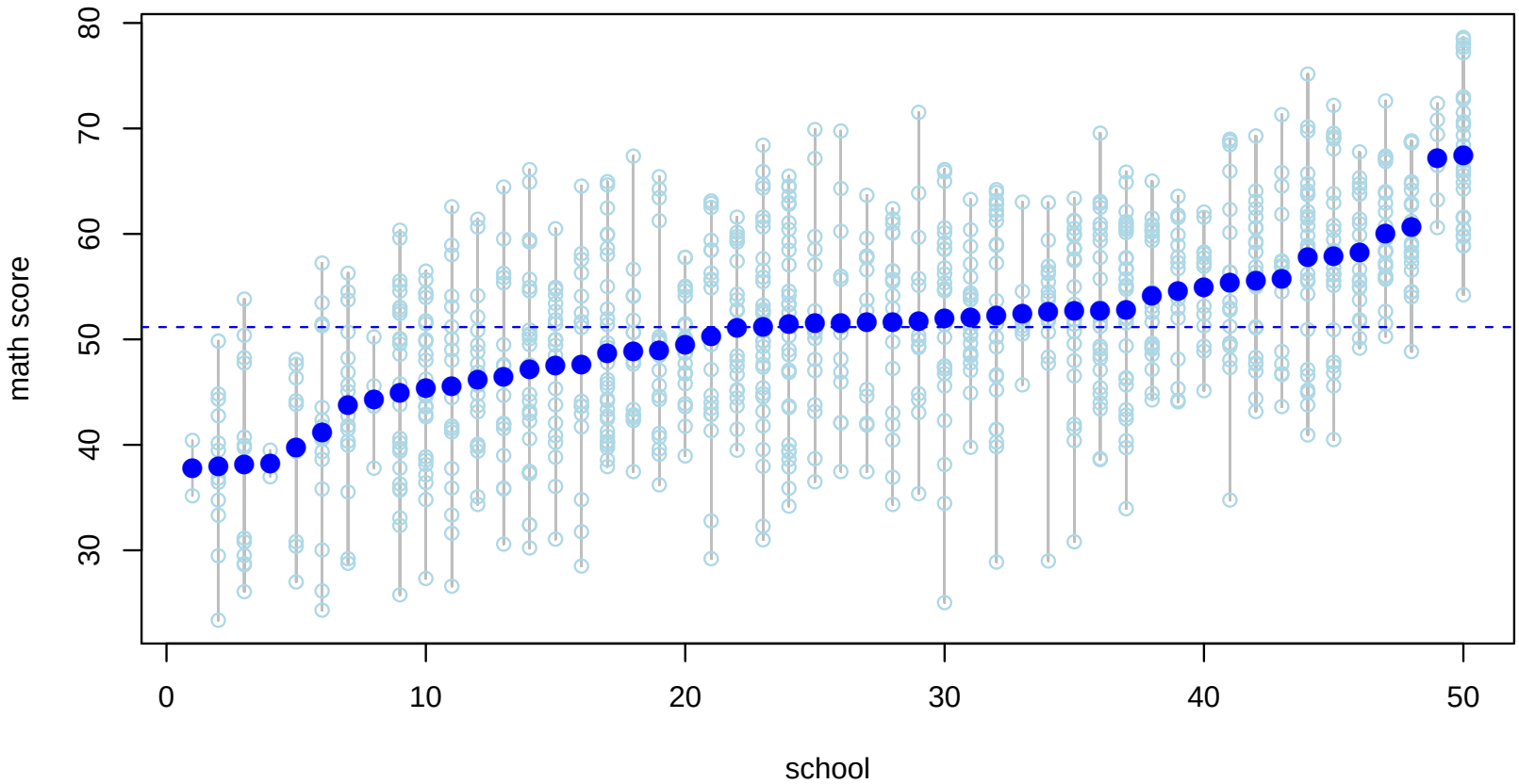
# Radon data

Log radon levels of 85 Minnesota counties.



# ELS data

Subset of 50 schools from 684 total.



# Multilevel data

**Data:** $y_{i,j}$ = outcome for unit $i$ in group $j$:

$$\mathbf{y}_1 = (y_{1,1}, \ldots, y_{n_1,1})$$
$$\vdots \qquad\qquad \vdots$$
$$\mathbf{y}_p = (y_{1,p}, \ldots, y_{n_p,p})$$

**Estimand:** $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, vector of group-specific means.

**Group-specific inference:** Obtain $\hat{\theta}_j$ and $C_j(\mathbf{y}) = [l_j(\mathbf{y}), u_j(\mathbf{y})]$ for each $j$ so

$$\mathrm{E}[(\hat{\theta}_j - \theta_j)^2 | \boldsymbol{\theta}] \text{ is small}$$
$$\Pr\left(\theta_j \in C_j(\mathbf{y}) \,|\boldsymbol{\theta}\right) = 1 - \alpha.$$

# Multilevel data

**Data:** $y_{i,j}$ = outcome for unit $i$ in group $j$:

$$\begin{aligned} \mathbf{y}_1 &= (y_{1,1}, \ldots, y_{n_1,1}) \\ &\vdots \qquad\qquad \vdots \\ \mathbf{y}_p &= (y_{1,p}, \ldots, y_{n_p,p}) \end{aligned}$$

**Estimand:** $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, vector of group-specific means.

**Group-specific inference:** Obtain $\hat{\theta}_j$ and $C_j(\mathbf{y}) = [l_j(\mathbf{y}), u_j(\mathbf{y})]$ for each $j$ so

$$\mathrm{E}[(\hat{\theta}_j - \theta_j)^2 | \boldsymbol{\theta}] \text{ is small}$$
$$\Pr\left(\theta_j \in C_j(\mathbf{y}) \,|\boldsymbol{\theta}\right) = 1 - \alpha.$$

# Multilevel data

**Data:** $y_{i,j}$ = outcome for unit $i$ in group $j$:

$$\mathbf{y}_1 = (y_{1,1}, \ldots, y_{n_1,1})$$
$$\vdots \qquad\qquad \vdots$$
$$\mathbf{y}_p = (y_{1,p}, \ldots, y_{n_p,p})$$

**Estimand:** $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, vector of group-specific means.

**Group-specific inference:** Obtain $\hat{\theta}_j$ and $C_j(\mathbf{y}) = [l_j(\mathbf{y}), u_j(\mathbf{y})]$ for each $j$ so

$$\mathrm{E}[(\hat{\theta}_j - \theta_j)^2 | \boldsymbol{\theta}] \text{ is small}$$
$$\Pr\left(\theta_j \in C_j(\mathbf{y}) \,|\boldsymbol{\theta}\right) = 1 - \alpha.$$

# Point estimation

How should we estimate $\theta_j$?

**Unbiased estimation:**

$$\hat{\theta}_j \stackrel{?}{=} \bar{y}_j = \sum_i y_{i,j}/n_j$$

- Guaranteed to be unbiased, $\mathrm{E}[\bar{y}_j] = \theta_j$.
- Minimum variance among linear unbiased estimators.
- Error depends on sample size:

$$\mathrm{E}[(\bar{y}_j - \theta_j)^2|\boldsymbol{\theta}] = \sigma^2/n_j.$$

**Bayes/Shrinkage estimation:**

$$\hat{\theta}_j \stackrel{?}{=} w\bar{y}_j + (1-w)\bar{y}_{-j}$$

- Bias is bigger than that of $\bar{y}_j$.
- Variance is lower than that of $\bar{y}_j$.

How to pick $w$?

# Point estimation

How should we estimate $\theta_j$?

**Unbiased estimation:**

$$\hat{\theta}_j \stackrel{?}{=} \bar{y}_j = \sum_i y_{i,j}/n_j$$

- Guaranteed to be unbiased, $\mathrm{E}[\bar{y}_j] = \theta_j$.
- Minimum variance among linear unbiased estimators.
- Error depends on sample size:

$$\mathrm{E}[(\bar{y}_j - \theta_j)^2|\theta] = \sigma^2/n_j.$$

**Bayes/Shrinkage estimation:**

$$\hat{\theta}_j \stackrel{?}{=} w\bar{y}_j + (1-w)\bar{y}_{-j}$$

- Bias is bigger than that of $\bar{y}_j$.
- Variance is lower than that of $\bar{y}_j$.

How to pick $w$?

# Point estimation

How should we estimate $\theta_j$?

**Unbiased estimation:**

$$\hat{\theta}_j \stackrel{?}{=} \bar{y}_j = \sum_i y_{i,j}/n_j$$

- Guaranteed to be unbiased, $\mathrm{E}[\bar{y}_j] = \theta_j$.
- Minimum variance among linear unbiased estimators.
- Error depends on sample size:

$$\mathrm{E}[(\bar{y}_j - \theta_j)^2 | \boldsymbol{\theta}] = \sigma^2/n_j.$$

**Bayes/Shrinkage estimation:**

$$\hat{\theta}_j \stackrel{?}{=} w\bar{y}_j + (1-w)\bar{y}_{-j}$$

- Bias is bigger than that of $\bar{y}_j$.
- Variance is lower than that of $\bar{y}_j$.

How to pick $w$?

# Point estimation

How should we estimate $\theta_j$?

**Unbiased estimation:**

$$\hat{\theta}_j \stackrel{?}{=} \bar{y}_j = \sum_i y_{i,j}/n_j$$

- Guaranteed to be unbiased, $\mathrm{E}[\bar{y}_j] = \theta_j$.
- Minimum variance among linear unbiased estimators.
- Error depends on sample size:

$$\mathrm{E}[(\bar{y}_j - \theta_j)^2 | \boldsymbol{\theta}] = \sigma^2/n_j.$$

**Bayes/Shrinkage estimation:**

$$\hat{\theta}_j \stackrel{?}{=} w\bar{y}_j + (1-w)\bar{y}_{-j}$$

- Bias is bigger than that of $\bar{y}_j$.
- Variance is lower than that of $\bar{y}_j$.

How to pick $w$?

# Point estimation

How should we estimate $\theta_j$?

**Unbiased estimation:**

$$\hat{\theta}_j \stackrel{?}{=} \bar{y}_j = \sum_i y_{i,j}/n_j$$

- Guaranteed to be unbiased, $\mathrm{E}[\bar{y}_j] = \theta_j$.
- Minimum variance among linear unbiased estimators.
- Error depends on sample size:

$$\mathrm{E}[(\bar{y}_j - \theta_j)^2 | \boldsymbol{\theta}] = \sigma^2/n_j.$$

**Bayes/Shrinkage estimation:**

$$\hat{\theta}_j \stackrel{?}{=} w\bar{y}_j + (1-w)\bar{y}_{-j}$$

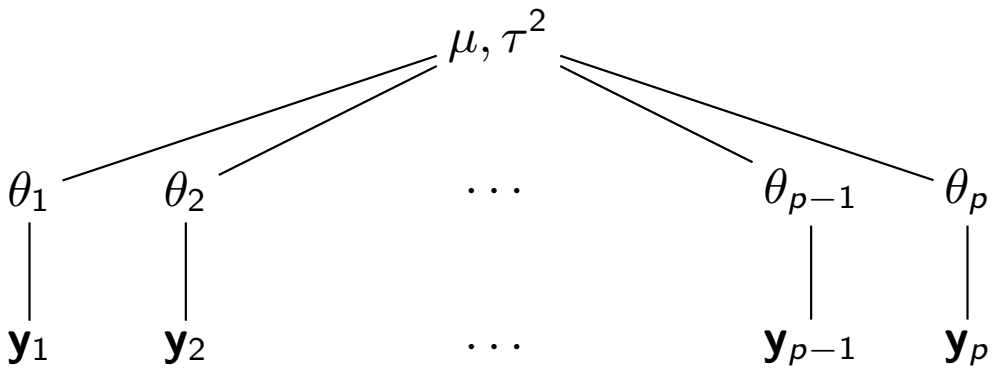
- Bias is bigger than that of $\bar{y}_j$.
- Variance is lower than that of $\bar{y}_j$.

How to pick $w$?

# Indirect information

Pick $w$ with **indirect information**:

- prior information - $\theta_j \sim N(\mu, \tau^2)$;
- information from other groups - hierarchical models.




$$y_{1,j}, \ldots, y_{n_j,j} | \boldsymbol{\theta} \sim \text{i.i.d. } N(\theta_j, \sigma^2)$$

$$\theta_1, \ldots, \theta_p \sim \text{i.i.d. } N(\mu, \tau^2)$$

Operationally very similar, but somewhat different interpretations.

