

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

Andrew Gelman, Jennifer Hill, and Masanao Yajima
Columbia University

8 Nov 2007

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

Contents

- ▶ What is the multiple comparisons problem?
- ▶ Why don't I (usually) care about it?
- ▶ Some stories
- ▶ Statistical framework and multilevel modeling

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

Contents

- ▶ What is the multiple comparisons problem?
- ▶ Why don't I (usually) care about it?
- ▶ Some stories
- ▶ Statistical framework and multilevel modeling

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

Contents

- ▶ What is the multiple comparisons problem?
- ▶ Why don't I (usually) care about it?
- ▶ Some stories
- ▶ Statistical framework and multilevel modeling

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

Contents

- ▶ What is the multiple comparisons problem?
- ▶ Why don't I (usually) care about it?
- ▶ Some stories
- ▶ Statistical framework and multilevel modeling

Contents

- ▶ What is the multiple comparisons problem?
- ▶ Why don't I (usually) care about it?
- ▶ Some stories
- ▶ Statistical framework and multilevel modeling

What is the multiple comparisons problem?

Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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?

Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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:

▶ The number of comparisons grows exponentially with the number of groups, but the number of comparisons that matter grows much more slowly

▶ The number of comparisons that matter is often much smaller than the number of comparisons that are possible

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 think being wrong 1% 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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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, σ_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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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, σ_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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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, σ_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 "Perfect in A" > effect in C" are false

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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, σ_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 $\Pr(\text{effect in A} > \text{effect in C}) = 0.7$

Statistical framework and multilevel modeling

Comparing test scores across states

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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, σ_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 $\Pr(\text{effect in A} > \text{effect in C}) = 0.7$

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

SAT coaching in 8 schools

Effects of electromagnetic fields at 38 frequencies

Teacher and school effects in NYC schools

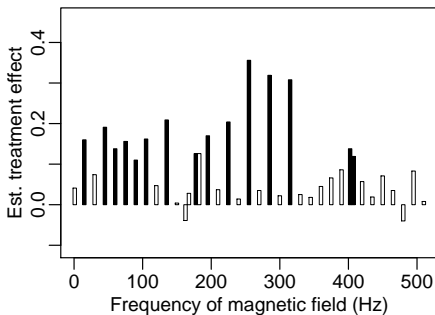
Grades and classroom seating

Beautiful parents have more daughters

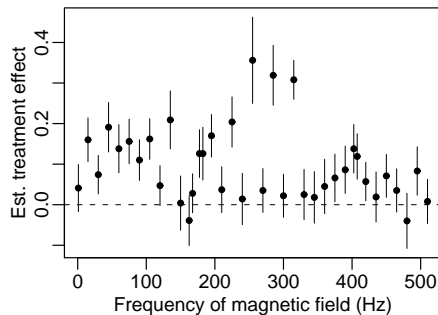
Comparing test scores across states

Effects of electromagnetic fields at 38 frequencies

Estimates with statistical significance



Estimates \pm standard errors



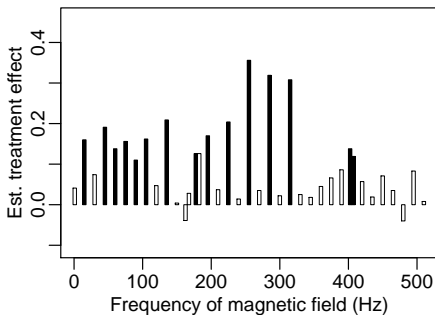
- ▶ Background: electromagnetic fields and cancer
- ▶ Original article summarized using p-values
- ▶ Confidence intervals show comparisons more clearly

