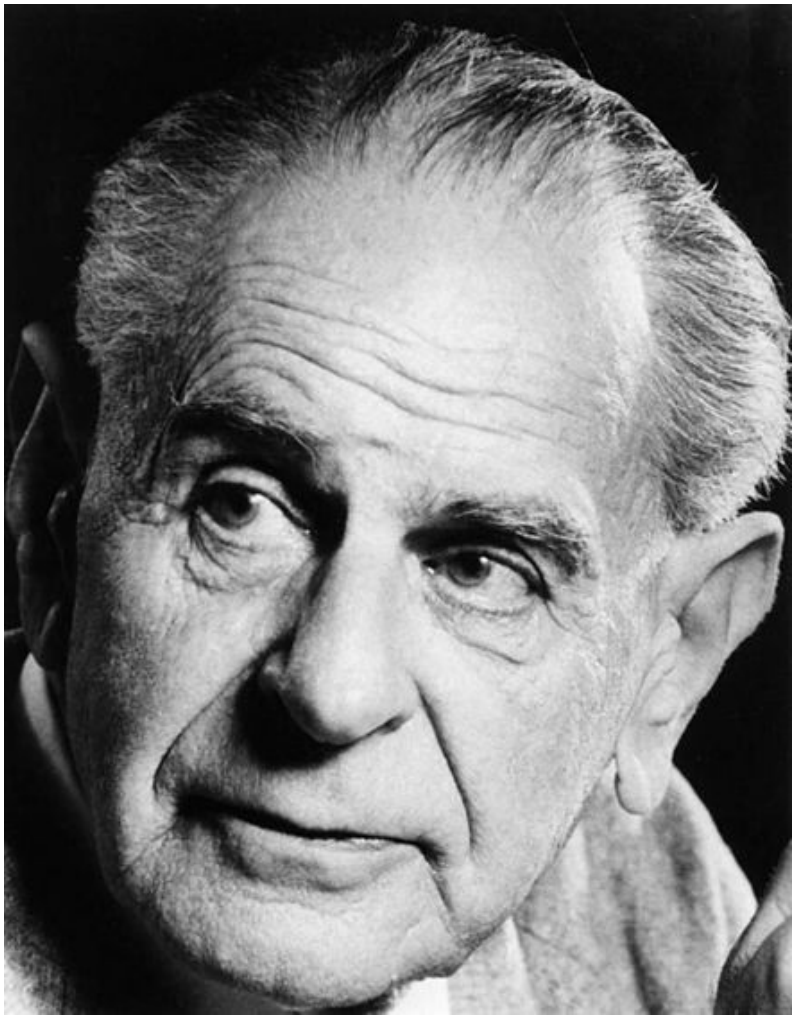
Hypothesis Testing is a Bad Idea

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University, New York

Contemporary Issues in Hypothesis Testing workshop, Warwick,
15 Sept 2016



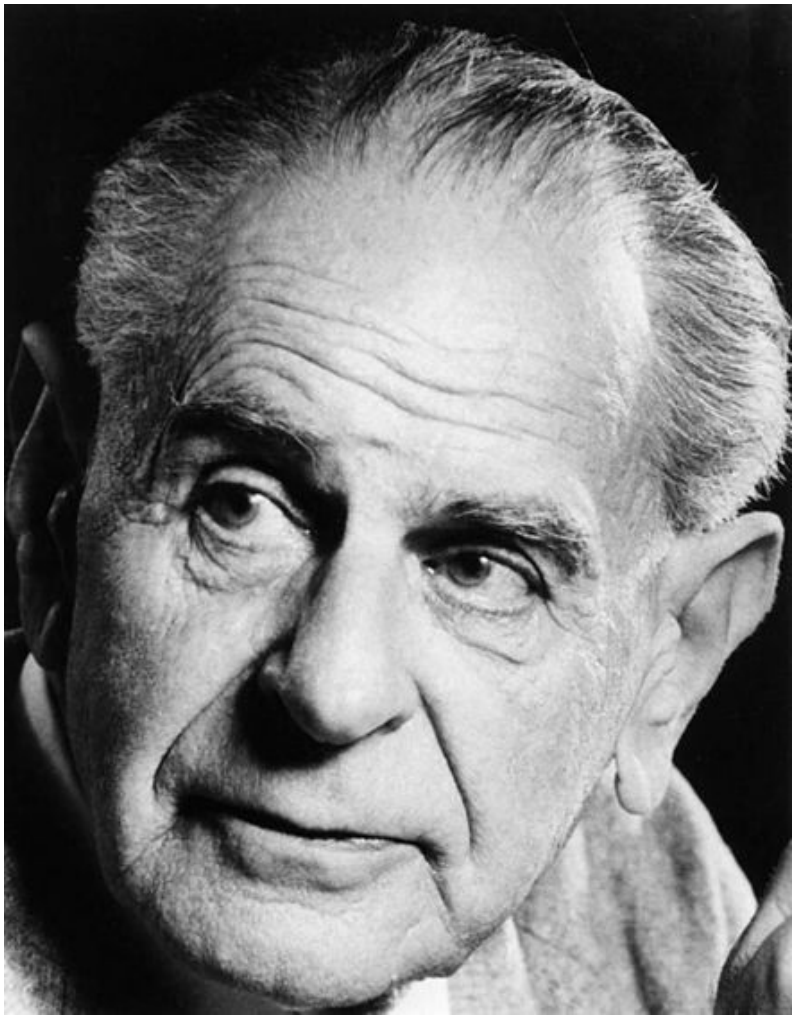
The logic of scientific discovery

"The application of the chosen test left little doubt that the lymphocytes from household contact persons are, by and 1arge, better breast cancer cell killers than those from the controls." — J. Neyman, 1976

