# Why we (usually) don't worry about multiple comparisons

Andrew Gelman, Jennifer Hill, and Masanao Yajima
Columbia University

8 Nov 2007

# Contents

# Contents

# Contents

# Contents

# Contents

# What is the multiple comparisons problem?

- Even if nothing is going on, you can find things
  - Data snooping
  - Overwhelmed by data and plausible "findings"
- "If not accounted for, false positive differences are very likely to be identified": 5% of our 95% intervals will be wrong

# What is the multiple comparisons problem?

- Even if nothing is going on, you can find things
  - Data snooping
  - Overwhelmed by data and plausible "findings"
- "If not accounted for, false positive differences are very likely to be identified": 5% of our 95% intervals will be wrong

# What is the multiple comparisons problem?

- Even if nothing is going on, you can find things
  - Data snooping
  - Overwhelmed by data and plausible "findings"
- "If not accounted for, false positive differences are very likely to be identified": 5% of our 95% intervals will be wrong

# What is the multiple comparisons problem?

- Even if nothing is going on, you can find things
  - Data snooping
  - Overwhelmed by data and plausible "findings"
- "If not accounted for, false positive differences are very likely to be identified": 5% of our 95% intervals will be wrong

# What is the multiple comparisons problem?

- Even if nothing is going on, you can find things
  - Data snooping
  - Overwhelmed by data and plausible "findings"
- "If not accounted for, false positive differences are very likely to be identified": 5% of our 95% intervals will be wrong

# Why don't I (usually) care about multiple comparisons?

- When looked at one way, multiple comparisons seem like a major worry
- But from another perspective, they don't matter at all:
  - I don't (usually) study study phenomena with zero effects
  - I don't (usually) study comparisons with zero differences
  - I don't mind being wrong 5% of the time

# Why don't I (usually) care about multiple comparisons?

- When looked at one way, multiple comparisons seem like a major worry
- But from another perspective, they don't matter at all:
  - I don't (usually) study study phenomena with zero effects
  - I don't (usually) study comparisons with zero differences
  - I don't mind being wrong 5% of the time

# Why don't I (usually) care about multiple comparisons?

- When looked at one way, multiple comparisons seem like a major worry
- But from another perspective, they don't matter at all:
  - I don't (usually) study study phenomena with zero effects
  - I don't (usually) study comparisons with zero differences
  - I don't mind being wrong 5% of the time

# Why don't I (usually) care about multiple comparisons?

- When looked at one way, multiple comparisons seem like a major worry
- But from another perspective, they don't matter at all:
  - I don't (usually) study study phenomena with zero effects
  - I don't (usually) study comparisons with zero differences
  - I don't mind being wrong 5% of the time

# Why don't I (usually) care about multiple comparisons?

- When looked at one way, multiple comparisons seem like a major worry
- But from another perspective, they don't matter at all:
  - I don't (usually) study study phenomena with zero effects
  - I don't (usually) study comparisons with zero differences
  - I don't mind being wrong 5% of the time

# Why don't I (usually) care about multiple comparisons?

- When looked at one way, multiple comparisons seem like a major worry
- But from another perspective, they don't matter at all:
  - I don't (usually) study study phenomena with zero effects
  - I don't (usually) study comparisons with zero differences
  - I don't mind being wrong 5% of the time

# Some stories

- SAT coaching in 8 schools
- Effects of electromagnetic fields at 38 frequencies
- Teacher and school effects in NYC schools
- Grades and classroom seating
- Beautiful parents have more daughters
- Comparing test scores across states

# Some stories

- SAT coaching in 8 schools
- Effects of electromagnetic fields at 38 frequencies
- Teacher and school effects in NYC schools
- Grades and classroom seating
- Beautiful parents have more daughters
- Comparing test scores across states

# Some stories

- SAT coaching in 8 schools
- Effects of electromagnetic fields at 38 frequencies
- Teacher and school effects in NYC schools
- Grades and classroom seating
- Beautiful parents have more daughters
- Comparing test scores across states

# Some stories

- SAT coaching in 8 schools
- Effects of electromagnetic fields at 38 frequencies
- Teacher and school effects in NYC schools
- Grades and classroom seating
- Beautiful parents have more daughters
- Comparing test scores across states

# Some stories

- SAT coaching in 8 schools
- Effects of electromagnetic fields at 38 frequencies
- Teacher and school effects in NYC schools
- Grades and classroom seating
- Beautiful parents have more daughters
- Comparing test scores across states

# Some stories

- SAT coaching in 8 schools
- Effects of electromagnetic fields at 38 frequencies
- Teacher and school effects in NYC schools
- Grades and classroom seating
- Beautiful parents have more daughters
- Comparing test scores across states

# Some stories

- SAT coaching in 8 schools
- Effects of electromagnetic fields at 38 frequencies
- Teacher and school effects in NYC schools
- Grades and classroom seating
- Beautiful parents have more daughters
- Comparing test scores across states

# SAT coaching in 8 schools

| School | Estimated treatment effect, $y_j$ | Standard error of effect estimate, $\sigma_j$ |
|---|---|---|
| A | 28 | 15 |
| B | 8 | 10 |
| C | −3 | 16 |
| D | 7 | 11 |
| E | −1 | 9 |
| F | 1 | 11 |
| G | 18 | 10 |
| H | 12 | 18 |

- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis

# SAT coaching in 8 schools

| School | Estimated treatment effect, $y_j$ | Standard error of effect estimate, $\sigma_j$ |
|---|---|---|
| A | 28 | 15 |
| B | 8 | 10 |
| C | −3 | 16 |
| D | 7 | 11 |
| E | −1 | 9 |
| F | 1 | 11 |
| G | 18 | 10 |
| H | 12 | 18 |

- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
  - Overlapping confidence intervals for the 8 schools
  - Statements such as: Effect in A is probably [illegible]

# SAT coaching in 8 schools

| School | Estimated treatment effect, $y_j$ | Standard error of effect estimate, $\sigma_j$ |
|---|---|---|
| A | 28 | 15 |
| B | 8 | 10 |
| C | −3 | 16 |
| D | 7 | 11 |
| E | −1 | 9 |
| F | 1 | 11 |
| G | 18 | 10 |
| H | 12 | 18 |

- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
  - Overlapping confidence intervals for the 8 school effects
  - Statements such as Pr (effect in A > effect in C) = 0.7

# SAT coaching in 8 schools

| School | Estimated treatment effect, $y_j$ | Standard error of effect estimate, $\sigma_j$ |
|---|---|---|
| A | 28 | 15 |
| B | 8 | 10 |
| C | −3 | 16 |
| D | 7 | 11 |
| E | −1 | 9 |
| F | 1 | 11 |
| G | 18 | 10 |
| H | 12 | 18 |

- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
  - Overlappling confidence intervals for the 8 school effects
  - Statements such as Pr (effect in A > effect in C)= 0.7

# SAT coaching in 8 schools

| School | Estimated treatment effect, $y_j$ | Standard error of effect estimate, $\sigma_j$ |
|---|---|---|
| A | 28 | 15 |
| B | 8 | 10 |
| C | −3 | 16 |
| D | 7 | 11 |
| E | −1 | 9 |
| F | 1 | 11 |
| G | 18 | 10 |
| H | 12 | 18 |

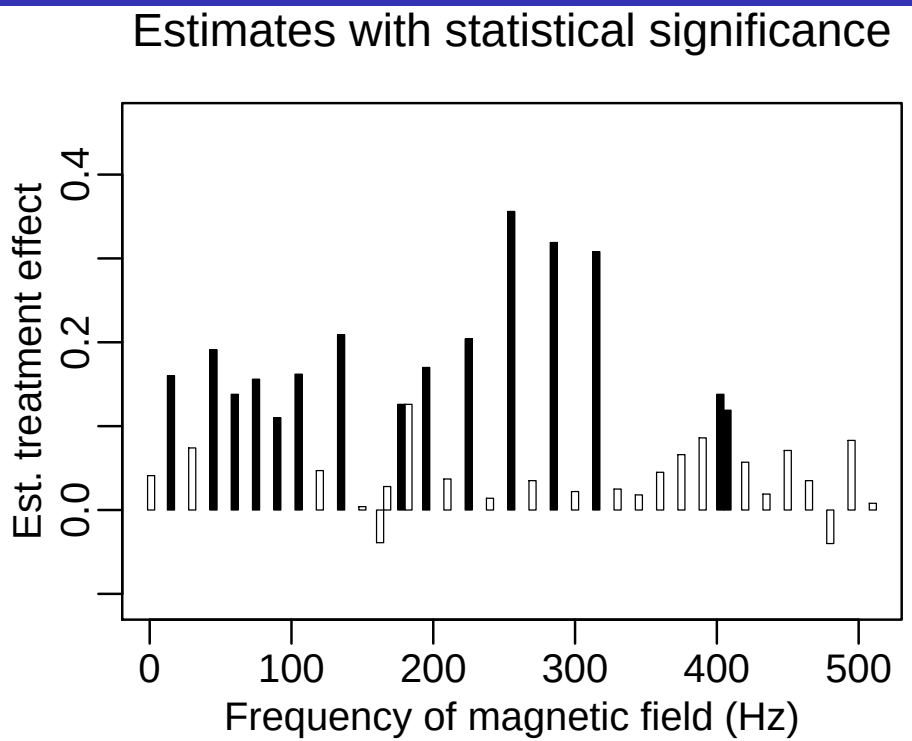
- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
  - Overlappling confidence intervals for the 8 school effects
  - Statements such as Pr (effect in A > effect in C)= 0.7