# Optimal shrinkage

If $\mu, \tau^2, \sigma^2$ were known, the optimal estimator is

$$\check{\theta}_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}\bar{y}_j + \frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}\mu.$$

Since $\mu, \tau^2, \sigma^2$ are generally unknown, the following estimator is typically used:

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

**Terminology:** shrinkage estimator, JS estimator, eBayes estimator, BLUP.

*Information from all groups helps with making inferences for a specific group.*

# Optimal shrinkage

If $\mu, \tau^2, \sigma^2$ were known, the optimal estimator is

$$\check{\theta}_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}\bar{y}_j + \frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}\mu.$$

Since $\mu, \tau^2, \sigma^2$ are generally unknown, the following estimator is typically used:

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

**Terminology:** shrinkage estimator, JS estimator, eBayes estimator, BLUP.

*Information from all groups helps with making inferences for a specific group.*

# Optimal shrinkage

If $\mu, \tau^2, \sigma^2$ were known, the optimal estimator is

$$\check{\theta}_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}\bar{y}_j + \frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}\mu.$$

Since $\mu, \tau^2, \sigma^2$ are generally unknown, the following estimator is typically used:

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

**Terminology:** shrinkage estimator, JS estimator, eBayes estimator, BLUP.

*Information from all groups helps with making inferences for a specific group.*

# Optimal shrinkage

If $\mu, \tau^2, \sigma^2$ were known, the optimal estimator is

$$\check{\theta}_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}\bar{y}_j + \frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}\mu.$$

Since $\mu, \tau^2, \sigma^2$ are generally unknown, the following estimator is typically used:

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

**Terminology:** shrinkage estimator, JS estimator, eBayes estimator, BLUP.

*Information from all groups helps with making inferences for a specific group.*

# Results on performance

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

**Across-group performance for the approximately true believer**

- $\hat{\theta}_j$ was derived from the assumption $\theta_1, \ldots \theta_m \sim$ i.i.d. $N(\mu, \tau^2)$.
- If approximately true, $\hat{\boldsymbol{\theta}}$ approximately optimal in terms of $\mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$.

**Across-group performance for the skeptic**

- $\hat{\boldsymbol{\theta}}$ guaranteed to be better than $\bar{\mathbf{y}}$ in terms of $\mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$.

*By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong.*

**Group-specific performance**

- $\hat{\theta}_j$ better than $\bar{y}_j$ if $\theta_j$ near $\mu$.
- $\hat{\theta}_j$ worse than $\bar{y}_j$ if $\theta_j$ far from $\mu$.

# Results on performance

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

**Across-group performance for the approximately true believer**

- $\hat{\theta}_j$ was derived from the assumption $\theta_1, \ldots \theta_m \sim$ i.i.d. $N(\mu, \tau^2)$.
- If approximately true, $\hat{\boldsymbol{\theta}}$ approximately optimal in terms of $\mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$.

**Across-group performance for the skeptic**

- $\hat{\boldsymbol{\theta}}$ guaranteed to be better than $\bar{\mathbf{y}}$ in terms of $\mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$.

*By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong.*

**Group-specific performance**

- $\hat{\theta}_j$ better than $\bar{y}_j$ if $\theta_j$ near $\mu$.
- $\hat{\theta}_j$ worse than $\bar{y}_j$ if $\theta_j$ far from $\mu$.

# Results on performance

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

**Across-group performance for the approximately true believer**

- $\hat{\theta}_j$ was derived from the assumption $\theta_1, \ldots \theta_m \sim$ i.i.d. $N(\mu, \tau^2)$.
- If approximately true, $\hat{\boldsymbol{\theta}}$ approximately optimal in terms of $\mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$.

**Across-group performance for the skeptic**

- $\hat{\boldsymbol{\theta}}$ guaranteed to be better than $\bar{\mathbf{y}}$ in terms of $\mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$.

*By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong*.

**Group-specific performance**

- $\hat{\theta}_j$ better than $\bar{y}_j$ if $\theta_j$ near $\mu$.
- $\hat{\theta}_j$ worse than $\bar{y}_j$ if $\theta_j$ far from $\mu$.

# Results on performance

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

**Across-group performance for the approximately true believer**

- $\hat{\theta}_j$ was derived from the assumption $\theta_1, \ldots \theta_m \sim$ i.i.d. $N(\mu, \tau^2)$.
- If approximately true, $\hat{\boldsymbol{\theta}}$ approximately optimal in terms of $\mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$.

**Across-group performance for the skeptic**

- $\hat{\boldsymbol{\theta}}$ guaranteed to be better than $\bar{\mathbf{y}}$ in terms of $\mathrm{E}[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$.

*By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong*.

**Group-specific performance**

- $\hat{\theta}_j$ better than $\bar{y}_j$ if $\theta_j$ near $\mu$.
- $\hat{\theta}_j$ worse than $\bar{y}_j$ if $\theta_j$ far from $\mu$.

# Intervals

**Confidence interval:**

$$C(\mathbf{y}) \stackrel{?}{=} \bar{y}_j \pm \frac{\hat{\sigma}}{\sqrt{n_j}} t_{1-\alpha/2}$$

- Exact constant coverage:

$$\Pr(\theta_j \in C(\mathbf{y})|\boldsymbol{\theta}) = 1-\alpha \text{ for all values of } \theta_j.$$

- Narrowest interval among "unbiased" intervals.
- Doesn't use indirect information.

Can indirect information be incorporated, while maintaining constant coverage?

**"Prediction" interval:**

$$C(\hat{\theta}_j) = \hat{\theta}_j \pm t_{1-\alpha/2}/\sqrt{1/\hat{\tau}^2 + n_j/\hat{\sigma}^2}$$

- $1-\alpha$ coverage *on average across groups*.
- Lower for some groups, higher for others, and you don't know which.

# Intervals

**Confidence interval:**

$$C(\mathbf{y}) \stackrel{?}{=} \bar{y}_j \pm \frac{\hat{\sigma}}{\sqrt{n_j}} t_{1-\alpha/2}$$

- Exact constant coverage:

$$\Pr(\theta_j \in C(\mathbf{y})|\boldsymbol{\theta}) = 1-\alpha \text{ for all values of } \theta_j.$$

- Narrowest interval among "unbiased" intervals.
- Doesn't use indirect information.

Can indirect information be incorporated, while maintaining constant coverage?

**"Prediction" interval:**

$$C(\hat{\theta}_j) = \hat{\theta}_j \pm t_{1-\alpha/2}/\sqrt{1/\hat{\tau}^2 + n_j/\hat{\sigma}^2}$$

- $1-\alpha$ coverage *on average across groups*.
- Lower for some groups, higher for others, and you don't know which.

# Intervals

**Confidence interval:**

$$C(\mathbf{y}) \stackrel{?}{=} \bar{y}_j \pm \frac{\hat{\sigma}}{\sqrt{n_j}} t_{1-\alpha/2}$$

- Exact constant coverage:

$$\Pr(\theta_j \in C(\mathbf{y})|\boldsymbol{\theta}) = 1-\alpha \text{ for all values of } \theta_j.$$

- Narrowest interval among "unbiased" intervals.
- Doesn't use indirect information.

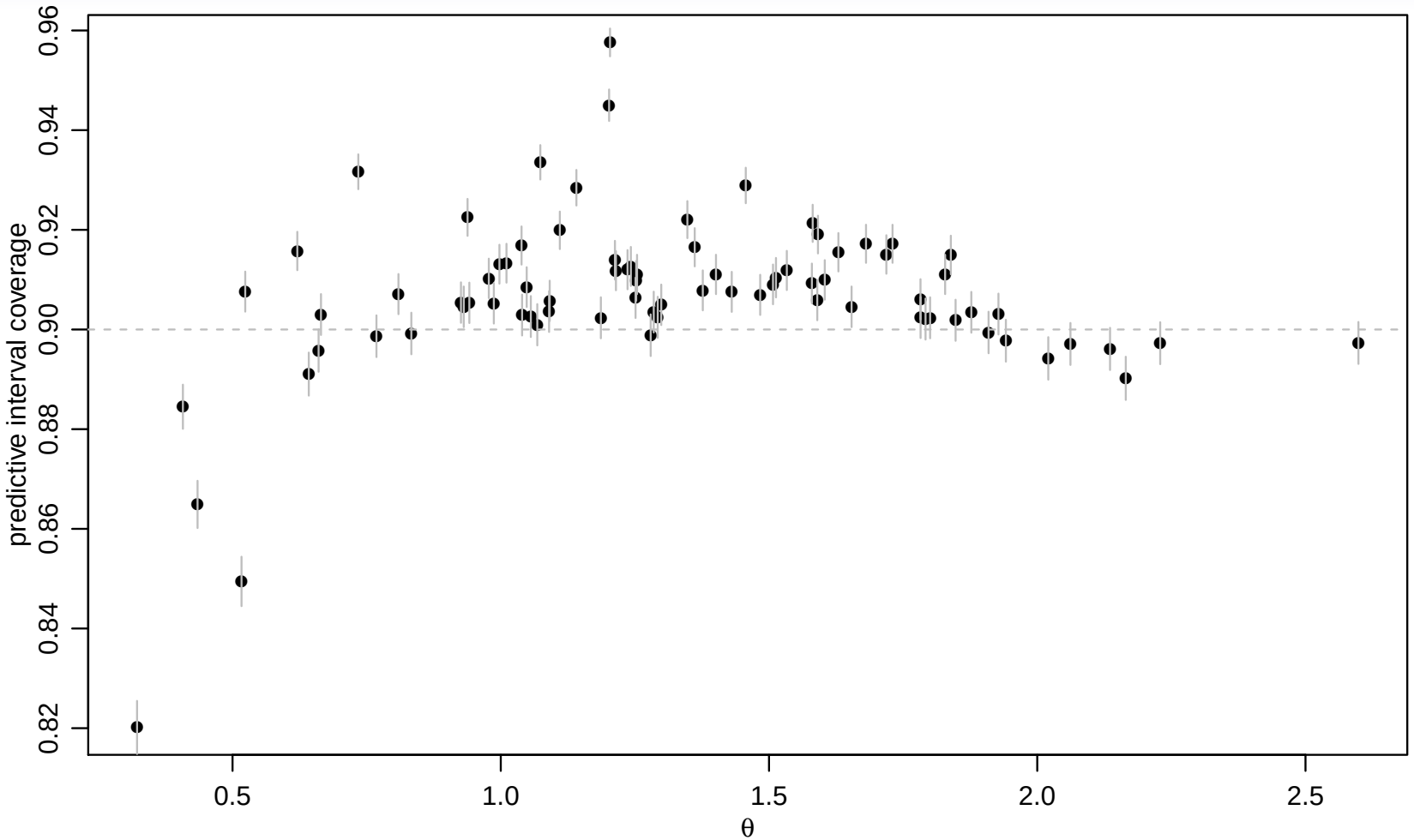
Can indirect information be incorporated, while maintaining constant coverage?

**"Prediction" interval:**

$$C(\hat{\theta}_j) = \hat{\theta}_j \pm t_{1-\alpha/2}/\sqrt{1/\hat{\tau}^2 + n_j/\hat{\sigma}^2}$$

- $1-\alpha$ coverage *on average across groups*.
- Lower for some groups, higher for others, and you don't know which.

