Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 2b: The challenge of estimating small effects

Millions of scientific papers published every year . . .

Each paper has its own story . . .

Journal's Paper on ESP Expected to Prompt Outrage

By BENEDICT CAREY
Published: January 5, 2011

One of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events.

Enlarge This Image

Heather Ainsworth for The New York Times

Work by Daryl J. Bem on extrasensory perception is scheduled to be published this year.

The decision may delight believers in so-called paranormal events, but it is already mortifying scientists. Advance copies of the paper, to be published this year in The Journal of Personality and Social Psychology, have circulated widely among psychological researchers in recent weeks and have generated a mixture of amusement and scorn.

The paper describes nine unusual lab experiments performed over the past decade by its author, Daryl J. Bem, an emeritus professor at Cornell, testing the ability of college students to accurately sense random events,

Under pressure . . .

Enlarge This Image

Joris Buijs/Pve

The psychologist, Diederik Stapel
committed academic fraud in "se
papers, many accepted in respect
in the news media, according to a
Monday by the three Dutch institu
worked: the University of Groning
Amsterdam, and Tilburg. The jou
published one of Dr. Stapel's pap
"editorial expression of concern"
online on Tuesday.

Under pressure . . .

Marc Hauser Resigns From Harvard

By Tom Bartlett

Marc D. Hauser, the Harvard psych
found responsible for eight counts
misconduct by the university, has
ending speculation about whether
embattled professor would return
this fall.

In a letter dated July 7, Mr. Hauser
Michael D. Smith, Harvard's dean

It's not just the silly stuff . . .

SPECIAL ARTICLE

The Spread of Obesity in a Large Social Network over 32 Years

Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

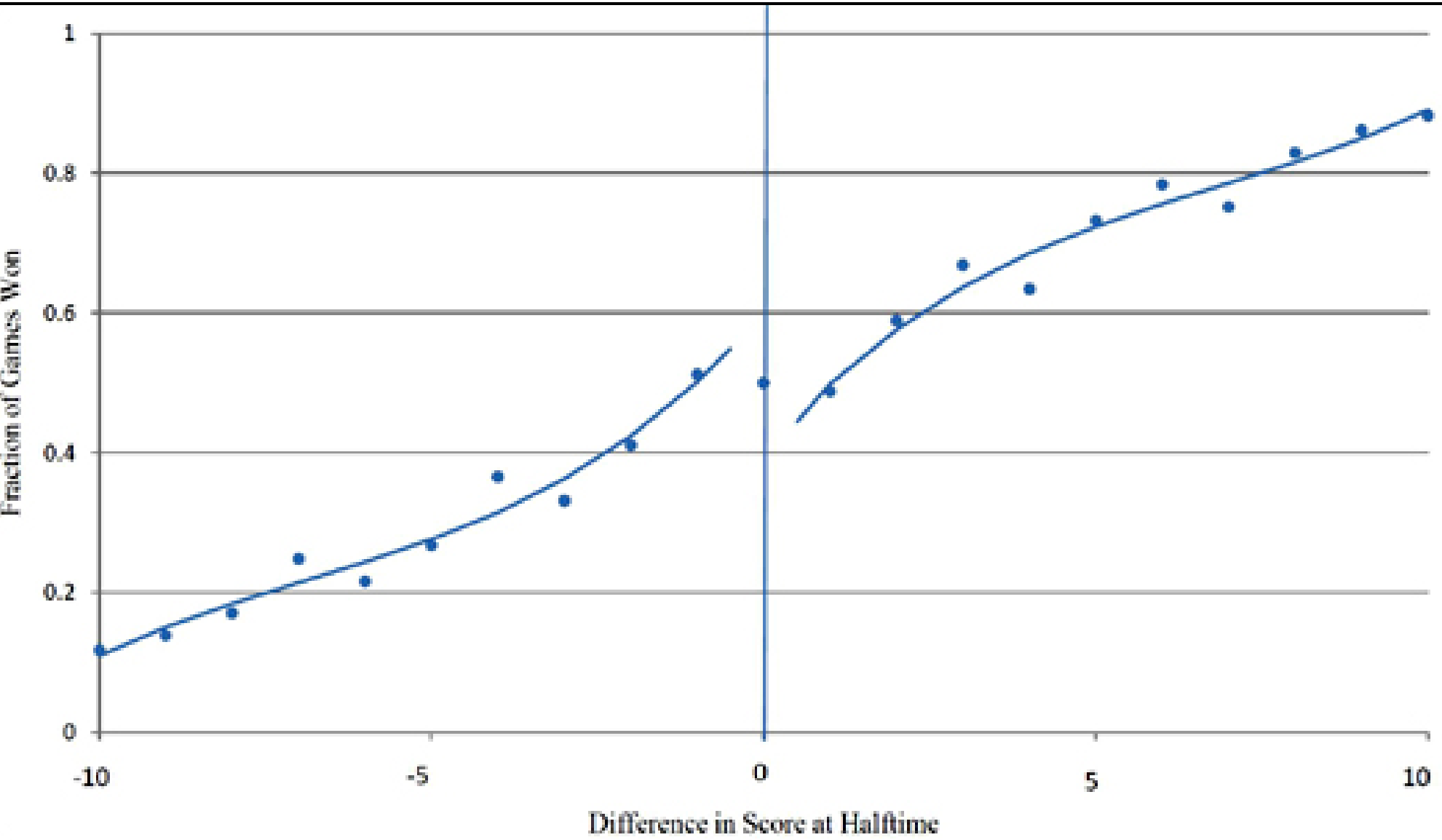
N Engl J Med 2007; 357:370-379 | July 26, 2007

Abstract | Article | References | Citing Articles (405) | Glossary | Letters

BACKGROUND

The prevalence of obesity has increased substantially over the past 30 years. We performed a quantitative analysis of the nature and extent of the person-to-person spread of obesity as a possible factor contributing to the obesity epidemic.

Business-relevant examples . . .



