Pay No Attention To the Model Behind the Curtain

Interdisciplinary Data Science Institute Group Meeting
UC Berkeley

Philip B. Stark

25 August 2025

University of California, Berkeley

Mark Twain:

> *In the space of one hundred and seventy-six years the Lower Mississippi has shortened itself two hundred and forty-two miles. That is an average of a trifle over one mile and a third per year.*
>
> *Therefore, any calm person, who is not blind or idiotic, can see that in the Old Oolitic Silurian Period, just a million years ago next November, the Lower Mississippi River was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of Mexico like a fishing-rod.*
>
> *And by the same token any person can see that seven hundred and forty-two years from now the lower Mississippi will be only a mile and three-quarters long.. . . .*
>
> *There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.*

George Box:

> *All models are wrong, but some are useful.*

George Box:

> *All models are wrong, but some are useful.*
>
> *Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.*

George Box:

> *All models are wrong, but some are useful.*
>
> *Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.*

Me:

> *In applied statistics, often the model is importantly wrong.*

For ~70 years, fashionable to assign numbers to things to make them “scientific.”

Qualitative arguments considered unscientific.

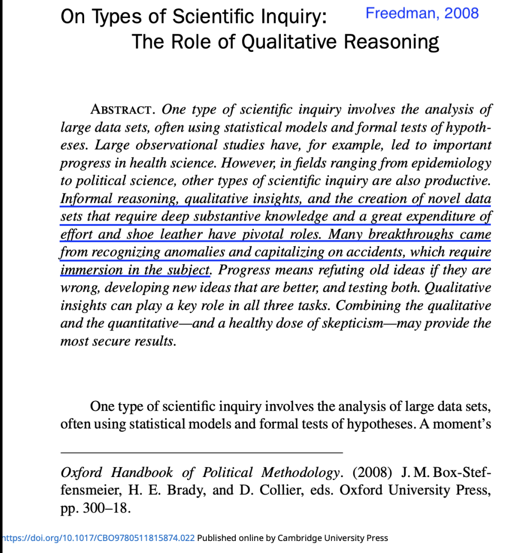
History, sociology, & economics exalt computation as scientific and objective.

For $\sim$70 years, fashionable to assign numbers to things to make them "scientific."

Qualitative arguments considered unscientific.

History, sociology, & economics exalt computation as scientific and objective.

> *Regression modeling is a dominant paradigm, and many investigators seem to consider that any piece of empirical research has to be equivalent to a regression model. Questioning the value of regression is then tantamount to denying the value of data. –Freedman, 1991*



On Types of Scientific Inquiry: The Role of Qualitative Reasoning

Freedman, 2008

ABSTRACT. *One type of scientific inquiry involves the analysis of large data sets, often using statistical models and formal tests of hypotheses. Large observational studies have, for example, led to important progress in health science. However, in fields ranging from epidemiology to political science, other types of scientific inquiry are also productive. Informal reasoning, qualitative insights, and the creation of novel data sets that require deep substantive knowledge and a great expenditure of effort and shoe leather have pivotal roles. Many breakthroughs came from recognizing anomalies and capitalizing on accidents, which require immersion in the subject. Progress means refuting old ideas if they are wrong, developing new ideas that are better, and testing both. Qualitative insights can play a key role in all three tasks. Combining the qualitative and the quantitative—and a healthy dose of skepticism—may provide the most secure results.*

One type of scientific inquiry involves the analysis of large data sets, often using statistical models and formal tests of hypotheses. A moment's

Oxford Handbook of Political Methodology. (2008) J. M. Box-Steffensmeier, H. E. Brady, and D. Collier, eds. Oxford University Press, pp. 300–18.

ttps://doi.org/10.1017/CBO9780511815874.022 Published online by Cambridge University Press

It is objective and rational to take account of imponderable factors. It is subjective, irrational, and dangerous not to take account of them. As that champion of rationality, the philosopher Bertrand Russell, would have argued, rationality involves the whole and balanced use of human faculty, not a rejection of that fraction of it that cannot be made numerical. –Nature editorial, 1978.

