Of beauty, sex, and power: Statistical challenges in estimating small effects

Andrew Gelman[1]
Department of Statistics and Department of Political Science
Columbia University

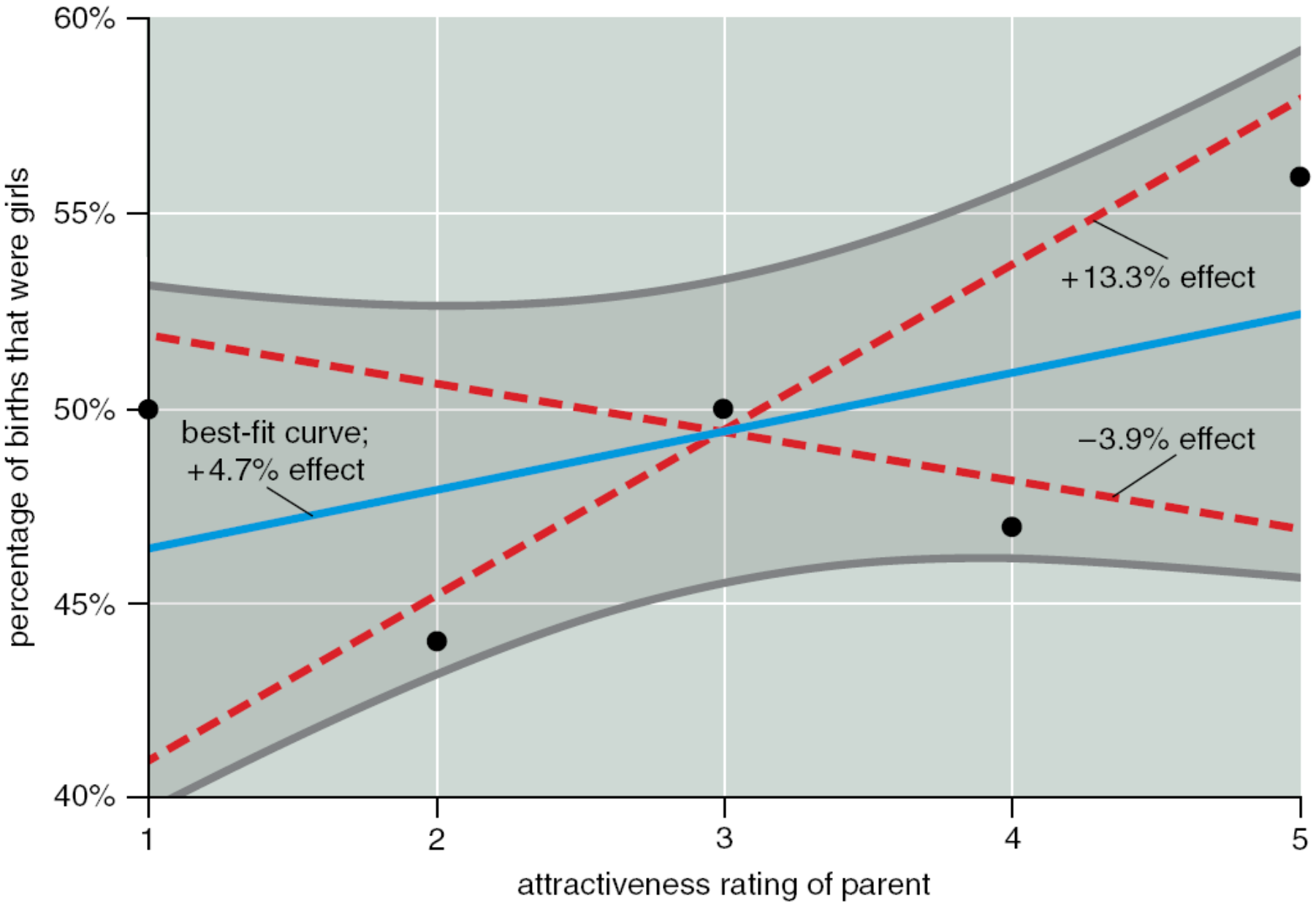
26 April 2011

[1]Collaborators on these projects include David Weakliem, David Park, Boris Shor, Yu-Sung Su, and Daniel Lee

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!

The data and fitted regression line



The larger statistical questions

- The questions
  - How to think about findings that are not "statistically significant"?
  - How to estimate small effects?
- The answers
  - Interpret the estimates in light of how large you think they might be (compared to your previous experience)
  - Estimate the pattern of effects rather than considering each individually

Background on sex ratios

- Pr (boy birth) $\approx$ 51.5%
  - Boys die at a higher rate than girls
  - At age 20, the number of boys and girls is about the same
  - Evolutionary story
- What can affect Pr (boy births)?
  - Race, parental age, birth order, maternal weight, season of birth: effects of about 1% or less
  - Extreme poverty and famine: effects as high as 3%
- We expect any effects of beauty to be less than 1%

Interpreting the Kanazawa study

- Data are consistent with effects ranging from -4% to $+13.3\%$
- More plausibly, consistent with effects less than 0.5% (in either direction!)
- You can take the evolutionary argument in either direction:
  - Beauty is more useful for women than for men, selection pressure, ...
  - Assessed "beauty" is associated with wealthy, dominant ethnic groups who have more power, a trait that is more useful for men than for women, ...
- Results are "more 'vampirical' than 'empirical'—unable to be killed by mere evidence" (Freese, 2007)
- Bottom line
  - Beautiful parents *in this one survey* have more daughters
  - Can't say much about the general population

Another try: data from *People* magazine

The children of each year's "50 most beautiful people"

- We collected data from 1995–2000
- 1995: 32 girls and 24 boys: 57.1% girls (standard error 8.6)
- 1996: 45 girls and 35 boys: $56.2\% \pm 7.8\%$
- 1995 + 1996: $56.6\% \pm 4.3\%$: almost statistically significant!
- 1997: 24 girls and 35 boys, . . .
- Pooling 1995–2000: $47.7\% \pm 2.8\%$: not statistically significantly different from 48.5%

Statistical inference for small effects

- Estimated effect of 4.7 percentage points (with standard error of 4.3):
  - 95% confidence interval is $[-4\%, 13\%]$
  - Given that true effect is most likely below 1%, the study provides essentially *no information*
- Theoretical analysis
  - Suppose the true effect was 0.3% and we gather data on 3000 people
  - 3% probability of a statistically-significant positive result
  - 2% probability of a statistically-significant *negative* result

Which headline sells more papers?