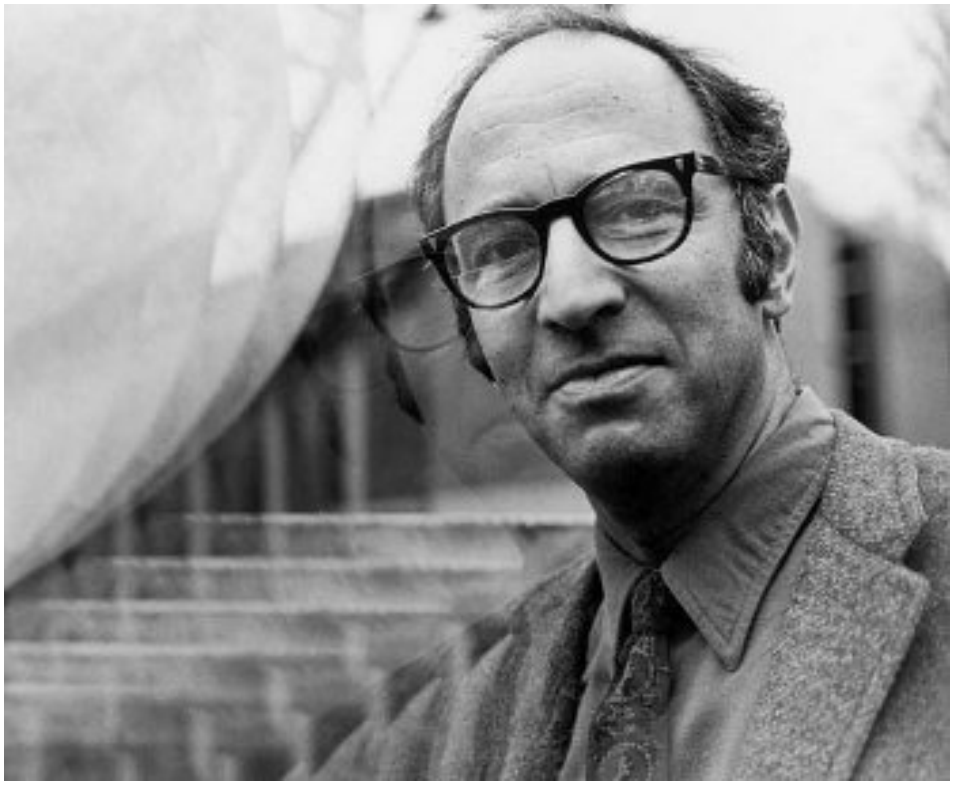
Confirmationist vs. falsificationist reasoning

The null hypothesis

Researchers are happy to reject this guy.

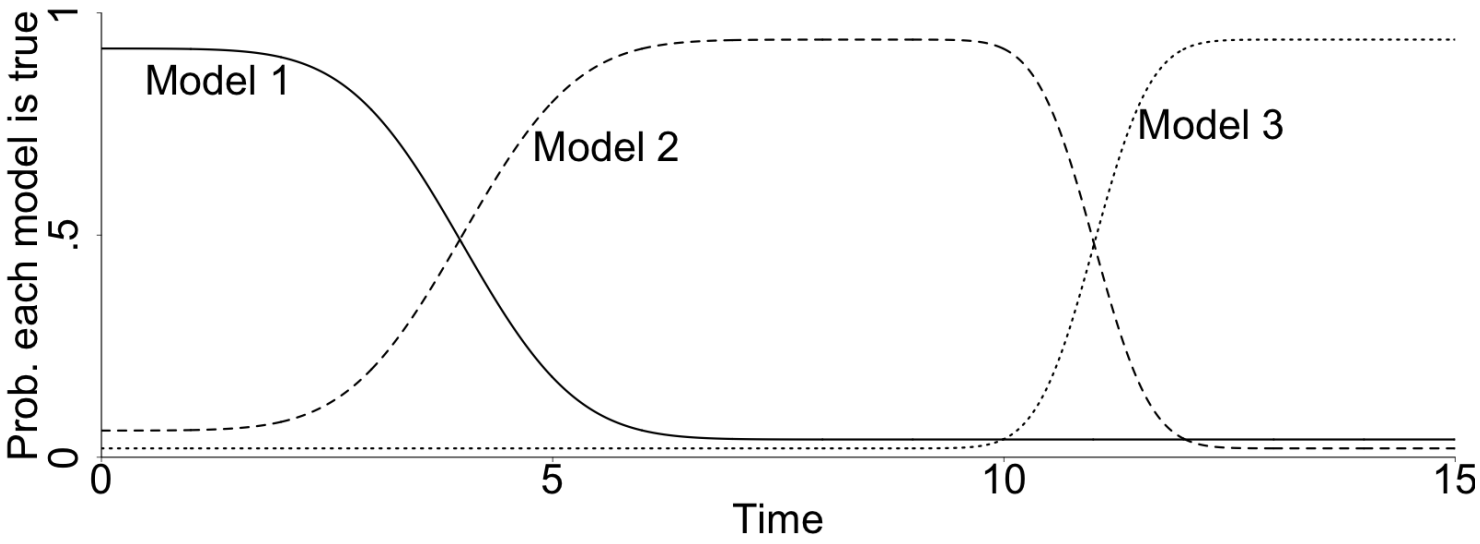


The logic of scientific discovery



The Structure of Scientific Revolutions

Here's a story I don't believe:



The "I've got statistical significance and I'm outta here" attitude

I hate discrete models!

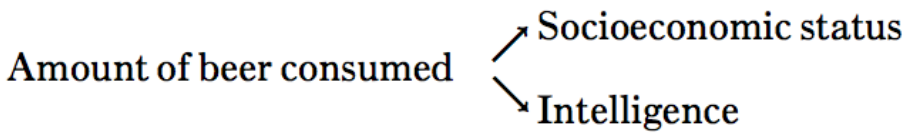
From a cognitive scientist:

If two of the variables are dependent, say, intelligence and socioeconomic status, but conditionally independent given the third variable [beer consumption], then either they are related by one of two chains:

