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

Andrew Gelman, Jennifer Hill, and Masanao Yajima

Department of Political Science, Columbia University School of Education and Human Development, NYU Department of Statistics, UCLA

15 July 2009



- ► My experiences 20 years ago
- A longstanding principle in statistics
- I wish I wasn't here

- ▶ My experiences 20 years ago
- ► A longstanding principle in statistics
- ▶ I wish I wasn't here

- My experiences 20 years ago
- A longstanding principle in statistics
- ▶ I wish I wasn't here

- ▶ My experiences 20 years ago
- A longstanding principle in statistics
- I wish I wasn't here

- ▶ PET scans and schizophrenia
- Two-way ANOVA: people × regions of interest
- ▶ Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism

- PET scans and schizophrenia
- Two-way ANOVA: people × regions of interest
- ▶ Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism

- ▶ PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- ▶ Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism

- ▶ PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism
  - ► Ranked as the #25 neuroscience journal
  - n ilipact factor 3.7
  - By comparison, impact factors of top statistics pormattees.
     1ASA 1.6, 1RSS 1.5, Ann Stat. 1.3, Ann Prob. 0.0, Blow
    - 1.8, Biometrics 1.1, Stat Sci 2.0, Technometrics 1.3

- PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism
  - Ranked as the #25 neuroscience journal
  - ▶ Impact factor 5.7
  - ▶ By comparison, impact factors of top statistics journals:
  - JASA 1.6, JRSS 1.5, Ann Stat 1.3, Ann Prob 0.9, Biometrika 1.8. Biometrics 1.1. Stat Sci 2.0. Technometrics 1.3



- PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism
  - ▶ Ranked as the #25 neuroscience journal
  - Impact factor 5.7
  - ▶ By comparison, impact factors of top statistics journals
  - JASA 1.6, JRSS 1.5, Ann Stat 1.3, Ann Prob 0.9, Biometrika 1.8, Biometrics 1.1, Stat Sci 2.0, Technometrics 1.3



- PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism
  - ▶ Ranked as the #25 neuroscience journal
  - Impact factor 5.7
  - By comparison, impact factors of top statistics journals
  - JASA 1.6, JRSS 1.5, Ann Stat 1.3, Ann Prob 0.9, Biometrika
    - 4 D > 4 A > 4 B > 4 B > B 9 9 9

- PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism
  - ▶ Ranked as the #25 neuroscience journal
  - Impact factor 5.7
  - By comparison, impact factors of top statistics journals:
  - JASA 1.6, JRSS 1.5, Ann Stat 1.3, Ann Prob 0.9, Biometrika 1.8, Biometrics 1.1, Stat Sci 2.0, Technometrics 1.3

- ▶ PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism
  - Ranked as the #25 neuroscience journal
  - Impact factor 5.7
  - By comparison, impact factors of top statistics journals:
  - JASA 1.6, JRSS 1.5, Ann Stat 1.3, Ann Prob 0.9, Biometrika 1.8, Biometrics 1.1, Stat Sci 2.0, Technometrics 1.3



- PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism
  - Ranked as the #25 neuroscience journal
  - Impact factor 5.7
  - By comparison, impact factors of top statistics journals:
  - ▶ JASA 1.6, JRSS 1.5, Ann Stat 1.3, Ann Prob 0.9, Biometrika 1.8, Biometrics 1.1, Stat Sci 2.0, Technometrics 1.3



- PET scans and schizophrenia
- ► Two-way ANOVA: people × regions of interest
- Within-person and between-person variability
- Published in the Journal of Cerebral Blood Flow and Metabolism
  - Ranked as the #25 neuroscience journal
  - ▶ Impact factor 5.7
  - By comparison, impact factors of top statistics journals:
  - ▶ JASA 1.6, JRSS 1.5, Ann Stat 1.3, Ann Prob 0.9, Biometrika 1.8, Biometrics 1.1, Stat Sci 2.0, Technometrics 1.3



- Difficulty in following Bill Cosby and Bob Newhart
- ▶ Where should be putting our technical thinking?
  - Scientific modeling
  - Mapping scientific models to data collection and statisticaed
    - #Beildies a secondation leader to des
  - Building a cumulative knowledge base
  - lias etc.