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

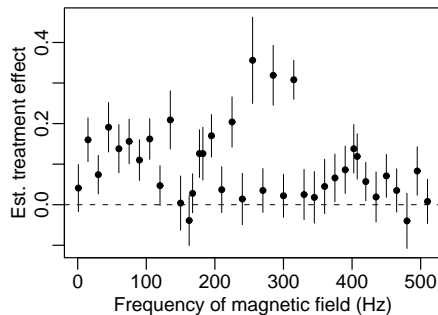
SAT coaching in 8 schools
Effects of electromagnetic fields at 38 frequencies
Teacher and school effects in NYC schools
Grades and classroom seating
Beautiful parents have more daughters
Comparing test scores across states

Effects of electromagnetic fields at 38 frequencies

Estimates with statistical significance



Estimates \pm standard errors



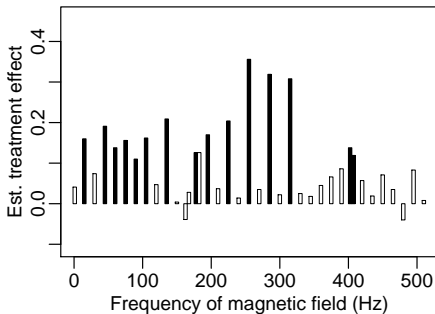
- Background: electromagnetic fields and cancer
- Original article summarized using p-values
- Confidence intervals show comparisons more clearly

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

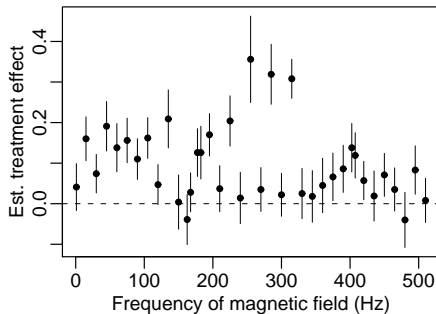
SAT coaching in 8 schools
Effects of electromagnetic fields at 38 frequencies
Teacher and school effects in NYC schools
Grades and classroom seating
Beautiful parents have more daughters
Comparing test scores across states

Effects of electromagnetic fields at 38 frequencies

Estimates with statistical significance



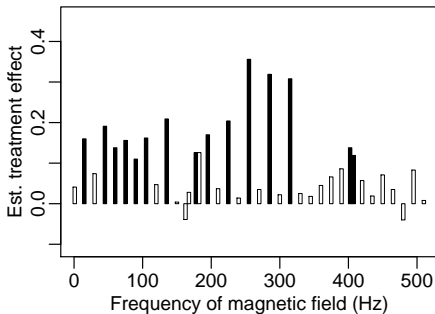
Estimates \pm standard errors



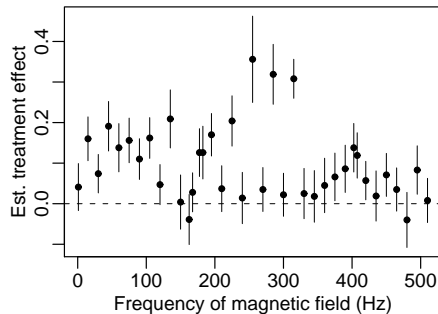
- ▶ Background: electromagnetic fields and cancer
- ▶ Original article summarized using p-values
- ▶ Confidence intervals show comparisons more clearly

Effects of electromagnetic fields at 38 frequencies

Estimates with statistical significance



Estimates \pm standard errors



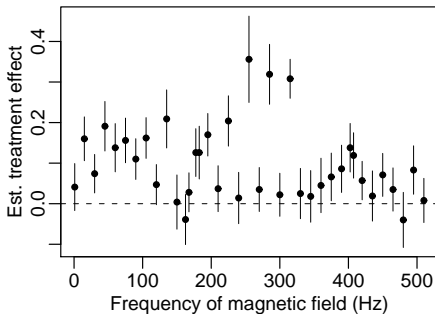
- ▶ Background: electromagnetic fields and cancer
- ▶ Original article summarized using p-values
- ▶ Confidence intervals show comparisons more clearly