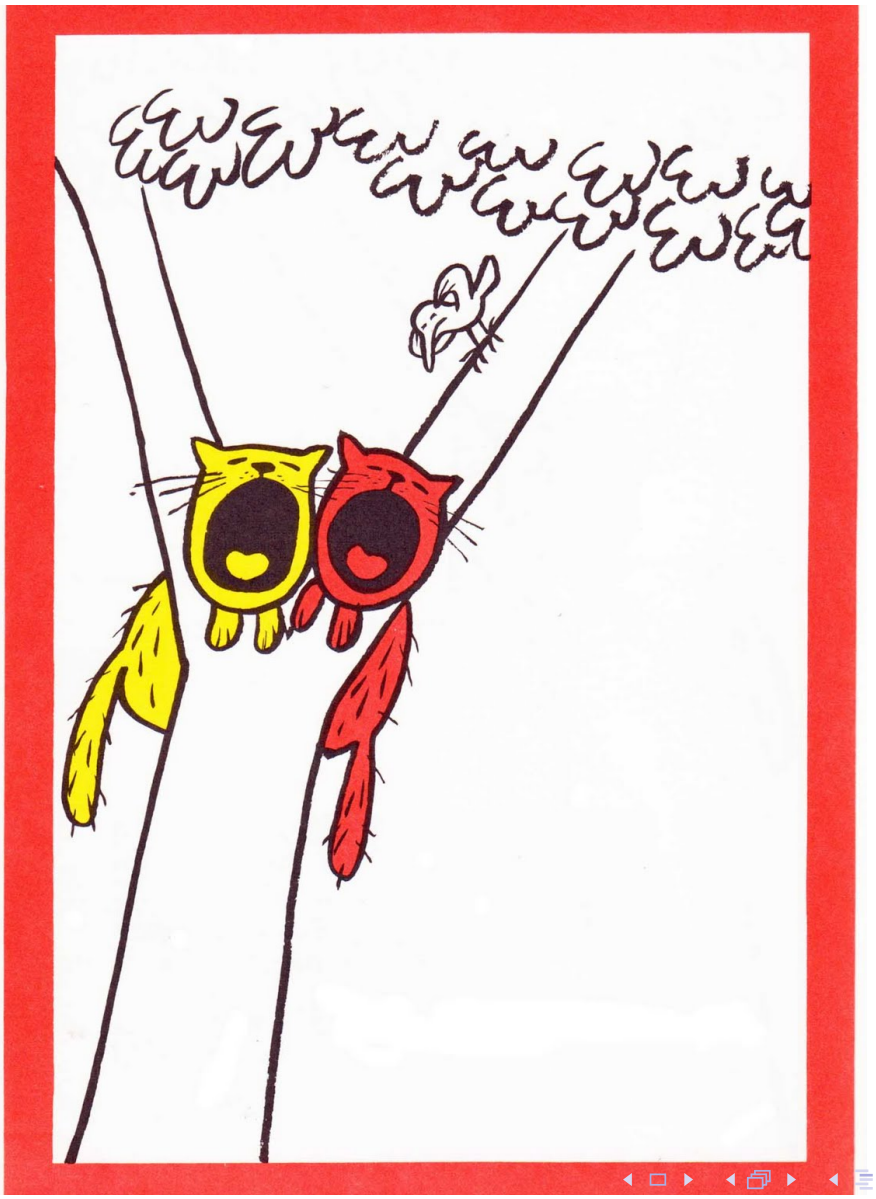
(Intelligence → Amount of beer consumed → Socioeconomic status)

(Socioeconomic status → Amount of beer consumed → Intelligence)

or by a fork:



and then we must use some other means [other than observational data] to decide between these three possibilities. In some cases, common sense may be sufficient, but we can also, if necessary, run an experiment. If we intervene and vary the amount of beer consumed and see that we affect intelligence, that implies that the second or third model is possible; the first one is not. Or course, all this assumes that there aren't other variables mediating between the ones shown that provide alternative explanations of the depen-



What makes me more upset than binary thinking?

The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle

Kristina M. Durante[1], Ashley Rae[1], and Vladas Griskevicius[2]

[1]College of Business, University of Texas, San Antonio, and [2]Carlson School of Management, University of Minnesota

Abstract

Each month, many women experience an ovulatory cycle that regulates fertility. Although re
cycle influences women's mating preferences, we proposed that it might also change wome
views. Building on theory suggesting that political and religious orientation are linked to repr
how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidentia
with large and diverse samples, ovulation had drastically different effects on single women a
relationships. Ovulation led single women to become more liberal, less religious, and more
Obama. In contrast, ovulation led women in committed relationships to become more con
and more likely to vote for Mitt Romney. In addition, ovulation-induced changes in polit
women's voting behavior. Overall, the ovulatory cycle not only influences women's politics
differently for single women than for women in relationships.

Psychological SCIENCE
A Journal of the Association for Psychological Science

The Ancestral Logic of Politics

Upper-Body Strength Regulates Men's Assertion of Self-Interest Over Economic Redistribution

Over human evolutionary history, upper-body strength has been a major component of fighting ability. Evolutionary models of animal conflict predict that actors with greater fighting ability will more actively attempt to acquire or defend resources than less formidable contestants will. Here, we applied these models to political decision making about redistribution of income and wealth among modern humans. In studies conducted in Argentina, Denmark, and the United States, men with greater upper-

Unfalsifiable!

But . . .
falsification of statistical models is not our goal.

1 quick tip to falsify any statistical model . . .

Birthdays!

Social Science & Medicine 73 (2011) 1246–1248

Contents lists available at SciVerse ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Short report

Influence of Valentine's Day and Halloween on Birth Timing

Becca R. Levy*, Pil H. Chung, Martin D. Slade

Yale University, School of Public Health, Division of Social & Behavioral Sciences, 60 College Street, New Haven, CT 06520-8034, United States