The Daily W

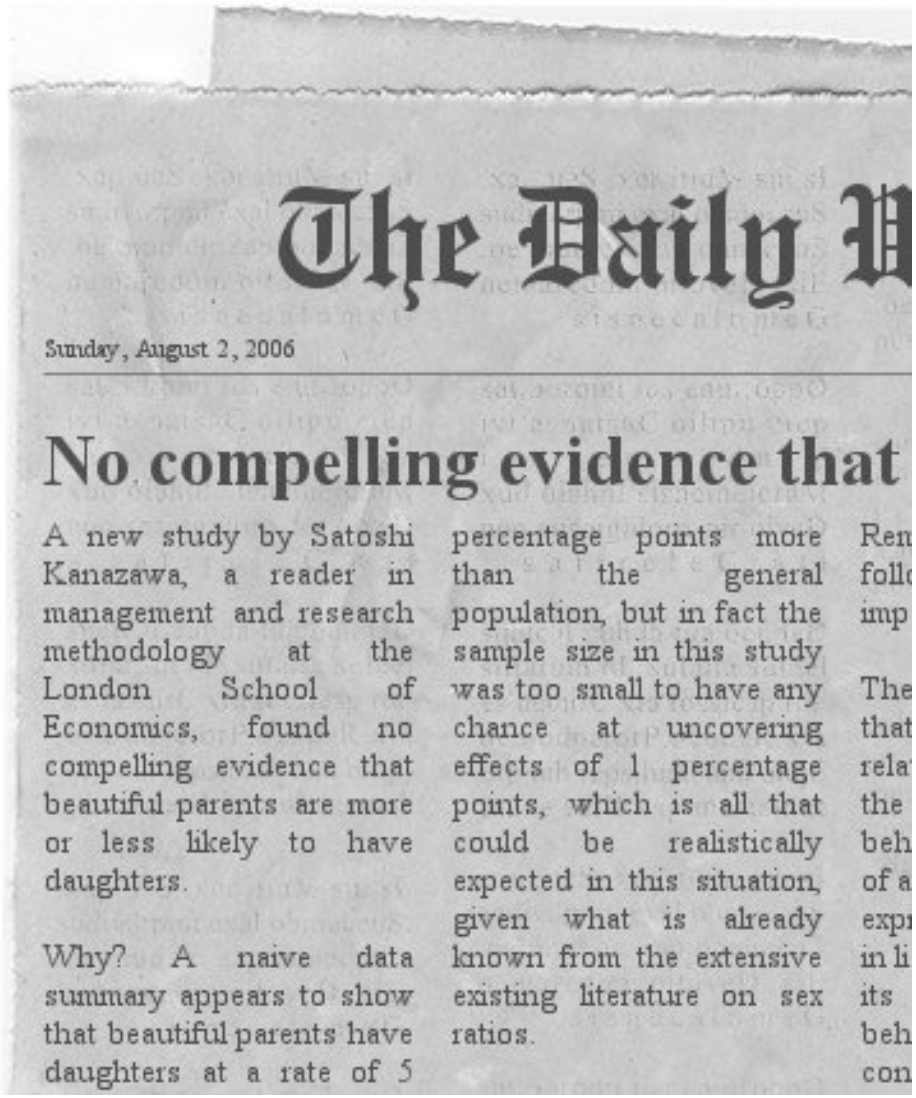
Sunday, August 2, 2006

Beautiful parents have more

A new study by Satoshi Kanazawa, an evolutionary psychologist at the London School of Economics, suggests there are more beautiful women in the world than there are handsome men.

Why? Kanazawa argues it¥s because good-looking parents are 36 percent more likely to have a baby daughter as their first child than a baby sonswhich suggests, evolutionarily speaking, that beauty is a trait more valuable for women than for men.

The study was conducted with data from 3,000 Americans, derived from the National Longitudinal Study of Adolescent Health, and was published in the Journal of Theoretical Biology.



The Daily W

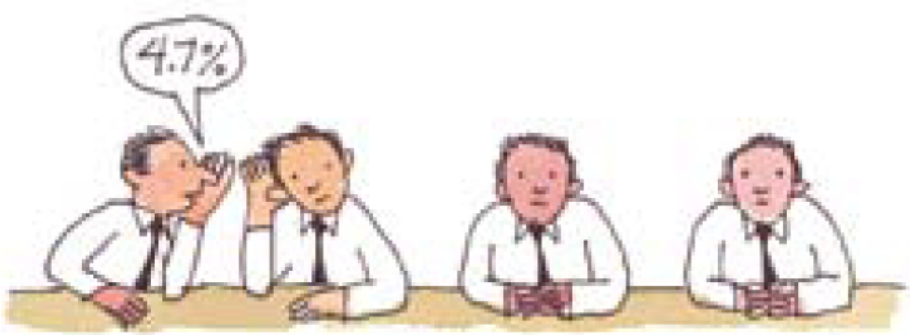
Sunday, August 2, 2006

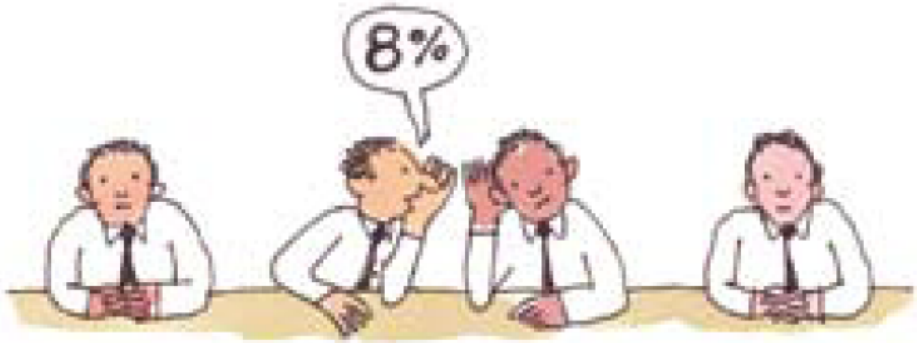
No compelling evidence that

A new study by Satoshi Kanazawa, a reader in management and research methodology at the London School of Economics, found no compelling evidence that beautiful parents are more or less likely to have daughters.

Why? A naive data summary appears to show that beautiful parents have daughters at a rate of 5 percentage points more than the general population, but in fact the sample size in this study was too small to have any chance at uncovering effects of 1 percentage points, which is all that could be realistically expected in this situation, given what is already known from the extensive existing literature on sex ratios.

Communication of the findings





How to evaluate such claims?

- From the *Freakonomics* blog:
  - "A new study by Satoshi Kanazawa, an evolutionary psychologist at the London School of Economics, suggests . . . there are more beautiful women in the world than there are handsome men. Why? Kanazawa argues it's because good-looking parents are 36 percent more likely to have a baby daughter as their first child than a baby son—which suggests, evolutionarily speaking, that beauty is a trait more valuable for women than for men. The study was conducted with data from 3,000 Americans, derived from the National Longitudinal Study of Adolescent Health, and was published in the Journal of Theoretical Biology."
- If Steven Levitt can't get this right, who can??

My reaction

- The claim of "36%" raised suspicion
  - 10 to 100 times larger than reported sex-ratio effects in the literature
- An avoidable error:
  - Small sample size . . .
  - Standard error of 4.3 percentage points . . .
  - To be "statistically significant," the estimate must be at least 2 standard errors away from 0 . . .
  - Any statistically significant finding is *necessarily* a huge overestimate!

