5. Why we don't (usually) have to worry about multiple comparisons

- What is the multiple comparisons problem?
- Why don't we (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

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 phenomena with zero effects
  - I don't (usually) study comparisons with zero differences
  - I don't mind being wrong 5% of the time

Some stories

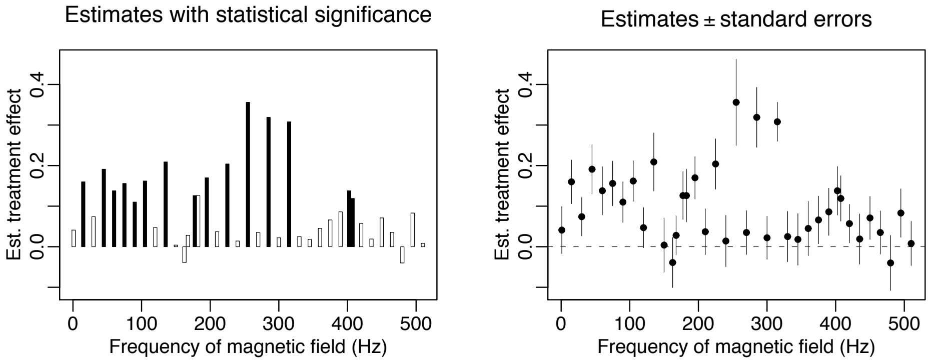
- "Beautiful parents have more daughters" (already discussed)
- SAT coaching in 8 schools
- Effects of electromagnetic fields at 38 frequencies
- Teacher and school effects in NYC schools
- Grades and classroom seating
- Comparing test scores across states

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

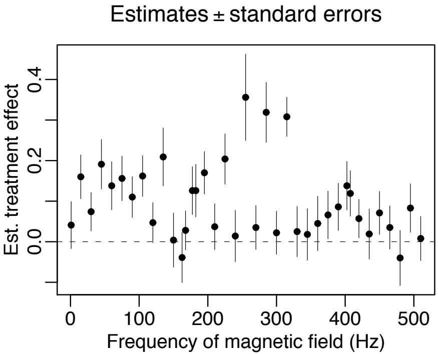
- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
    - Overlapping confidence intervals for the 8 school effects
    - Statements such as $\Pr(\text{effect in A} > \text{effect in C}) = 0.7$

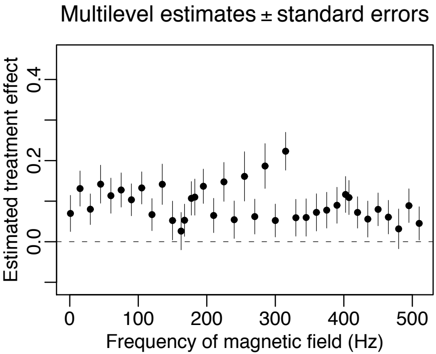
Effects of electromagnetic fields at 38 frequencies