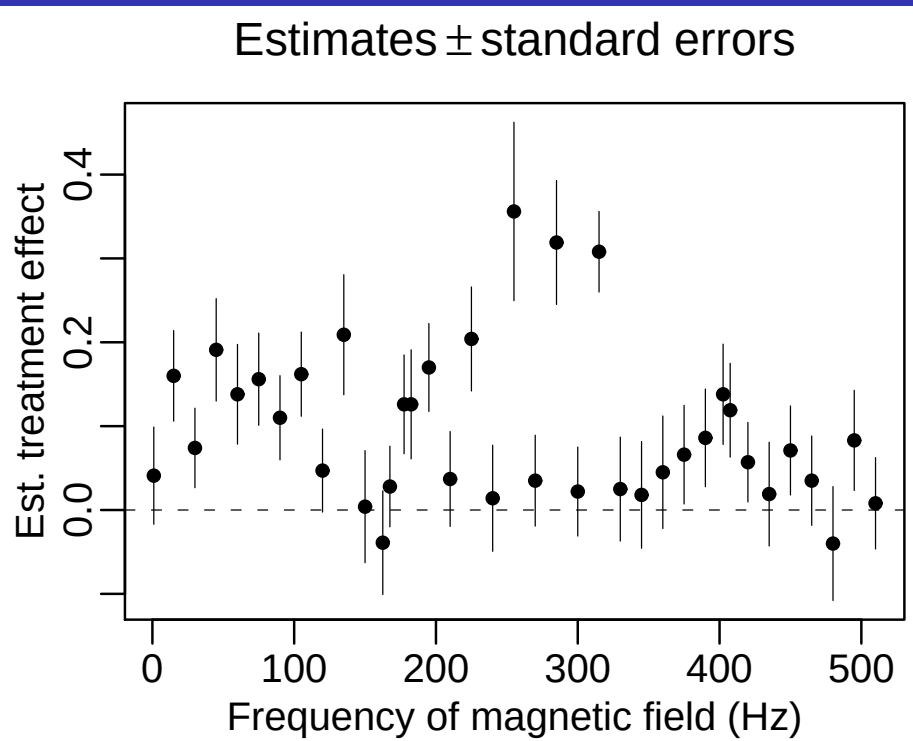
# SAT coaching in 8 schools

| School | Estimated treatment effect, $y_j$ | Standard error of effect estimate, $\sigma_j$ |
|---|---|---|
| A | 28 | 15 |
| B | 8 | 10 |
| C | −3 | 16 |
| D | 7 | 11 |
| E | −1 | 9 |
| F | 1 | 11 |
| G | 18 | 10 |
| H | 12 | 18 |

- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
  - Overlappling confidence intervals for the 8 school effects
  - Statements such as Pr (effect in A $>$ effect in C)$= 0.7$

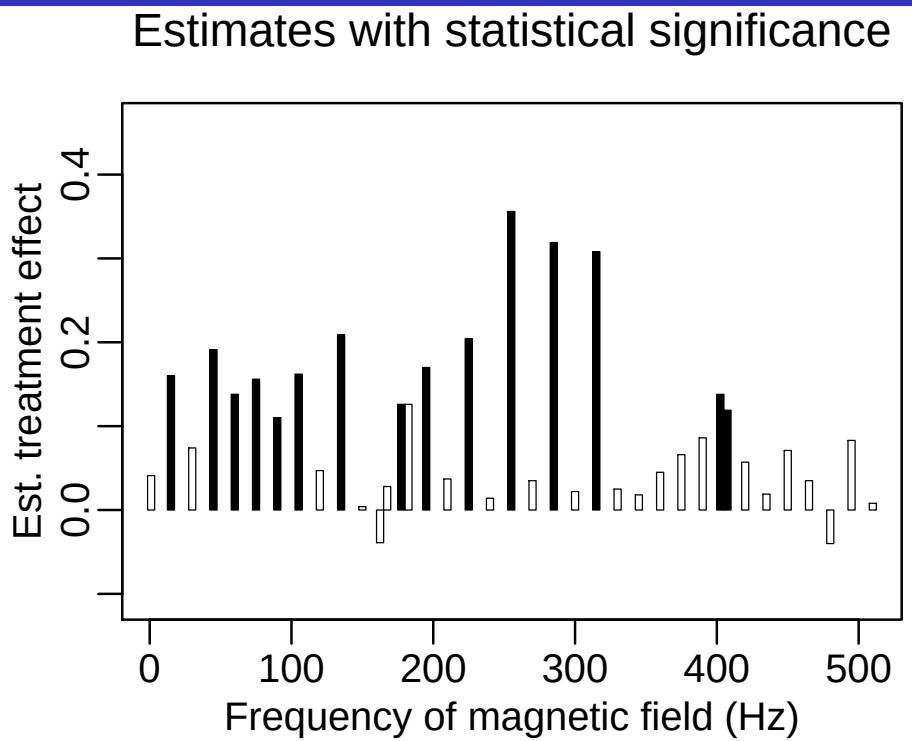
# Effects of electromagnetic fields at 38 frequencies

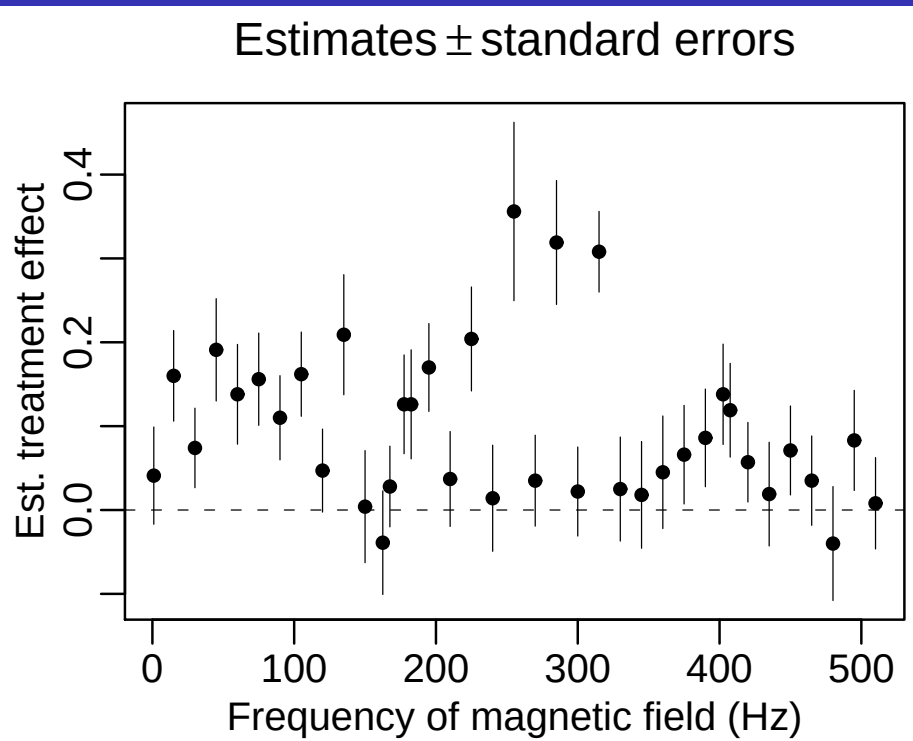



- Background: electromagnetic fields and cancer
- Original article summarized using p-values
- Confidence intervals show comparisons more clearly

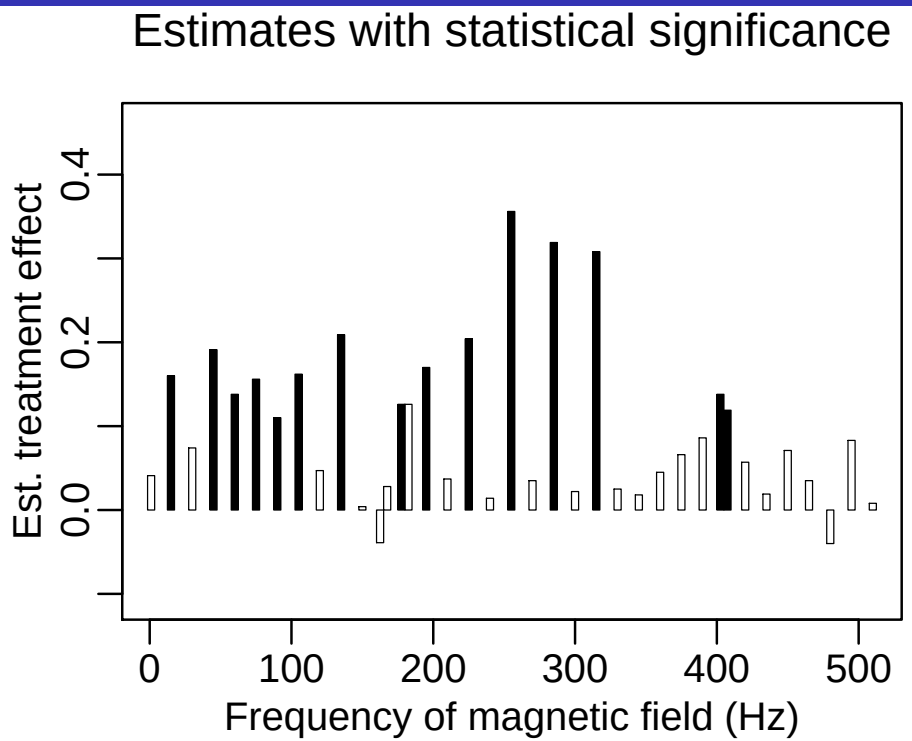
# Effects of electromagnetic fields at 38 frequencies