- ▶ Difficulty in following Bill Cosby and Bob Newhart
- ▶ Where should be putting our technical thinking?
- Scientific modeling
  - Mapping scientific models to data collection and statistical
    - modeling
    - "Building a cumulative knowledge base"
  - But not ... discussions of p-values, cross-validation, selectionments.

- ▶ Difficulty in following Bill Cosby and Bob Newhart
- ▶ Where should be putting our technical thinking?
  - Scientific modeling
  - Mapping scientific models to data collection and statistical modeling
  - "Building a cumulative knowledge base"
  - But not .... discussions of p-values, cross-validation, selection bias, etc.

- ▶ Difficulty in following Bill Cosby and Bob Newhart
- Where should be putting our technical thinking?
  - Scientific modeling
  - Mapping scientific models to data collection and statistical modeling
  - "Building a cumulative knowledge base"
  - ▶ But not ... discussions of p-values, cross-validation, selection bias, etc.

- ▶ Difficulty in following Bill Cosby and Bob Newhart
- Where should be putting our technical thinking?
  - Scientific modeling
  - Mapping scientific models to data collection and statistical modeling
  - "Building a cumulative knowledge base"
  - ▶ But not ... discussions of p-values, cross-validation, selection bias, etc.

- ▶ Difficulty in following Bill Cosby and Bob Newhart
- Where should be putting our technical thinking?
  - Scientific modeling
  - Mapping scientific models to data collection and statistical modeling
  - "Building a cumulative knowledge base"
  - But not ... discussions of p-values, cross-validation, selection bias, etc.



- ▶ Difficulty in following Bill Cosby and Bob Newhart
- Where should be putting our technical thinking?
  - Scientific modeling
  - Mapping scientific models to data collection and statistical modeling
  - "Building a cumulative knowledge base"
  - But not ... discussions of p-values, cross-validation, selection bias, etc.



- ▶ 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 about it?
- Some stories
- ► Statistical framework and multilevel modeling

- ▶ 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 about it?
- Some stories
- Statistical framework and multilevel modeling

- ▶ 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 about it?
- Some stories
- Statistical framework and multilevel modeling

- ▶ 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 about it?
- Some stories
- Statistical framework and multilevel modeling

- ▶ 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 about it?
- 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

# 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

- When looked at one way, multiple comparisons seem like a major worry
- ▶ But from another perspective, they don't matter at all:

- 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

- 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

- 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

- 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

- 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

- ► 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
- 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
- ▶ 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
- ▶ 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
- 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
- ▶ 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
- 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

	Estimated	Standard error
	treatment	of effect
School	effect, $y_j$	estimate, $\sigma_j$
Α	28	15
В	8	10
C	-3	16
D	7	11
Ε	-1	9
F	1	11
G	18	10
Н	12	18

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



	Estimated treatment	Standard error of effect
School	effect, $y_j$	estimate, $\sigma_j$
Α	28	15
В	8	10
C	-3	16
D	7	11
Ε	-1	9
F	1	11
G	18	10
Н	12	18

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



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

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



	Estimated	Standard error
	treatment	of effect
School	effect, $y_j$	estimate, $\sigma_j$
Α	28	15
В	8	10
C	-3	16
D	7	11
Ε	-1	9
F	1	11
G	18	10
Н	12	18

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



	Estimated	Standard error
	treatment	of effect
School	effect, <i>y<sub>j</sub></i>	estimate, $\sigma_j$
А	28	15
В	8	10
C	-3	16
D	7	11
Е	-1	9
F	1	11
G	18	10
Н	12	18

