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

Andrew Gelman, Jennifer Hill, and Masanao Yajima
Columbia University, New York University, University of California

30 Aug 2011

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?

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

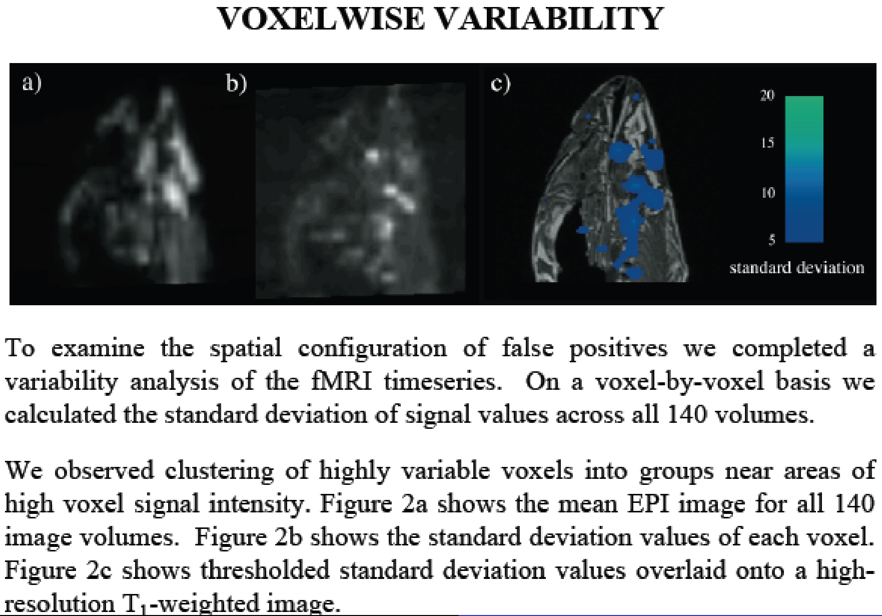
Why don't I (usually) care about multiple comparisons?

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

Some stories

- Medical imaging
- 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

Neural activity in a dead fish



VOXELWISE VARIABILITY

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

“Voodoo correlations” in neuroscience

- Vul, Harris, Winkelman, Pashler:
  - Correlations reported in medical imaging studies are commonly overstated because researchers select the highest values
  - These *statistical* problems are leading to *scientific* errors
- Multiple comparisons corrections do *not* solve the problem:
  - Adjustment of significance levels does not stop the selected correlations themselves from being too high
  - Multiple comparisons methods control the rate of false alarms in a setting where true effects are zero
  - But this is not so relevant in imaging, where differences are not in fact zero
- Discussion in *Perspectives in Psychological Science* (2009)

Where should be putting our technical thinking?

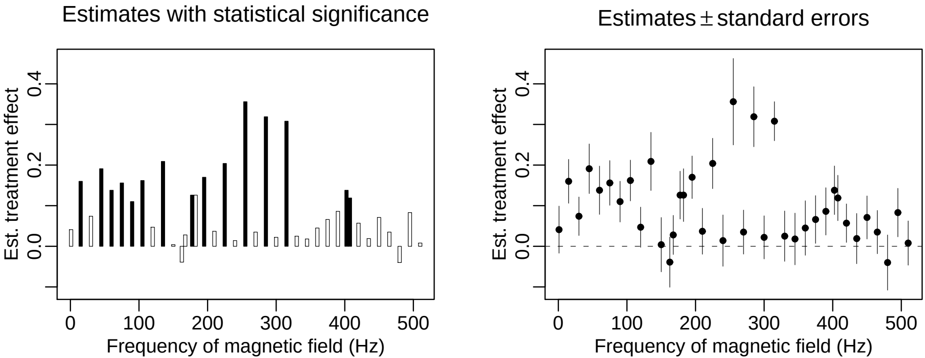
- Scientific modeling
- Mapping scientific models to data collection and statistical modeling
- "Building a cumulative knowledge base"
- But not . . . discussions of p-values, cross-validation, selection bias, etc.

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

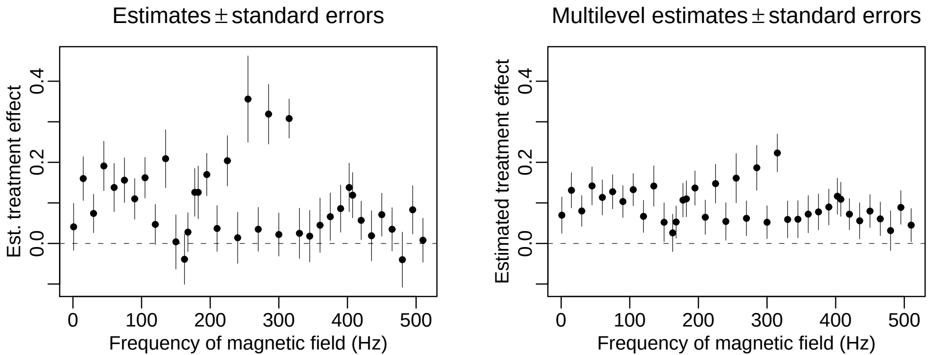
- Separate experiment in each school
- Variation in treatment effects is indistinguishable from 0
- Multilevel Bayes analysis
    - Overlappling confidence intervals for the 8 school effects
    - Statements such as Pr (effect in A $>$ effect in C)$= 0.7$

Effects of electromagnetic fields at 38 frequencies



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

Separate estimates and hierarchical Bayes estimates


Estimates ± standard errors

Multilevel estimates ± 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!