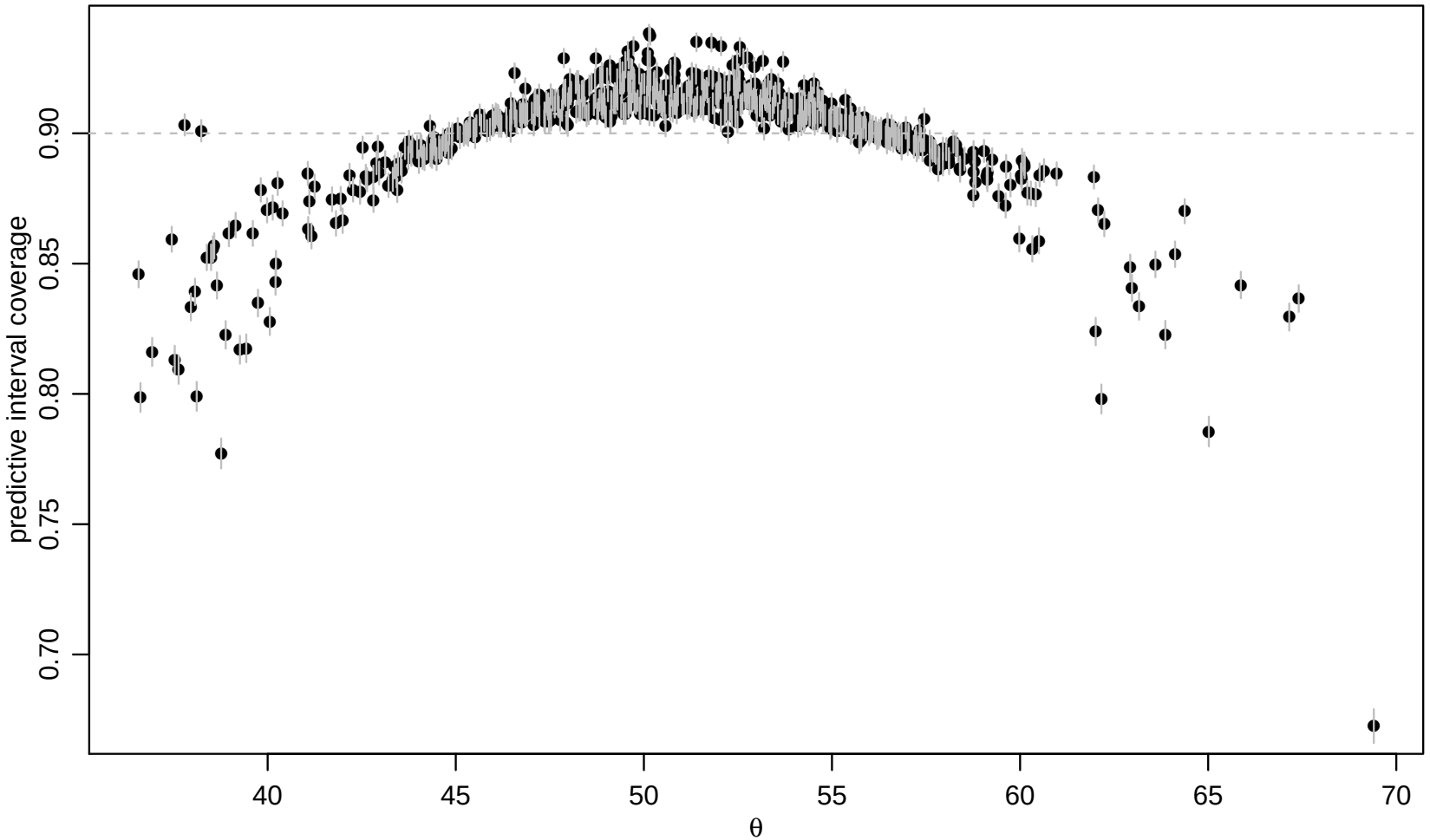
# Nonconstant coverage: Radon data




$$\Pr(\theta_j \in C(\hat{\theta}_j)) \approx 1 - \alpha$$

$$\Pr(\theta_j \in C(\hat{\theta}_j)|\boldsymbol{\theta}) \quad \text{depends on } j, \boldsymbol{\theta}.$$

# Nonconstant coverage: ELS data




$$\Pr(\theta_j \in C(\hat{\theta}_j)) \approx 1 - \alpha$$

$$\Pr(\theta_j \in C(\hat{\theta}_j)|\boldsymbol{\theta}) \quad \text{depends on } j, \boldsymbol{\theta}.$$

# Valid confidence intervals that share information

**Goal:** Construct confidence intervals $C^1, \ldots, C^p$ having

- constant coverage: $\Pr(\theta_j \in C^j(\mathbf{y})|\boldsymbol{\theta}) = 1-\alpha$ for all groups/$\boldsymbol{\theta}$'s.
- improved precision: $\mathrm{E}[|C^j(\boldsymbol{y})|] < 2t_{1-\alpha/2}$ on average across groups/$\boldsymbol{\theta}$'s.

The first criterion is group-specific/frequentist - conditional on $\theta_j$.

The second is study-specific/Bayes - on average across $\theta_1, \ldots, \theta_p$.

# Valid confidence intervals that share information

**Goal:** Construct confidence intervals $C^1, \ldots, C^p$ having

- constant coverage: $\Pr(\theta_j \in C^j(\mathbf{y})|\boldsymbol{\theta}) = 1-\alpha$ for all groups/$\boldsymbol{\theta}$'s.
- improved precision: $\mathrm{E}[|C^j(\boldsymbol{y})|] < 2t_{1-\alpha/2}$ on average across groups/$\boldsymbol{\theta}$'s.

The first criterion is group-specific/frequentist - conditional on $\theta_j$.

The second is study-specific/Bayes - on average across $\theta_1, \ldots, \theta_p$.

# All CIPs

**Standard procedure:**

$$C_{1/2}(y) = \{\theta : y + \sigma z_{\alpha/2} < \theta < y + \sigma z_{1-\alpha/2}\}$$

**Any procedure:**

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w)} < \theta < y + \sigma z_{1-\alpha w}\}$$

**In fact, $w$ may depend on $\theta$:** If $w : \mathbb{R} \to [0, 1]$ then

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w(\theta))} < \theta < y + \sigma z_{1-\alpha w(\theta)}\}$$

satisfies $\Pr(\theta \in C_w(y)|\theta) = 1 - \alpha$

- Examples in Bartholomew [1971], Stein [1962].
- Essentially complete class result in Yu and Hoff [2018].

# All CIPs

**Standard procedure:**

$$C_{1/2}(y) = \{\theta : y + \sigma z_{\alpha/2} < \theta < y + \sigma z_{1-\alpha/2}\}$$

**Any procedure:**

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w)} < \theta < y + \sigma z_{1-\alpha w}\}$$

**In fact, $w$ may depend on $\theta$:** If $w : \mathbb{R} \to [0, 1]$ then

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w(\theta))} < \theta < y + \sigma z_{1-\alpha w(\theta)}\}$$

satisfies $\Pr(\theta \in C_w(y)|\theta) = 1 - \alpha$

- Examples in Bartholomew [1971], Stein [1962].
- Essentially complete class result in Yu and Hoff [2018].

# All CIPs

**Standard procedure:**

$$C_{1/2}(y) = \{\theta : y + \sigma z_{\alpha/2} < \theta < y + \sigma z_{1-\alpha/2}\}$$

**Any procedure:**

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w)} < \theta < y + \sigma z_{1-\alpha w}\}$$

**In fact, $w$ may depend on $\theta$:** If $w : \mathbb{R} \to [0, 1]$ then

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w(\theta))} < \theta < y + \sigma z_{1-\alpha w(\theta)}\}$$

satisfies $\Pr(\theta \in C_w(y)|\boldsymbol{\theta}) = 1 - \alpha$

- Examples in Bartholomew [1971], Stein [1962].
- Essentially complete class result in Yu and Hoff [2018].

# FAB: Bayes-optimal frequentist interval

**Simplified model**:

- $y|\theta \sim N(\theta, \sigma^2)$, $\sigma^2$ known.
- $\pi(\theta)$ is prior information about $\theta$.

**Idea:** Find the $w$-function that minimizes the prior expected width

$$\int \int |C_w(y)|\, p(dy|\theta)\pi(d\theta) < \int \int |C(y)|\, p(dy|\theta)\pi(d\theta)$$

Such an interval will have

- constant coverage, because $C_w$ has constant coverage for any $w$-function;
- optimal precision on average with respect to $\pi$, by construction.

We call it FAB - frequentist, assisted by Bayes.

# FAB: Bayes-optimal frequentist interval

**Simplified model**:

- $y|\theta \sim N(\theta, \sigma^2)$, $\sigma^2$ known.
- $\pi(\theta)$ is prior information about $\theta$.

**Idea:** Find the $w$-function that minimizes the prior expected width

$$\int \int |C_w(y)|\, p(dy|\theta)\pi(d\theta) < \int \int |C(y)|\, p(dy|\theta)\pi(d\theta)$$

Such an interval will have

- constant coverage, because $C_w$ has constant coverage for any $w$-function;
- optimal precision on average with respect to $\pi$, by construction.

We call it FAB - frequentist, assisted by Bayes.

# FAB: Bayes-optimal frequentist interval

**Simplified model**:

- $y|\theta \sim N(\theta, \sigma^2)$, $\sigma^2$ known.
- $\pi(\theta)$ is prior information about $\theta$.

**Idea:** Find the $w$-function that minimizes the prior expected width

$$\int \int |C_w(y)|\, p(dy|\theta)\pi(d\theta) < \int \int |C(y)|\, p(dy|\theta)\pi(d\theta)$$

Such an interval will have

- constant coverage, because $C_w$ has constant coverage for any $w$-function;
- optimal precision on average with respect to $\pi$, by construction.

We call it FAB - frequentist, assisted by Bayes.

# Optimal $w$-function

If $\pi(\theta)$ is the $N(\mu, \tau^2)$ density, then

$$E[|C_w|] = \int\int |C_w(y)|\, p(dy|\theta)\, \pi(d\theta)$$

is minimized by

$$w(\theta) = g^{-1}(2\sigma(\theta - \mu)/\tau^2)$$
$$g(w) = \Phi^{-1}(\alpha w) - \Phi^{-1}(\alpha(1-w))$$

This $w$-function yields Pratt's (1963) $z$-interval.

- Yu and Hoff (2018): Extension to $t$-intervals, multigroup inference.
- Hoff and Yu (2019): Linear regression coefficients.
- Burris and Hoff (2019): Small area estimation.

# Optimal $w$-function

If $\pi(\theta)$ is the $N(\mu, \tau^2)$ density, then

$$E[|C_w|] = \int \int |C_w(y)|\, p(dy|\theta)\, \pi(d\theta)$$

is minimized by

$$w(\theta) = g^{-1}(2\sigma(\theta - \mu)/\tau^2)$$
$$g(w) = \Phi^{-1}(\alpha w) - \Phi^{-1}(\alpha(1-w))$$

This $w$-function yields Pratt's (1963) $z$-interval.

- Yu and Hoff (2018): Extension to $t$-intervals, multigroup inference.
- Hoff and Yu (2019): Linear regression coefficients.
- Burris and Hoff (2019): Small area estimation.

# Optimal $w$-function

If $\pi(\theta)$ is the $N(\mu, \tau^2)$ density, then

$$E[|C_w|] = \int\int |C_w(y)|\, p(dy|\theta)\, \pi(d\theta)$$

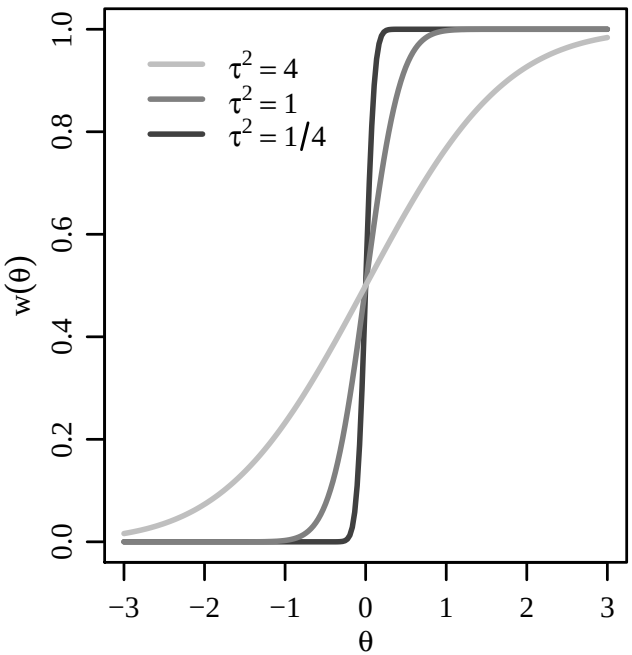
is minimized by

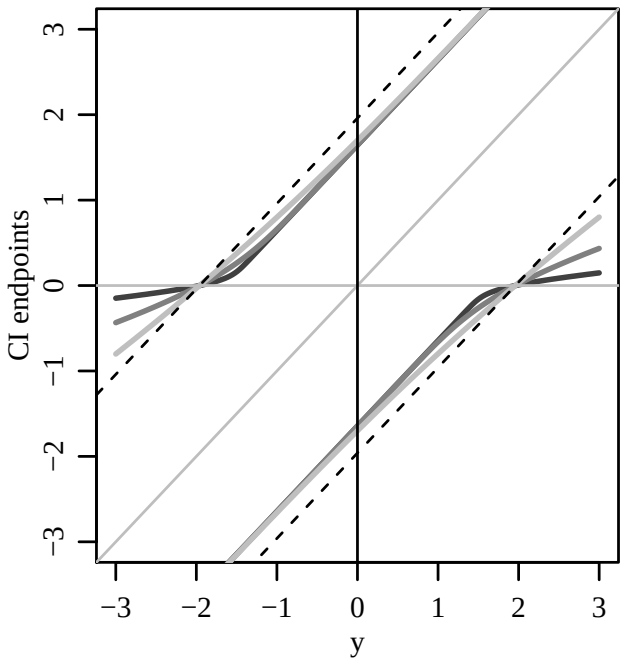
$$w(\theta) = g^{-1}(2\sigma(\theta - \mu)/\tau^2)$$
$$g(w) = \Phi^{-1}(\alpha w) - \Phi^{-1}(\alpha(1-w))$$

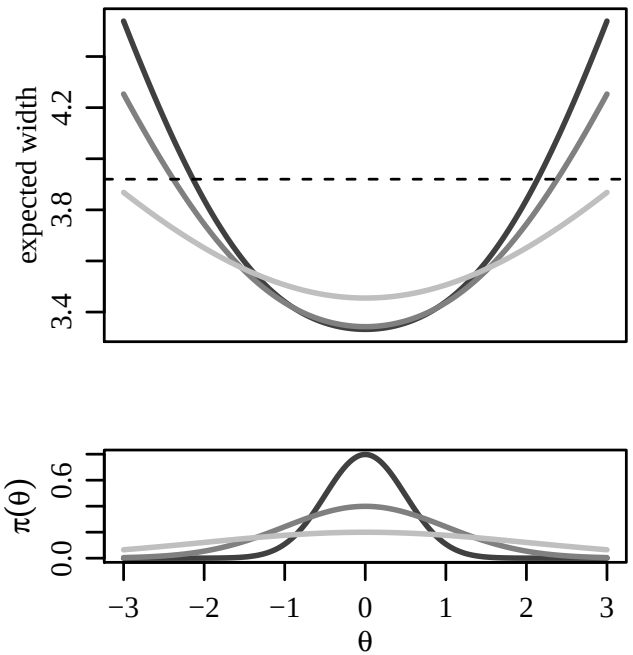
This $w$-function yields Pratt's (1963) $z$-interval.