Bad-news feedback

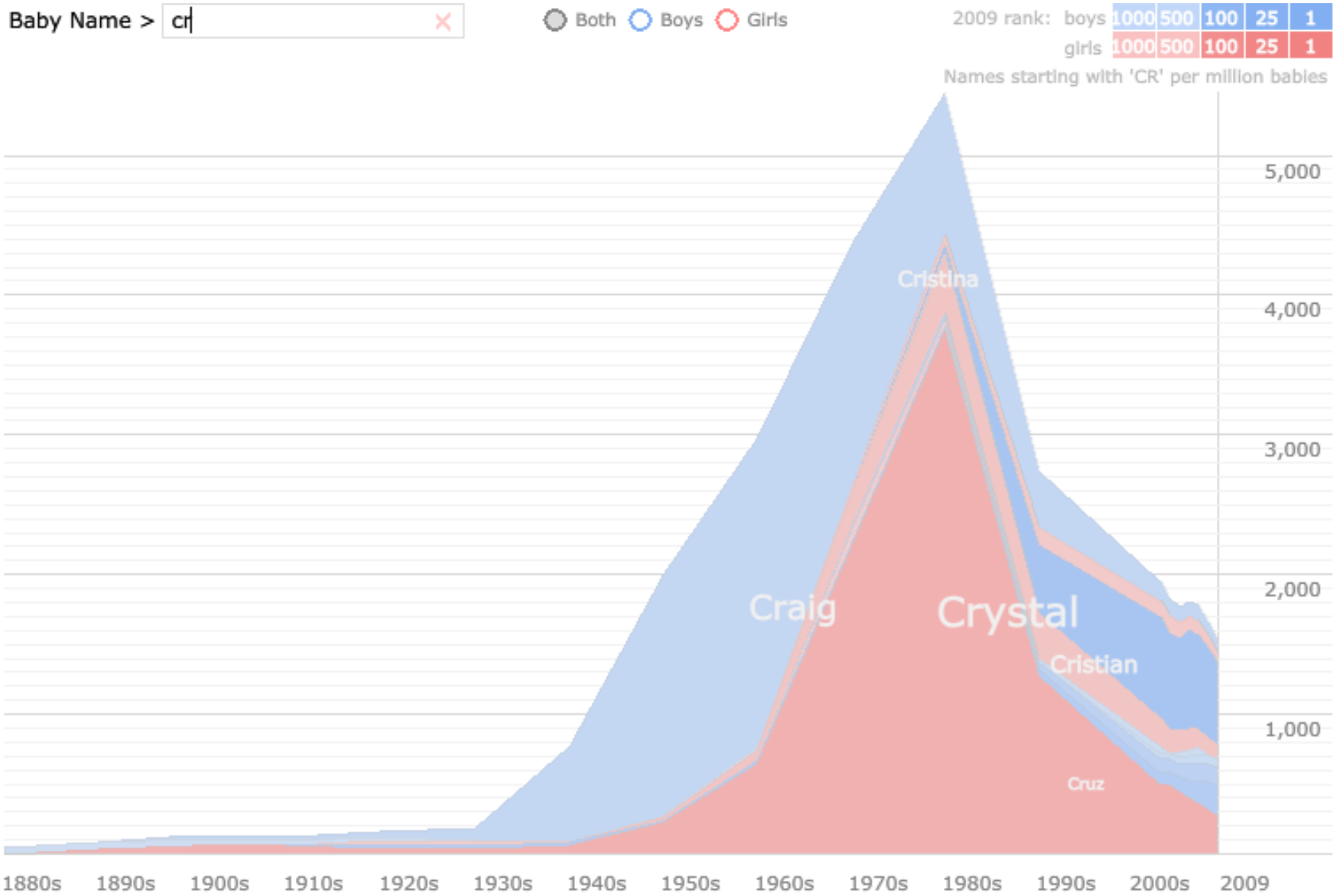
- Proponents of sex differences and other "politically incorrect" results produce papers that, for reasons of inadequate statistical power, produce essentially random results
- Opponents can only reply, "the data are insufficient"
- From *Freakonomics* blog:
  - "It is good that Kanazawa is only a researcher and not, say, the president of Harvard. If he were, that last finding about scientists may have gotten him fired."
- Kanazawa continues to promote his claims in a column for *Psychology Today*

Why is this not obvious?

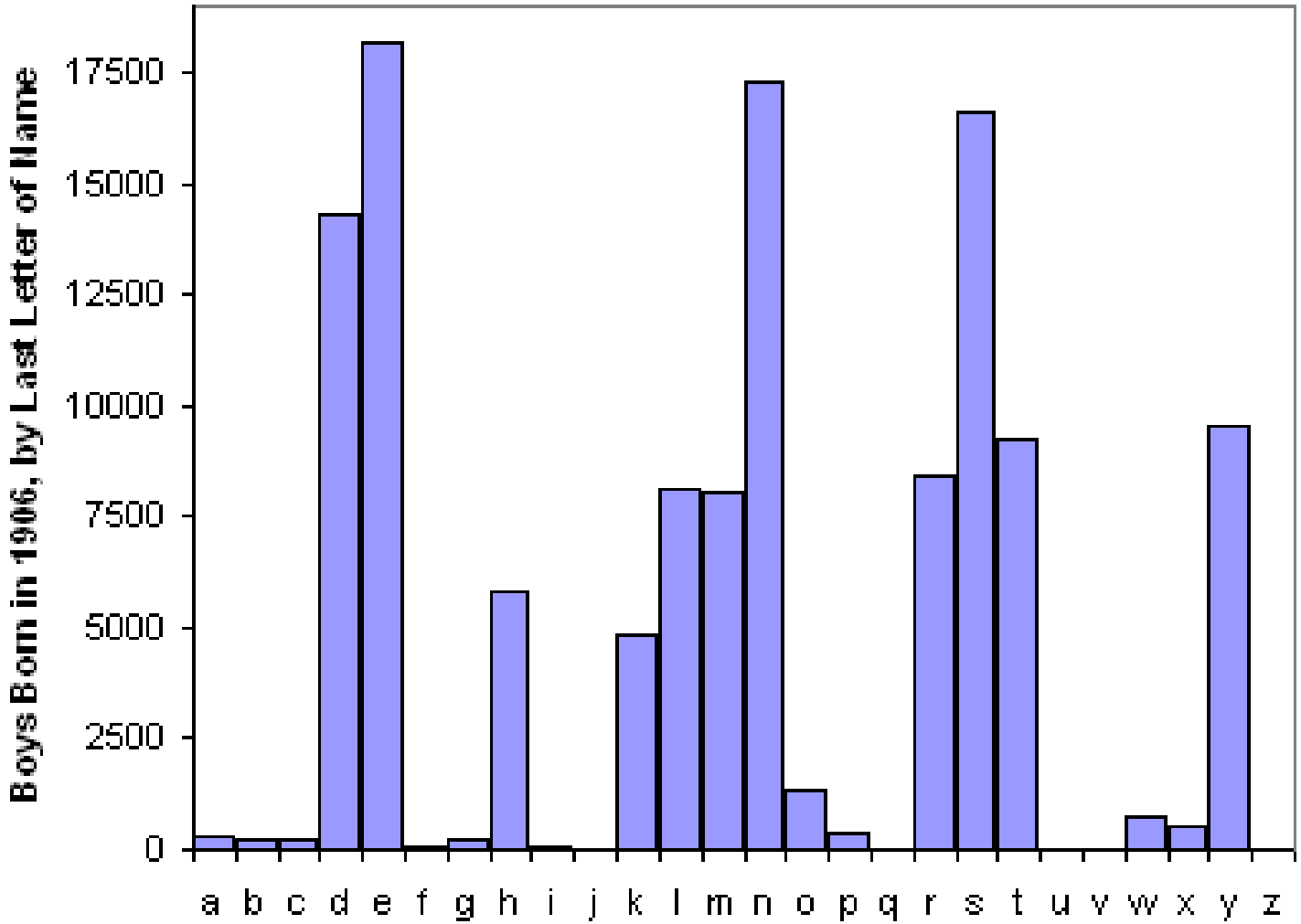
- Statistical theory and education are focused on estimating one effect at a time
- "Statistical significance" is a useful idea, but it doesn't work when studying very small effects
- Methods exist for including prior knowledge of effect sizes, but these methods are not well integrated into statistical practice

Not all effects are small!

Laura and Martin Wattenberg's Baby Name Wizard:

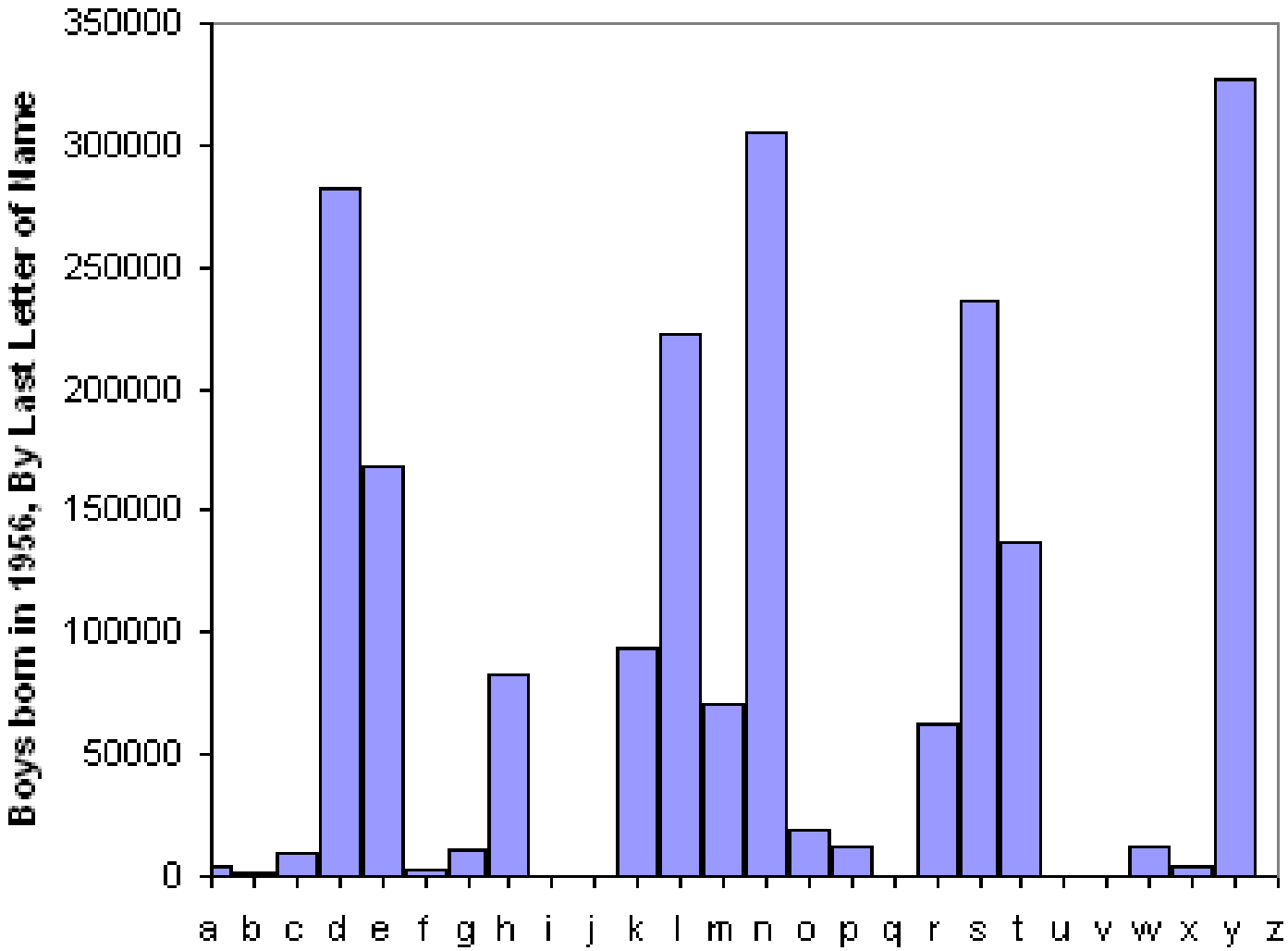


Last letters of boys' names, 100 years ago



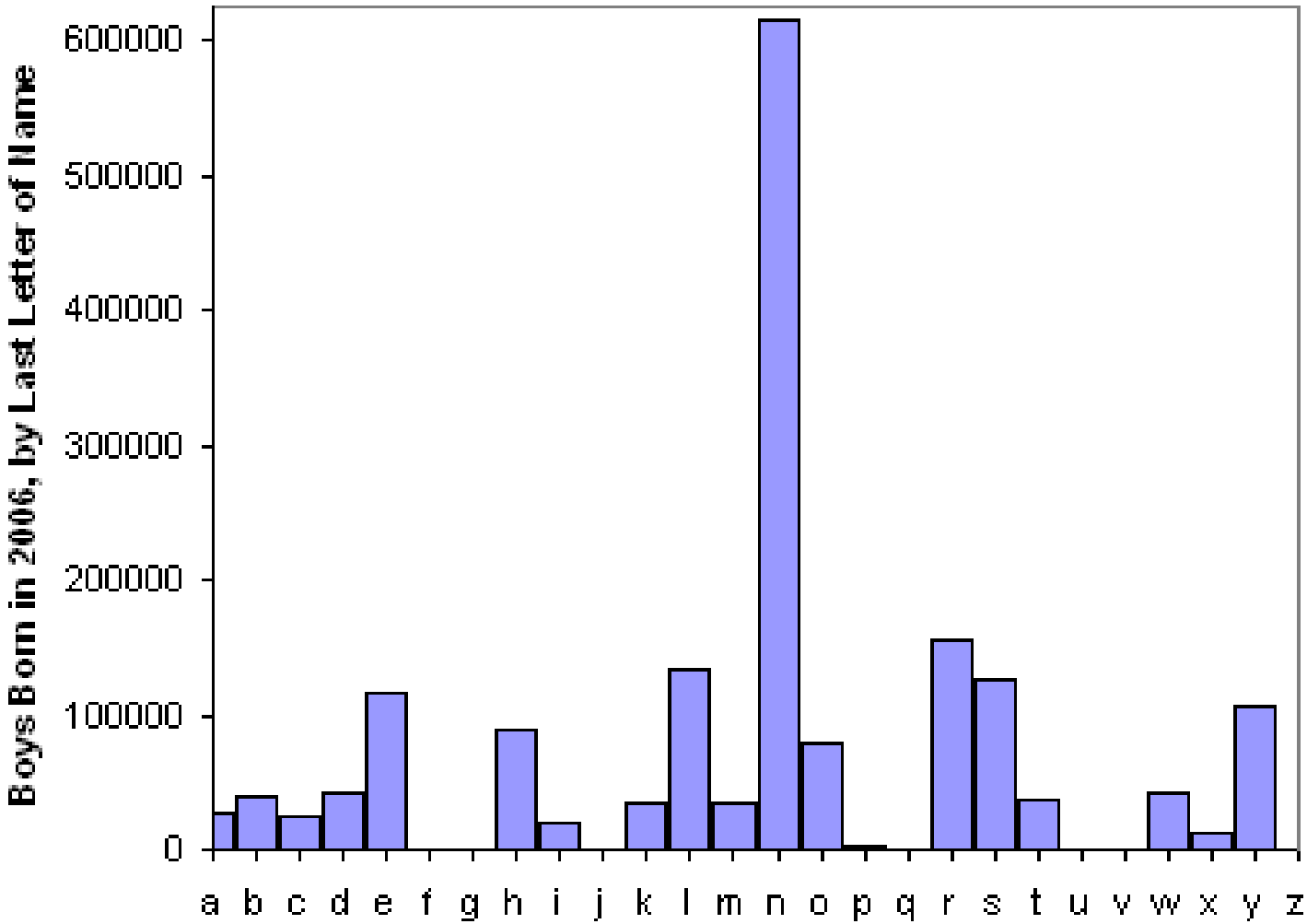
John, James, George, Edward, Henry, ...

Last letters of boys' names, 50 years ago



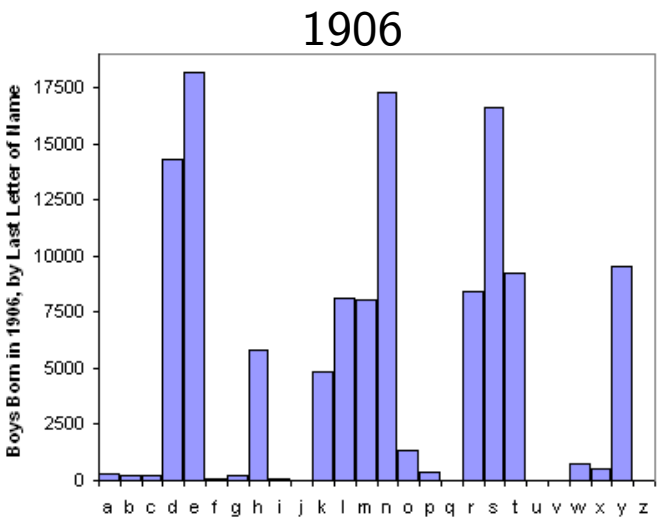
Michael, Thomas, Larry, Jeffrey, ...