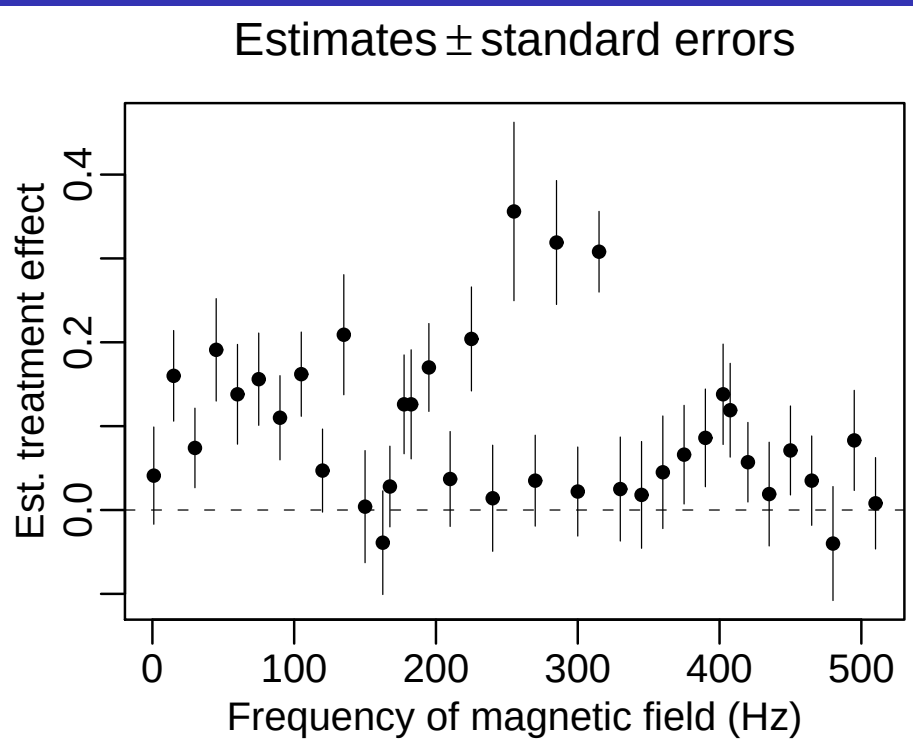



- Background: electromagnetic fields and cancer
- Original article summarized using p-values
- Confidence intervals show comparisons more clearly

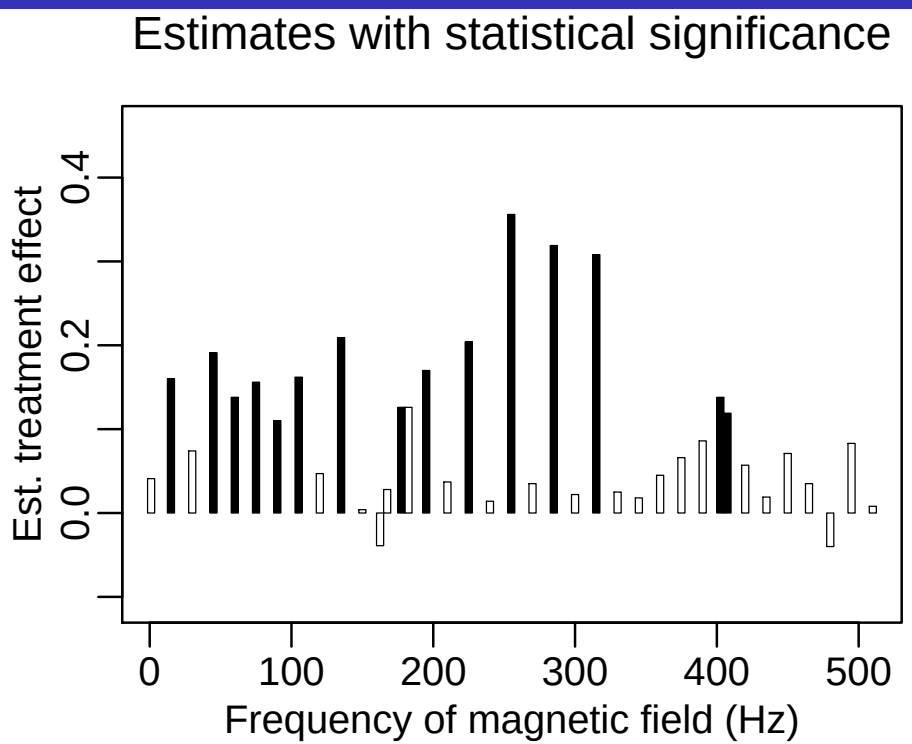
# Effects of electromagnetic fields at 38 frequencies

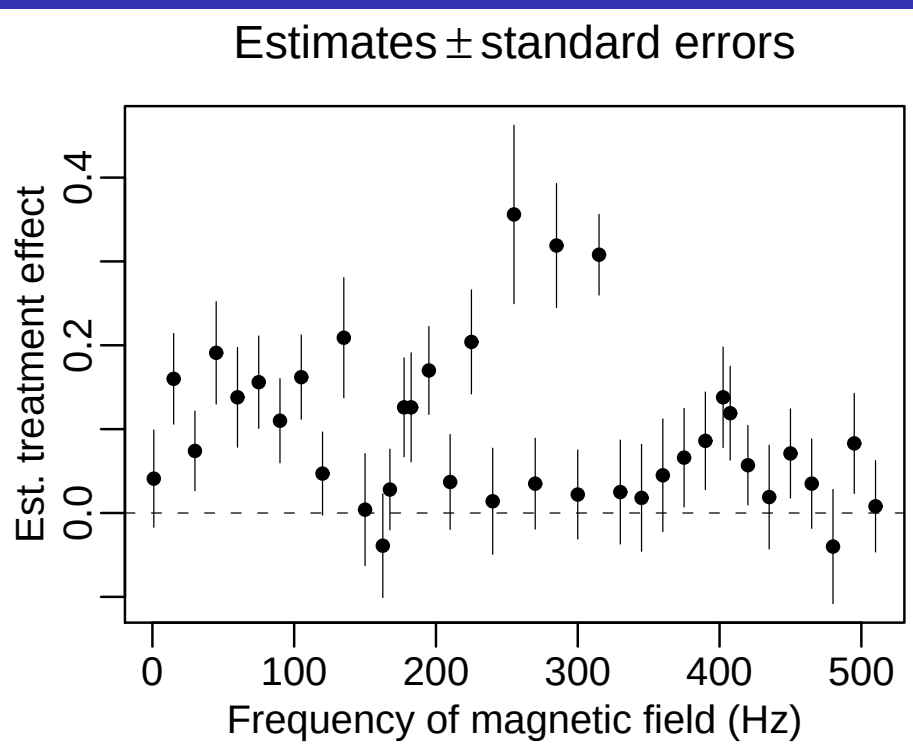



- Background: electromagnetic fields and cancer
- Original article summarized using p-values
- Confidence intervals show comparisons more clearly

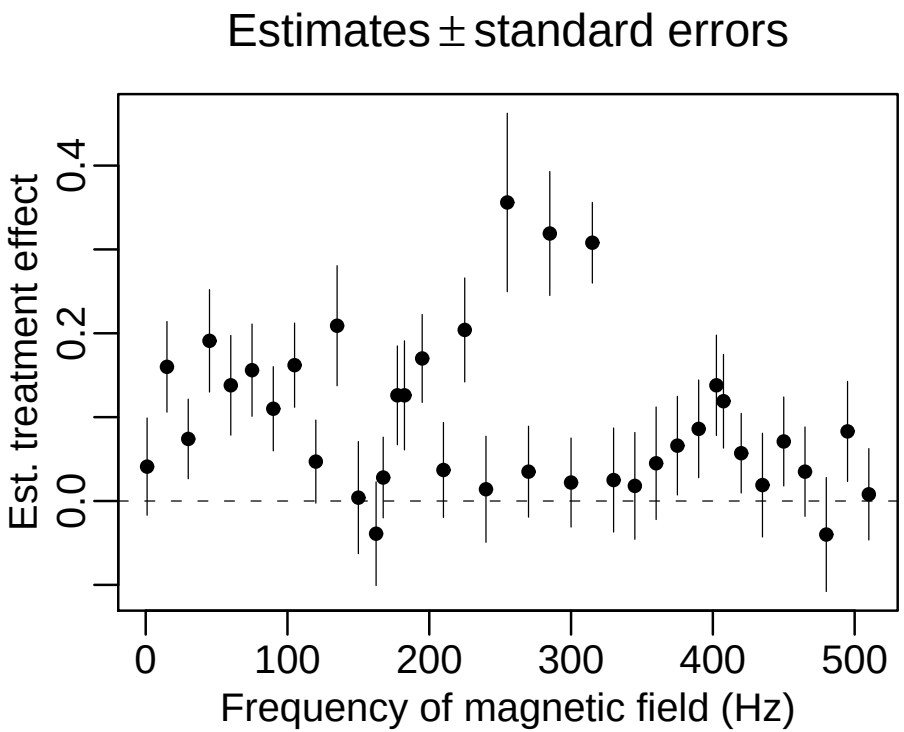
# Effects of electromagnetic fields at 38 frequencies