- Yu and Hoff (2018): Extension to $t$-intervals, multigroup inference.
- Hoff and Yu (2019): Linear regression coefficients.
- Burris and Hoff (2019): Small area estimation.

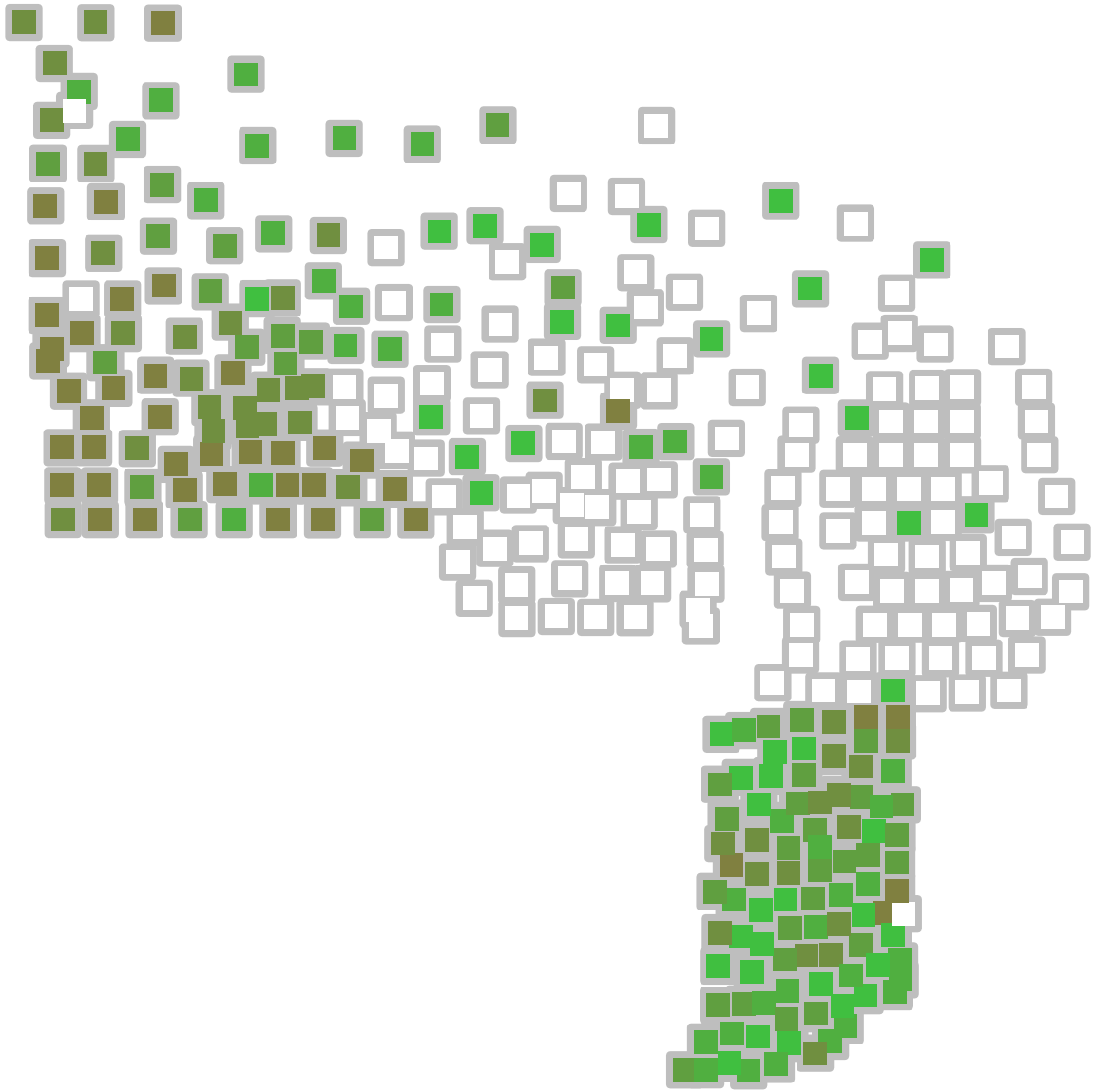
# Bayes-optimal procedure



# Radon data



# Small area estimation

Sampling model: $\bar{y}_j \sim N(\theta_j, \sigma_j^2)$ independently across groups.

Linking Model: $\theta_j = \boldsymbol{\beta}^\top \mathbf{x}_j + e_j$ , $\text{Cov}[\boldsymbol{\theta}] = \Sigma$ (spatial FH model).

Direct interval: $\bar{y}_j \pm \hat{\sigma}_j t_{1-\alpha/2}$

AFAB interval: For each area $j = 1, \ldots, p$

1. using areas other than $j$, obtain estimates of $\boldsymbol{\theta}_{-j}$, $\boldsymbol{\beta}$ and $\Sigma$;
2. obtain "prior" distribution for $\theta_j$ from estimates and working model;
3. compute optimal $w$-function and construct FAB interval for $\theta_j$.

- Both intervals have $1 - \alpha$ area-specific coverage, under random sampling within each area.The linking model need not be correct.
- FAB intervals make use of information from neighboring areas and known area-level characteristics (surficial radium).

*By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong*.

# Small area estimation

Sampling model: $\bar{y}_j \sim N(\theta_j, \sigma_j^2)$ independently across groups.

Linking Model: $\theta_j = \boldsymbol{\beta}^\top \mathbf{x}_j + e_j$ , $\text{Cov}[\boldsymbol{\theta}] = \Sigma$ (spatial FH model).

Direct interval: $\bar{y}_j \pm \hat{\sigma}_j t_{1-\alpha/2}$

AFAB interval: For each area $j = 1, \ldots, p$

1. using areas other than $j$, obtain estimates of $\boldsymbol{\theta}_{-j}$, $\boldsymbol{\beta}$ and $\Sigma$;
2. obtain "prior" distribution for $\theta_j$ from estimates and working model;
3. compute optimal $w$-function and construct FAB interval for $\theta_j$.

- Both intervals have $1 - \alpha$ area-specific coverage, under random sampling within each area.The linking model need not be correct.
- FAB intervals make use of information from neighboring areas and known area-level characteristics (surficial radium).

*By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong*.

# Small area estimation

Sampling model: $\bar{y}_j \sim N(\theta_j, \sigma_j^2)$ independently across groups.

Linking Model: $\theta_j = \boldsymbol{\beta}^\top \mathbf{x}_j + e_j$ , $\text{Cov}[\boldsymbol{\theta}] = \Sigma$ (spatial FH model).

Direct interval: $\bar{y}_j \pm \hat{\sigma}_j t_{1-\alpha/2}$

AFAB interval: For each area $j = 1, \ldots, p$

1. using areas other than $j$, obtain estimates of $\boldsymbol{\theta}_{-j}$, $\boldsymbol{\beta}$ and $\Sigma$;
2. obtain "prior" distribution for $\theta_j$ from estimates and working model;
3. compute optimal $w$-function and construct FAB interval for $\theta_j$.

- Both intervals have $1 - \alpha$ area-specific coverage, under random sampling within each area.The linking model need not be correct.
- FAB intervals make use of information from neighboring areas and known area-level characteristics (surficial radium).

*By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong.*

# Small area estimation

Sampling model: $\bar{y}_j \sim N(\theta_j, \sigma_j^2)$ independently across groups.

Linking Model: $\theta_j = \boldsymbol{\beta}^\top \mathbf{x}_j + e_j$ , $\text{Cov}[\boldsymbol{\theta}] = \Sigma$ (spatial FH model).

Direct interval: $\bar{y}_j \pm \hat{\sigma}_j t_{1-\alpha/2}$

AFAB interval: For each area $j = 1, \ldots, p$

1. using areas other than $j$, obtain estimates of $\boldsymbol{\theta}_{-j}$, $\boldsymbol{\beta}$ and $\Sigma$;
2. obtain "prior" distribution for $\theta_j$ from estimates and working model;
3. compute optimal $w$-function and construct FAB interval for $\theta_j$.

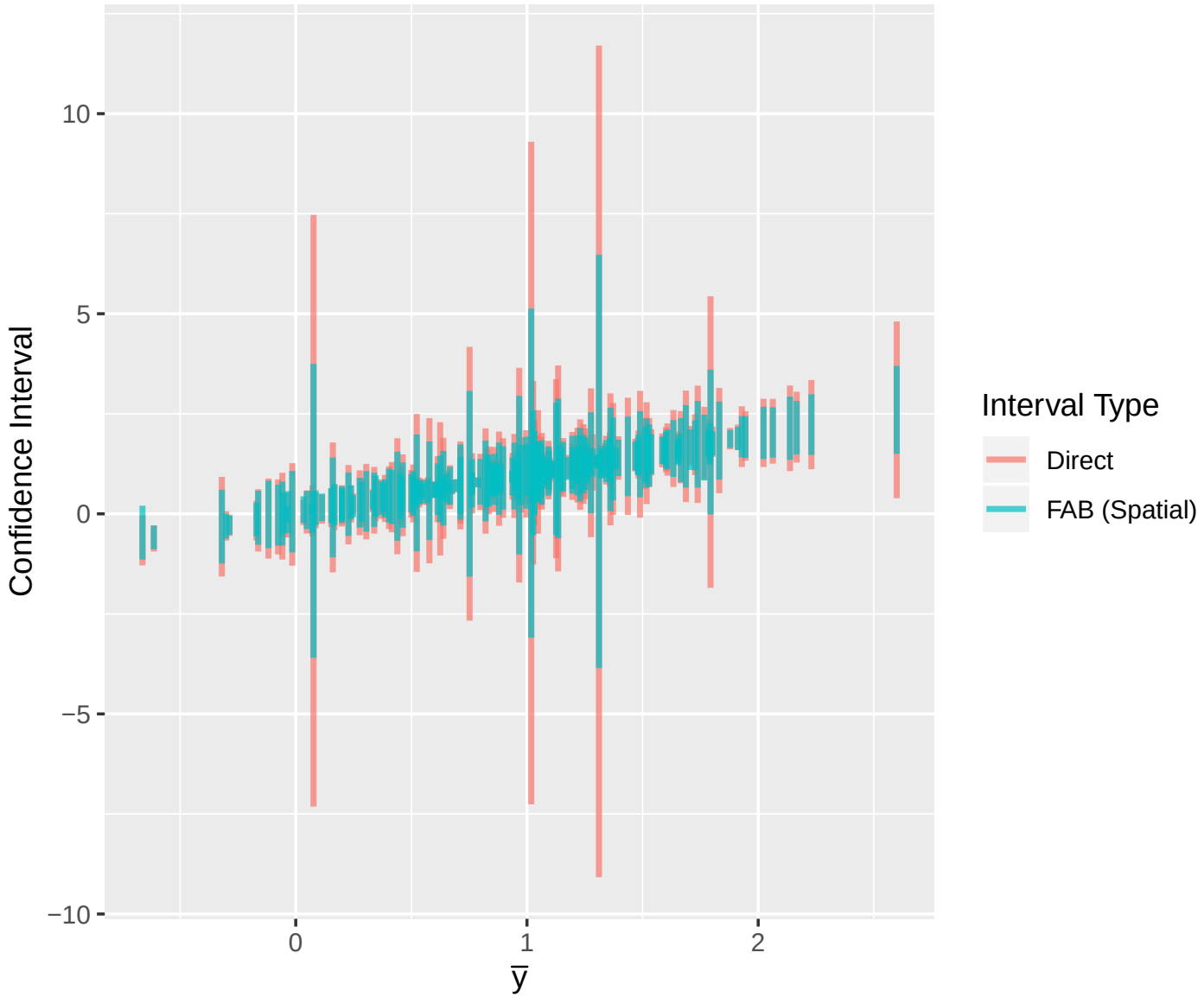
- Both intervals have $1 - \alpha$ area-specific coverage, under random sampling within each area.The linking model need not be correct.
- FAB intervals make use of information from neighboring areas and known area-level characteristics (surficial radium).