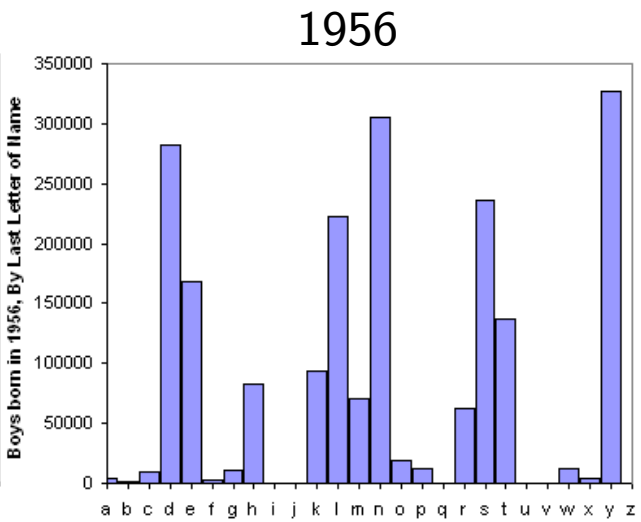
Last letters of boys' names, now

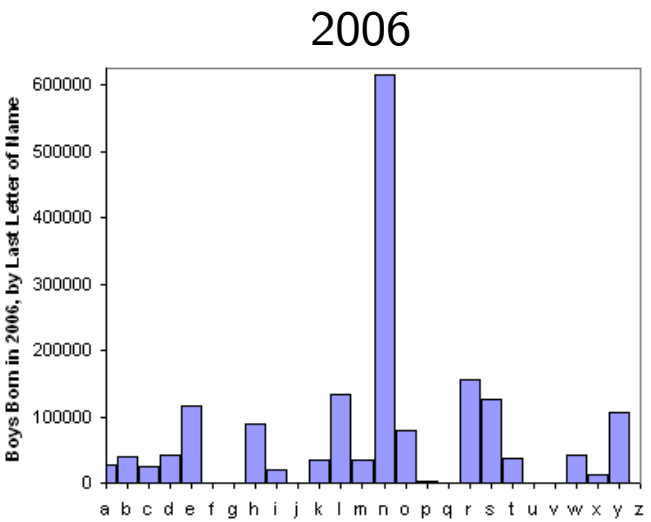


Ethan (#8), John (18), Jonathan (19), Brandon (21), Christian (22), Dylan (23), Benjamin (25), Nathan (27), Logan (28), Justin (29), . . .

The trend in last letters of boys' names







The long tail . . .

. . . and the paradox of freedom

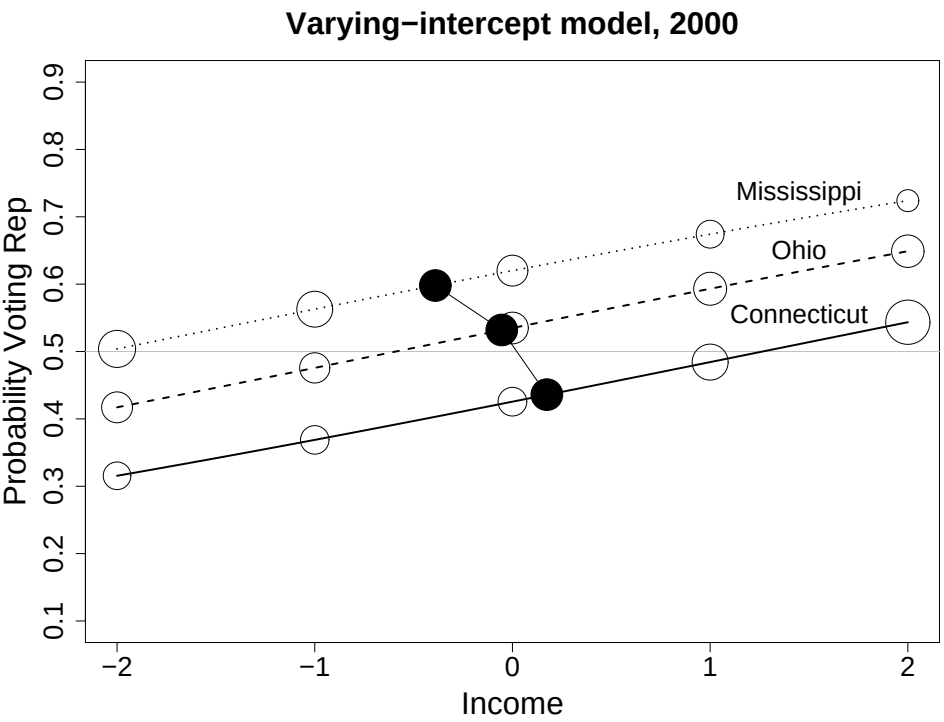
What should we do instead?

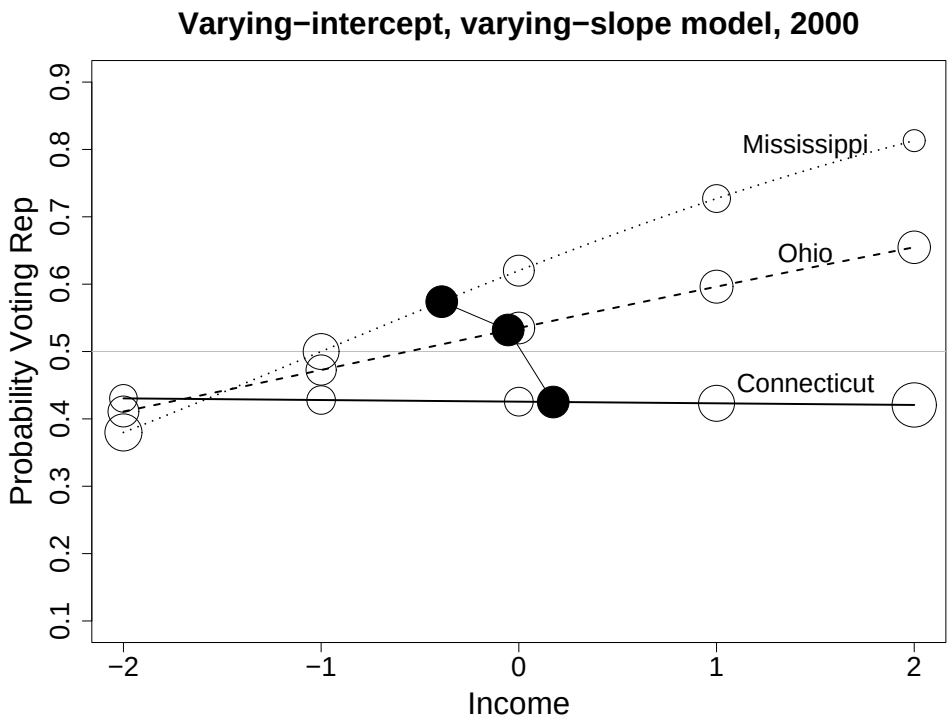
- Don't estimate effects in isolation
- Instead, build a model
- A couple examples from my own research . . .

Red state, blue state, rich state, poor state

- Richer *voters* favor the Republicans, *but*
- Richer *states* favor the Democrats
- Hierarchical logistic regression: predict your vote given your income and your state ("varying-intercept model")