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

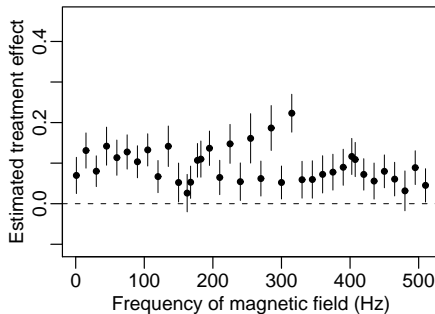
SAT coaching in 8 schools
Effects of electromagnetic fields at 38 frequencies
Teacher and school effects in NYC schools
Grades and classroom seating
Beautiful parents have more daughters
Comparing test scores across states

Separate estimates and hierarchical Bayes estimates

Estimates \pm standard errors



Multilevel estimates \pm standard errors



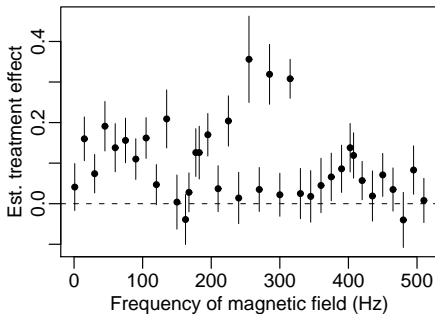
- ▶ Most comparisons are no longer statistically significant
- ▶ “Multiple comparisons” is less of a concern
- ▶ We moved the intervals together instead of widening them!

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

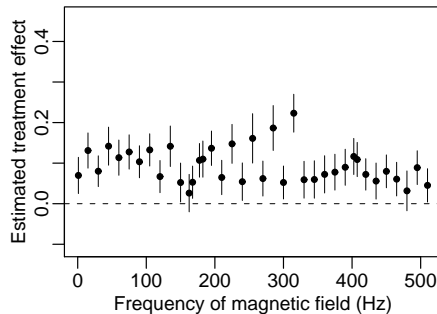
SAT coaching in 8 schools
Effects of electromagnetic fields at 38 frequencies
Teacher and school effects in NYC schools
Grades and classroom seating
Beautiful parents have more daughters
Comparing test scores across states

Separate estimates and hierarchical Bayes estimates

Estimates \pm standard errors



Multilevel estimates \pm standard errors



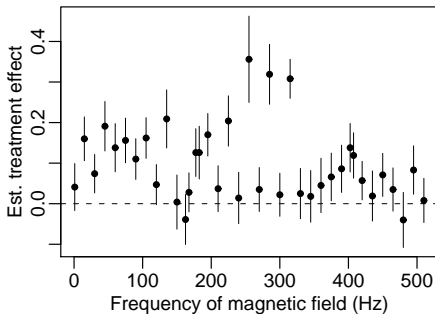
- ▶ Most comparisons are no longer statistically significant
- ▶ “Multiple comparisons” is less of a concern
- ▶ We moved the intervals together instead of widening them!

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

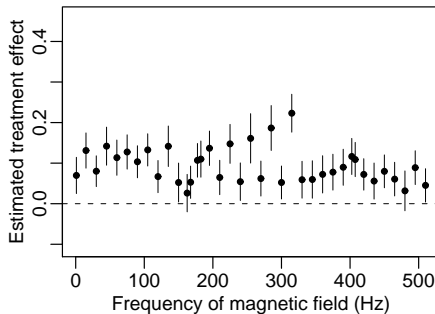
SAT coaching in 8 schools
Effects of electromagnetic fields at 38 frequencies
Teacher and school effects in NYC schools
Grades and classroom seating
Beautiful parents have more daughters
Comparing test scores across states

Separate estimates and hierarchical Bayes estimates

Estimates \pm standard errors



Multilevel estimates \pm standard errors



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

What is the multiple comparisons problem?

Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

SAT coaching in 8 schools

Effects of electromagnetic fields at 38 frequencies

Teacher and school effects in NYC schools

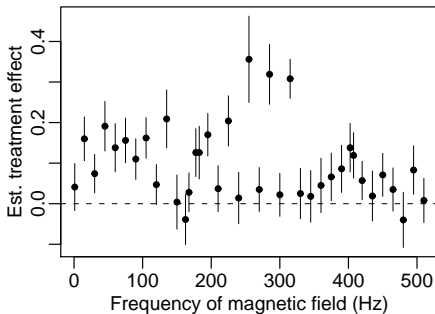
Grades and classroom seating

Beautiful parents have more daughters

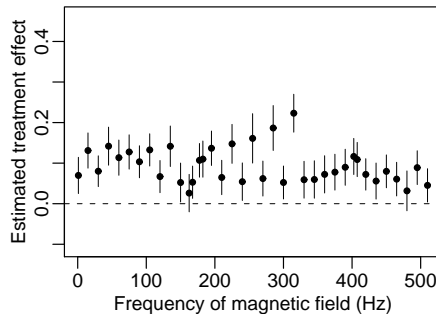
Comparing test scores across states

Separate estimates and hierarchical Bayes estimates

Estimates \pm standard errors



Multilevel estimates \pm standard errors



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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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”)
 - ▶ 1000 children in 1000 families
 - ▶ 50% of children of parents in top 10% of attractiveness
- ▶ 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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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!

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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 comparisons problem!
- ▶ Our multilevel inferences

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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 comparisons problem!
- ▶ Our multilevel inferences

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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 comparisons problem!
- ▶ Our multilevel inferences

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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 comparisons problem!
- ▶ Our multilevel inferences

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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 comparisons problem!
- ▶ Our multilevel inferences