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 (Gelman and Nolan, 2002)
- Assign students random numbers as "grades"
- Ask students with "grades" 0–25 to raise one finger, students with "grades" 75–100 to raise one hand
- Instructor scans the room to find a statistically significant comparison (e.g., "boys on the left side of the classroom have higher grades than girls in the back row")
- This is a pure multiple comparisons problem!

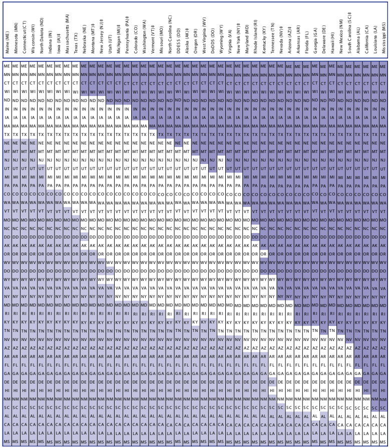
Beautiful parents have more daughters

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

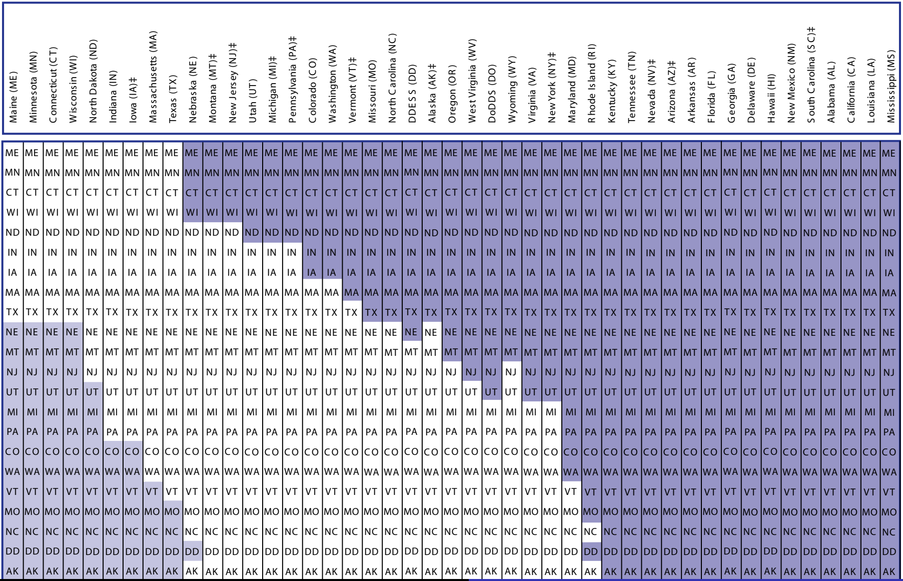
Comparing test scores across states

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

Classical multiple comparisons inferences for NAEP



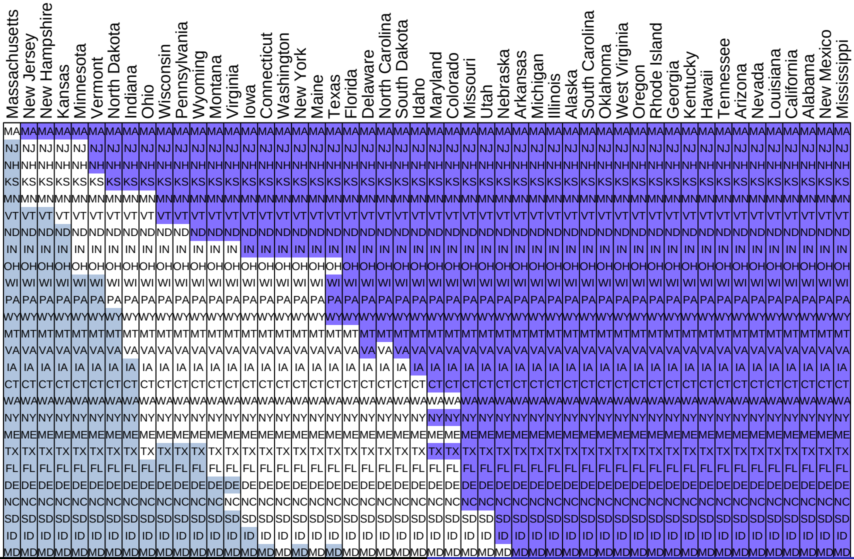
Classical inferences for NAEP: close-up



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

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

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

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

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

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
- "Adjustments" are a dead end; "modeling" is forward-looking
- Open questions
  - Structured models (classrooms, grades, teachers, ages, . . .)
  - Small number of groups