Quantifauxcation. Assign a meaningless number, then conclude that because it is quantitative, it must mean something.

Type III errors. Answering the wrong question, e.g., testing a statistical hypothesis that is untethered from the scientific hypothesis.

Procrustes' quantifauxcation: forcing incommensurables to one scale

- form an "index" that combines a variety of things into a single number, e.g., adding or averaging "points" on Likert scales
  - clinical outcomes for conditions like PTSD: add "severity" of symptoms
  - university rankings
  - student evaluations
- combine different kinds of uncertainty as if they were all probabilities
- cost-benefit analyses when costs and benefits aren't all monetary

Example: Cost-benefit analysis

- often claimed to be the only rational basis for policy
- costs & consequences hard to anticipate, enumerate, or estimate.
- assumes all costs and all benefits can be put on same 1-d scale, e.g., “utility”

Conjoint analysis (Luce and Tukey, 1964)

Combining multiple attributes on a single scale involves nontrivial conditions.

Attribute	Possible attribute values		
Filling	peanut butter	turkey	lamb
Condiment	grape jelly	mustard	mint sauce

Conjoint analysis (Luce and Tukey, 1964)

Combining multiple attributes on a single scale involves nontrivial conditions.

Attribute	Possible attribute values		
Filling	peanut butter	turkey	lamb
Condiment	grape jelly	mustard	mint sauce

Double-cancellation axiom:

> *If you prefer peanut butter and jelly to turkey and mint sauce, and you prefer turkey and mustard to lamb and grape jelly, then you* ***must*** *prefer peanut butter and mustard to lamb and cranberry sauce.*

Example: Risk preferences

“Risk = probability × consequences”

- Human preference orderings are not based on expected returns.
- Preference for sure things over bets: many would prefer \$1 million for sure over a 10% chance of receiving \$20 million
- Loss aversion: many would prefer a 10% chance of winning \$1 million over a 50% chance of winning \$2 million with a 50% chance of losing \$100 thousand (9.5x expected return)

Lucien LeCam (1977) on treating all uncertainties alike

. . . [W]e can be uncertain for many reasons. For instance, we may be uncertain because (1) we lack definite information, (2) the events involved will occur according to the results of the spin of a roulette wheel, (3) we could find out by pure logic but it is too hard. The first type of uncertainty occurs in practically every question. The second assumes a well-defined mechanism. However, the neo-Bayesian theory seems to make no real distinction between probabilities attached to the three types. It answers in the same manner the following questions.

(1) What is the probability that Eudoxus had bigger feet than Euclid?

(2) What is the probability that a toss of a 'fair' coin will result in tails?

(3) What is the probability that the $10^{137} + 1$ *digit of* π *is a 7?*

. . . Thus, presumably, when neo-Bayesians state that a certain event A has probability one-half, this may mean either that he did not bother to think about it, or that he has no information on the subject, or that whether A occurs or not will be decided by the toss of a fair coin. The number 1/2 itself does not contain any information about the process by which it was obtained . . . .

Subjective estimates/beliefs can be manipulated through *anchoring* and *priming* (Tversky & Kahneman, 1975).

Anchoring affects entire disciplines: Millikan oil drop experiment, speed of light, iron in spinach

- We're poor judges of probability: biases from *representativeness* and *availability*
- We're poor judges of weights, even of objects in our hands: bias by the density and shape of the object—and even its color.
- We're bad judges of randomness: *apophenia* and *pareidolia*
- We're over-confident about our estimates and predictions
- Our confidence is unrelated to our actual accuracy