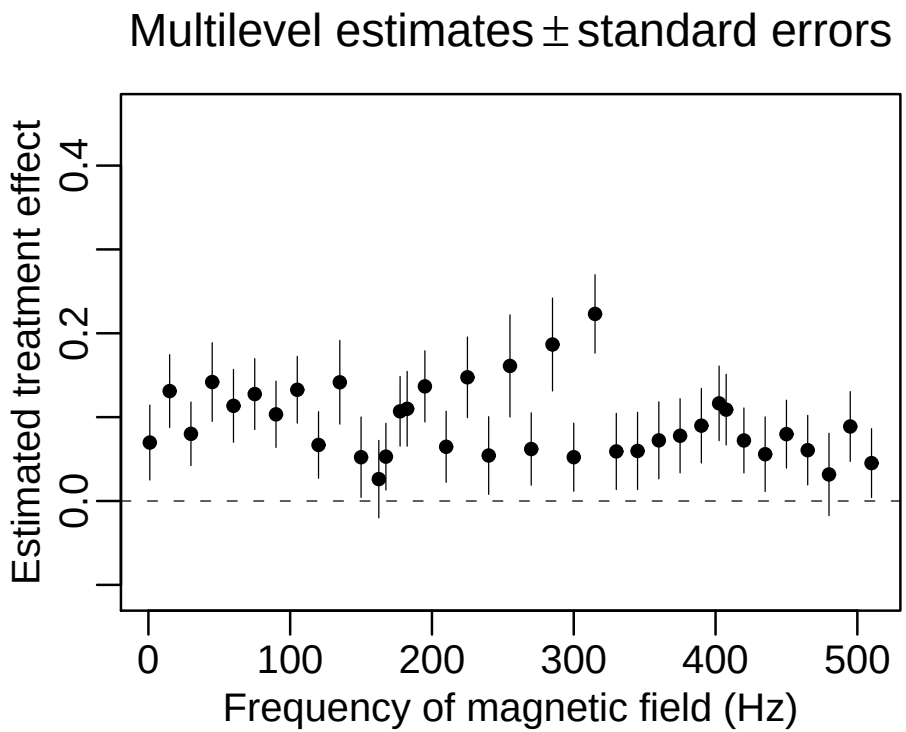



- Background: electromagnetic fields and cancer
- Original article summarized using p-values
- Confidence intervals show comparisons more clearly

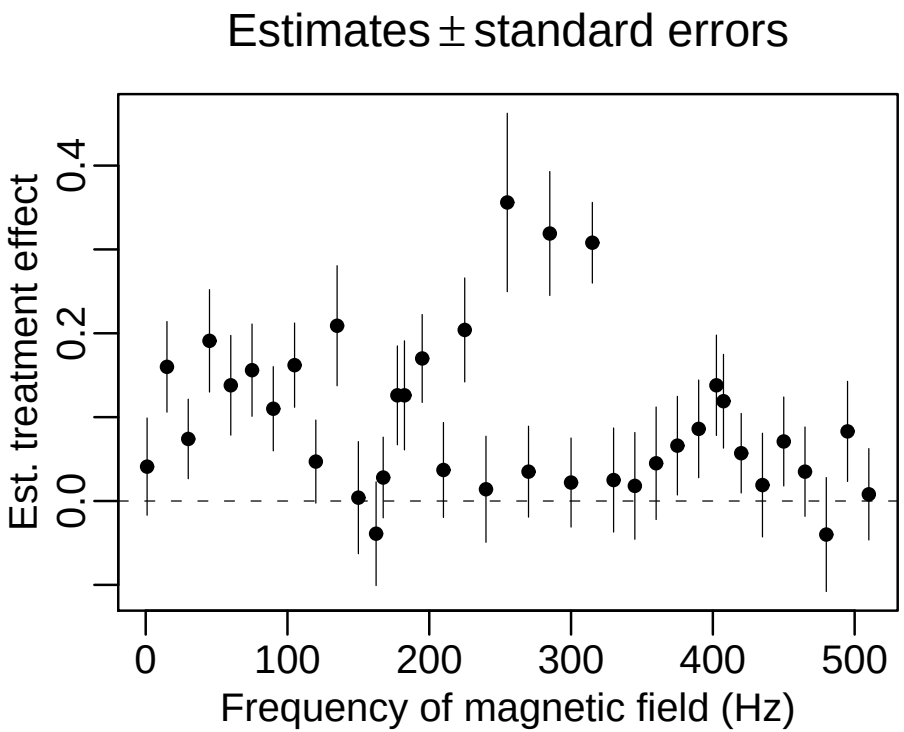
# Separate estimates and hierarchical Bayes estimates

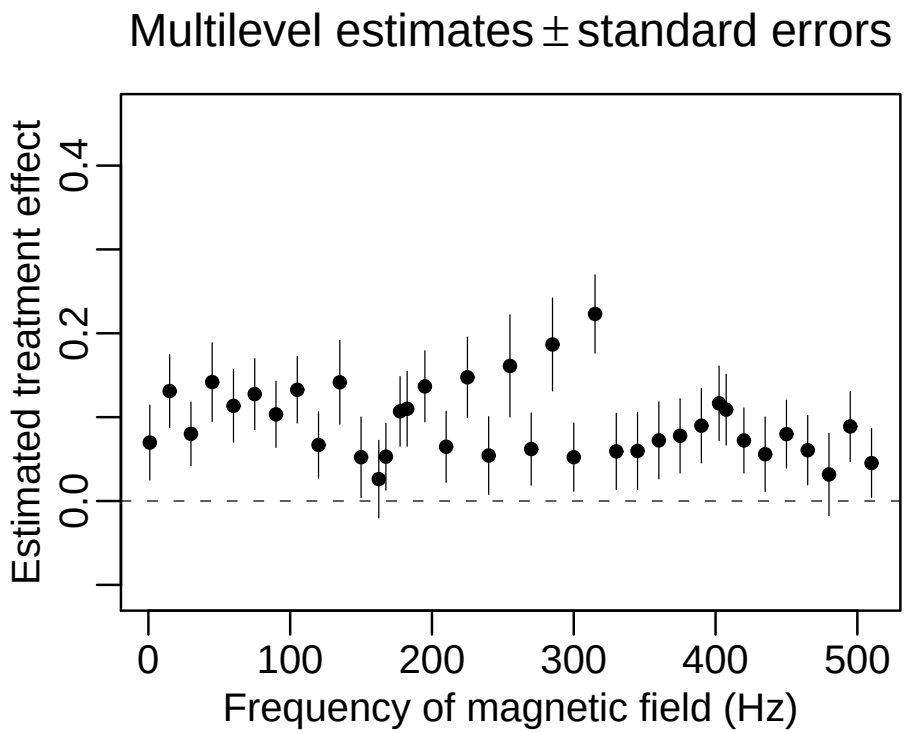



- Most comparisons are no longer statistically significant
- "Multiple comparisons" is less of a concern
- We moved the intervals together instead of widening them!

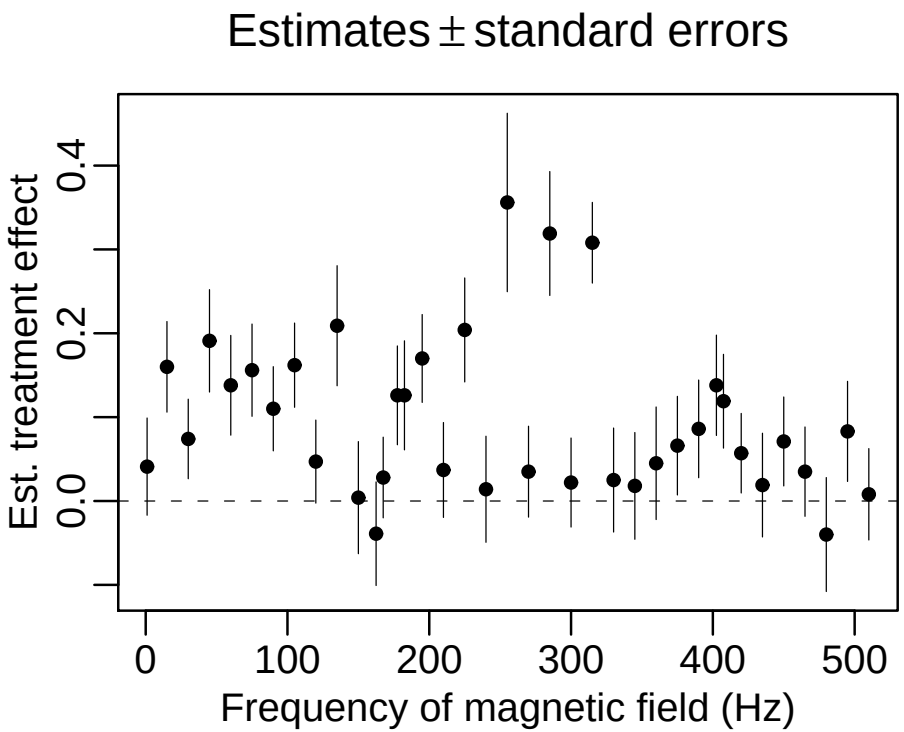
# Separate estimates and hierarchical Bayes estimates