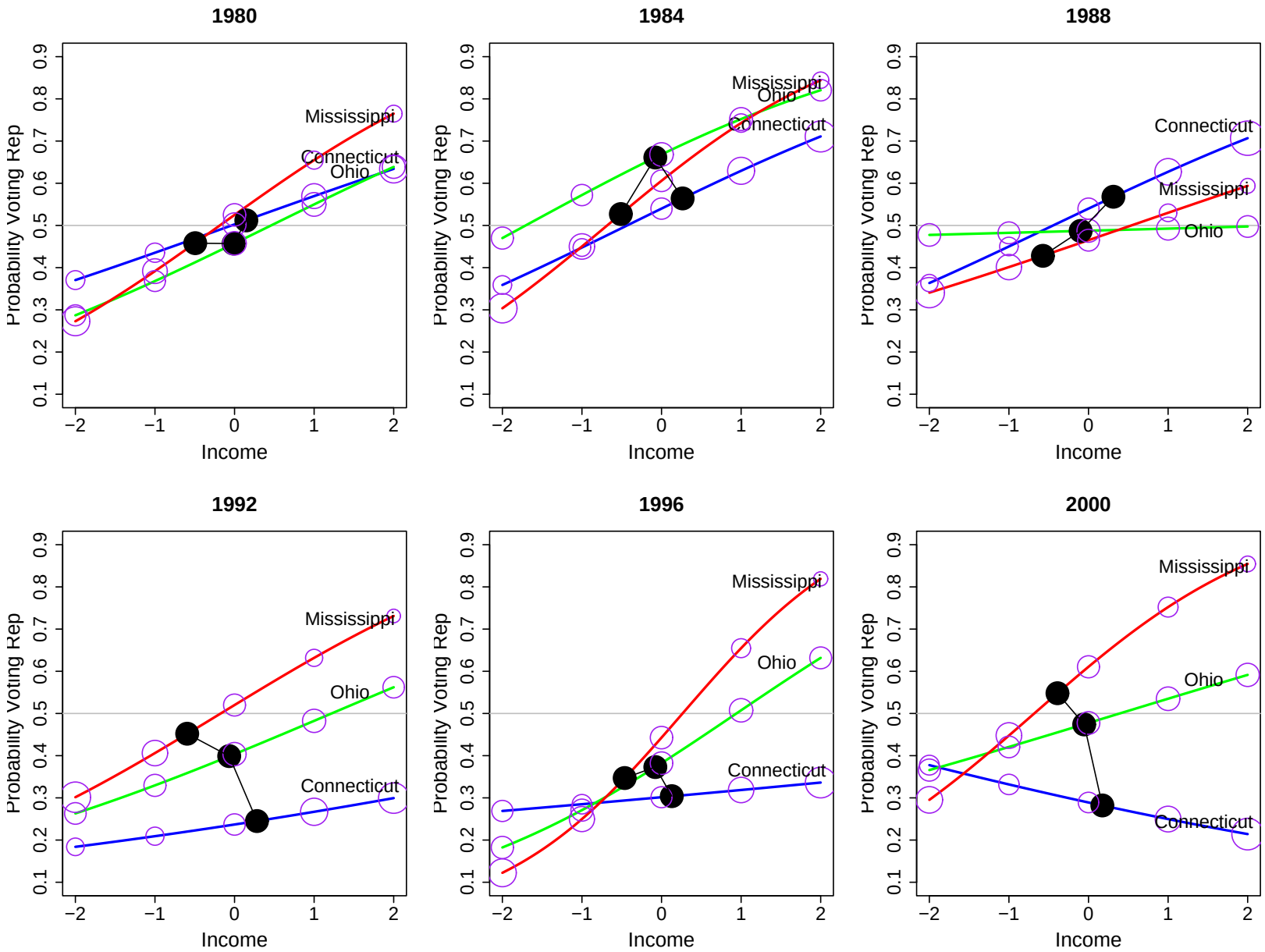
Varying-intercept model, then model criticism, then varying-slope model





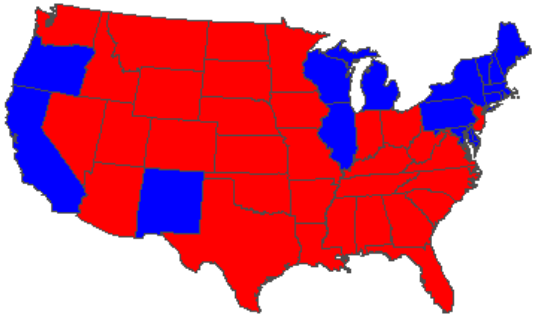
In any given state, the estimates would not be statistically significant!

3-way interactions!

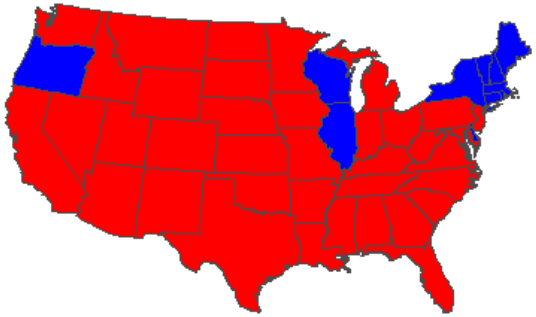


Adding another factor: The inference . . .

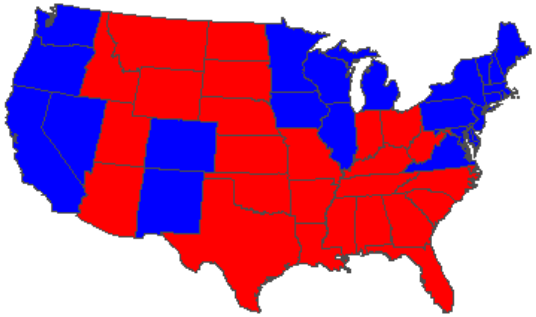
State winners in 2008 (rich voters only)



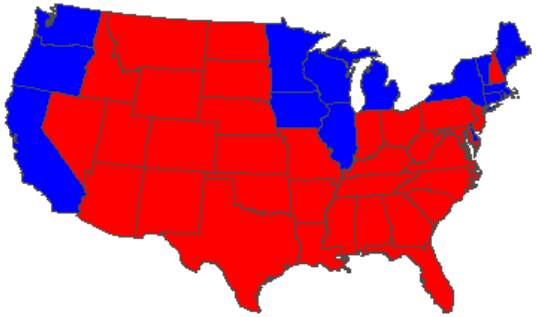
State winners in 2008 (rich Whites only)



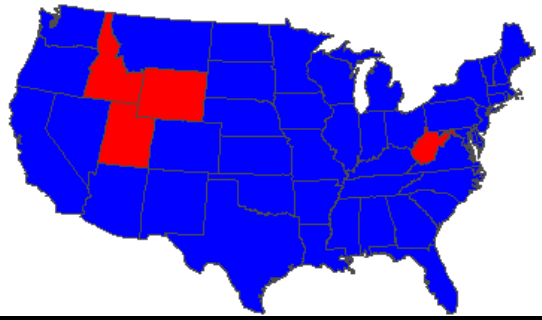
State winners in 2008 (middle-income voters)



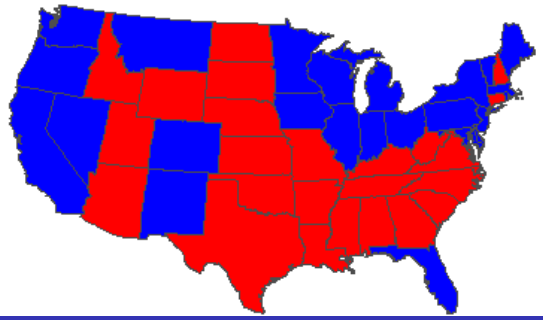
State winners in 2008 (middle-income Whites)



State winners in 2008 (poor voters only)