ARTICLE INFO

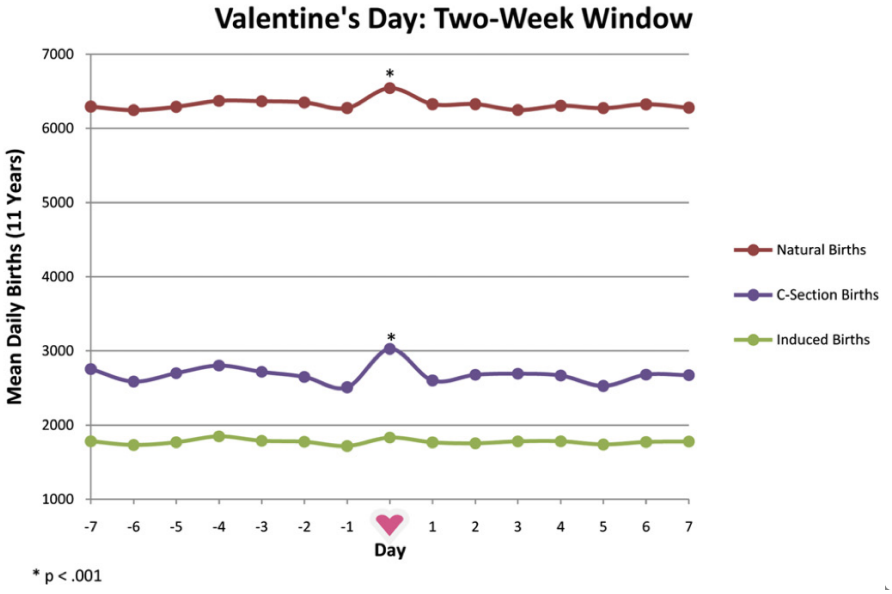
Article history:
Available online 28 July 2011

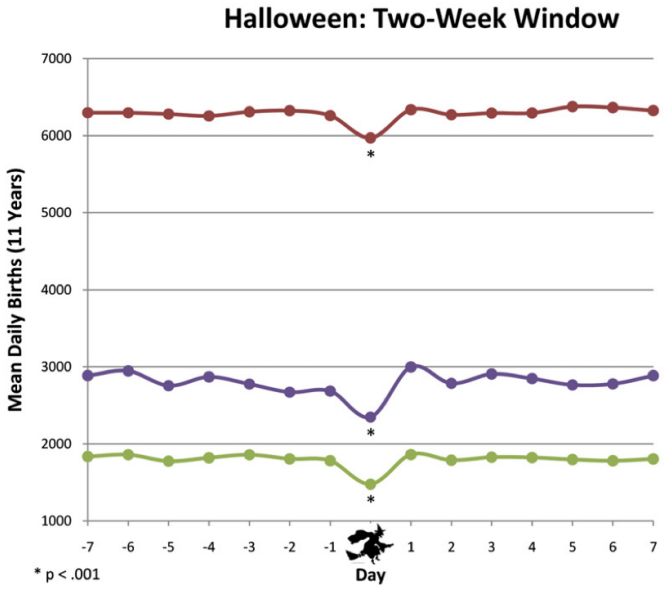
Keywords:
United States
Culture
Birth timing
Holidays
Pregnancy
Biocultural
Birth