Nature editorial, 1978

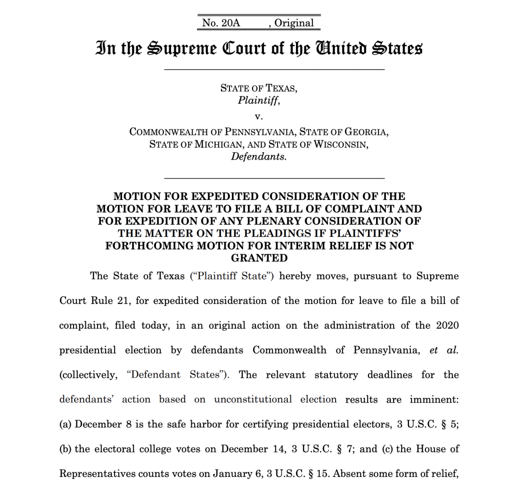
LORD ROTHSCHILD, speaking on British television last week, argued that we should develop a table of risks so we could compare, say, the risk of our dying in an automobile accident with the risk of Baader-Meinhoff guerillas taking over the nuclear reactor next door. Then we would know how seriously to take our risks, be they nuclear power, damage to the environment or whatever.

. . .

It is fine for Rothschild to demonstrate his agility with arithmetic, converting probabilities from one form to another (and implying that the viewers could not do it) but this is only the kindergarten of risk.

. . . Rothschild confused two fundamental distinct kinds of risk in his table: known risks-such as car accidents-where the risk is simply calculated from past events; and unknown risks—such as the terrorists taking over a fast breeder—which are matters of estimating the future. The latter risks inevitably depend on theory. Whether the theory is a social theory of terrorism or a risk-tree analysis of fast breeder failure, it will be open to conjecture. And it ought to be remembered that the history of engineering is largely a history of unforeseen accidents. Risk estimates can be proved only by events. Thus it is easy for groups, consciously or unconsciously, to bend their calculations to suit their own objectives or prejudices. With unknown risks it is as important to take these into account as to come up with a number.

Application 1: election fraud



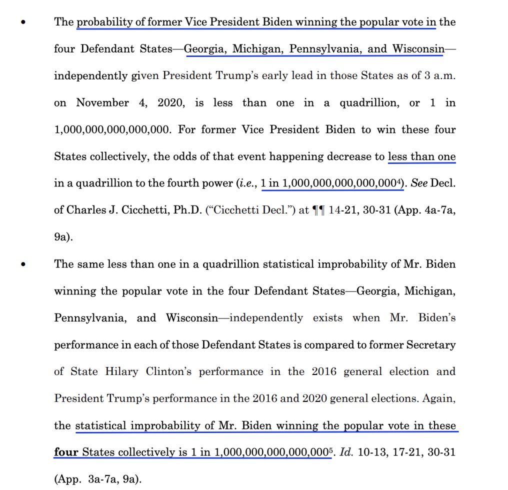
No. 20A , Original

In the Supreme Court of the United States

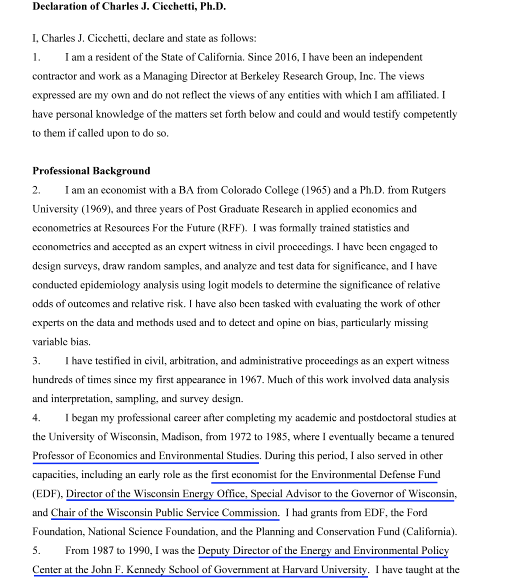
STATE OF TEXAS,
Plaintiff,
v.
COMMONWEALTH OF PENNSYLVANIA, STATE OF GEORGIA,
STATE OF MICHIGAN, AND STATE OF WISCONSIN,
Defendants.