*By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong*.

# Interval comparisons

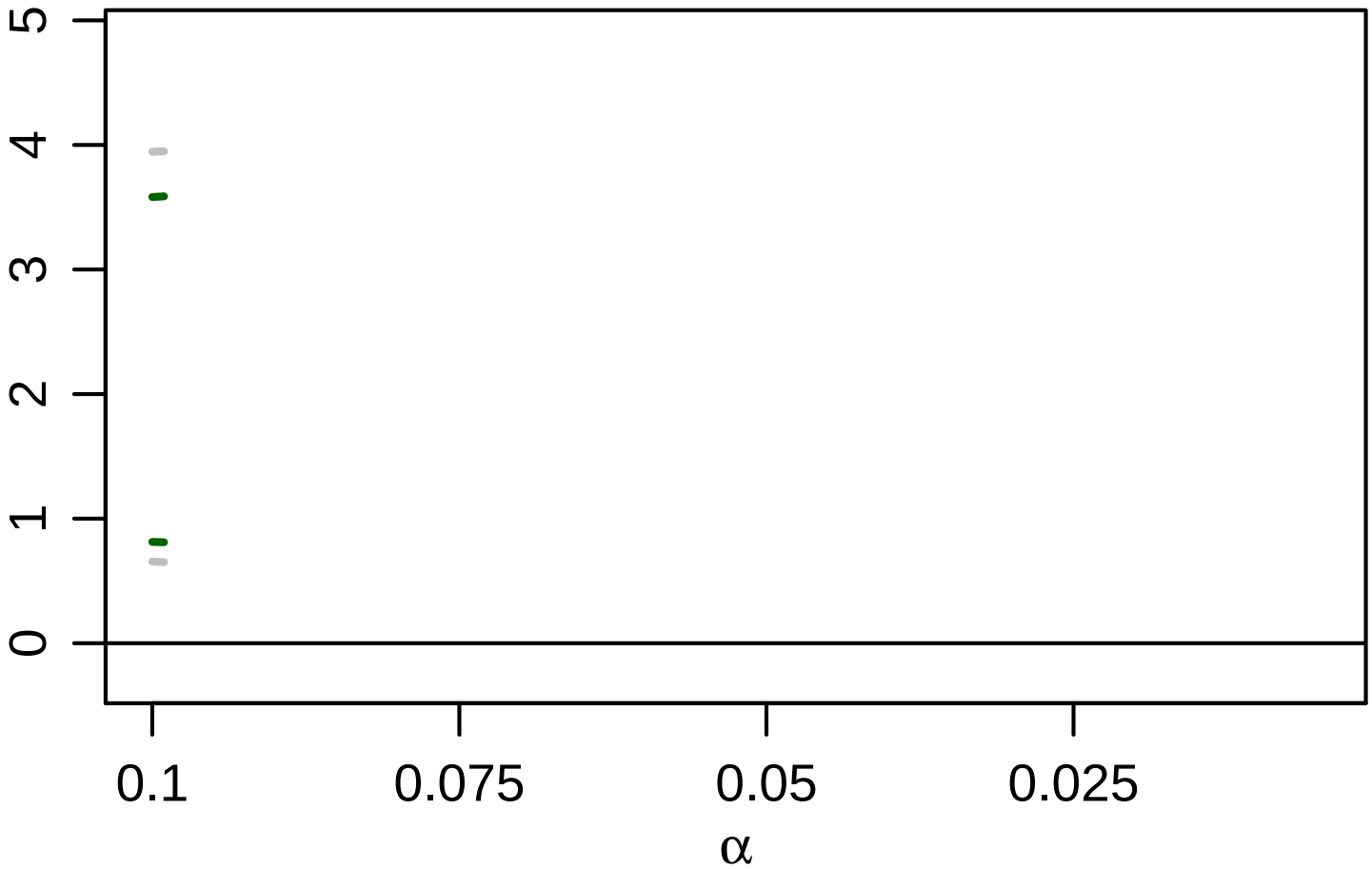
| Type | Hierarchical model | relative width | fraction intervals improved |
|---|---|---|---|
| Direct | - | 1.0 | - |
| FAB | exchangeable | .77 | .898 |
| FAB | covariate | .77 | .888 |
| FAB | spatial | .74 | .964 |
| FAB | spatial, covariate | .74 | .955 |