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



# SAT coaching in 8 schools 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

	Estimated	Standard error
	treatment	of effect
School	effect, <i>y<sub>j</sub></i>	estimate, $\sigma_j$
А	28	15
В	8	10
C	-3	16
D	7	11
Ε	-1	9
F	1	11
G	18	10
Н	12	18

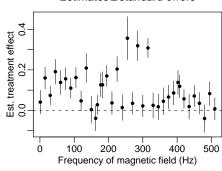
- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
  - Overlappling confidence intervals for the 8 school effects
  - ▶ Statements such as Pr(effect in A > effect in C) = 0.7



# Effects of electromagnetic fields at 38 frequencies

#### Estimates with statistical significance

# 



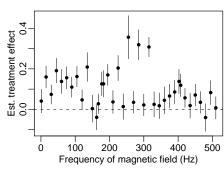
- ▶ Background: electromagnetic fields and cancer
- Original article summarized using p-values
- ► Confidence intervals show comparisons more clearly



# Effects of electromagnetic fields at 38 frequencies

#### Estimates with statistical significance

# 

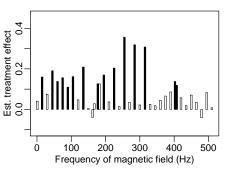


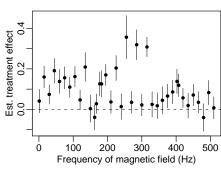
- ▶ Background: electromagnetic fields and cancer
- Original article summarized using p-values
- ► Confidence intervals show comparisons more clearly



# Effects of electromagnetic fields at 38 frequencies

#### Estimates with statistical significance





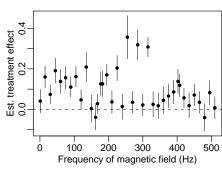
- ▶ Background: electromagnetic fields and cancer
- Original article summarized using p-values
- ► Confidence intervals show comparisons more clearly



# Effects of electromagnetic fields at 38 frequencies

#### Estimates with statistical significance

# 



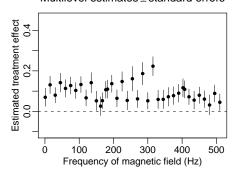
- ▶ Background: electromagnetic fields and cancer
- Original article summarized using p-values
- Confidence intervals show comparisons more clearly



# Separate estimates and hierarchical Bayes estimates



# Estimates Estandard errors By The Figure 1 of the Control of the

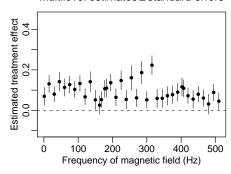


- Most comparisons are no longer statistically significant
- ► "Multiple comparisons" is less of a concern
- ▶ We moved the intervals together instead of widening them!

### Separate estimates and hierarchical Bayes estimates



# 

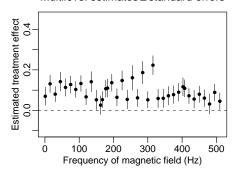


- ▶ Most comparisons are no longer statistically significant
- "Multiple comparisons" is less of a concern
- ▶ We moved the intervals together instead of widening them!

# Separate estimates and hierarchical Bayes estimates



# Est: treatment effect of the control of the control

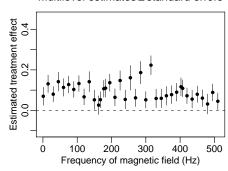


- Most comparisons are no longer statistically significant
- "Multiple comparisons" is less of a concern
- ▶ We moved the intervals together instead of widening them!

### Separate estimates and hierarchical Bayes estimates



# Estimated Estandard errors The standard errors Th



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



- ► 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!

- ► 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!

- 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!

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



# Grades and classroom seating

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

# Grades and classroom seating

#### ► Classroom demonstration

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

# Grades and classroom seating

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

# Grades and classroom seating

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



# Grades and classroom seating

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



## Grades and classroom seating

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



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

- 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")

- ▶ 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 × 4 possible time summaries!