MOTION FOR EXPEDITED CONSIDERATION OF THE MOTION FOR LEAVE TO FILE A BILL OF COMPLAINT AND FOR EXPEDITION OF ANY PLENARY CONSIDERATION OF THE MATTER ON THE PLEADINGS IF PLAINTIFFS' FORTHCOMING MOTION FOR INTERIM RELIEF IS NOT GRANTED

The State of Texas ("Plaintiff State") hereby moves, pursuant to Supreme Court Rule 21, for expedited consideration of the motion for leave to file a bill of complaint, filed today, in an original action on the administration of the 2020 presidential election by defendants Commonwealth of Pennsylvania, *et al.* (collectively, "Defendant States"). The relevant statutory deadlines for the defendants' action based on unconstitutional election results are imminent: (a) December 8 is the safe harbor for certifying presidential electors, 3 U.S.C. § 5; (b) the electoral college votes on December 14, 3 U.S.C. § 7; and (c) the House of Representatives counts votes on January 6, 3 U.S.C. § 15. Absent some form of relief,



- The probability of former Vice President Biden winning the popular vote in the four Defendant States—Georgia, Michigan, Pennsylvania, and Wisconsin—independently given President Trump's early lead in those States as of 3 a.m. on November 4, 2020, is less than one in a quadrillion, or 1 in 1,000,000,000,000,000. For former Vice President Biden to win these four States collectively, the odds of that event happening decrease to less than one in a quadrillion to the fourth power (*i.e.*, 1 in 1,000,000,000,000,000^4). *See* Decl. of Charles J. Cicchetti, Ph.D. ("Cicchetti Decl.") at ¶¶ 14-21, 30-31 (App. 4a-7a, 9a).
- The same less than one in a quadrillion statistical improbability of Mr. Biden winning the popular vote in the four Defendant States—Georgia, Michigan, Pennsylvania, and Wisconsin—independently exists when Mr. Biden's performance in each of those Defendant States is compared to former Secretary of State Hilary Clinton's performance in the 2016 general election and President Trump's performance in the 2016 and 2020 general elections. Again, the statistical improbability of Mr. Biden winning the popular vote in these four States collectively is 1 in 1,000,000,000,000,000^5. *Id.* 10-13, 17-21, 30-31 (App. 3a-7a, 9a).



Declaration of Charles J. Cicchetti, Ph.D.

I, Charles J. Cicchetti, declare and state as follows:

1. I am a resident of the State of California. Since 2016, I have been an independent contractor and work as a Managing Director at Berkeley Research Group, Inc. The views expressed are my own and do not reflect the views of any entities with which I am affiliated. I have personal knowledge of the matters set forth below and could and would testify competently to them if called upon to do so.

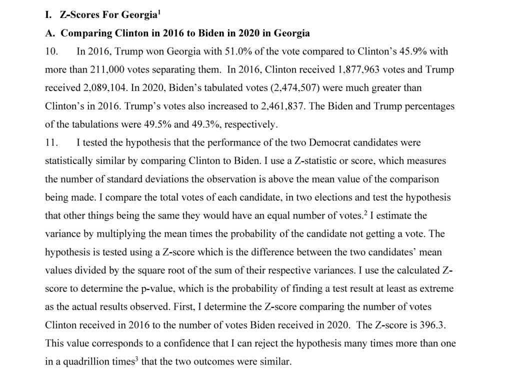
Professional Background

2. I am an economist with a BA from Colorado College (1965) and a Ph.D. from Rutgers University (1969), and three years of Post Graduate Research in applied economics and econometrics at Resources For the Future (RFF). I was formally trained statistics and econometrics and accepted as an expert witness in civil proceedings. I have been engaged to design surveys, draw random samples, and analyze and test data for significance, and I have conducted epidemiology analysis using logit models to determine the significance of relative odds of outcomes and relative risk. I have also been tasked with evaluating the work of other experts on the data and methods used and to detect and opine on bias, particularly missing variable bias.

3. I have testified in civil, arbitration, and administrative proceedings as an expert witness hundreds of times since my first appearance in 1967. Much of this work involved data analysis and interpretation, sampling, and survey design.