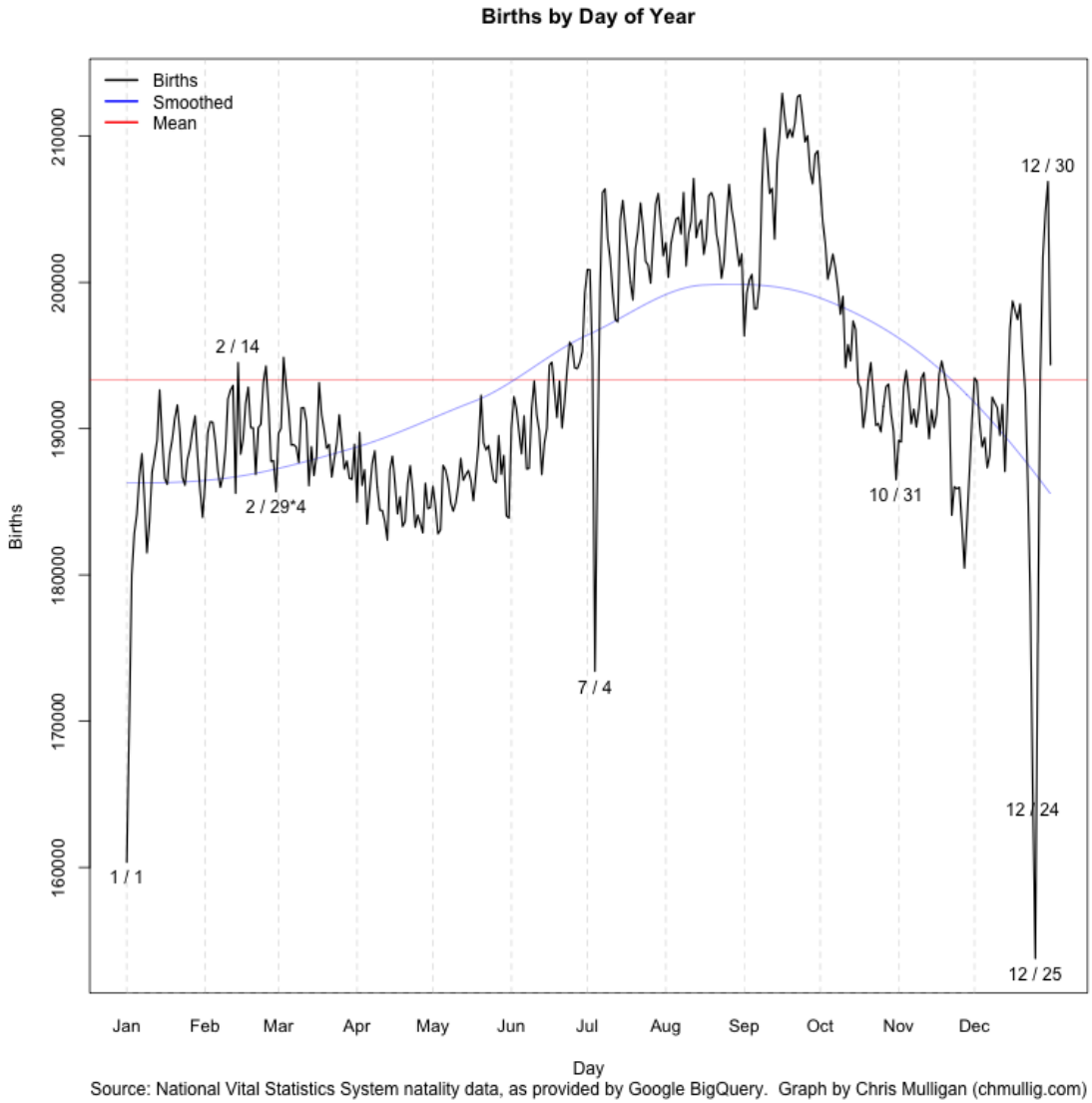
ABSTRACT

It is known that cultural representations, in the form of stereotypes, can influence functional health. We predicted that the influence of cultural representations, in the form of salient holidays, would extend to birth timing. On Valentine's Day, which conveys positive symbolism, there was a 3.6% increase in spontaneous births and a 12.1% increase in cesarean births. Whereas, on Halloween, which conveys negative symbolism, there was a 5.3% decrease in spontaneous births and a 16.9% decrease in cesarean births. These effects reached significance at $p < .0001$, after adjusting for year and day of the week. The sample was based on birth-certificate information for all births in the United States within one week on either side of each holiday across 11 years. The Valentine's-Day window included 1,676,217 births and the Halloween window included 1,809,304 births. Our findings raise the possibility that pregnant women may be able to control the timing of spontaneous births, in contrast to the traditional assumption, and that scheduled births are also influenced by the cultural representations of the two holidays.

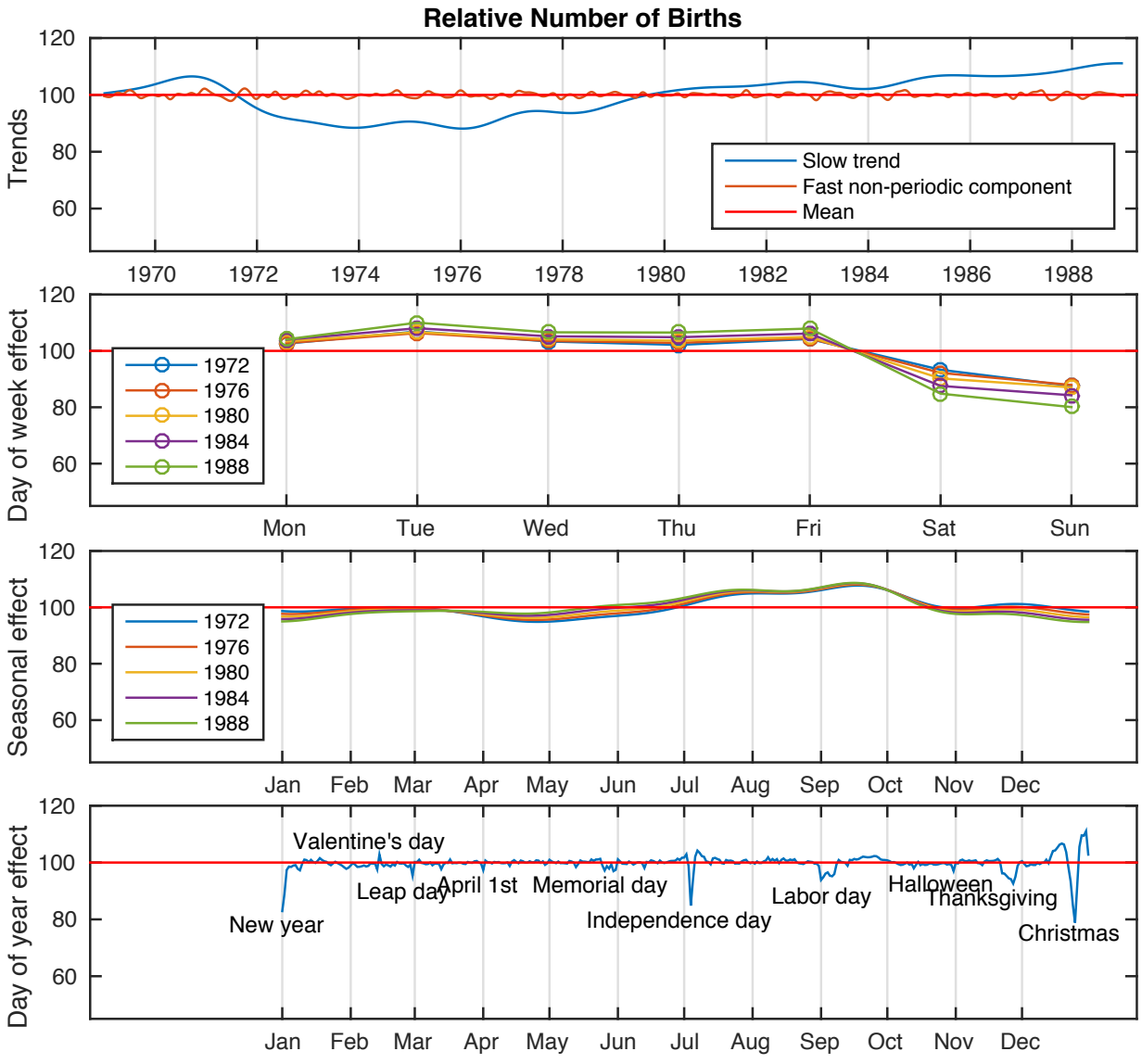
The published graphs show data from 30 days in the year