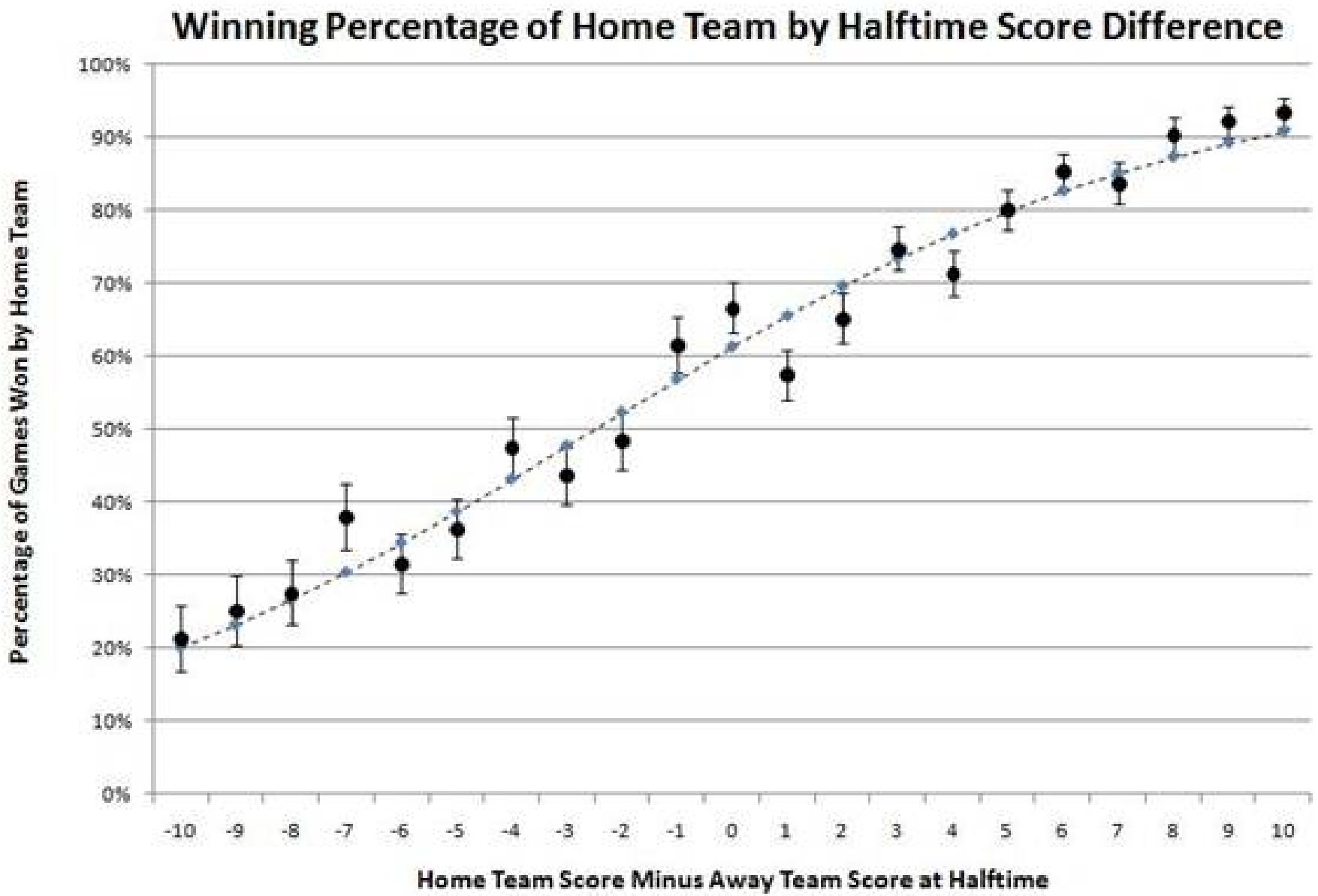
Halftime motivation in basketball

- Economists Jonah Berger and Devin Pope:
  "Analysis of over 6,000 collegiate basketball games illustrates that being slightly behind increases a team's chance of winning. Teams behind by a point at halftime, for example, actually win more often than teams ahead by one. This increase is between 5.5 and 7.7 percentage points . . ."
- But . . . in their data, teams that were behind at halftime by 1 point won 51.3% of the time
- Approx 600 such games; thus, std. error is $0.5/\sqrt{600} = 0.02$
- Estimate ± 1 se is $[0.513 \pm 0.02] = [0.49, 0.53]$
- So where did they get "5.5 and 7.7 percentage points"??

Methodological conservatism

- What about that 5th-degree polynomial?
- Berger and Pope write:
  "While the regression discontinuity methods we use in the paper (including the 5th degree polynomial) are standard in economics (see for example the 2009 working paper on R&D implementation by David Lee and Thomas Lemiuex) we respect that you may find a different approach to the problem to be more useful. . . ."

The data without the 5th-degree polynomial



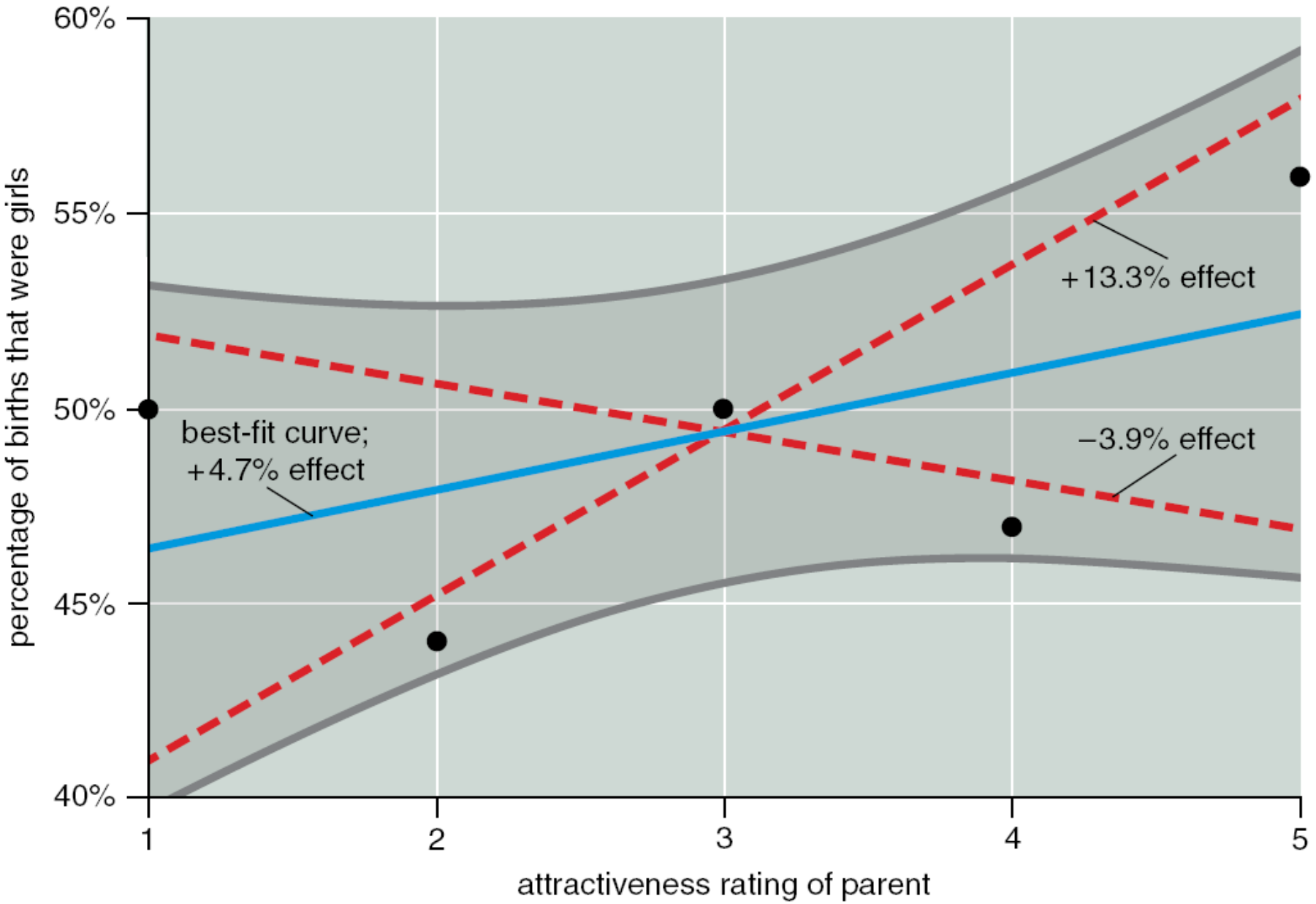
"A Raise Won't Make You Work Harder"