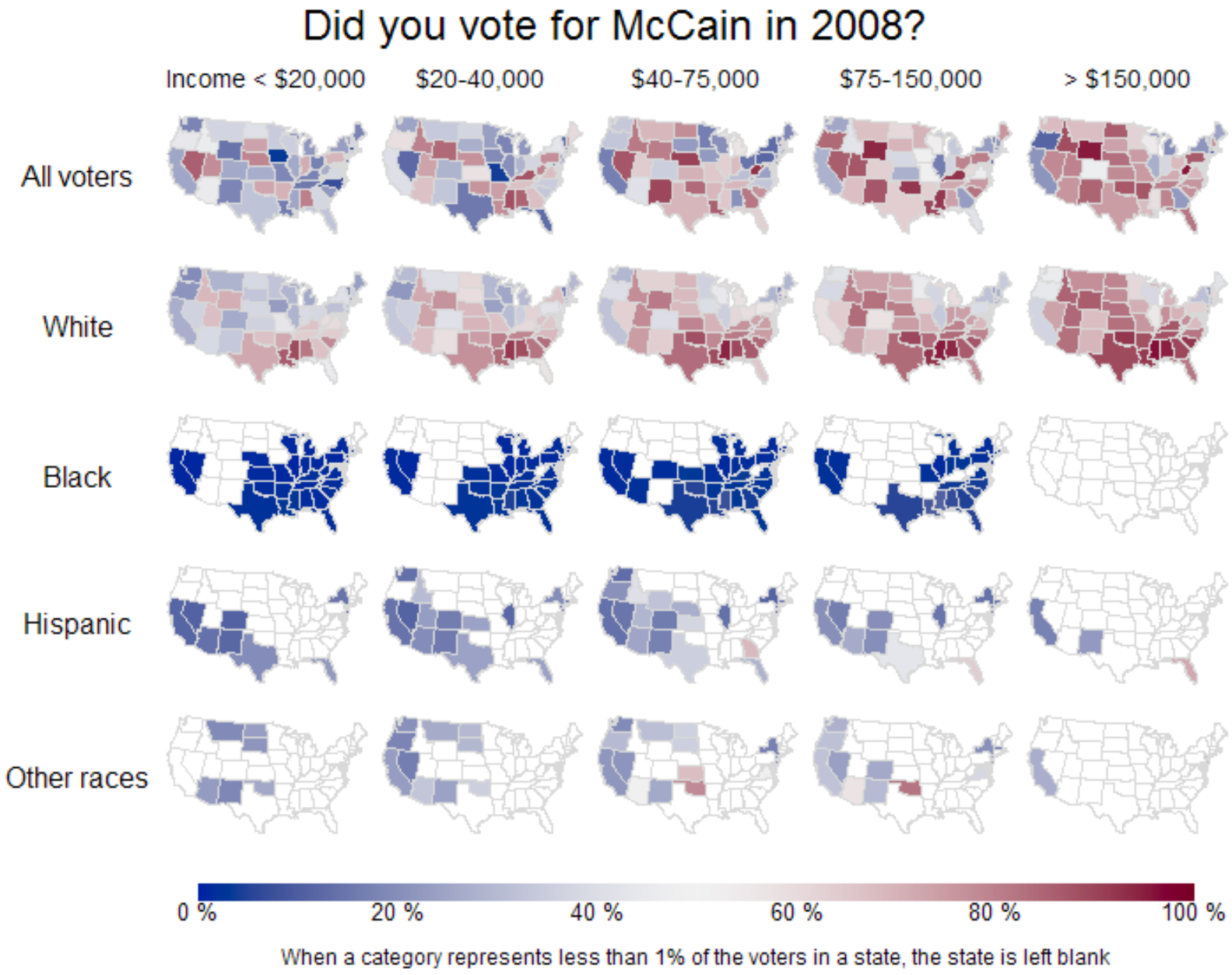
State winners in 2008 (poor Whites only)



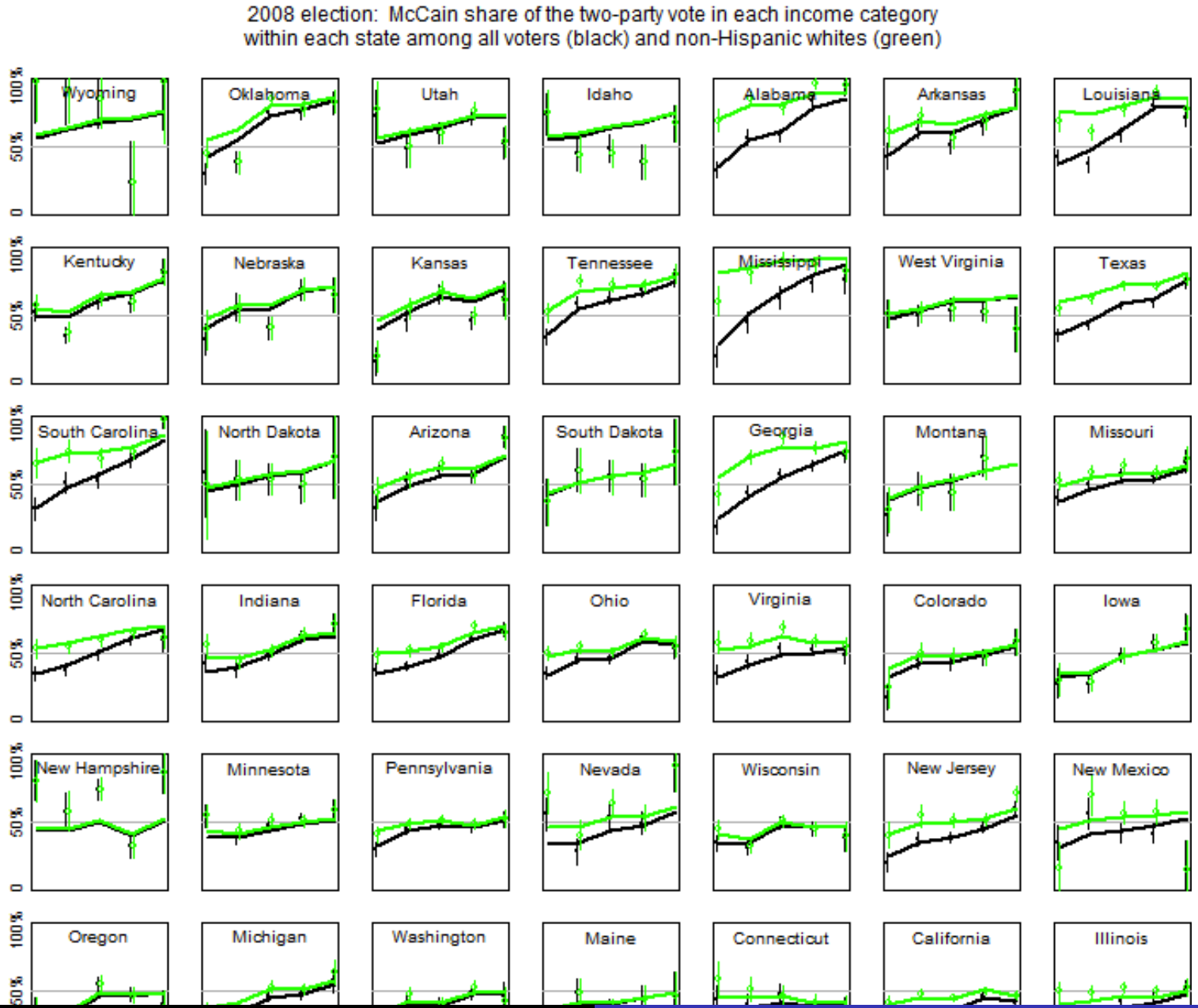
. . . and the refutation!

- Criticisms from the blogger "Daily Kos":
  - Criticisms of the inferences:
    "While Gelman claims only the under-$20K white demo went for Obama, the results were far different. Per the exit poll – real voters – Obama won all whites: 54-45 percent for those making under $50K, and 51-47% for those making over $50K. . . . New Hampshire is solidly Blue unlike Gelman's maps, 58-40 – one of the most obvious misses in Gelman's analysis. . . ."
  - Criciticms of the method:
    "Gelman inexplicably avoids using exit poll data . . . while exit polls have their own margin of errors and sample composition problems, they sure as heck beat anything done over the telephone."
  - Traditional statistical "conservatism" will be no defense here!