Some stories

Statistical framework and multilevel modeling

- 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

Classical multiple comparisons inferences for NAEP

Region (R)		Country (C)		City (CITY)		State (S)		County (CO)		Municipality (MUN)		District (D)		Ward (W)		Precinct (P)		Block (B)		Lot (L)		Parcel (PAR)		Address (ADD)		Zip (ZIP)		Phone (PHN)		Email (EML)		Website (WEB)		Social (SOC)		Media (MED)		Other (OTH)		Notes (NOTES)		Comments (COMMENTS)		Status (STAT)		Type (TYPE)		Category (CAT)		Subcategory (SUBCAT)		Priority (PRI)		Severity (SEV)		Impact (IMP)		Risk (RISK)		Score (SCR)		Rating (RAT)		Grade (GRD)		Level (LEV)		Class (CLS)		Group (GRP)		Team (TEAM)		Project (PROJ)		Phase (PHASE)		Task (TASK)		Step (STEP)		Action (ACTN)		Result (RESLT)		Feedback (FBK)		Review (REVW)		Approval (APPRVL)		Signature (SIGN)		Date (DATE)		Time (TIME)		Location (LOCN)		Weather (WEATH)		Season (SEASON)		Month (MONTH)		Day (DAY)		Week (WEEK)		Year (YEAR)		Decade (DECADE)		Century (CENTURY)		Era (ERA)		Period (PERIOD)		Epoch (EPOCH)		Age (AGE)		Generation (GEN)		Cohort (COHORT)		Generation (GEN2)		Cohort (COHORT2)		Generation (GEN3)		Cohort (COHORT3)		Generation (GEN4)		Cohort (COHORT4)		Generation (GEN5)		Cohort (COHORT5)		Generation (GEN6)		Cohort (COHORT6)		Generation (GEN7)		Cohort (COHORT7)		Generation (GEN8)		Cohort (COHORT8)		Generation (GEN9)		Cohort (COHORT9)		Generation (GEN10)		Cohort (COHORT10)		Generation (GEN11)		Cohort (COHORT11)		Generation (GEN12)		Cohort (COHORT12)		Generation (GEN13)		Cohort (COHORT13)		Generation (GEN14)		Cohort (COHORT14)		Generation (GEN15)		Cohort (COHORT15)		Generation (GEN16)		Cohort (COHORT16)		Generation (GEN17)		Cohort (COHORT17)		Generation (GEN18)		Cohort (COHORT18)		Generation (GEN19)		Cohort (COHORT19)		Generation (GEN20)		Cohort (COHORT20)		Generation (GEN21)		Cohort (COHORT21)		Generation (GEN22)		Cohort (COHORT22)		Generation (GEN23)		Cohort (COHORT23)		Generation (GEN24)		Cohort (COHORT24)		Generation (GEN25)		Cohort (COHORT25)		Generation (GEN26)		Cohort (COHORT26)		Generation (GEN27)		Cohort (COHORT27)		Generation (GEN28)		Cohort (COHORT28)		Generation (GEN29)		Cohort (COHORT29)		Generation (GEN30)		Cohort (COHORT30)		Generation (GEN31)		Cohort (COHORT31)		Generation (GEN32)		Cohort (COHORT32)		Generation (GEN33)		Cohort (COHORT33)		Generation (GEN34)		Cohort (COHORT34)		Generation (GEN35)		Cohort (COHORT35)		Generation (GEN36)		Cohort (COHORT36)		Generation (GEN37)		Cohort (COHORT37)		Generation (GEN38)		Cohort (COHORT38)		Generation (GEN39)		Cohort (COHORT39)		Generation (GEN40)		Cohort (COHORT40)		Generation (GEN41)		Cohort (COHORT41)		Generation (GEN42)		Cohort (COHORT42)		Generation (GEN43)		Cohort (COHORT43)		Generation (GEN44)		Cohort (COHORT44)		Generation (GEN45)		Cohort (COHORT45)		Generation (GEN46)		Cohort (COHORT46)		Generation (GEN47)		Cohort (COHORT47)		Generation (GEN48)		Cohort (COHORT48)		Generation (GEN49)		Cohort (COHORT49)		Generation (GEN50)		Cohort (COHORT50)		Generation (GEN51)		Cohort (COHORT51)		Generation (GEN52)		Cohort (COHORT52)		Generation (GEN53)		Cohort (COHORT53)		Generation (GEN54)		Cohort (COHORT54)		Generation (GEN55)		Cohort (COHORT55)		Generation (GEN56)		Cohort (COHORT56)		Generation (GEN57)		Cohort (COHORT57)		Generation (GEN58)		Cohort (COHORT58)		Generation (GEN59)		Cohort (COHORT59)		Generation (GEN60)		Cohort (COHORT60)		Generation (GEN61)		Cohort (COHORT61)		Generation (GEN62)		Cohort (COHORT62)		Generation (GEN63)		Cohort (COHORT63)		Generation (GEN64)		Cohort (COHORT64)		Generation (GEN65)		Cohort (COHORT65)		Generation (GEN66)		Cohort (COHORT66)		Generation (GEN67)		Cohort (COHORT67)		Generation (GEN68)		Cohort (COHORT68)		Generation (GEN69)		Cohort (COHORT69)		Generation (GEN70)		Cohort (COHORT70)		Generation (GEN71)		Cohort (COHORT71)		Generation (GEN72)		Cohort (COHORT72)		Generation (GEN73)		Cohort (COHORT73)		Generation (GEN74)		Cohort (COHORT74)		Generation (GEN75)		Cohort (COHORT75)		Generation (GEN76)		Cohort (COHORT76)		Generation (GEN77)		Cohort (COHORT77)		Generation (GEN78)		Cohort (COHORT78)		Generation (GEN79)		Cohort (COHORT79)		Generation (GEN80)		Cohort (COHORT80)		Generation (GEN81)		Cohort (COHORT81)		Generation (GEN82)		Cohort (COHORT82)		Generation (GEN83)		Cohort (COHORT83)		Generation (GEN84)		Cohort (COHORT84)		Generation (GEN85)		Cohort (COHORT85)	
Western (W)	Eastern (E)	Central (C)	South (S)	Northern (N)	Midwestern (M)	Southeastern (SE)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW)	Northcentral (NC)	Southcentral (SC)	Northwestern (NW)	Southwestern (SW																																																																																																																																																																																															