- Economist Ray Fisman writing in *Slate*:
  - Students were employed in a six-hour data-entry job for \$12/hour. Half the students were actually paid this amount. The other half were paid \$20/hour.
  - At first, the \$20-per-hour employees were more productive than the \$12-an-hour employees. But by the end the two groups were working at the same pace.
- Conclusions:
  - "The goodwill of high wages took less than three hours to evaporate completely—hardly a prescription for boosting long-term productivity."
  - "A raise won't make you work harder."
- Conflict between internal and external validity:
  - "All participants were told that this was a one-time job—otherwise, the higher-paid group might work harder in hopes of securing another overpaying library gig."

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: differences of about 1% or less
  - Extreme poverty and famine: differences as high as 3%
- We expect any differences associated with 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

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

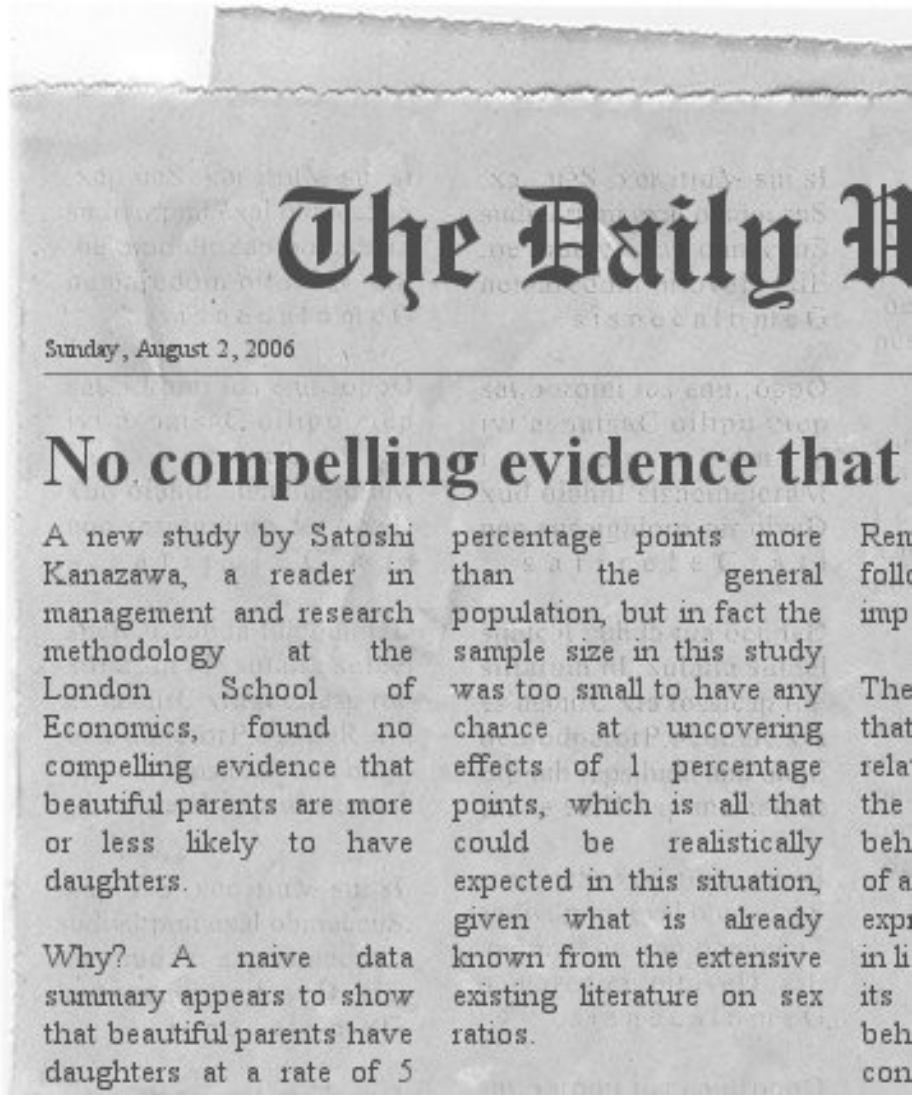
Sunday, August 2, 2006

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

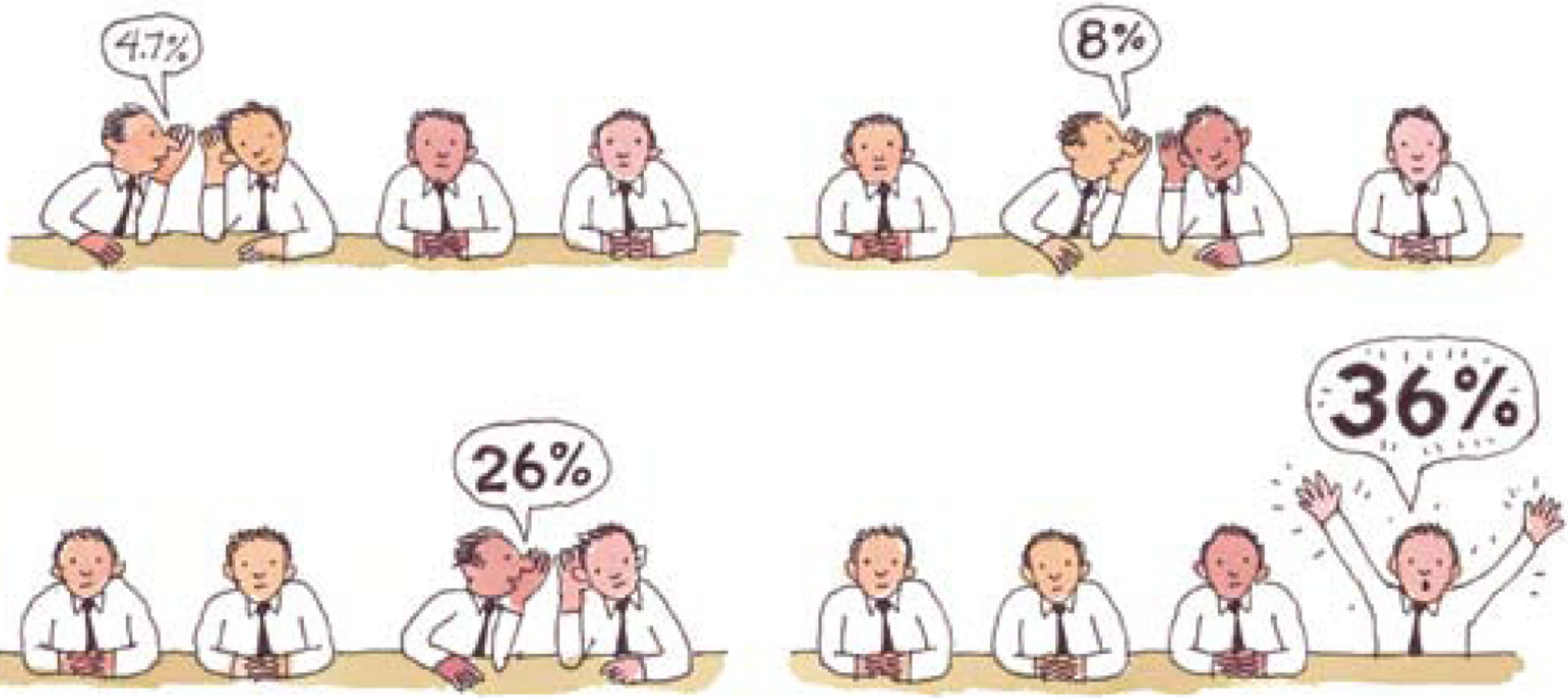
Sunday, August 2, 2006

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

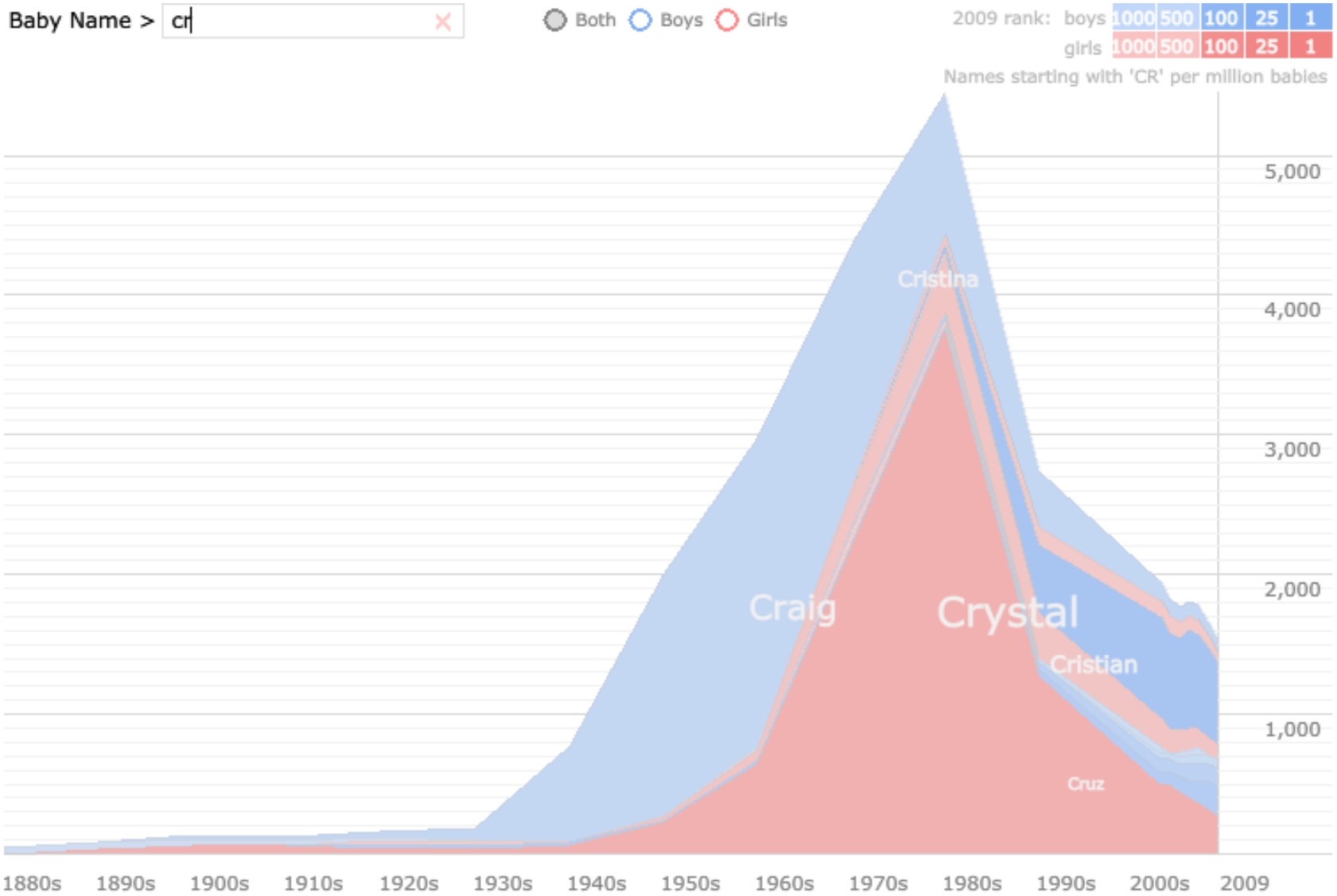
- The claim of "36%" raised suspicion
  - 10 to 100 times larger than reported sex-ratio effects in the literature
- An avoidable error:
  - Small sample size . . .
  - Standard error of 4.3 percentage points . . .
  - To be "statistically significant," the estimate must be at least 2 standard errors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

Why is this not obvious?