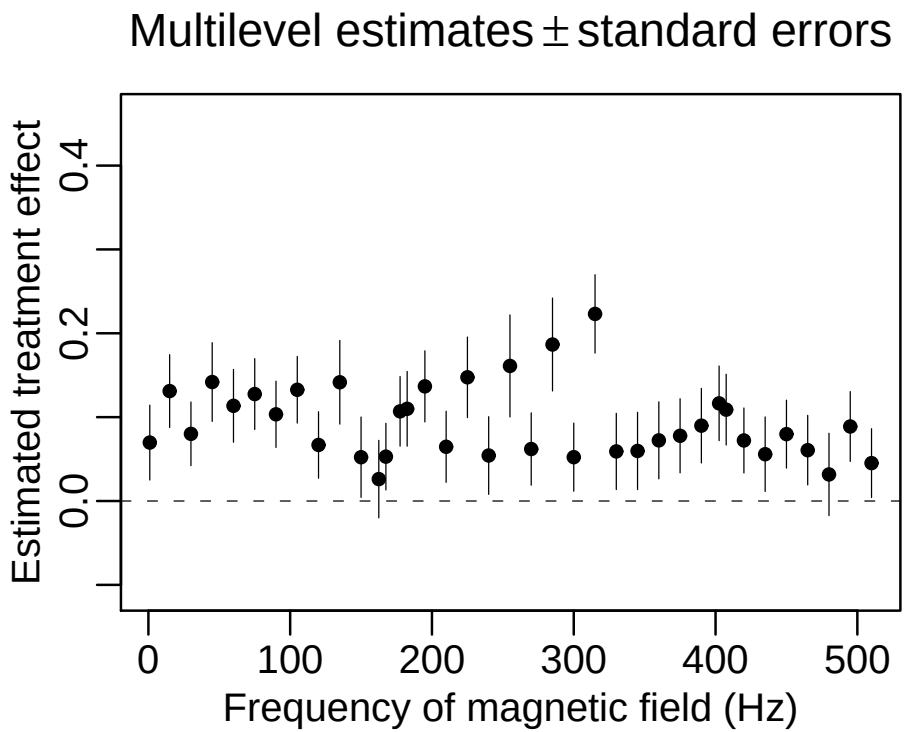



- Most comparisons are no longer statistically significant
- "Multiple comparisons" is less of a concern
- We moved the intervals together instead of widening them!

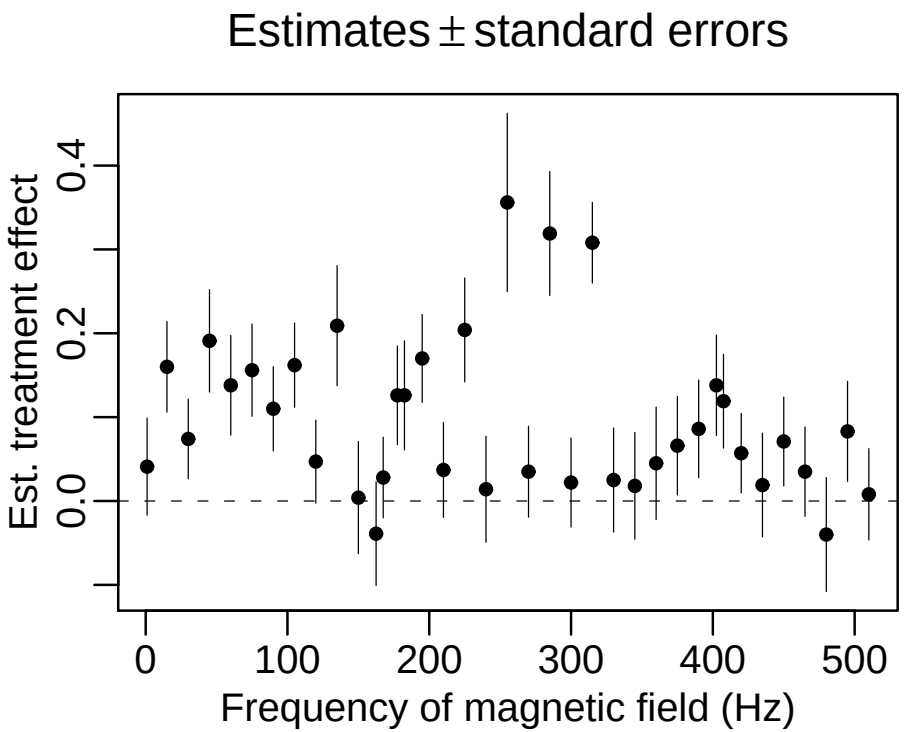
# Separate estimates and hierarchical Bayes estimates

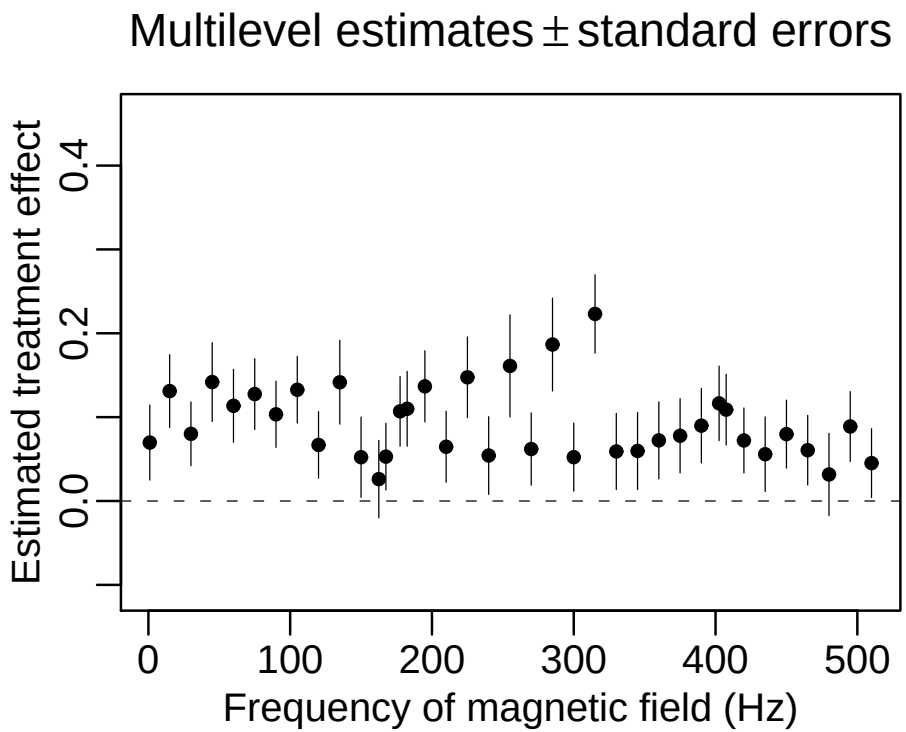



- Most comparisons are no longer statistically significant
- "Multiple comparisons" is less of a concern
- We moved the intervals together instead of widening them!

# Separate estimates and hierarchical Bayes estimates




- Most comparisons are no longer statistically significant
- "Multiple comparisons" is less of a concern
- We moved the intervals together instead of widening them!

# Teacher and school effects in NYC schools

- Goal is to estimate range of variation
  (How important are teachers? Schools?)
- Key statistic is year-to-year persistence (e.g., for teachers ranked in top 25% one year, how well do they do the next?)
- The "multiple comparisons" issue never arises!

# Teacher and school effects in NYC schools

- Goal is to estimate range of variation
  (How important are teachers? Schools?)
- Key statistic is year-to-year persistence (e.g., for teachers ranked in top 25% one year, how well do they do the next?)
- The "multiple comparisons" issue never arises!

# Teacher and school effects in NYC schools

- Goal is to estimate range of variation
  (How important are teachers? Schools?)
- Key statistic is year-to-year persistence (e.g., for teachers ranked in top 25% one year, how well do they do the next?)
- The "multiple comparisons" issue never arises!

# Teacher and school effects in NYC schools

- Goal is to estimate range of variation
  (How important are teachers? Schools?)
- Key statistic is year-to-year persistence (e.g., for teachers ranked in top 25% one year, how well do they do the next?)
- The "multiple comparisons" issue never arises!

# Grades and classroom seating

- Classroom demonstration
- Assign students random numbers as "grades"
- Ask students with "grades" 0–25 to raise one finger, students with "grades" 75–100 to raise one hand
- Instructor scans the room to find a statistically significant comparison (e.g., "boys on the left side of the classroom have higher grades than girls in the back row")
- This is a pure multiple comparisons problem!

# Grades and classroom seating

- Classroom demonstration
- Assign students random numbers as "grades"
- Ask students with "grades" 0–25 to raise one finger, students with "grades" 75–100 to raise one hand
- Instructor scans the room to find a statistically significant comparison (e.g., "boys on the left side of the classroom have higher grades than girls in the back row")
- This is a pure multiple comparisons problem!

# Grades and classroom seating

- Classroom demonstration
- Assign students random numbers as "grades"
- Ask students with "grades" 0–25 to raise one finger, students with "grades" 75–100 to raise one hand
- Instructor scans the room to find a statistically significant comparison (e.g., "boys on the left side of the classroom have higher grades than girls in the back row")
- This is a pure multiple comparisons problem!

# Grades and classroom seating

- Classroom demonstration
- Assign students random numbers as "grades"
- Ask students with "grades" 0–25 to raise one finger, students with "grades" 75–100 to raise one hand
- Instructor scans the room to find a statistically significant comparison (e.g., "boys on the left side of the classroom have higher grades than girls in the back row")
- This is a pure multiple comparisons problem!

# Grades and classroom seating

- Classroom demonstration
- Assign students random numbers as "grades"
- Ask students with "grades" 0–25 to raise one finger, students with "grades" 75–100 to raise one hand
- Instructor scans the room to find a statistically significant comparison (e.g., "boys on the left side of the classroom have higher grades than girls in the back row")
- This is a pure multiple comparisons problem!

# Grades and classroom seating

- Classroom demonstration
- Assign students random numbers as "grades"
- Ask students with "grades" 0–25 to raise one finger, students with "grades" 75–100 to raise one hand
- Instructor scans the room to find a statistically significant comparison (e.g., "boys on the left side of the classroom have higher grades than girls in the back row")
- This is a pure multiple comparisons problem!

# Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology.
- Attractiveness was measured on a 1–5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1–4 were girls
- Statistically significant (2.44 s.e.'s from zero, $p = 1.5\%$)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons $\times$ 4 possible time summaries!

# Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology.
- Attractiveness was measured on a 1–5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1–4 were girls
- Statistically significant (2.44 s.e.'s from zero, $p = 1.5\%$)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons $\times$ 4 possible time summaries!

# Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology.
- Attractiveness was measured on a 1–5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1–4 were girls
- Statistically significant (2.44 s.e.'s from zero, $p = 1.5\%$)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons $\times$ 4 possible time summaries!

# Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology.
- Attractiveness was measured on a 1–5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1–4 were girls
- Statistically significant (2.44 s.e.'s from zero, $p = 1.5\%$)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons $\times$ 4 possible time summaries!

# Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology.
- Attractiveness was measured on a 1–5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1–4 were girls
- Statistically significant (2.44 s.e.'s from zero, $p = 1.5\%$)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons $\times$ 4 possible time summaries!

# Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology.
- Attractiveness was measured on a 1–5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1–4 were girls
- Statistically significant (2.44 s.e.'s from zero, $p = 1.5\%$)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons $\times$ 4 possible time summaries!

# Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology.
- Attractiveness was measured on a 1–5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1–4 were girls
- Statistically significant (2.44 s.e.'s from zero, $p = 1.5\%$)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons $\times$ 4 possible time summaries!

# Beautiful parents have more daughters

- S. Kanazawa (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. Journal of Theoretical Biology.
- Attractiveness was measured on a 1–5 scale ("very unattractive" to "very attractive")
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1–4 were girls
- Statistically significant (2.44 s.e.'s from zero, $p = 1.5\%$)
- But the simple regression of sex ratio on attractiveness is not significant (estimate is 1.5 with s.e. of 1.4)
- Multiple comparisons problem: 5 natural comparisons $\times$ 4 possible time summaries!

# Comparing test scores across states

- National Assessment of Educational Progress (NAEP)
- Comparing states: which comparisons are statistically significant?
- $50 \times 49/2$: a classic multiple comparions problem!
- Our multilevel inferences

# Comparing test scores across states

- National Assessment of Educational Progress (NAEP)
- Comparing states: which comparisons are statistically significant?
- $50 \times 49/2$: a classic multiple comparions problem!
- Our multilevel inferences

# Comparing test scores across states

- National Assessment of Educational Progress (NAEP)
- Comparing states: which comparisons are statistically significant?
- $50 \times 49/2$: a classic multiple comparions problem!
- Our multilevel inferences

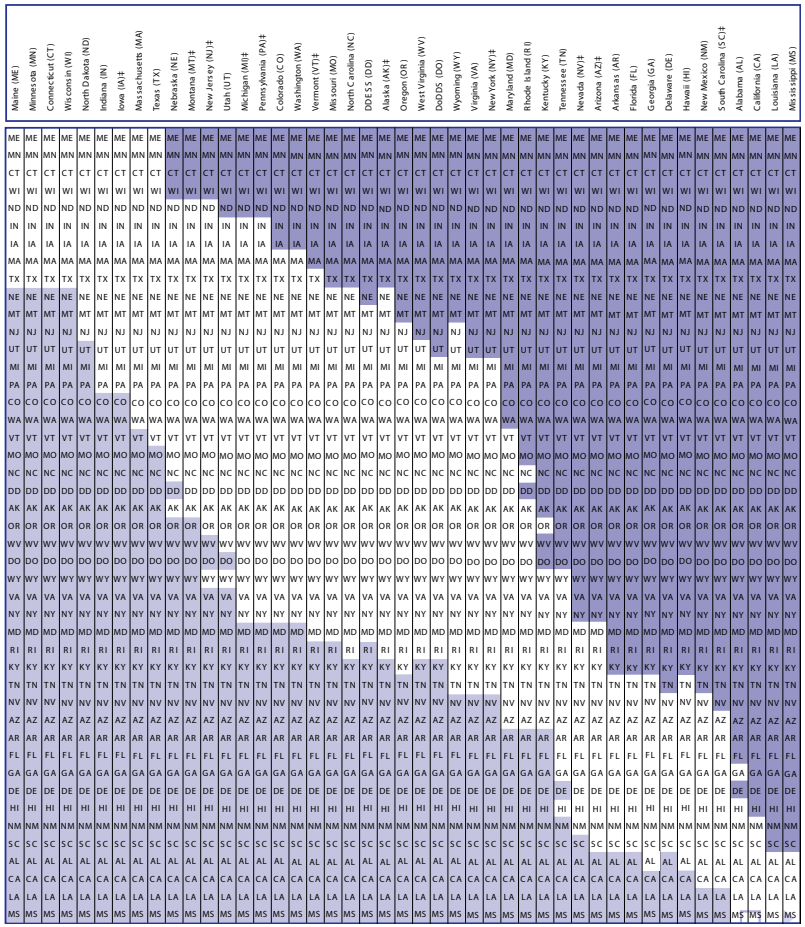
# Comparing test scores across states

- National Assessment of Educational Progress (NAEP)
- Comparing states: which comparisons are statistically significant?
- $50 \times 49/2$: a classic multiple comparions problem!
- Our multilevel inferences

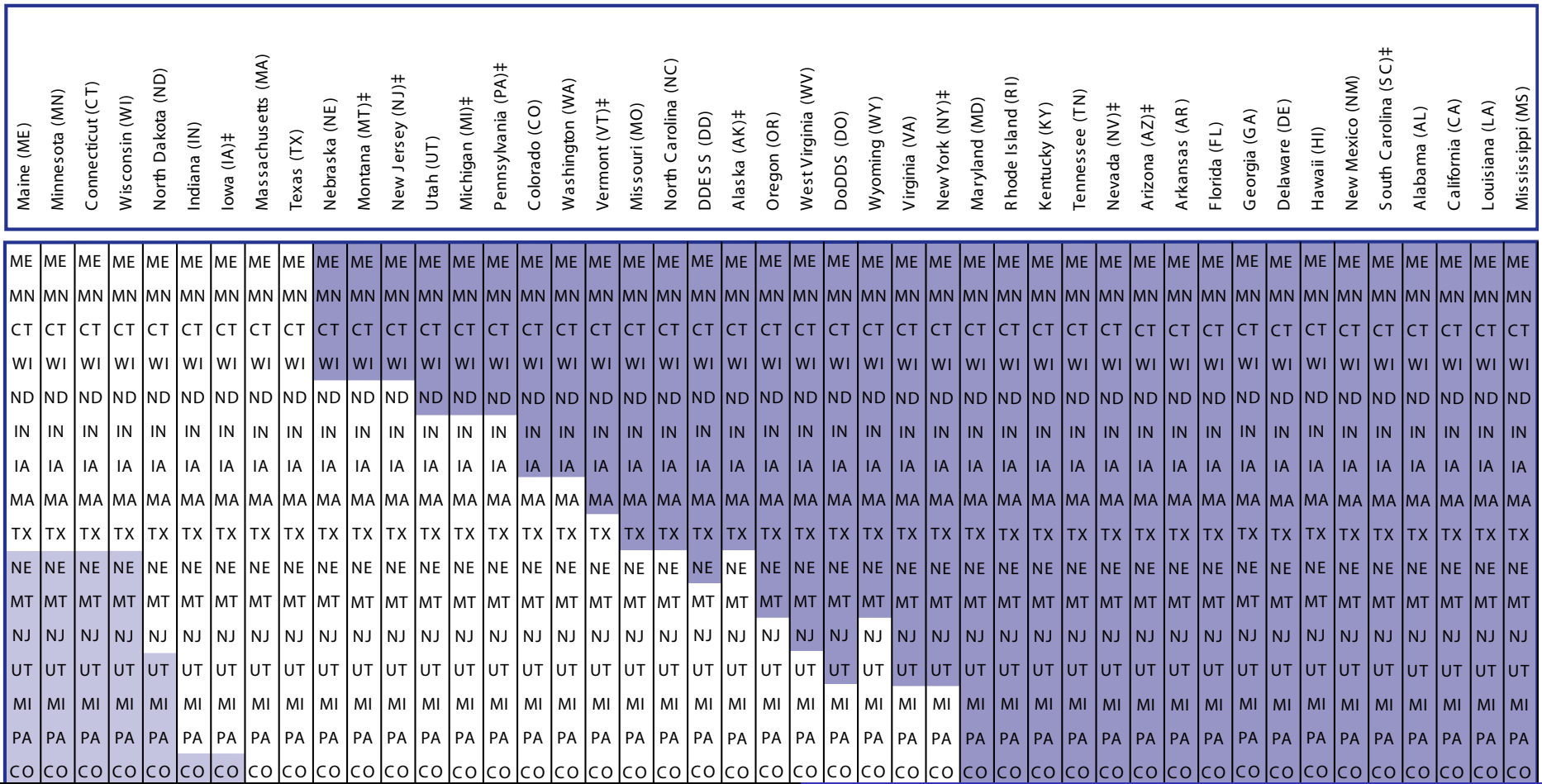
# Comparing test scores across states

- National Assessment of Educational Progress (NAEP)
- Comparing states: which comparisons are statistically significant?
- $50 \times 49/2$: a classic multiple comparions problem!
- Our multilevel inferences