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

Classical inferences for NAEP: close-up

United States by State and District																																																	
Maine (ME)																																																	
Minnesota (MN)																																																	
Connecticut (CT)																																																	
Wisconsin (WI)																																																	
North Dakota (ND)																																																	
Indiana (IN)																																																	
Iowa (IA) #																																																	
Massachusetts (MA)																																																	
Texas (TX)																																																	
Nebraska (NE)																																																	
Montana (MT) #																																																	
New Jersey (NJ) #																																																	
Utah (UT)																																																	
Michigan (MI) #																																																	
Pennsylvania (PA) #																																																	
Colorado (CO)																																																	
Washington (WA)																																																	
Vermont (VT) #																																																	
Missouri (MO)																																																	
North Carolina (NC)																																																	
DDE SS (DD)																																																	
Alaska (AK) #																																																	
Oregon (OR)																																																	
West Virginia (WV)																																																	
DODDS (DO)																																																	
Wyoming (WY)																																																	
Virginia (VA)																																																	
New York (NY) #																																																	
Maryland (MD)																																																	
Rhode Island (RI)																																																	
Kentucky (KY)																																																	
Tennessee (TN)																																																	
Nevada (NV) #																																																	
Arizona (AZ) #																																																	
Arkansas (AR)																																																	
Florida (FL)																																																	
Georgia (GA)																																																	
Delaware (DE)																																																	
Hawaii (HI)																																																	
New Mexico (NM)																																																	
South Carolina (SC) #																																																	
Alabama (AL)																																																	
California (CA)																																																	
Louisiana (LA)																																																	
Mississippi (MS)																																																	

ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	MT	NJ	UT	MI	PA	CO	WA	VT	MO	NC	DD	AK	OR	WV	DO	WY	VA	NY	MD	RI	KY	TN	NV	AZ	AR	FL	GA	DE	HI	NM	SC	AL	CA	LA	MS		
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	MT	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA	TX	TX	TX	NE	NE	NJ	UT	MI	PA	PA	PA	PA	PA	PA	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX
ME	MN	CT	WI	ND	IN	IA	MA	TX	NE	NE	NJ	UT	MI	PA	PA	CO	WA</																													

What is the multiple comparisons problem?
 Why don't I (usually) care?
 Some stories
 Statistical framework and multilevel modeling

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

Multilevel inferences for NAEP: close-up

Comparisons of Average Mathematics Scale Scores for
 Grade 4 Public Schools in Participating Jurisdictions

Maine	Minnesota	Connecticut	Wisconsin	North Dakota	Indiana	Iowa	Massachusetts	Texas	Nebraska	Montana	New Jersey	Utah	Michigan	Pennsylvania	Colorado	Washington	Vermont	Missouri	North Carolina	Alaska	Oregon	West Virginia	Wyoming	Virginia	New York	Maryland	Rhode Island	Kentucky	Tennessee	Nevada	Arizona	Arkansas	Florida	Georgia	Delaware	Hawaii	New Mexico	South Carolina	Alabama	California	Louisiana	Mississippi					
ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME		
MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN		
CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	
WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	WI	
ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	
IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	
IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	IA	
MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	
TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	TX	
NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	
MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	
NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ		
UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT	UT
MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	
PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	PA	
CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	CO	



What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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?

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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?

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

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?

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

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

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?

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

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

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?

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

Something for nothing? A free lunch?

- ▶ Classical multiple comparisons worries about $\theta_1 = \theta_2 = \dots = \theta_{50}$
- ▶ 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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

Something for nothing? A free lunch?