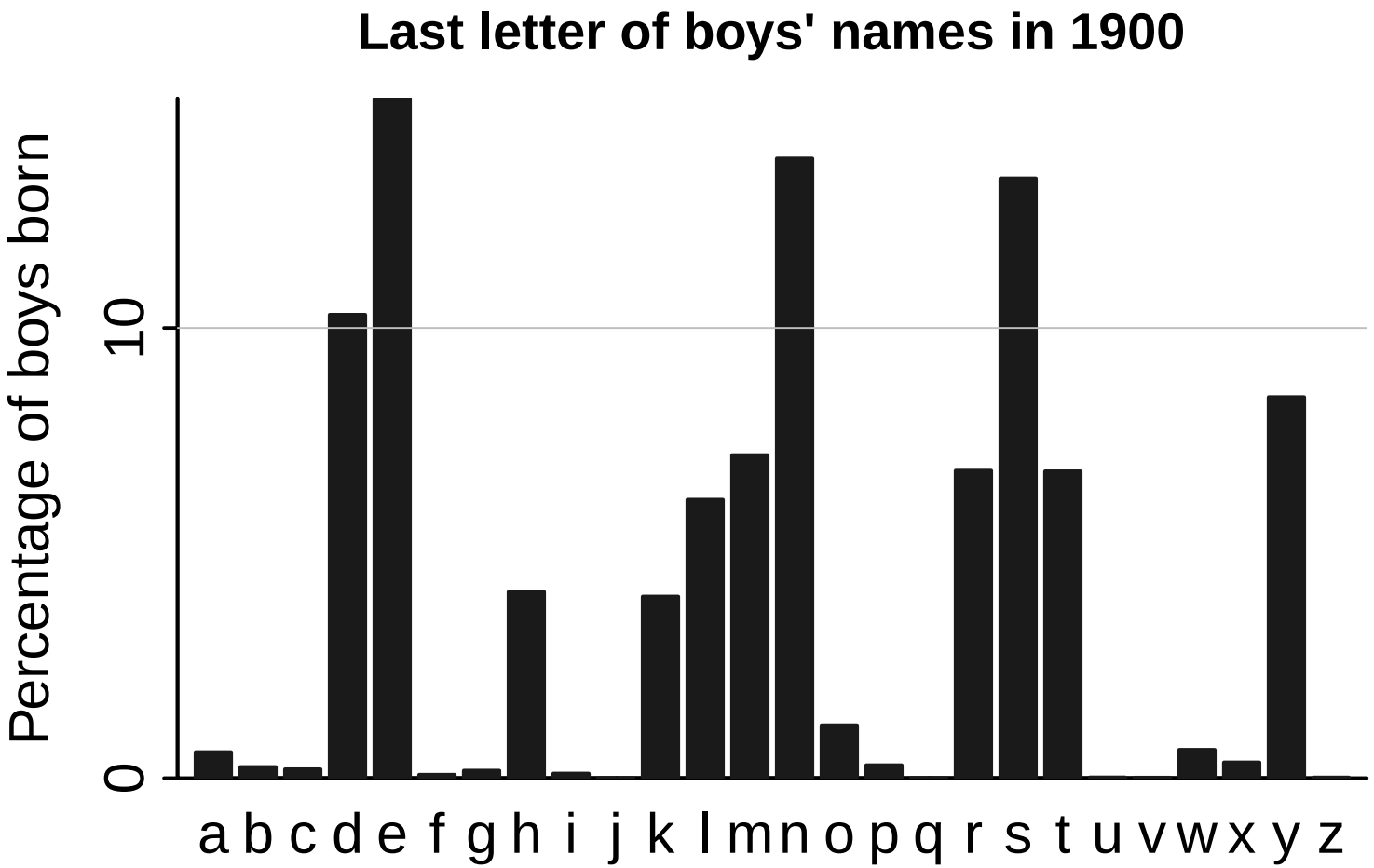
- Statistical theory and education are focused on estimating one effect at a time
- "Statistical significance" is a useful idea, but it doesn't work when studying very small effects
- Methods exist for including prior knowledge of effect sizes, but these methods are not well integrated into statistical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

Not all effects are small!

Laura and Martin Wattenberg's Baby Name Wizard:

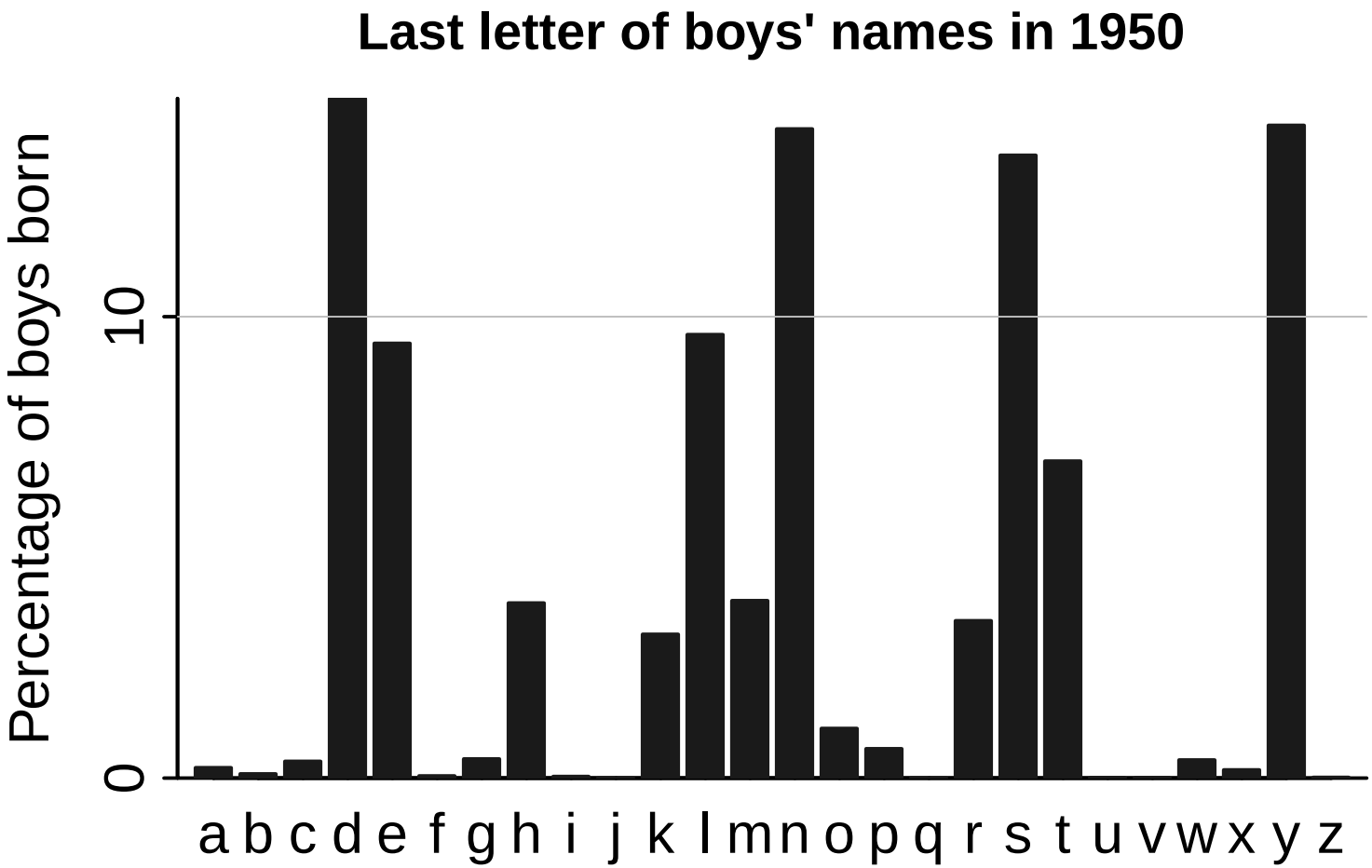


Last letters of boys' names, 1900



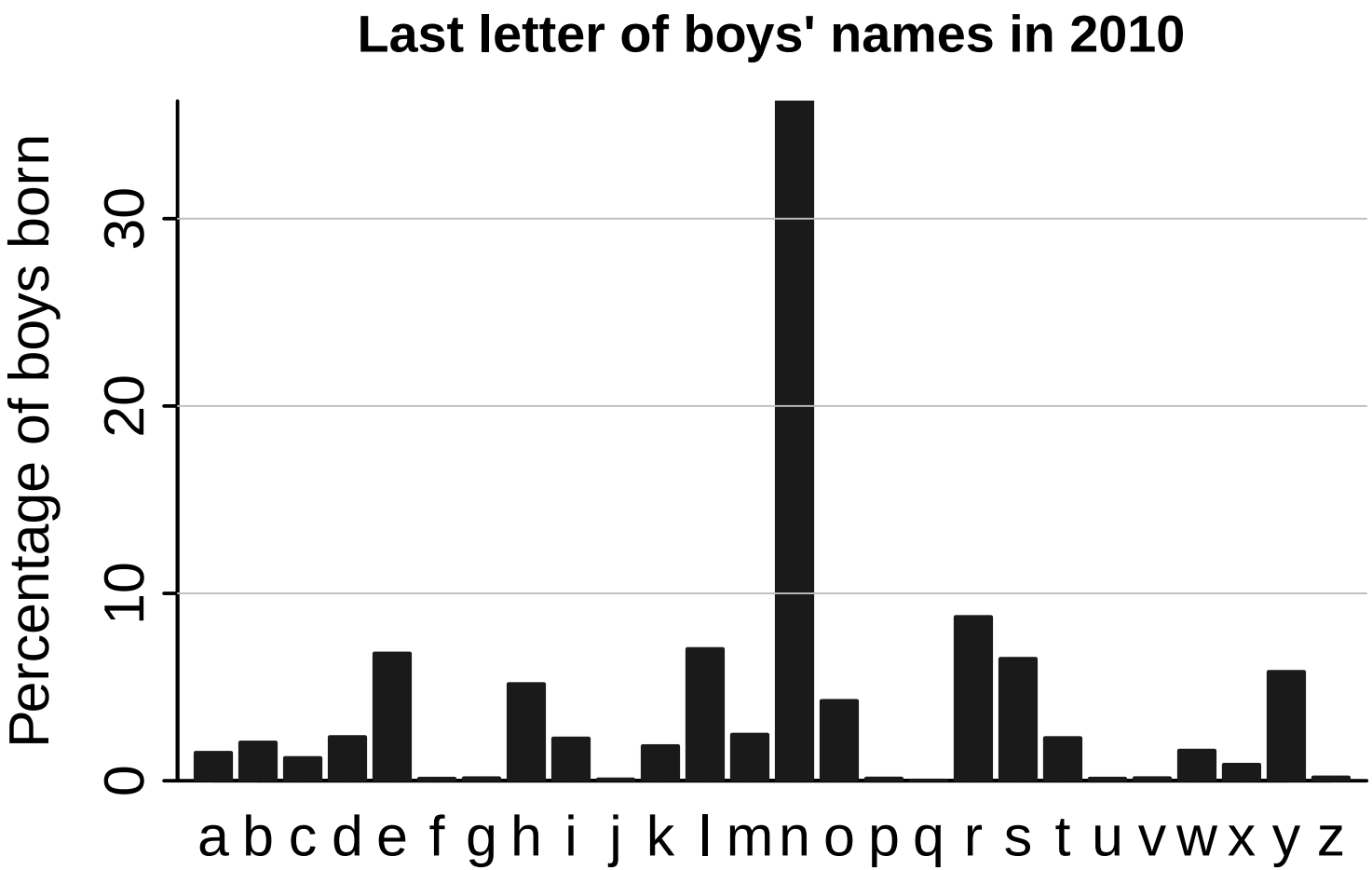
John, James, George, Charles, Edward, ...

Last letters of boys' names, 1950



Michael, David, Thomas, Larry, ...

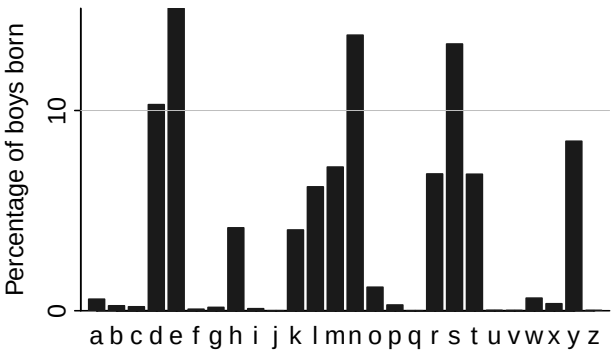
Last letters of boys' names, now



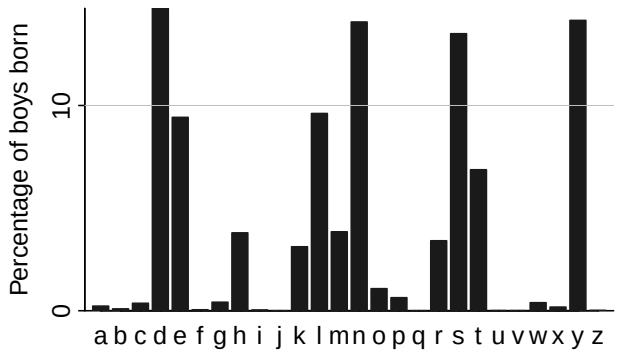
Ethan (#2), Jayden (4), Aiden (9), Mason (12), Logan (17), Benjamin (22), Ryan (23), Jackson (25), John (26), Nathan (27), Jonathan (28), Christian (29), 31, 32, 36, 37, 40, ...