- 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")

- ▶ 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 × 4 possible time summaries!



- 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 × 4 possible time summaries!



- 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
- lacktriangle 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 × 4 possible time summaries!



- 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 × 4 possible time summaries!



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

- 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 × 4 possible time summaries!

- 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 × 4 possible time summaries!

- 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 × 4 possible time summaries!



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

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

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

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

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

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

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

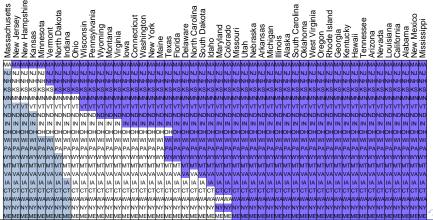
# Classical inferences for NAEP: close-up

Major (ME)	Minnesota (MN)	Connecticut (CT)	Wisconsin (WI)	North Dakota (ND)	Indiana (IN)	lowa (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)	DDESS (DD)	Alaska (AK)#	Oregon (OR)	WestVirginia (WV)	Dodds (DO)	Wyoming (WY)	Virginia (VA)	New York (NY)#	Maryland (MD)	Rhode Island (R1)	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)
М	Е МЕ	М	МЕ	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
м	ими	MN N	MN N	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
c	т ст	г	гст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст	ст
W	n w	ı w	ı wı	WI	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı	wı
N	D NE	NE	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
11	NI 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
1/	A 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
М	A MA	A MA	A MA	MA	MA	MA	MA	MA	MA	MA	MA	МА	MA	MA	МА	МА	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	МА	MA	MA
T:	х Тх	(TX	( TX	TX	ТX	ТX	ТX	ТX	ТX	TX	ТX	TX	TX	ТX	тх	TX	ТХ	ТX	ТX	ТX	ΤX	ТX	TX	тх	ΤX	ΤX	ΤX	ТX	ТХ	ΤX	ΤX	тх	ТX	ТX	ΤX	тх	ТХ	TX	тх	тх	тх	тх	тх	тх
N	E 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
м	ТМТ	гМ	ГМТ	МТ	МТ	МТ	МТ	МТ	МТ	МТ	мт	мт	мт	MT	МТ	МТ	МТ	МТ	MT	МТ	МТ	МТ	МТ	МТ	MT	МТ	MT	MT	MT	МТ	МТ	МТ	MT	МТ	МТ	МТ	MT	МТ	мт	мт	мт	мт	мт	мт
N	ΙΝ	I NJ	ΙΝ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	ΝJ	ΝJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ	ИЛ	NJ	NJ	NJ	NJ
U	T UT	דט	т	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
N	ш м	М	і мі	МІ	MI	MI	MI	МІ	MI	М	М	МІ	МІ	МІ	МІ	МІ	МІ	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	МІ	MI	MI	MI	MI	MI	MI	MI	МІ	МІ	МІ	МІ	МІ	МІ	MI
P	A 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
C	o cc	cc	co	СО	co	со	co	со	со	со	co	co	co	со	со	со	СO	СO	CO	CO	CO	СO	СO	со	сo	co	co	co	со	со	со	со	co	CO	CO	co	со	со	co	со	со	co	со	co

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 Schores for Grade 4 Public Schools in Participating Jurisdictions



- ▶ 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?

- ▶ 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?

- ▶ 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?

- ▶ 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?



- ▶ 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?