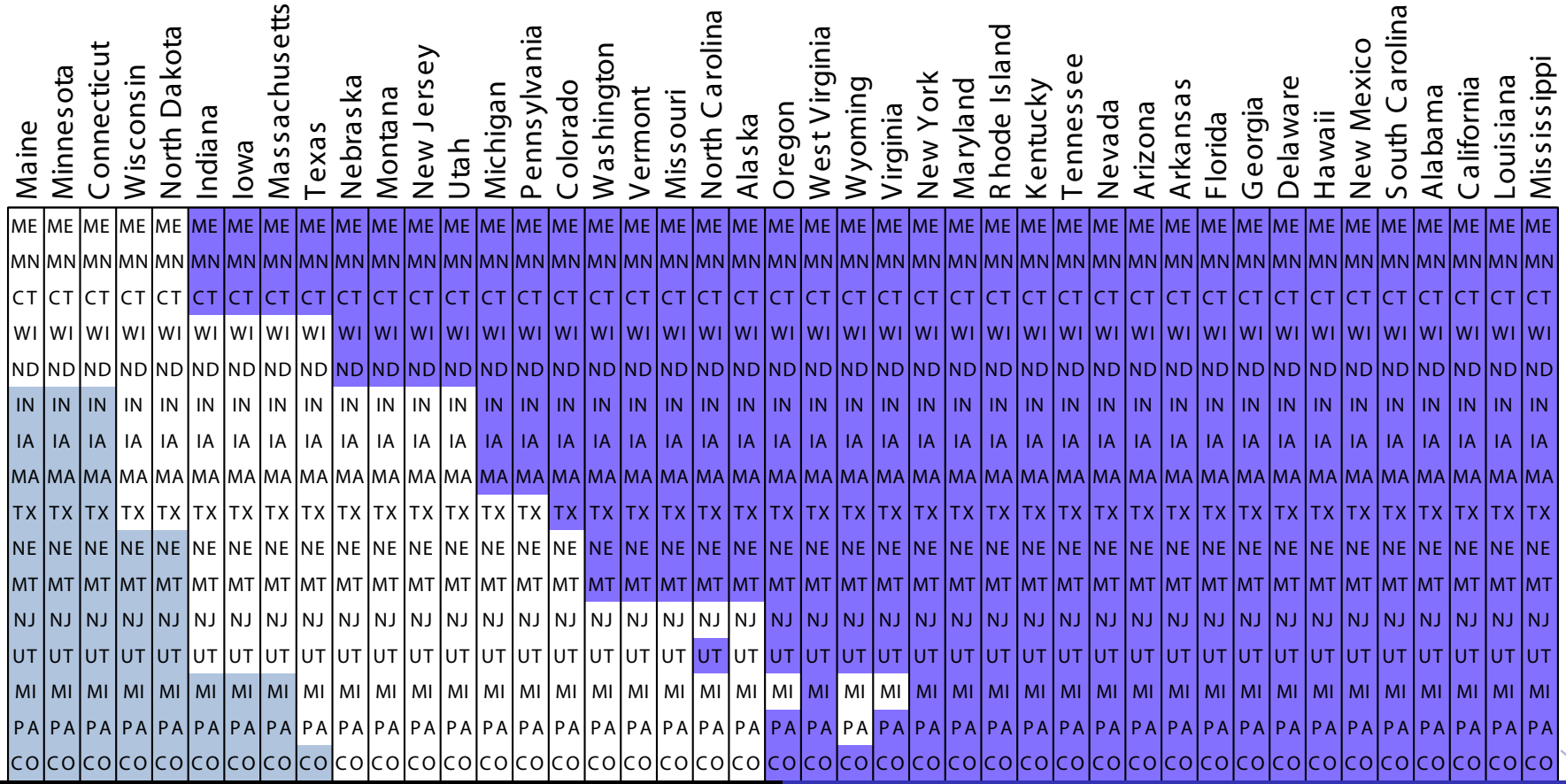
# Classical multiple comparisons inferences for NAEP



# Classical inferences for NAEP: close-up



# Multilevel inferences for NAEP: close-up



# NAEP: classical vs. multilevel

- Both procedures are algorithmic ("push a button")
- Both procedures treat 50 states exchangeably
- Multilevel inferences are sharper (more comparisons are "statistically significant")
- How can this be?

# NAEP: classical vs. multilevel

- Both procedures are algorithmic ("push a button")
- Both procedures treat 50 states exchangeably
- Multilevel inferences are sharper (more comparisons are "statistically significant")
- How can this be?

# NAEP: classical vs. multilevel

- Both procedures are algorithmic ("push a button")
- Both procedures treat 50 states exchangeably
- Multilevel inferences are sharper (more comparisons are "statistically significant")
- How can this be?

# NAEP: classical vs. multilevel

- Both procedures are algorithmic ("push a button")
- Both procedures treat 50 states exchangeably
- Multilevel inferences are sharper (more comparisons are "statistically significant")
- How can this be?

# NAEP: classical vs. multilevel

- Both procedures are algorithmic ("push a button")
- Both procedures treat 50 states exchangeably
- Multilevel inferences are sharper (more comparisons are "statistically significant")
- How can this be?

# Something for nothing? A free lunch?

- Classical multiple comparisons worries about $\theta_1 = \theta_2 = \cdots = \theta_{50}$
- Not an issue with NAEP
- Multilevel model estimates the group-level variance, decides based on the data how much to adjust
- Classical procedure does not learn from the data

# Something for nothing? A free lunch?

- Classical multiple comparisons worries about $\theta_1 = \theta_2 = \cdots = \theta_{50}$
- Not an issue with NAEP
- Multilevel model estimates the group-level variance, decides based on the data how much to adjust
- Classical procedure does not learn from the data

# Something for nothing? A free lunch?

- Classical multiple comparisons worries about $\theta_1 = \theta_2 = \cdots = \theta_{50}$
- Not an issue with NAEP
- Multilevel model estimates the group-level variance, decides based on the data how much to adjust
- Classical procedure does not learn from the data

# Something for nothing? A free lunch?

- Classical multiple comparisons worries about $\theta_1 = \theta_2 = \cdots = \theta_{50}$
- Not an issue with NAEP
- Multilevel model estimates the group-level variance, decides based on the data how much to adjust
- Classical procedure does not learn from the data

# Something for nothing? A free lunch?

- Classical multiple comparisons worries about $\theta_1 = \theta_2 = \cdots = \theta_{50}$
- Not an issue with NAEP
- Multilevel model estimates the group-level variance, decides based on the data how much to adjust
- Classical procedure does not learn from the data

# Message from the examples

- Classical multiple comparisons corrections don't seem so important when we fit hierarchical models
- But they can be crucial for classical comparisons

# Message from the examples

- Classical multiple comparisons corrections don't seem so important when we fit hierarchical models
- But they can be crucial for classical comparisons

# Message from the examples

- Classical multiple comparisons corrections don't seem so important when we fit hierarchical models
- But they can be crucial for classical comparisons

# Statistical framework

- Goal is to estimate $\theta_j$, for $j = 1, \ldots, J$ (for example, effects of $J$ schools)
- Comparisons have the form, $\theta_j - \theta_k$.
- For simplicity, suppose data come from $J$ separate experiments
- Type S errors
- Multilevel modeling as a solution to the multiple comparisons issue

# Statistical framework

- Goal is to estimate $\theta_j$, for $j = 1, \ldots, J$ (for example, effects of $J$ schools)
- Comparisons have the form, $\theta_j - \theta_k$.
- For simplicity, suppose data come from $J$ separate experiments
- Type S errors
- Multilevel modeling as a solution to the multiple comparisons issue

# Statistical framework

- Goal is to estimate $\theta_j$, for $j = 1, \ldots, J$ (for example, effects of $J$ schools)
- Comparisons have the form, $\theta_j - \theta_k$.
- For simplicity, suppose data come from $J$ separate experiments
- Type S errors
- Multilevel modeling as a solution to the multiple comparisons issue

# Statistical framework

- Goal is to estimate $\theta_j$, for $j = 1, \ldots, J$ (for example, effects of $J$ schools)
- Comparisons have the form, $\theta_j - \theta_k$.
- For simplicity, suppose data come from $J$ separate experiments
- Type S errors
- Multilevel modeling as a solution to the multiple comparisons issue

# Statistical framework

- Goal is to estimate $\theta_j$, for $j = 1, \ldots, J$ (for example, effects of $J$ schools)
- Comparisons have the form, $\theta_j - \theta_k$.
- For simplicity, suppose data come from $J$ separate experiments
- Type S errors
- Multilevel modeling as a solution to the multiple comparisons issue

# Statistical framework

- Goal is to estimate $\theta_j$, for $j = 1, \ldots, J$ (for example, effects of $J$ schools)
- Comparisons have the form, $\theta_j - \theta_k$.
- For simplicity, suppose data come from $J$ separate experiments
- Type S errors
- Multilevel modeling as a solution to the multiple comparisons issue

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life

- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - Type 2 error is $\theta_j \neq \theta_k$, but I claim they're the same
  - I've never claimed that $\theta_j = \theta_k$
- But I make errors all the time!
- *Type S error*: $\theta_1 > \theta_2$, but I claim that $\theta_2 > \theta_1$ (or vice versa)
- Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for $\theta_j - \theta_k$ that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the $\theta_j$'s

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
  - $\sigma_\theta$ = not large variation
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
  - $\sigma_\theta$ = not large variation
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \to \infty$: large variation
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \to \infty$: large variation

- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \longrightarrow \infty$: large variation

- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \to \infty$: large variation
    - Multilevel model is equivalent to estimating each $\theta_j$ separately
    - "Multiple comparisons" corrections are not needed
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \to \infty$: large variation
    - Multilevel model is equivalent to estimating each $\theta_j$ separately
    - "Multiple comparisons" corrections are not needed
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \to \infty$: large variation
    - Multilevel model is equivalent to estimating each $\theta_j$ separately
    - "Multiple comparisons" corrections are not needed
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \to \infty$: large variation
    - Multilevel model is equivalent to estimating each $\theta_j$ separately
    - "Multiple comparisons" corrections are not needed
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \to \infty$: large variation
    - Multilevel model is equivalent to estimating each $\theta_j$ separately
    - "Multiple comparisons" corrections are not needed
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Multilevel (hierarchical) modeling

- Key parameter: $\sigma_\theta$, the sd of the true $\theta_j$'s
- Understand through special cases:
  - $\sigma_\theta \approx 0$: no variation
    - Multilevel model pools the estimated $\theta_j$'s toward each other
    - "Multiple comparisons" correction is done by shrinking comparisons
    - Very few claims with confidence (far fewer than 5%)
  - $\sigma_\theta \to \infty$: large variation
    - Multilevel model is equivalent to estimating each $\theta_j$ separately
    - "Multiple comparisons" corrections are not needed
- Bayesian multilevel modeling bounds the Type S error rate by automatically restricting the rate of claims with confidence

# Conclusions

- "Multiple comparisons" is a real concern, but . . .
- Don't "fix" by altering p-values or (equivalently) by making confidence intervals wider
- Instead, multilevel modeling does partial pooling where necessary (especially when much of the variation in the data can be explained by noise), so that few claims can be made with confidence
- Learn from the data how much to adjust
- By comparison, classical intervals can have Type S error rates as high as 50%: then, multiple comparisons can be a big deal!

# Conclusions

- "Multiple comparisons" is a real concern, but . . .
- Don't "fix" by altering p-values or (equivalently) by making confidence intervals wider
- Instead, multilevel modeling does partial pooling where necessary (especially when much of the variation in the data can be explained by noise), so that few claims can be made with confidence
- Learn from the data how much to adjust
- By comparison, classical intervals can have Type S error rates as high as 50%: then, multiple comparisons can be a big deal!

# Conclusions

- "Multiple comparisons" is a real concern, but . . .
- Don't "fix" by altering p-values or (equivalently) by making confidence intervals wider
- Instead, multilevel modeling does partial pooling where necessary (especially when much of the variation in the data can be explained by noise), so that few claims can be made with confidence
- Learn from the data how much to adjust
- By comparison, classical intervals can have Type S error rates as high as 50%: then, multiple comparisons can be a big deal!

# Conclusions

- "Multiple comparisons" is a real concern, but . . .
- Don't "fix" by altering p-values or (equivalently) by making confidence intervals wider
- Instead, multilevel modeling does partial pooling where necessary (especially when much of the variation in the data can be explained by noise), so that few claims can be made with confidence
- Learn from the data how much to adjust
- By comparison, classical intervals can have Type S error rates as high as 50%: then, multiple comparisons can be a big deal!

# Conclusions

- "Multiple comparisons" is a real concern, but . . .
- Don't "fix" by altering p-values or (equivalently) by making confidence intervals wider
- Instead, multilevel modeling does partial pooling where necessary (especially when much of the variation in the data can be explained by noise), so that few claims can be made with confidence
- Learn from the data how much to adjust
- By comparison, classical intervals can have Type S error rates as high as 50%: then, multiple comparisons can be a big deal!

# Conclusions

- "Multiple comparisons" is a real concern, but . . .
- Don't "fix" by altering p-values or (equivalently) by making confidence intervals wider
- Instead, multilevel modeling does partial pooling where necessary (especially when much of the variation in the data can be explained by noise), so that few claims can be made with confidence
- Learn from the data how much to adjust
- By comparison, classical intervals can have Type S error rates as high as 50%: then, multiple comparisons can be a big deal!

# Further thoughts inspired by the workshop: part 1

- The key step in getting everything to work is to move the coefficient estimates: not just to adjust the p-values or widen intervals, but to actually move the estimates toward each other (or, more generally, toward a group-level regression model)
- But, to be honest, multilevel modeling can be difficult for complicated structures. For example, analyzing 16 schools or 16 outcomes is simple enough, but a $2^4$ structure (e.g., 2 groups $\times$ 2 outcomes $\times$ 2 flavors of treatment $\times$ 2 time points) is more of a research problem.
- There was some discussion of main analyses as compared to minor analyses, with the thought that different levels of significance can be used for these two areas. I would prefer to use the full model for all of this and then report all the results graphically. In no case do I think would it be necessary to do a multiple comparisons adjustment. As noted in the first bullet point above, the multilevel inference should do fine at combining the information.

# Further thoughts inspired by the workshop: part 1

- The key step in getting everything to work is to move the coefficient estimates: not just to adjust the p-values or widen intervals, but to actually move the estimates toward each other (or, more generally, toward a group-level regression model)
- But, to be honest, multilevel modeling can be difficult for complicated structures. For example, analyzing 16 schools or 16 outcomes is simple enough, but a $2^4$ structure (e.g., 2 groups $\times$ 2 outcomes $\times$ 2 flavors of treatment $\times$ 2 time points) is more of a research problem.
- There was some discussion of main analyses as compared to minor analyses, with the thought that different levels of significance can be used for these two areas. I would prefer to use the full model for all of this and then report all the results graphically. In no case do I think would it be necessary to do a multiple comparisons adjustment. As noted in the first bullet point above, the multilevel inference should do fine at combining the information.

# Further thoughts inspired by the workshop: part 1

- The key step in getting everything to work is to move the coefficient estimates: not just to adjust the p-values or widen intervals, but to actually move the estimates toward each other (or, more generally, toward a group-level regression model)
- But, to be honest, multilevel modeling can be difficult for complicated structures. For example, analyzing 16 schools or 16 outcomes is simple enough, but a $2^4$ structure (e.g., 2 groups $\times$ 2 outcomes $\times$ 2 flavors of treatment $\times$ 2 time points) is more of a research problem.
- There was some discussion of main analyses as compared to minor analyses, with the thought that different levels of significance can be used for these two areas. I would prefer to use the full model for all of this and then report all the results graphically. In no case do I think would it be necessary to do a multiple comparisons adjustment. As noted in the first bullet point above, the multilevel inference should do fine at combining the information.

# Further thoughts inspired by the workshop: part 1

- The key step in getting everything to work is to move the coefficient estimates: not just to adjust the p-values or widen intervals, but to actually move the estimates toward each other (or, more generally, toward a group-level regression model)
- But, to be honest, multilevel modeling can be difficult for complicated structures. For example, analyzing 16 schools or 16 outcomes is simple enough, but a $2^4$ structure (e.g., 2 groups $\times$ 2 outcomes $\times$ 2 flavors of treatment $\times$ 2 time points) is more of a research problem.
- There was some discussion of main analyses as compared to minor analyses, with the thought that different levels of significance can be used for these two areas. I would prefer to use the full model for all of this and then report all the results graphically. In no case do I think would it be necessary to do a multiple comparisons adjustment. As noted in the first bullet point above, the multilevel inference should do fine at combining the information.

# Further thoughts inspired by the workshop: part 2

- The key concern about multiple comparisons is when the noise level is high, so that classical Type S error rates will be high. Having *many* comparisons (100, or 1000, or whatever) is not a problem at all in the context of multilevel modeling. This gets back to our willingness to be wrong occasionally. If we are testing 1000 comparisons, and there's really nothing much going on, then the multilevel analysis will do lots of pooling, and the result is to make very few claims with confidence.
- For this presentation, I'd been thinking about the sort of examples where the differences are unquestionably real, and the key concern is Type S errors: for example, saying that School A is better than School B, when School B is really better than School A. But at the meeting, there was lots of discussion of examples where the differences might actually be very close to zero: for example, comparing different educational interventions, none of which might be very effective. Here we would want to be thinking about "Type M" (magnitude) errors: saying that a treatment effect is near zero when it is actually large, or saying that it's

# Further thoughts inspired by the workshop: part 2

- The key concern about multiple comparisons is when the noise level is high, so that classical Type S error rates will be high. Having *many* comparisons (100, or 1000, or whatever) is not a problem at all in the context of multilevel modeling. This gets back to our willingness to be wrong occasionally. If we are testing 1000 comparisons, and there's really nothing much going on, then the multilevel analysis will do lots of pooling, and the result is to make very few claims with confidence.
- For this presentation, I'd been thinking about the sort of examples where the differences are unquestionably real, and the key concern is Type S errors: for example, saying that School A is better than School B, when School B is really better than School A. But at the meeting, there was lots of discussion of examples where the differences might actually be very close to zero: for example, comparing different educational interventions, none of which might be very effective. Here we would want to be thinking about "Type M" (magnitude) errors: saying that a treatment effect is near zero when it is actually large, or saying that it's

# Further thoughts inspired by the workshop: part 2

- The key concern about multiple comparisons is when the noise level is high, so that classical Type S error rates will be high. Having *many* comparisons (100, or 1000, or whatever) is not a problem at all in the context of multilevel modeling. This gets back to our willingness to be wrong occasionally. If we are testing 1000 comparisons, and there's really nothing much going on, then the multilevel analysis will do lots of pooling, and the result is to make very few claims with confidence.
- For this presentation, I'd been thinking about the sort of examples where the differences are unquestionably real, and the key concern is Type S errors: for example, saying that School A is better than School B, when School B is really better than School A. But at the meeting, there was lots of discussion of examples where the differences might actually be very close to zero: for example, comparing different educational interventions, none of which might be very effective. Here we would want to be thinking about "Type M" (magnitude) errors: saying that a treatment effect is near zero when it is actually large, or saying that it's