The trend in last letters of boys' names

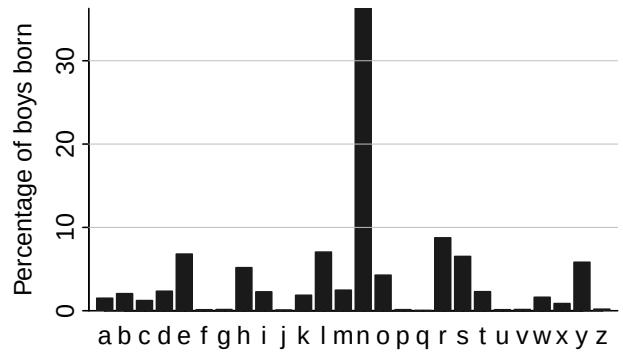
1900



1950



2010



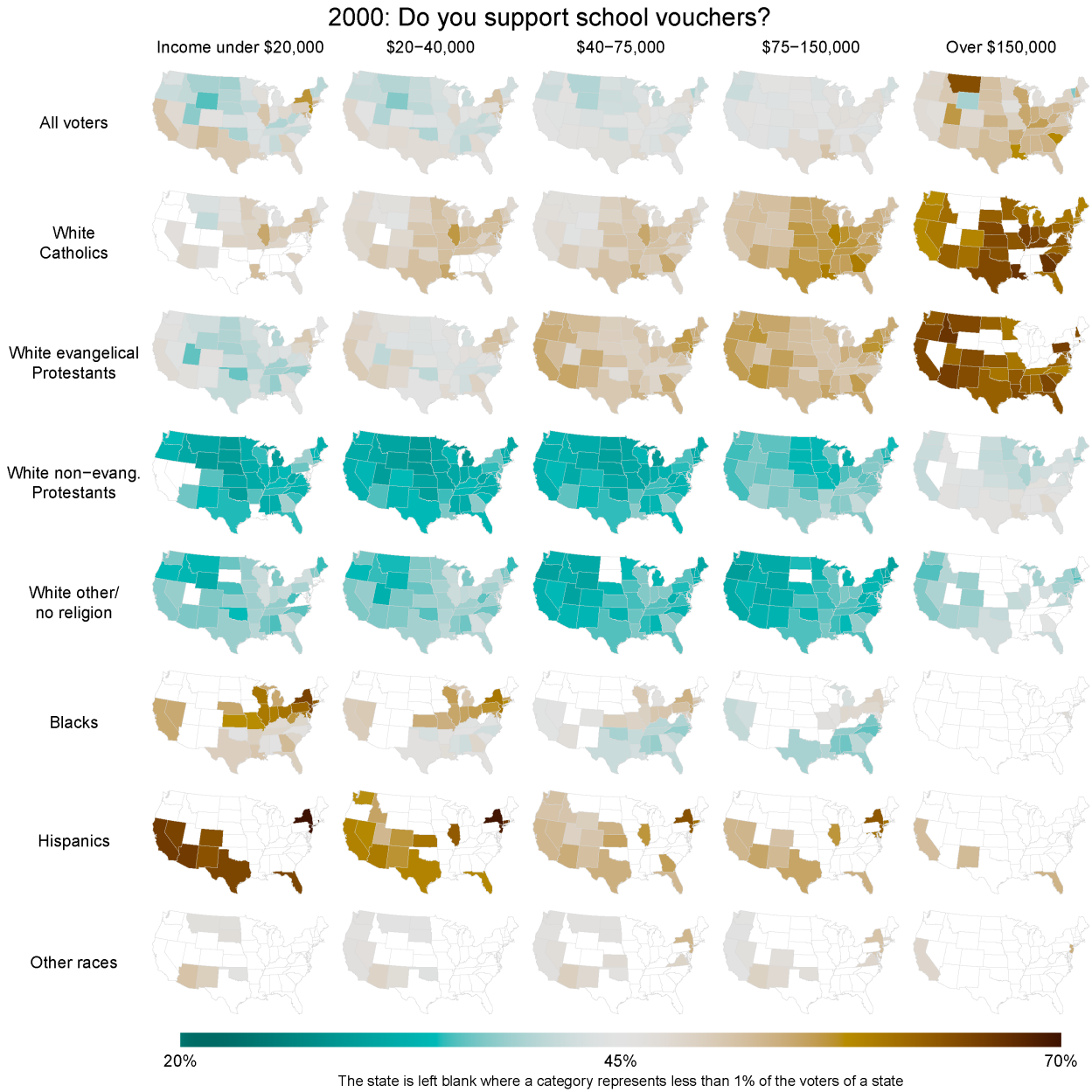
The long tail . . .

. . . and the paradox of freedom

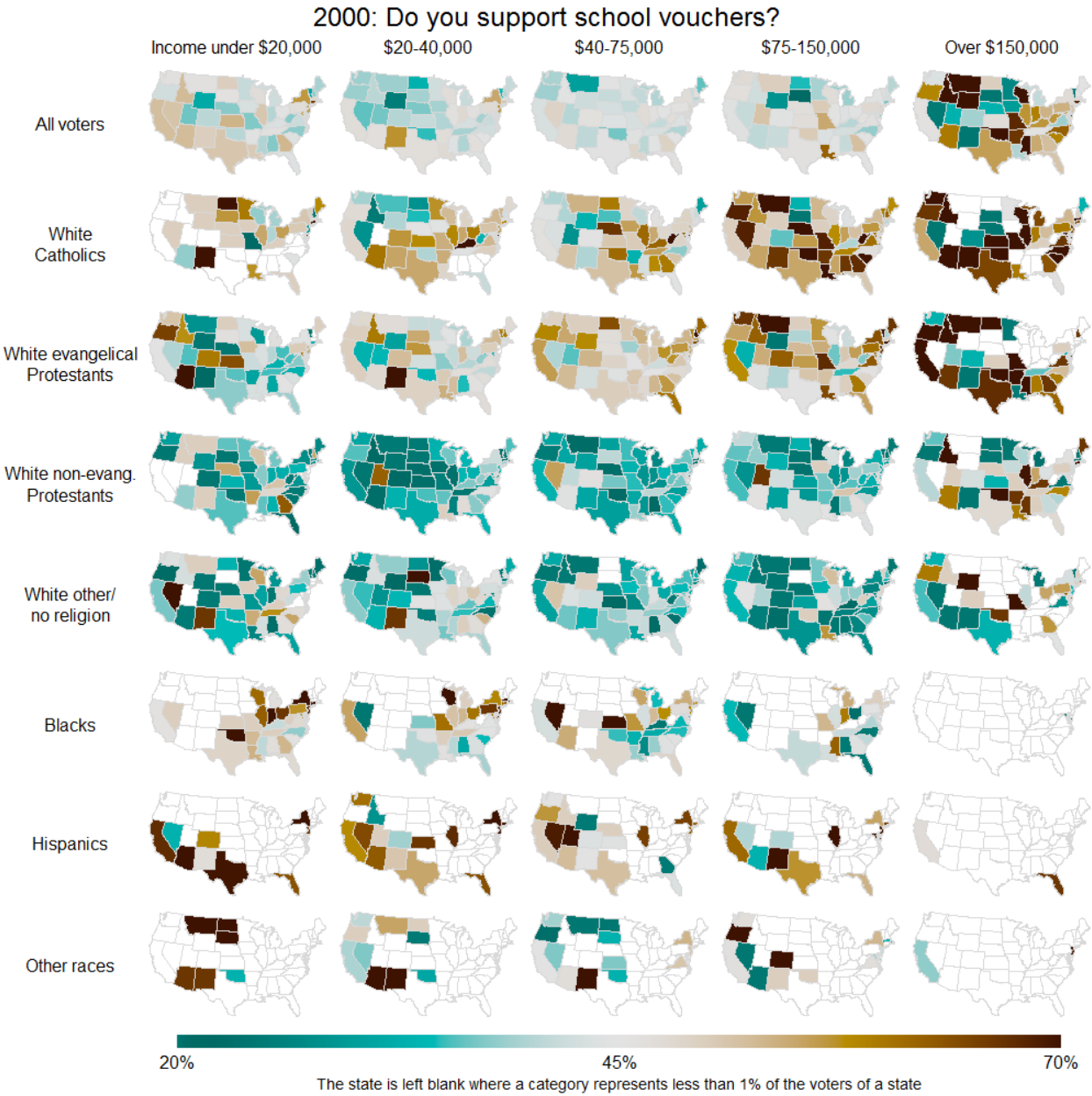
What should we do instead?