# Interval comparisons



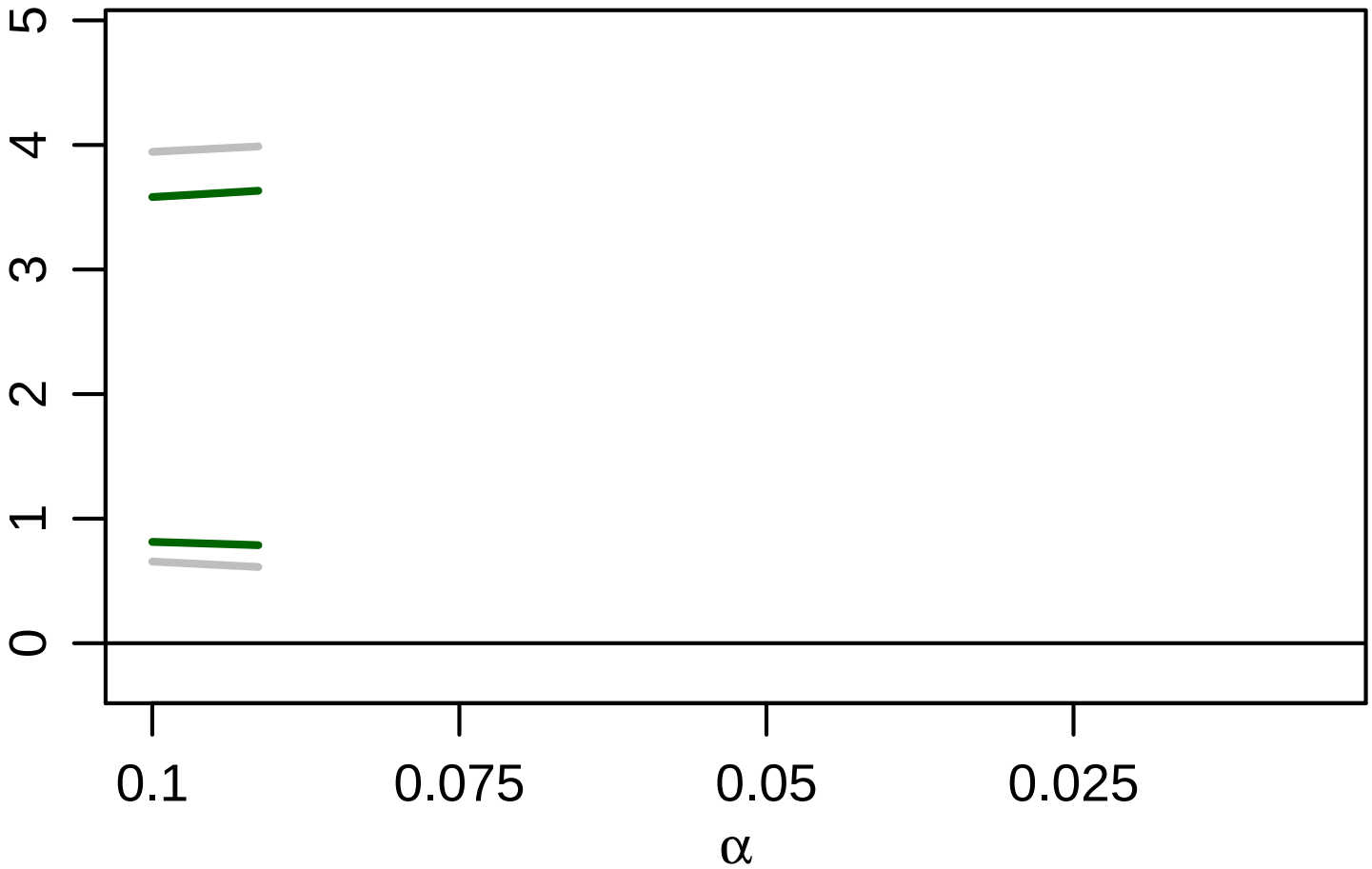
# FAB $p$-values

$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$$



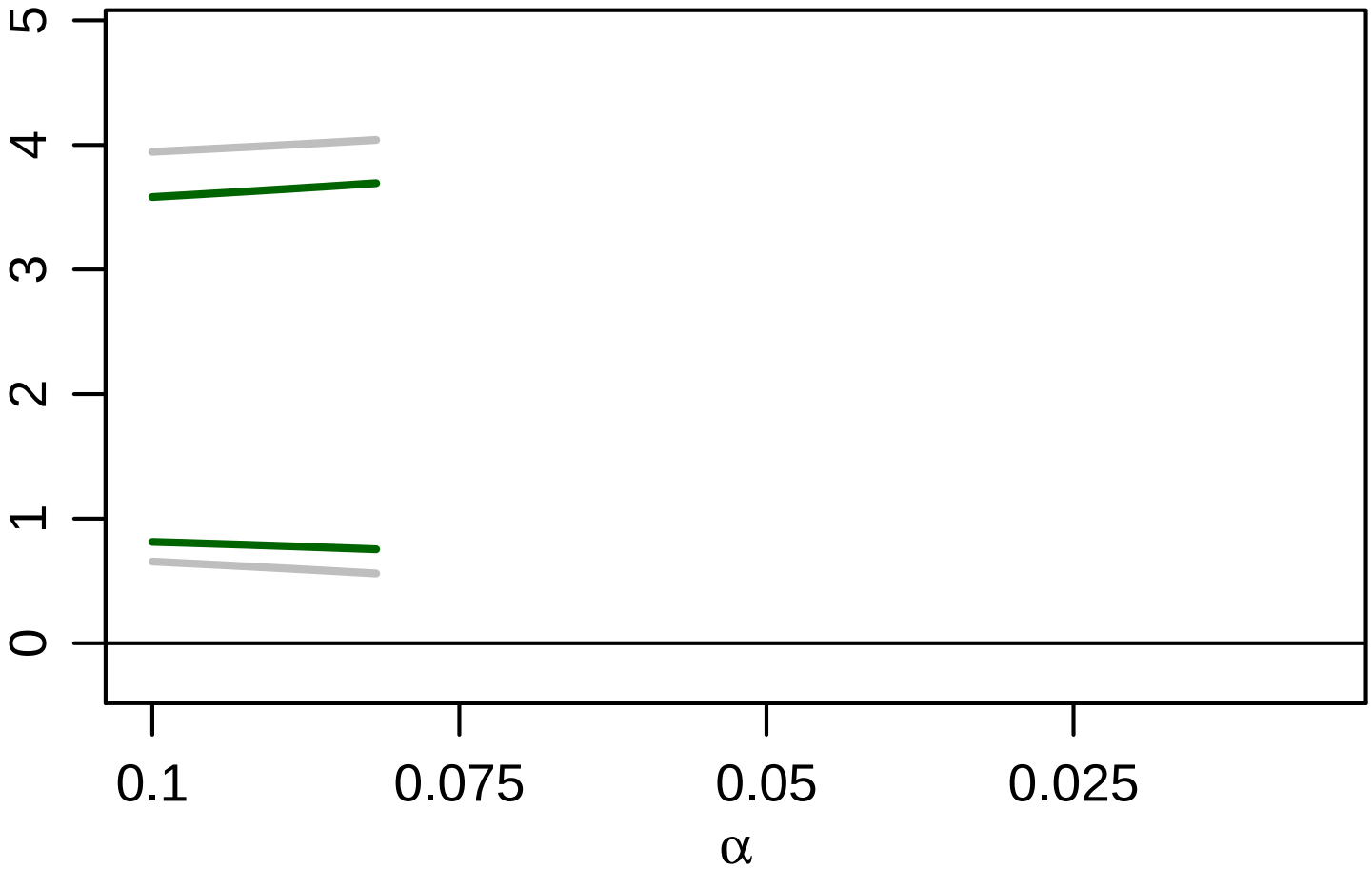
# FAB $p$-values

$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$$



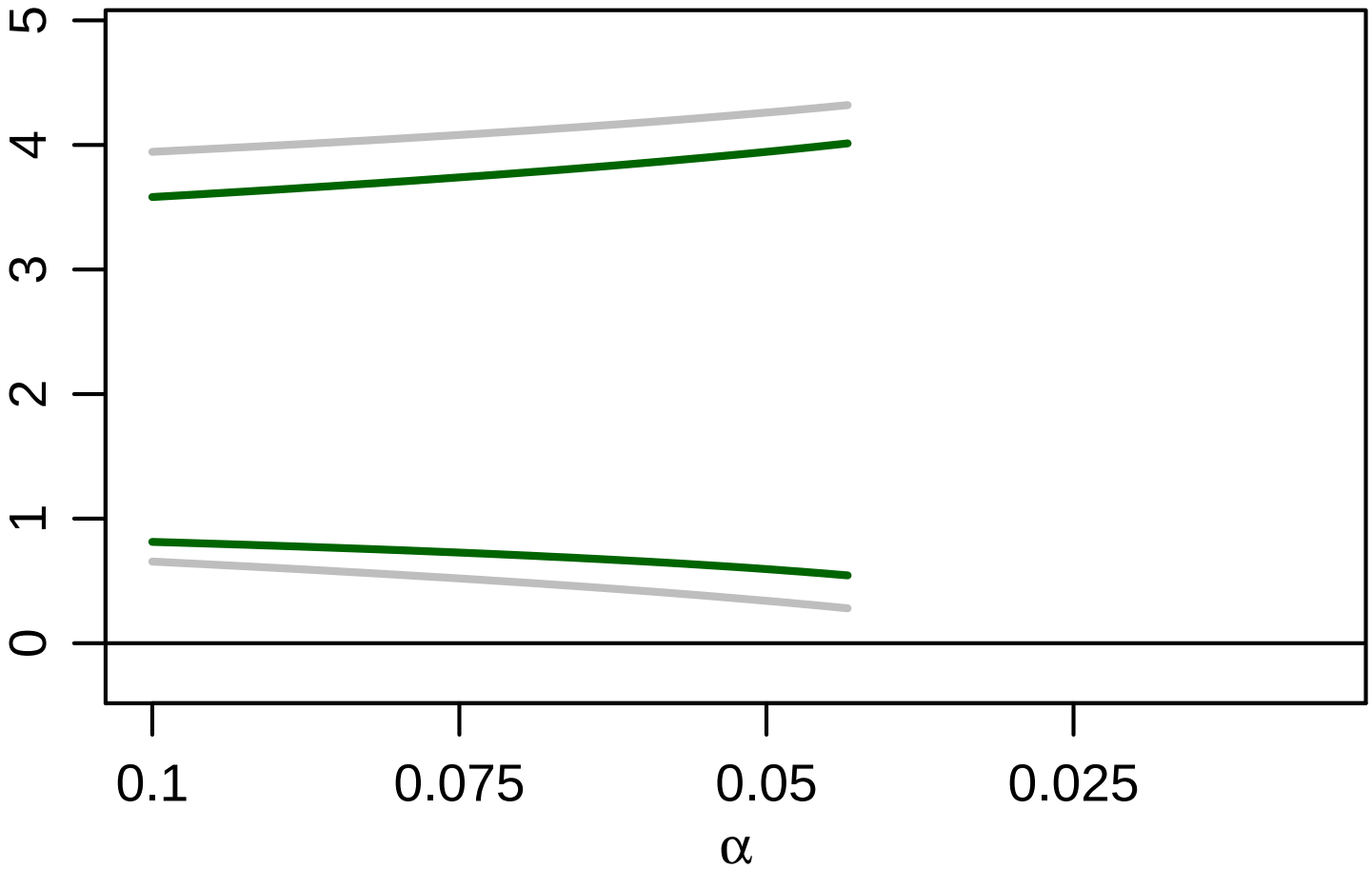
# FAB $p$-values

$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$$



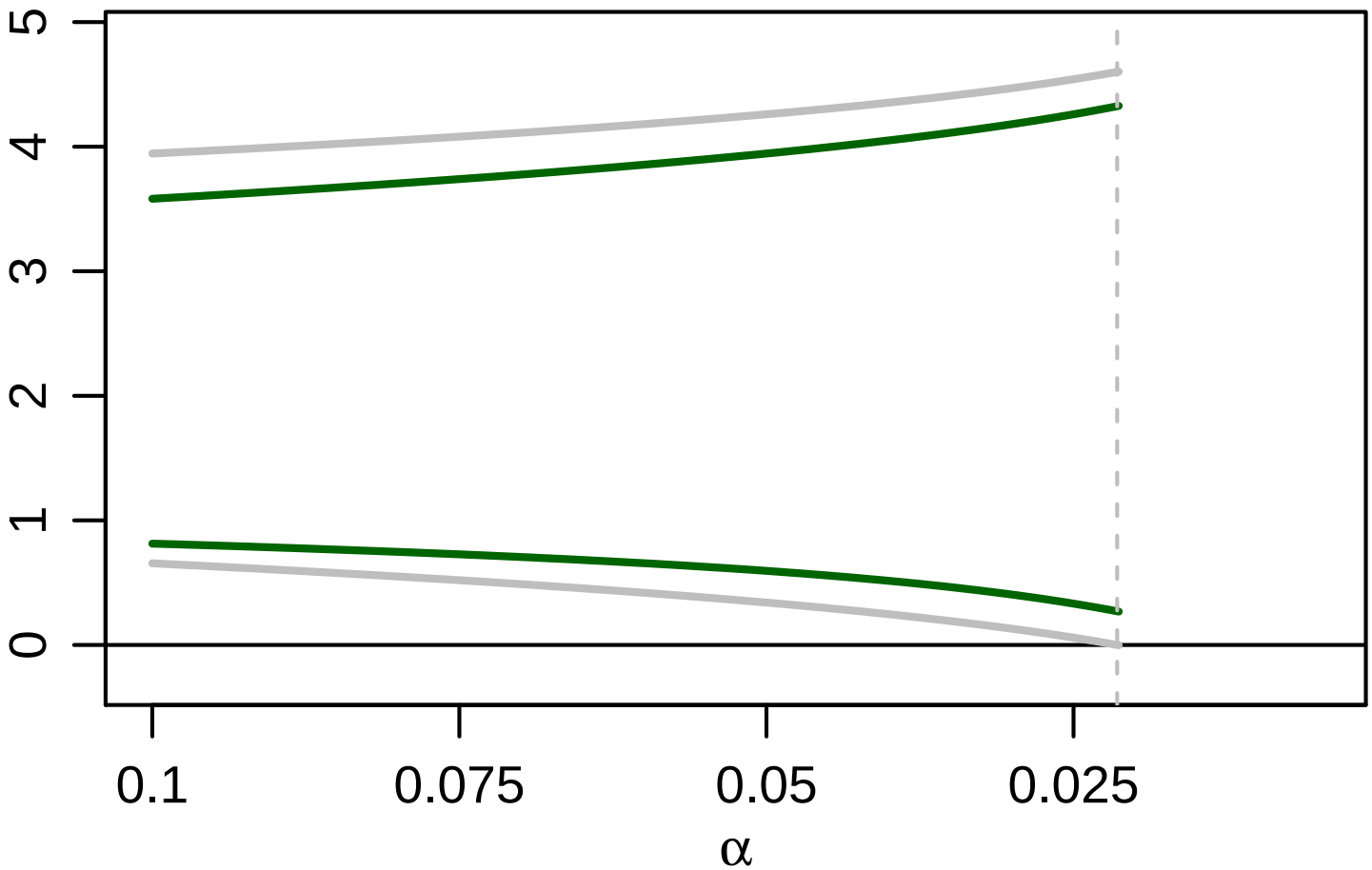
# FAB $p$-values

$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$$



# FAB $p$-values

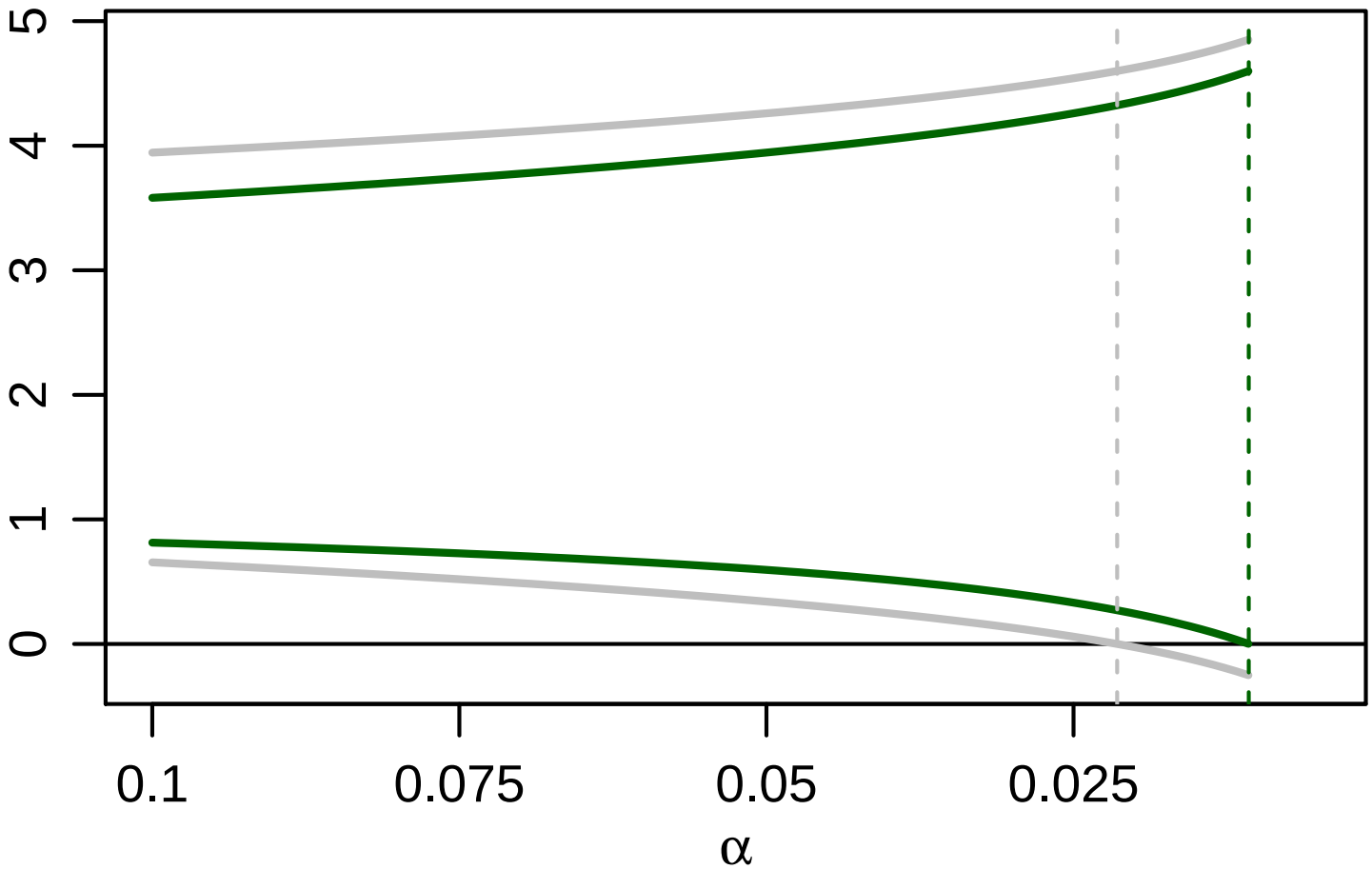
$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$