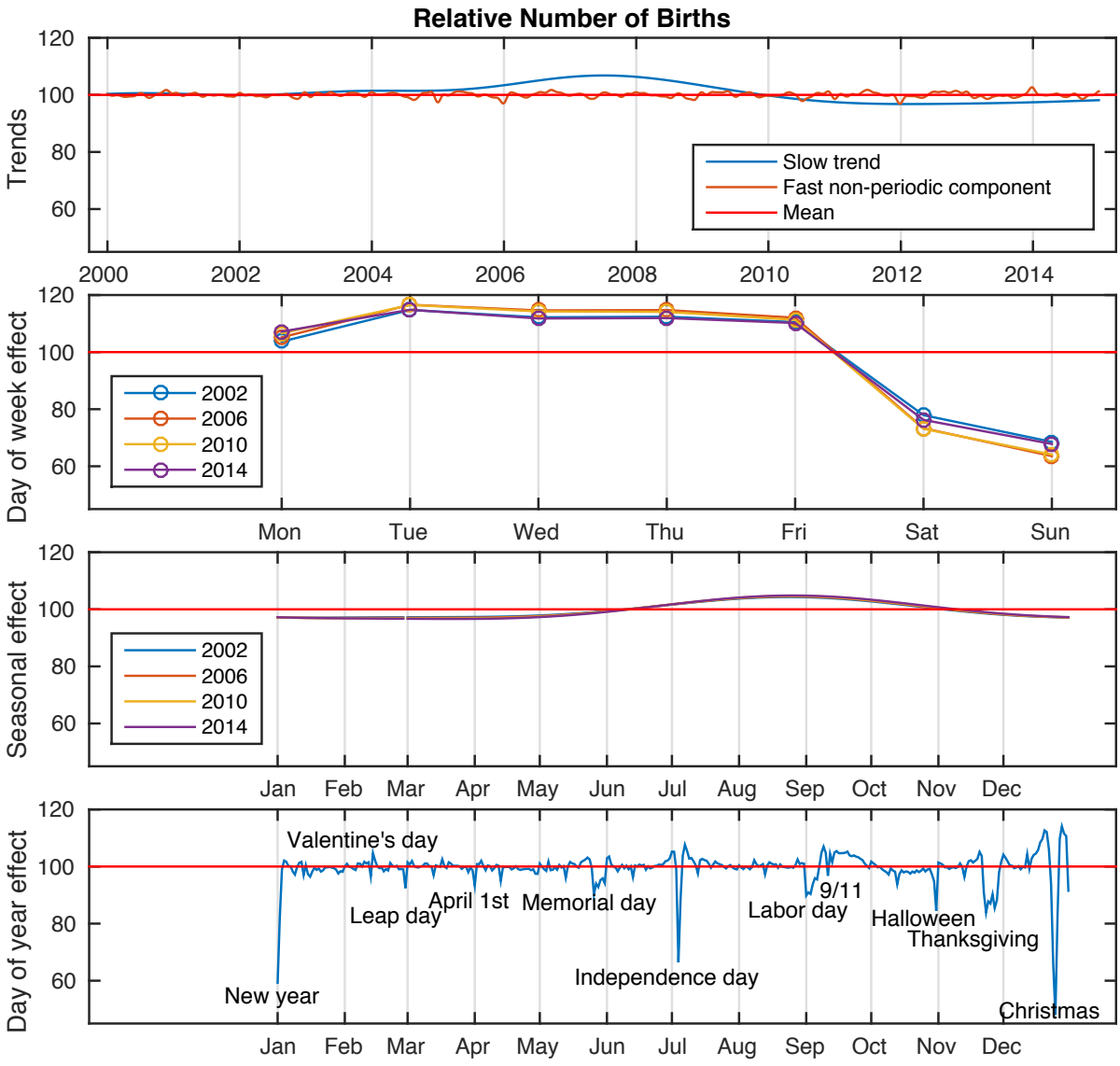


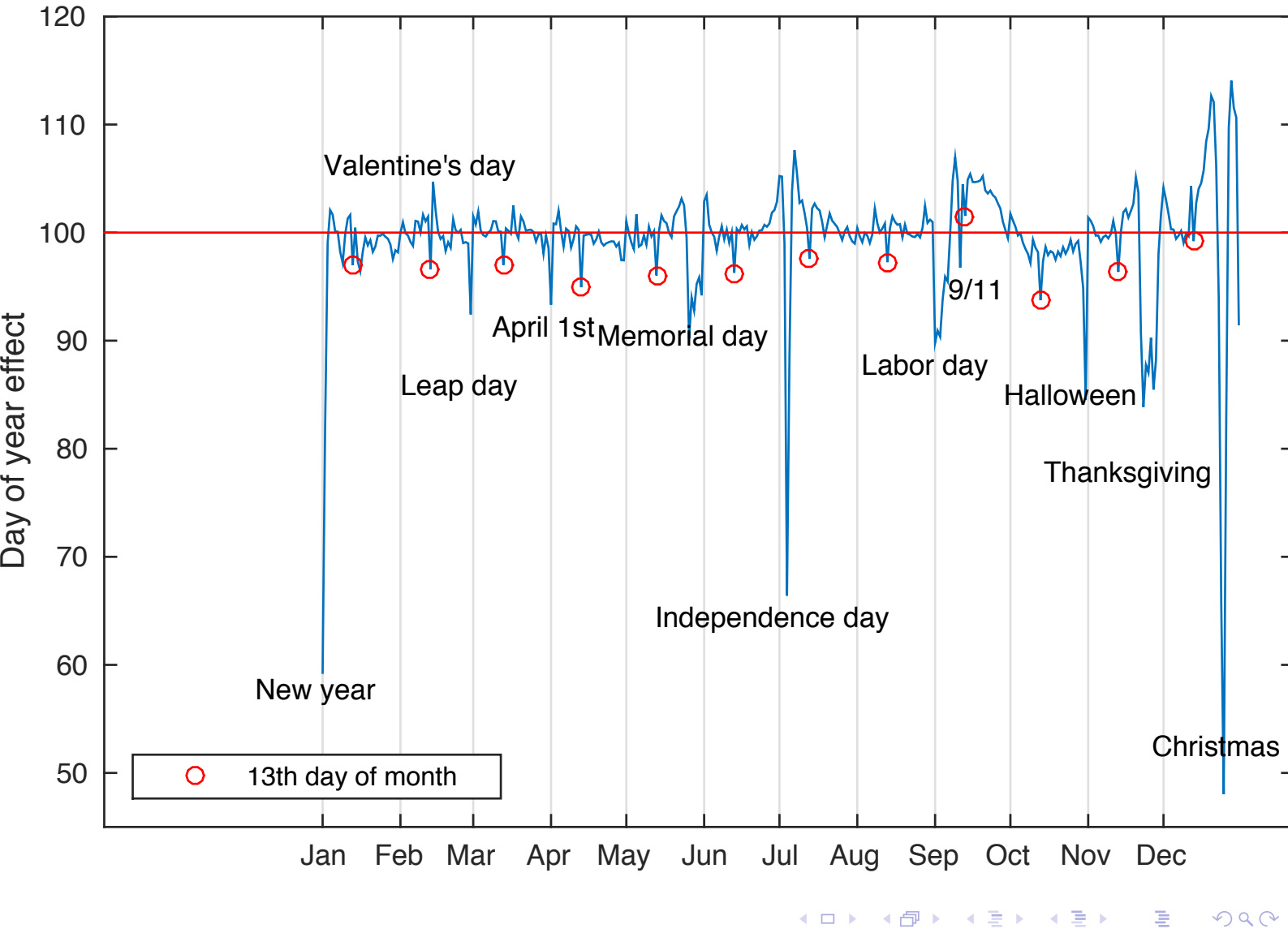


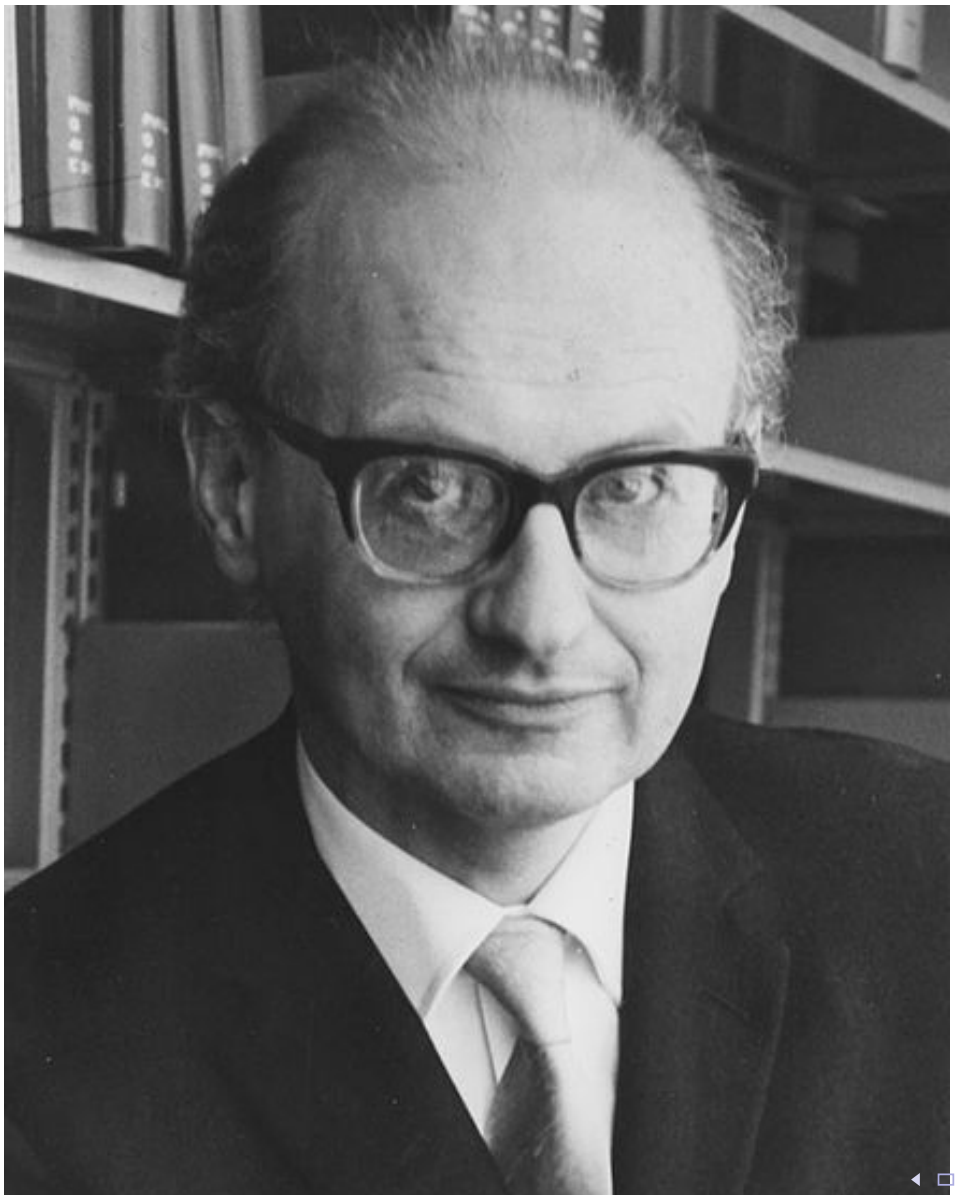


Source: National Vital Statistics System natality data, as provided by Google BigQuery. Graph by Chris Mulligan (chmullig.com)


Relative Number of Births
Trends
120
100
80
60
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
Slow trend
Fast non-periodic component
Mean
Day of week effect
1972
1976
1980
1984
1988
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Seasonal effect
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Day of year effect
Valentine's day
Leap day
April 1st
Memorial day
New year
Independence day
Labor day
Halloween
Thanksgiving
Christmas


Relative Number of Births
Trends
120
100
80
60
2000
2002
2004
2006
2008
2010
2012
2014
Slow trend
Fast non-periodic component
Mean
Day of week effect
2002
2006
2010
2014
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Seasonal effect
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Day of year effect
Valentine's day
Leap day
April 1st
Memorial day
Independence day
Labor day
9/11
Halloween
Thanksgiving
New year
Christmas


120
110
100
90
80
70
60
50
Day of year effect
Valentine's day
Leap day
April 1st
Memorial day
Independence day
Labor day
9/11
Halloween
Thanksgiving
New year
Christmas
13th day of month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