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

Separate estimates and hierarchical Bayes estimates





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

Teacher and school effects in NYC schools

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

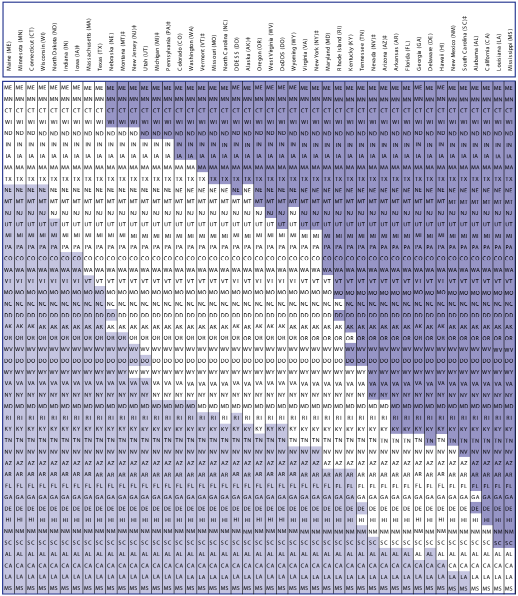
Grades and classroom seating

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

Comparing test scores across states

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

Classical multiple comparisons inferences for NAEP

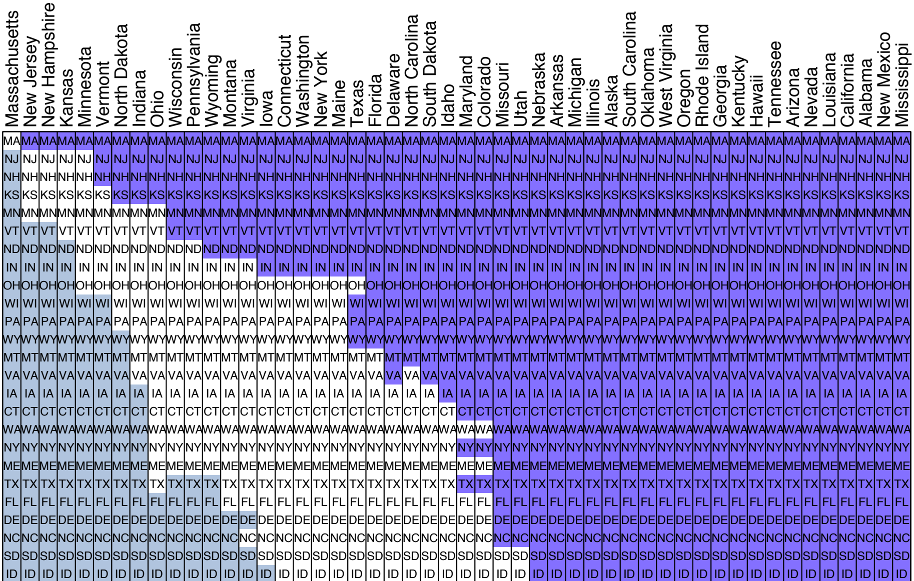


Classical inferences for NAEP: close-up

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)	DDESS (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	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
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
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
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
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
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
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
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
WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA	WA
VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT
MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO	MO
NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD	DD
AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK	AK
OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR

Multilevel inferences for NAEP: close-up

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



NAEP: classical vs. multilevel

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

Something for nothing? A free lunch?

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

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

Type S (sign) errors

- I've never made a Type 1 error in my life
  - Type 1 error is $\theta_j = \theta_k$, but I claim they're different
  - I've never studied anything where $\theta_j = \theta_k$
- I've never made a Type 2 error in my life
  - 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:
  - $\sigma_\theta \approx 0$: 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%)
  - $\sigma_\theta \to \infty$: 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

Bayes shrinks the comparisons toward 0

- Partial pooling tends to reduce the number of statistically significant comparisons:

$$
\begin{aligned}
\text{posterior } \mathrm{E}(\theta_j - \theta_k) &= \frac{\sigma_\theta^2}{\sigma_{\bar{y}}^2 + \sigma_\theta^2}(\bar{y}_j - \bar{y}_k) \\
\text{posterior } \mathrm{sd}(\theta_j - \theta_k) &= \sqrt{2}\sigma_{\bar{y}}\sigma_\theta \Big/ \sqrt{\sigma_{\bar{y}}^2 + \sigma_\theta^2} \\
\text{posterior } z\text{-score of } \theta_j - \theta_k : &\quad \frac{(\bar{y}_j - \bar{y}_k)}{\sqrt{2}\sigma_{\bar{y}}} \cdot \frac{1}{\sqrt{1 + \sigma_{\bar{y}}^2/\sigma_\theta^2}}
\end{aligned}
$$

- Posterior mean of the difference is pulled toward 0, faster than the posterior sd decreases

Summary on multiple comparisons

- 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