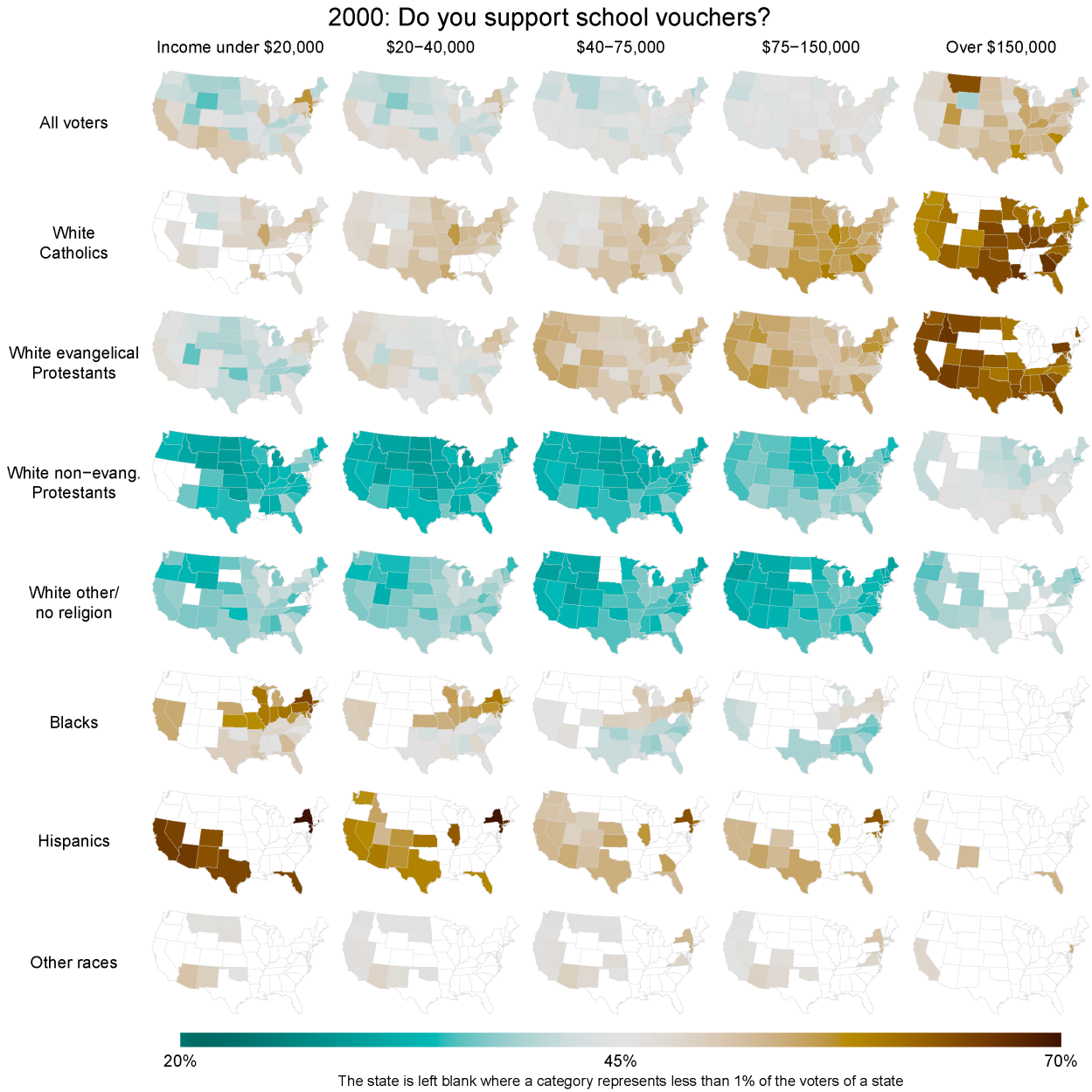
After improving the model



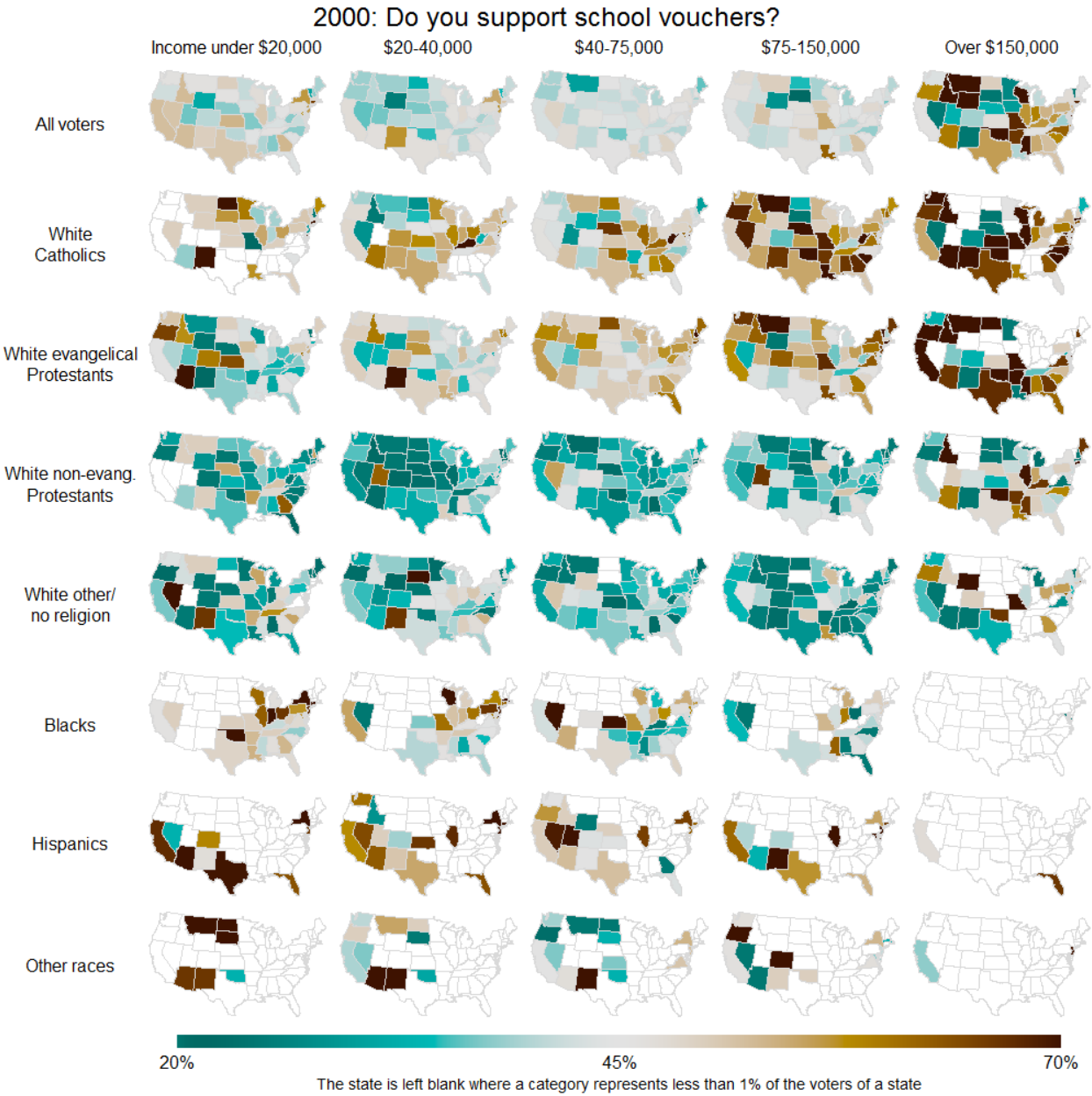
A graph we made to study and criticize our inferences



Ethnicity/religion, income, and school vouchers



The raw data


2000: Do you support school vouchers?
Income under $20,000
$20-40,000
$40-75,000
$75-150,000
Over $150,000
All voters
White Catholics
White evangelical Protestants
White non-evang. Protestants
White other/ no religion
Blacks
Hispanics
Other races
20%
45%
70%
The state is left blank where a category represents less than 1% of the voters of a state

Take-home points

- Inherent problems with "underpowered" (small-sample) studies of small effects
- Three kinds of selection bias:
  - False "statistical significance" via multiple comparisons
  - When using small samples to study small effects, any statistically significant finding is *necessarily* a huge overestimate
  - Incentives (in science and the media) to report dramatic claims
- How to do it right?
  - Don't study factors (e.g., beauty) in isolation
  - Place them in a larger model
  - Mulitlevel modeling as an exploratory tool