► Classical multiple comparisons worries about

- 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

▶ Classical multiple comparisons worries about

$$\theta_1 = \theta_2 = \cdots = \theta_{50}$$

- ▶ Not an issue with NAEP
- Multilevel model estimates the group-level variance, decides based on the data how much to adjust
- Classical procedure does not learn from the data



Classical multiple comparisons worries about

$$\theta_1 = \theta_2 = \cdots = \theta_{50}$$

- ▶ Not an issue with NAEP
- Multilevel model estimates the group-level variance, decides based on the data how much to adjust
- Classical procedure does not learn from the data



► 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



► Classical multiple comparisons worries about  $\theta_1 = \theta_2 = \cdots = \theta_{50}$ 

- Not an issue with NAEP
- Multilevel model estimates the group-level variance, decides based on the data how much to adjust
- ► Classical procedure does not learn from the data



# Message from the examples

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

# Message from the examples

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

# Message from the examples

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

- ▶ Goal is to estimate  $\theta_j$ , for j = 1, ..., 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

- ▶ Goal is to estimate  $\theta_j$ , for j = 1, ..., 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

- ▶ Goal is to estimate  $\theta_j$ , for j = 1, ..., 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

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

### Statistical framework

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

### Statistical framework

- ▶ Goal is to estimate  $\theta_j$ , for j = 1, ..., 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

- ▶ I've never made a Type 1 error in my life
- Type 1 error is θ<sub>j</sub> = θ<sub>k</sub>, but I claim they're different
  I've never studied anything where θ<sub>i</sub> = θ<sub>k</sub>
- ▶ 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<sub>i</sub>'s



- ▶ I've never made a Type 1 error in my life
  - ▶ Type 1 error is  $\theta_i = \theta_k$ , 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!
- lacktriangleq Type S erron:  $heta_1> heta_2$ , but I claim that  $heta_2> heta_1$  (or vice versa)
- ▶ Type S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for  $\theta_j \theta_k$  that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the  $\theta_i$ 's



- ▶ I've never made a Type 1 error in my life
  - ▶ Type 1 error is  $\theta_i = \theta_k$ , 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
- Implies Years I ype S errors can occur when we make *claims with confidence* (i.e., have confidence intervals for  $\theta_i \theta_k$  that exclude zero)
- We want to make fewer claims with confidence when we are less certain about the ranking of the θε's



- ▶ I've never made a Type 1 error in my life
  - ▶ Type 1 error is  $\theta_i = \theta_k$ , 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)
- (i.e., have confidence intervals for  $\theta_i \theta_k$  that exclude zero)
- ► We want to make fewer claims with confidence when we are less certain about the ranking of the θ./s.

- ▶ I've never made a Type 1 error in my life
  - ▶ Type 1 error is  $\theta_i = \theta_k$ , 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  $\theta_i \neq \theta_k$ , 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)
- (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.'s



- ▶ I've never made a Type 1 error in my life
  - ▶ Type 1 error is  $\theta_i = \theta_k$ , 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  $\theta_i \neq \theta_k$ , 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 \(\theta\_t = \theta\_t\) that exclude zero.)
- ► We want to make fewer claims with confidence when we are less certain about the ranking of the θ./s

- ▶ I've never made a Type 1 error in my life
  - ▶ Type 1 error is  $\theta_i = \theta_k$ , 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  $\theta_i \neq \theta_k$ , 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's

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



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



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

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



- ▶ Key parameter:  $\sigma_{\theta}$ , the sd of the true  $\theta_j$ 's
- Understand through special cases:

 $\sigma_{\theta} 
ightarrow \infty$ : large variation

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



- ▶ Key parameter:  $\sigma_{\theta}$ , the sd of the true  $\theta_j$ 's
- Understand through special cases:
  - lacksquare  $\sigma_{ heta} pprox 0$ : no variation

 $ho = \sigma_ heta 
ightarrow \infty$  cor large variation

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



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



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



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



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



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



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



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



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



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



- ▶ "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
- "Adjustments" are a dead end; "modeling" is forward-looking

- ▶ "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
- ▶ "Adjustments" are a dead end; "modeling" is forward-looking

- ▶ "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
- ▶ "Adjustments" are a dead end; "modeling" is forward-looking

- "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
- "Adjustments" are a dead end; "modeling" is forward-looking

- "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
- "Adjustments" are a dead end; "modeling" is forward-looking