$p_U = 0.021$

# FAB $p$-values

$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$




$p_U = 0.021 \quad p_F = 0.011$

# FAB $p$-value function

**FAB CI endpoints:**

$$\theta^U(\alpha) = \frac{y + \sigma\Phi^{-1}\{1-\alpha+\Phi(\frac{y-\theta^U}{\sigma})\}}{1+2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1+2\sigma^2/\tau^2},$$

$$\theta^L(\alpha) = \frac{y + \sigma\Phi^{-1}\{\alpha-\Phi(\frac{\theta^L-y}{\sigma})\}}{1+2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1+2\sigma^2/\tau^2}.$$

**FAB $p$-value:**

$$p = \min\{\alpha : \theta^U(\alpha) \times \theta^U(\alpha) > 0\}$$

**FAB $p$-value function:**

$$p_F(y) = 1 - |\Phi(y/\sigma + 2\mu\tfrac{\sigma}{\tau^2}) - \Phi(-y/\sigma)|$$

# FAB $p$-value function

**FAB CI endpoints:**

$$\theta^U(\alpha) = \frac{y + \sigma\Phi^{-1}\{1-\alpha+\Phi(\frac{y-\theta^U}{\sigma})\}}{1+2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1+2\sigma^2/\tau^2},$$

$$\theta^L(\alpha) = \frac{y + \sigma\Phi^{-1}\{\alpha-\Phi(\frac{\theta^L-y}{\sigma})\}}{1+2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1+2\sigma^2/\tau^2}.$$

**FAB $p$-value:**

$$p = \min\{\alpha : \theta^U(\alpha) \times \theta^U(\alpha) > 0\}$$

**FAB $p$-value function:**

$$p_F(y) = 1 - |\Phi(y/\sigma + 2\mu\tfrac{\sigma}{\tau^2}) - \Phi(-y/\sigma)|$$

# FAB $p$-value function

**FAB CI endpoints:**

$$\theta^U(\alpha) = \frac{y + \sigma\Phi^{-1}\{1-\alpha+\Phi(\frac{y-\theta^U}{\sigma})\}}{1+2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1+2\sigma^2/\tau^2},$$

$$\theta^L(\alpha) = \frac{y + \sigma\Phi^{-1}\{\alpha-\Phi(\frac{\theta^L-y}{\sigma})\}}{1+2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1+2\sigma^2/\tau^2}.$$

**FAB $p$-value:**

$$p = \min\{\alpha : \theta^U(\alpha) \times \theta^U(\alpha) > 0\}$$

**FAB $p$-value function:**

$$p_F(y) = 1 - |\Phi(y/\sigma + 2\mu\tfrac{\sigma}{\tau^2}) - \Phi(-y/\sigma)|$$

# More generally

**Thm.** $1 - |F(Z + b) - F(-Z)| \sim U[0, 1]$ if

- $Z \sim F$, $F$ symmetric about zero;
- $b$ is constant or otherwise independent of $Z$.

**Cor.** $1 - |\Phi(Y/\sigma + b) - \Phi(-Y/\sigma)| \sim U[0, 1]$ if

- $Y \sim N(0, \sigma^2)$;
- $b$ is constant or otherwise independent of $Y$.

**Bayes optimal test**

- $Y \sim N(\theta, \sigma^2)$;
- $H : \theta = 0$ versus $K : \theta \neq 0$;
- prior $\theta \sim N(\mu, \tau^2)$.

Thm. The $p$-value associated with the Bayes-optimal tests is given by

$$p_F(Y) = 1 - |\Phi(Y/\sigma + 2\mu\tfrac{\sigma}{\tau^2}) - \Phi(-Y/\sigma)|.$$

# More generally

**Thm.** $1 - |F(Z + b) - F(-Z)| \sim U[0, 1]$ if

- $Z \sim F$, $F$ symmetric about zero;
- $b$ is constant or otherwise independent of $Z$.

**Cor.** $1 - |\Phi(Y/\sigma + b) - \Phi(-Y/\sigma)| \sim U[0, 1]$ if

- $Y \sim N(0, \sigma^2)$;
- $b$ is constant or otherwise independent of $Y$.

**Bayes optimal test**

- $Y \sim N(\theta, \sigma^2)$;
- $H : \theta = 0$ versus $K : \theta \neq 0$;
- prior $\theta \sim N(\mu, \tau^2)$.

Thm. The $p$-value associated with the Bayes-optimal tests is given by

$$p_F(Y) = 1 - |\Phi(Y/\sigma + 2\mu\tfrac{\sigma}{\tau^2}) - \Phi(-Y/\sigma)|.$$

# More generally

**Thm.** $1 - |F(Z + b) - F(-Z)| \sim U[0, 1]$ if

- $Z \sim F$, $F$ symmetric about zero;
- $b$ is constant or otherwise independent of $Z$.

**Cor.** $1 - |\Phi(Y/\sigma + b) - \Phi(-Y/\sigma)| \sim U[0, 1]$ if

- $Y \sim N(0, \sigma^2)$;
- $b$ is constant or otherwise independent of $Y$.

**Bayes optimal test**

- $Y \sim N(\theta, \sigma^2)$;
- $H : \theta = 0$ versus $K : \theta \neq 0$;
- prior $\theta \sim N(\mu, \tau^2)$.

Thm. The $p$-value associated with the Bayes-optimal tests is given by

$$p_F(Y) = 1 - |\Phi(Y/\sigma + 2\mu\tfrac{\sigma}{\tau^2}) - \Phi(-Y/\sigma)|.$$

# Adaptive FAB $p$-values

**Sampling model:** $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

**Linking model:** $\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Psi})$ $\quad \mu = \mu(\gamma, \mathbf{X}),\ \boldsymbol{\Psi} = \boldsymbol{\Psi}(\gamma, \mathbf{X}).$

**Adaptive FAB $p$-value:** For each parameter $\theta_j$,

1. Construct *indirect information* $\mathbf{G}_j^\top \mathbf{Y}$ so that $Y_j \perp (\mathbf{G}_j^\top \mathbf{Y})$.
2. Find $(\mu_j, \tau_j^2) = (E[\theta_j | \mathbf{G}_j^\top \mathbf{Y}]\ ,\ V[\theta_j | \mathbf{G}_j^\top \mathbf{Y}])$.
3. Compute FAB $p$-value $p_F(Y_j) = 1 - |\Phi(Y_j/\sigma_j + 2\mu_j \frac{\sigma_j}{\tau_j^2}) - \Phi(-Y_j/\sigma_j)|$.

- If $\theta_j = 0$ then $p_F(Y_j) \sim U[0,1]$.
- Exact $p$-values if $\mathrm{Cor}[Y]$ is available (independence, or linear regression).

# Adaptive FAB $p$-values

**Sampling model:** $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

**Linking model:** $\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Psi}) \quad \boldsymbol{\mu} = \boldsymbol{\mu}(\gamma, \mathbf{X}),\ \boldsymbol{\Psi} = \boldsymbol{\Psi}(\gamma, \mathbf{X}).$

**Adaptive FAB $p$-value:** For each parameter $\theta_j$,

1. Construct *indirect information* $\mathbf{G}_j^\top \mathbf{Y}$ so that $Y_j \perp (\mathbf{G}_j^\top \mathbf{Y})$.
2. Find $(\mu_j, \tau_j^2) = (E[\theta_j | \mathbf{G}_j^\top \mathbf{Y}]\ ,\ V[\theta_j | \mathbf{G}_j^\top \mathbf{Y}])$.
3. Compute FAB $p$-value $p_F(Y_j) = 1 - |\Phi(Y_j/\sigma_j + 2\mu_j \frac{\sigma_j}{\tau_j^2}) - \Phi(-Y_j/\sigma_j)|$.

- If $\theta_j = 0$ then $p_F(Y_j) \sim U[0,1]$.
- Exact $p$-values if $\mathrm{Cor}[Y]$ is available (independence, or linear regression).

# Adaptive FAB $p$-values

**Sampling model:** $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

**Linking model:** $\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Psi}) \quad \boldsymbol{\mu} = \boldsymbol{\mu}(\gamma, \mathbf{X}),\ \boldsymbol{\Psi} = \boldsymbol{\Psi}(\gamma, \mathbf{X}).$

**Adaptive FAB $p$-value:** For each parameter $\theta_j$,

1. Construct *indirect information* $\mathbf{G}_j^\top \mathbf{Y}$ so that $Y_j \perp (\mathbf{G}_j^\top \mathbf{Y})$.
2. Find $(\mu_j, \tau_j^2) = (E[\theta_j | \mathbf{G}_j^\top \mathbf{Y}]\ ,\ V[\theta_j | \mathbf{G}_j^\top \mathbf{Y}])$.
3. Compute FAB $p$-value $p_F(Y_j) = 1 - |\Phi(Y_j/\sigma_j + 2\mu_j \frac{\sigma_j}{\tau_j^2}) - \Phi(-Y_j/\sigma_j)|$.

- If $\theta_j = 0$ then $p_F(Y_j) \sim U[0,1]$.
- Exact $p$-values if $\mathrm{Cor}[Y]$ is available (independence, or linear regression).

# Adaptive FAB $p$-values

**Sampling model:** $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

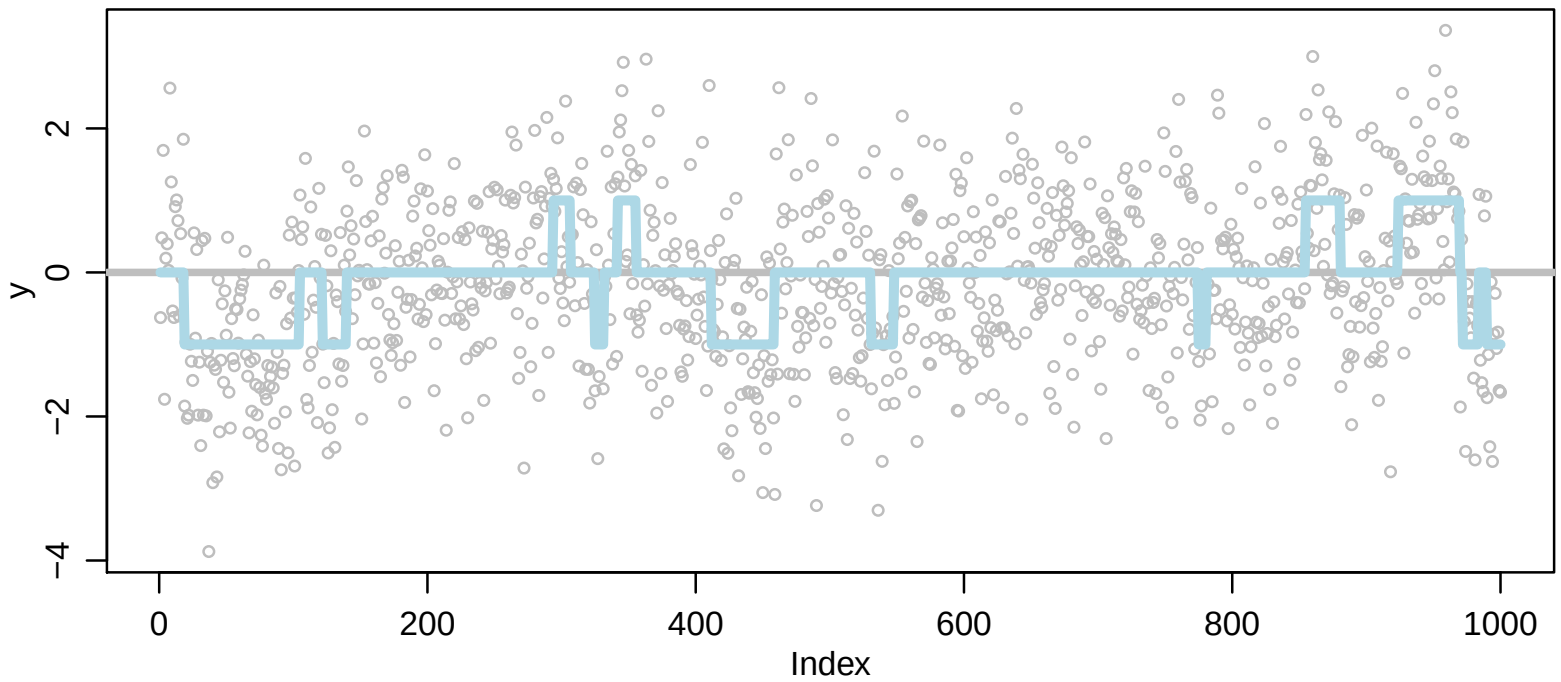
**Linking model:** $\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Psi}) \quad \boldsymbol{\mu} = \boldsymbol{\mu}(\gamma, \mathbf{X}),\ \boldsymbol{\Psi} = \boldsymbol{\Psi}(\gamma, \mathbf{X}).$

**Adaptive FAB $p$-value:** For each parameter $\theta_j$,

1. Construct *indirect information* $\mathbf{G}_j^\top \mathbf{Y}$ so that $Y_j \perp (\mathbf{G}_j^\top \mathbf{Y})$.
2. Find $(\mu_j, \tau_j^2) = (E[\theta_j | \mathbf{G}_j^\top \mathbf{Y}]\ ,\ V[\theta_j | \mathbf{G}_j^\top \mathbf{Y}])$.
3. Compute FAB $p$-value $p_F(Y_j) = 1 - |\Phi(Y_j/\sigma_j + 2\mu_j \frac{\sigma_j}{\tau_j^2}) - \Phi(-Y_j/\sigma_j)|$.