- ▶ Classical multiple comparisons worries about $\theta_1 = \theta_2 = \dots = \theta_{50}$
- ▶ 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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

Something for nothing? A free lunch?

- ▶ Classical multiple comparisons worries about
 $\theta_1 = \theta_2 = \dots = \theta_{50}$
- ▶ 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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

Something for nothing? A free lunch?

- ▶ Classical multiple comparisons worries about $\theta_1 = \theta_2 = \dots = \theta_{50}$
- ▶ 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

What is the multiple comparisons problem?
Why don't I (usually) care?

Some stories

Statistical framework and multilevel modeling

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

Something for nothing? A free lunch?

- ▶ Classical multiple comparisons worries about $\theta_1 = \theta_2 = \dots = \theta_{50}$
- ▶ 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

What is the multiple comparisons problem?
Why don't I (usually) care?
Some stories
Statistical framework and multilevel modeling

Message from the examples
Statistical framework
Conclusions
Further thoughts

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 θ_j , for $j = 1, \dots, J$ (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 θ_j , for $j = 1, \dots, J$ (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 θ_j , for $j = 1, \dots, J$ (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 θ_j , for $j = 1, \dots, J$ (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 θ_j , for $j = 1, \dots, J$ (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 θ_j , for $j = 1, \dots, J$ (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 θ_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 θ_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 $\theta_j > \theta_k$, but I claim they're the same
 - ▶ I've never claimed $\theta_j = \theta_k$
- ▶ 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 θ_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 θ_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 θ_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 θ_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 θ_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 θ_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 θ_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 θ_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 θ_j 's

Multilevel (hierarchical) modeling

- ▶ Key parameter: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:

▶ *Shrinkage* (regression)

▶ *Regression* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

▶ *ANOVA* (ANOVA)

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

Multilevel (hierarchical) modeling

- ▶ Key parameter: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation

▶ σ_θ can be large or small

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

Multilevel (hierarchical) modeling

- ▶ Key parameter: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ "Very few claims with confidence (big error, but 0.1)"
 - ▶ $\sigma_\theta \rightarrow \infty$: 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: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ Very few claims with confidence (far fewer than 5%)
 - ▶ $\sigma_\theta \rightarrow \infty$: 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: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ Very few claims with confidence (far fewer than 5%)
 - ▶ $\sigma_\theta \rightarrow \infty$: 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: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ Very few claims with confidence (far fewer than 5%)
 - ▶ $\sigma_\theta \rightarrow \infty$: 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: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ Very few claims with confidence (far fewer than 5%)
 - ▶ $\sigma_\theta \rightarrow \infty$: large variation
 - ▶ Multilevel model is equivalent to estimating each θ_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: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ Very few claims with confidence (far fewer than 5%)
 - ▶ $\sigma_\theta \rightarrow \infty$: large variation
 - ▶ Multilevel model is equivalent to estimating each θ_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: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ Very few claims with confidence (far fewer than 5%)
 - ▶ $\sigma_\theta \rightarrow \infty$: large variation
 - ▶ Multilevel model is equivalent to estimating each θ_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: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ Very few claims with confidence (far fewer than 5%)
 - ▶ $\sigma_\theta \rightarrow \infty$: large variation
 - ▶ Multilevel model is equivalent to estimating each θ_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: σ_θ , the sd of the true θ_j 's
- ▶ Understand through special cases:
 - ▶ $\sigma_\theta \approx 0$: no variation
 - ▶ Multilevel model pools the estimated θ_j 's toward each other
 - ▶ "Multiple comparisons" correction is done by shrinking comparisons
 - ▶ Very few claims with confidence (far fewer than 5%)
 - ▶ $\sigma_\theta \rightarrow \infty$: large variation
 - ▶ Multilevel model is equivalent to estimating each θ_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