- Don't estimate effects in isolation
- Instead, build a model
- Consider some examples from my own research

Ethnicity/religion, income, and school vouchers



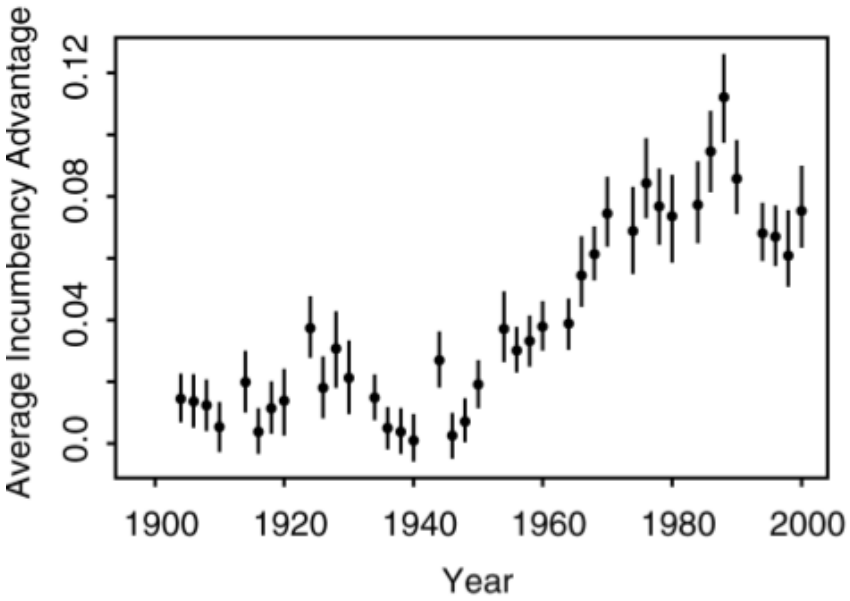
The raw data


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

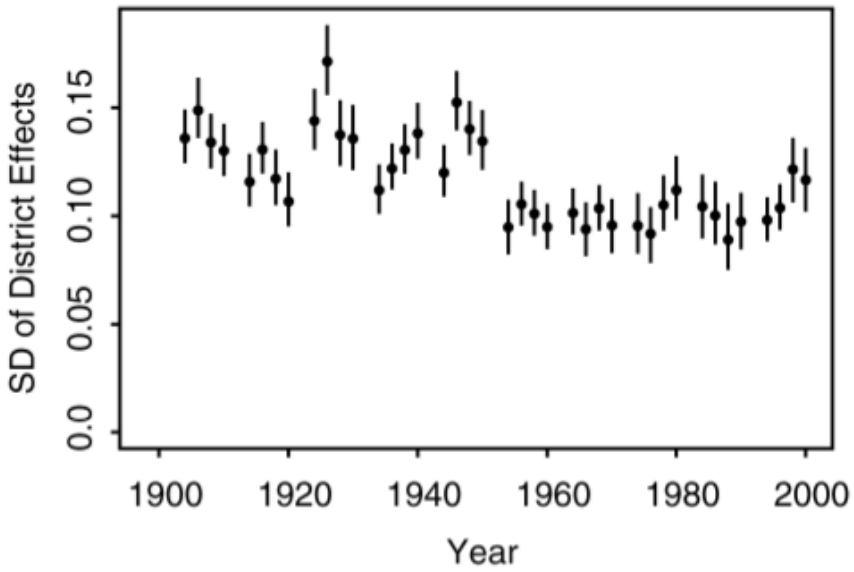
Doing it without the technology

- Display data or simple estimates in a grid of graphs
- Implicit multilevel modeling by eye

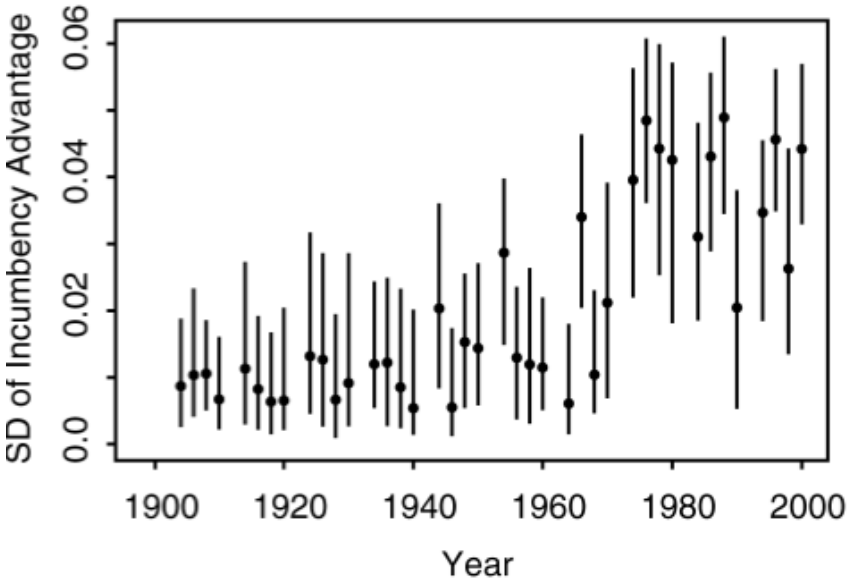
Example: incumbency advantage over time

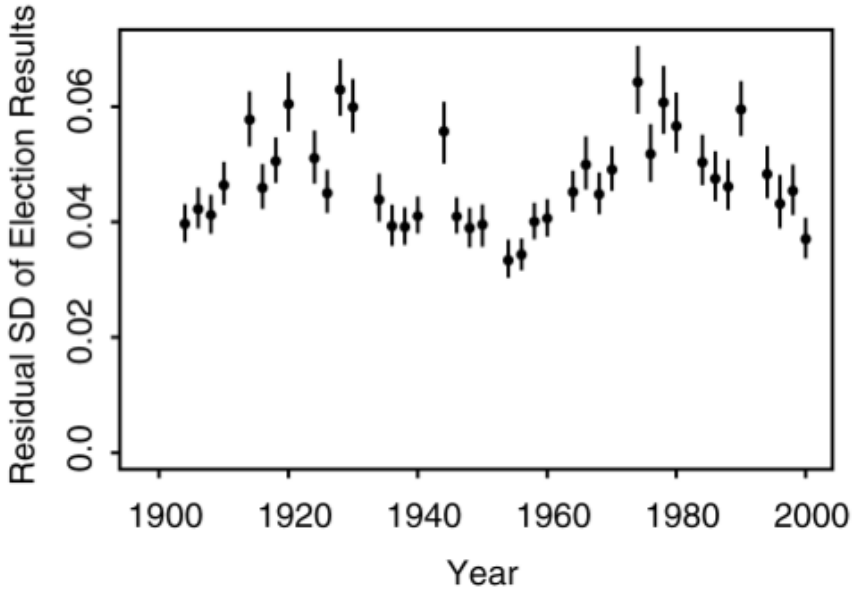


(c)



(d)





Take-home points

- 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
  - Multilevel modeling as an exploratory tool