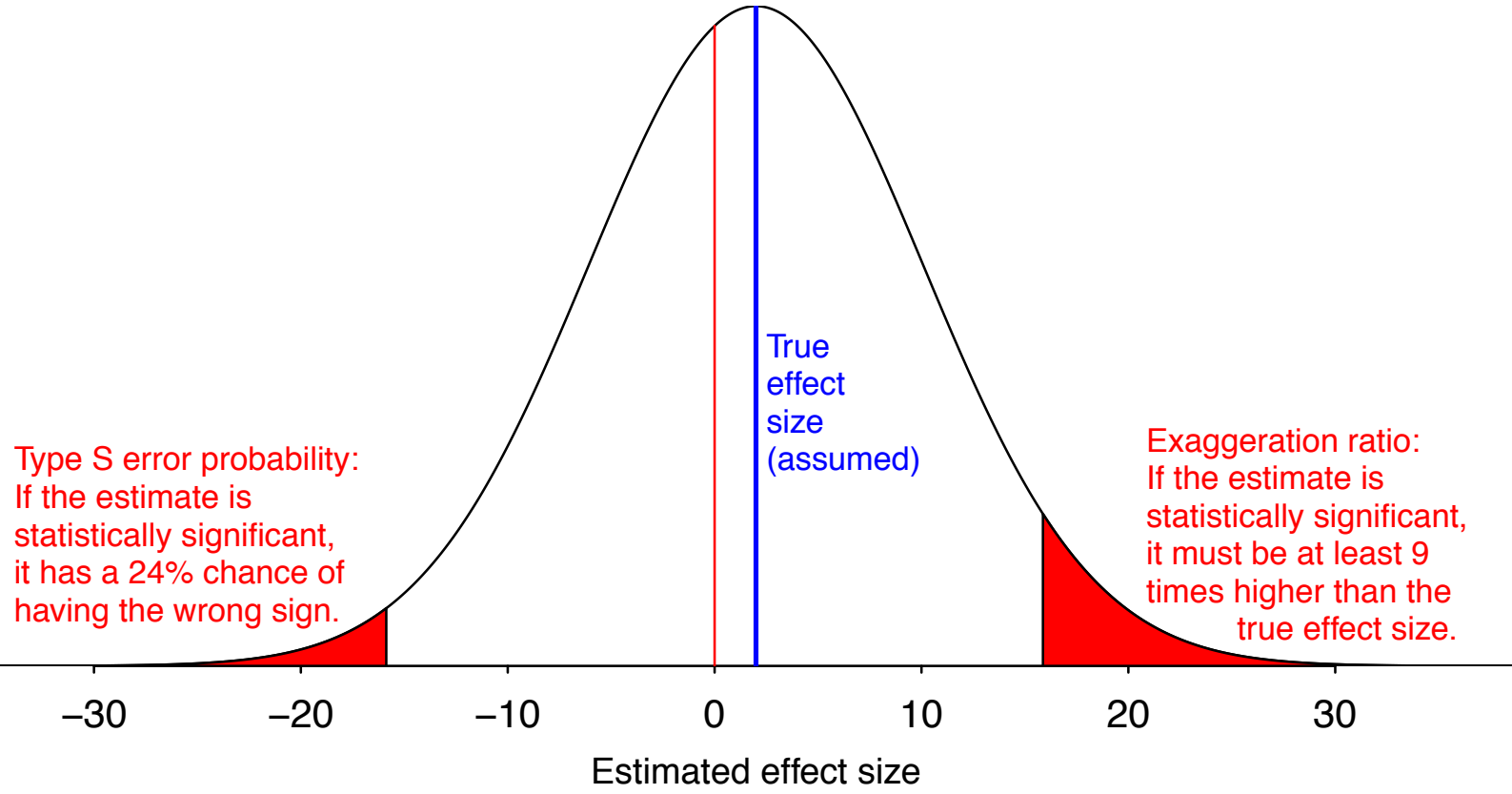


Methodology of Scientific Research Programmes

How can things go wrong?

The statistical significance filter and the resulting feedback loop

This is what "power = 0.06" looks like.
Get used to it.



Policy!

Labor Market Returns to Early Childhood Stimulation: a 20-year Followup to an Experimental Intervention in Jamaica

Paul Gertler, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M. Chang, Sally Grantham-McGregor

We find large effects on the earnings of participants from a randomized intervention that gave psychosocial stimulation to stunted Jamaican toddlers living in poverty. The intervention consisted of one-hour weekly visits from community Jamaican health workers over a 2-year period that taught parenting skills and encouraged mothers to interact and play with their children in ways that would develop their children's cognitive and personality skills. We re-interviewed the study participants 20 years after the intervention. Stimulation increased the average earnings of participants by 42 percent. Treatment group earnings caught up to the earnings of a matched non-stunted comparison group. These findings show that psychosocial stimulation early in childhood in disadvantaged settings can have substantial effects on labor market outcomes and reduce later life inequality.

Bad Bayes

Bad Bayes

- Model:

Bad Bayes

- Model:
  - $y|\theta \sim \mathsf{N}(\theta, 1)$

Bad Bayes

- Model:
  - $y|\theta \sim \mathrm{N}(\theta, 1)$
  - $p(\theta) \propto 1$

Bad Bayes

- Model:
  - $y|\theta \sim \mathsf{N}(\theta, 1)$
  - $p(\theta) \propto 1$
- Data:

Bad Bayes

- Model:
  - $y|\theta \sim \mathsf{N}(\theta, 1)$
  - $p(\theta) \propto 1$
- Data:
  - $y = 1$

Bad Bayes

- Model:
  - $y|\theta \sim \mathsf{N}(\theta, 1)$
  - $p(\theta) \propto 1$
- Data:
  - $y = 1$
- Inference:

Bad Bayes

- Model:
  - $y|\theta \sim \mathsf{N}(\theta, 1)$
  - $p(\theta) \propto 1$
- Data:
  - $y = 1$
- Inference:
  - $\theta|y \sim \mathsf{N}(y, 1)$

Bad Bayes

- Model:
  - $y|\theta \sim \mathrm{N}(\theta, 1)$
  - $p(\theta) \propto 1$
- Data:
  - $y = 1$
- Inference:
  - $\theta|y \sim \mathrm{N}(y, 1)$
  - $\Pr(\theta > 0|y) = .84$

Bad Bayes

- Model:
  - $y|\theta \sim \mathrm{N}(\theta, 1)$
  - $p(\theta) \propto 1$
- Data:
  - $y = 1$
- Inference:
  - $\theta|y \sim \mathrm{N}(y, 1)$
  - $\Pr(\theta > 0|y) = .84$
- Wanna bet??

Hypothesis testing is a bad idea:

Through a series of examples, we consider problems with classical hypothesis testing, whether performed using classical p-values or confidence intervals, Bayes factors, or Bayesian inference using noninformative priors. We locate the problem not in the use of any particular statistical method but rather with larger problems of deterministic thinking and a misguided version of Popperianism in which the rejection of a straw-man null hypothesis is taken as confirmation of a preferred alternative. We suggest solutions involving multilevel modeling and informative Bayesian inference.