- If $\theta_j = 0$ then $p_F(Y_j) \sim U[0, 1]$.
- Exact $p$-values if $\mathrm{Cor}[Y]$ is available (independence, or linear regression).

# Example: Spatial linking model




**Truth:** $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \mathbf{I})$, $\boldsymbol{\theta} \in \{-1, 0, +1\}^p$ follows HMM.

**Assume:** $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \mathbf{I})$, $\boldsymbol{\theta} \sim N_p(\mathbf{0}, \tau^2 \boldsymbol{\Psi}(\rho))$

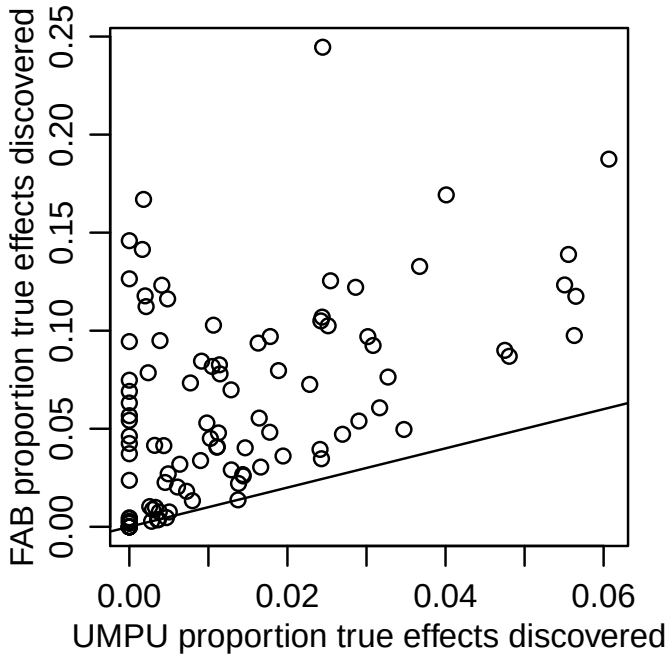
# Example: Spatial linking model

For each of 100 simulations,

1. compute $p$-value for each $\theta_j, j \in \{1, \ldots, 1000\}$;
2. compute BH threshold for FDR $< 0.2$;
3. record number of true, false discoveries.

**Results**

| | | |
|---|---|---|
| UMPU | FDP=0.085 | TDP = 0.013 |
| FAB | FDP=0.088 | TDP = 0.061 |

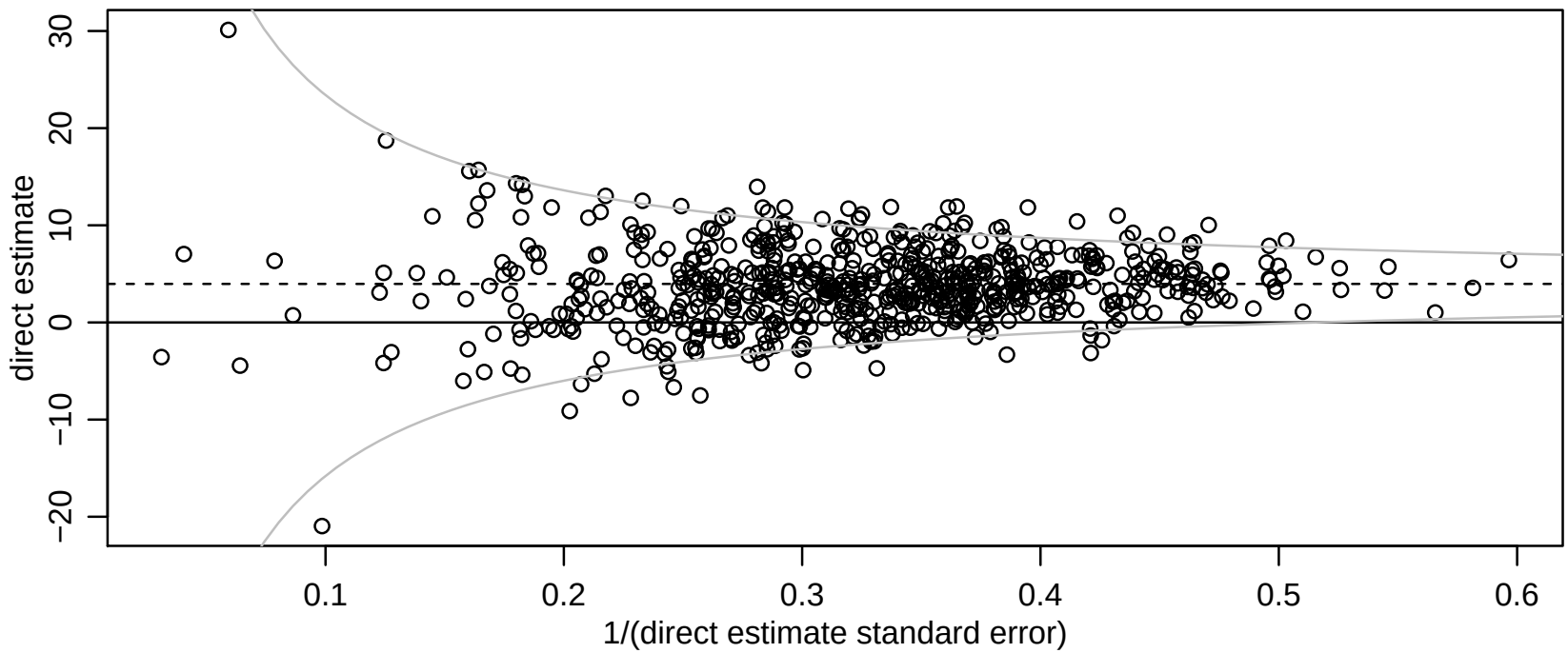


# Example: Interactions in linear models

**Educational Longitudinal Study:**

- Student data and reading scores from $p = 684$ schools.
- Evaluate relationship between SES and reading score in each school.

**Sampling model:** $y_{i,j} = \mu_j + \boldsymbol{\alpha}^\top \mathbf{w}_i + \beta_j \times x_{i,j} + \epsilon_{i,j}$



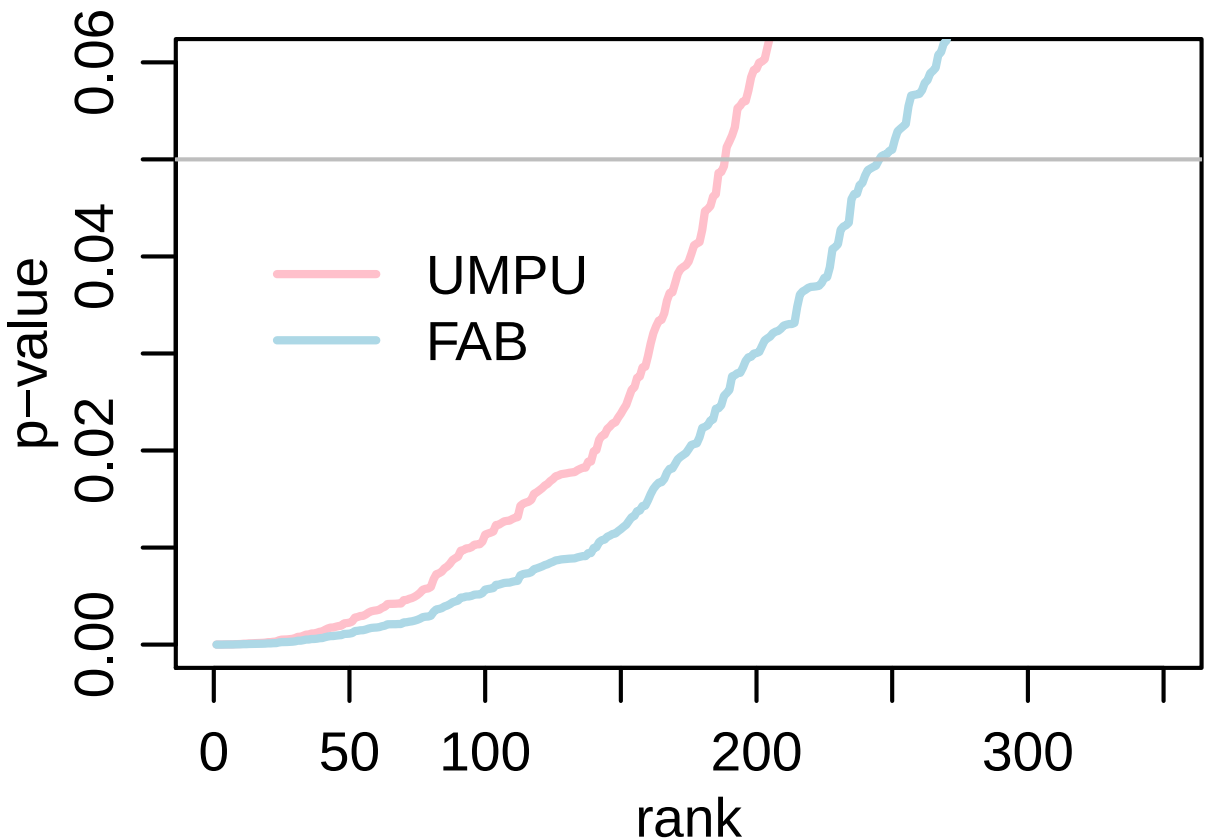
# Example: Interactions in linear models

**Adaptive FAB $p$-values:** $\hat{\boldsymbol{\beta}}_{OLS} \sim N_p(\boldsymbol{\beta}, \sigma^2 \boldsymbol{V})$, $\boldsymbol{V}$ is known.

1. Construct *indirect information* $\mathbf{G}_j^\top \hat{\boldsymbol{\beta}}$ so that $\hat{\beta}_j \perp (\mathbf{G}_j^\top \hat{\boldsymbol{\beta}})$.
2. Obtain $\hat{\mu}$, $\hat{\tau}^2$ from $\mathbf{G}_j^\top \hat{\boldsymbol{\beta}}$ and linking model $\boldsymbol{\beta} \sim N_p(\mu \mathbf{1}, \tau^2 \mathbf{I})$.
3. Compute FAB $p$-value $p_F(\hat{\beta}_j) = 1 - |F(\hat{\beta}_j/s_j + 2\hat{\mu}\tilde{s}_j/\hat{\tau}^2) - F(-\hat{\beta}_j/s_j)|$.

**Results:**

245 and 188 FAB and UMPU $p$-values less than 0.05.



# Summary

Group-level inference motivates group-level error rate guarantees.

Such error rate guarantees do not preclude inclusion of indirect information.

FAB inference

- maintains group-level error rates;
- improves performance on-average across groups;
- rates maintained even if the linking model is wrong.

Extensions and future work

- non-Guassian sampling and linking models;
- selection adjusted inference;
- FDR control.

# Summary

Group-level inference motivates group-level error rate guarantees.

Such error rate guarantees do not preclude inclusion of indirect information.

FAB inference

- maintains group-level error rates;
- improves performance on-average across groups;
- rates maintained even if the linking model is wrong.

Extensions and future work

- non-Guassian sampling and linking models;
- selection adjusted inference;
- FDR control.

# Summary

Group-level inference motivates group-level error rate guarantees.

Such error rate guarantees do not preclude inclusion of indirect information.

FAB inference

- maintains group-level error rates;
- improves performance on-average across groups;
- rates maintained even if the linking model is wrong.

Extensions and future work

- non-Guassian sampling and linking models;
- selection adjusted inference;
- FDR control.

# Summary

Group-level inference motivates group-level error rate guarantees.

Such error rate guarantees do not preclude inclusion of indirect information.

FAB inference

- maintains group-level error rates;
- improves performance on-average across groups;
- rates maintained even if the linking model is wrong.

Extensions and future work

- non-Guassian sampling and linking models;
- selection adjusted inference;
- FDR control.