4. I began my professional career after completing my academic and postdoctoral studies at the University of Wisconsin, Madison, from 1972 to 1985, where I eventually became a tenured Professor of Economics and Environmental Studies. During this period, I also served in other capacities, including an early role as the first economist for the Environmental Defense Fund (EDF), Director of the Wisconsin Energy Office, Special Advisor to the Governor of Wisconsin, and Chair of the Wisconsin Public Service Commission. I had grants from EDF, the Ford Foundation, National Science Foundation, and the Planning and Conservation Fund (California).

5. From 1987 to 1990, I was the Deputy Director of the Energy and Environmental Policy Center at the John F. Kennedy School of Government at Harvard University. I have taught at the

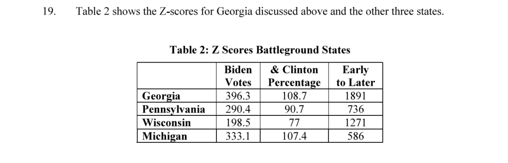


I. Z-Scores For Georgia[1]

A. Comparing Clinton in 2016 to Biden in 2020 in Georgia

10. In 2016, Trump won Georgia with 51.0% of the vote compared to Clinton's 45.9% with more than 211,000 votes separating them. In 2016, Clinton received 1,877,963 votes and Trump received 2,089,104. In 2020, Biden's tabulated votes (2,474,507) were much greater than Clinton's in 2016. Trump's votes also increased to 2,461,837. The Biden and Trump percentages of the tabulations were 49.5% and 49.3%, respectively.

11. I tested the hypothesis that the performance of the two Democrat candidates were statistically similar by comparing Clinton to Biden. I use a Z-statistic or score, which measures the number of standard deviations the observation is above the mean value of the comparison being made. I compare the total votes of each candidate, in two elections and test the hypothesis that other things being the same they would have an equal number of votes.[2] I estimate the variance by multiplying the mean times the probability of the candidate not getting a vote. The hypothesis is tested using a Z-score which is the difference between the two candidates' mean values divided by the square root of the sum of their respective variances. I use the calculated Z-score to determine the p-value, which is the probability of finding a test result at least as extreme as the actual results observed. First, I determine the Z-score comparing the number of votes Clinton received in 2016 to the number of votes Biden received in 2020. The Z-score is 396.3. This value corresponds to a confidence that I can reject the hypothesis many times more than one in a quadrillion times[3] that the two outcomes were similar.

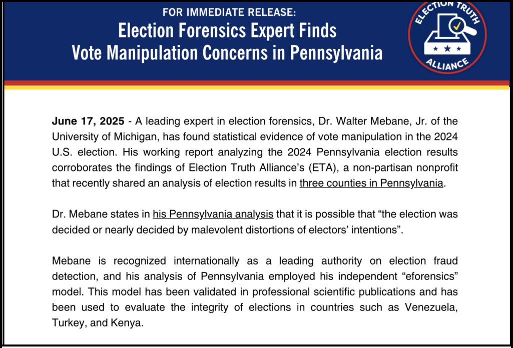


19. Table 2 shows the Z-scores for Georgia discussed above and the other three states.

Table 2: Z Scores Battleground States

	Biden	& Clinton	Early
	Votes	Percentage	to Later
Georgia	396.3	108.7	1891
Pennsylvania	290.4	90.7	736
Wisconsin	198.5	77	1271
Michigan	333.1	107.4	586

Type III error: answers wrong question.



FOR IMMEDIATE RELEASE:

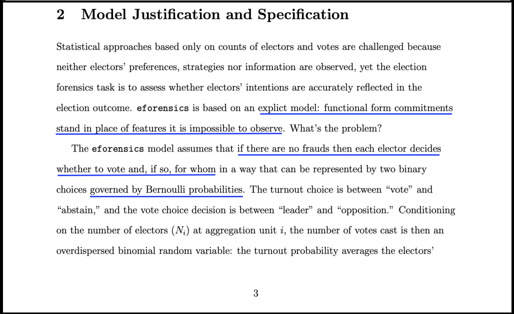
Election Forensics Expert Finds Vote Manipulation Concerns in Pennsylvania

ELECTION TRUTH ALLIANCE

June 17, 2025 - A leading expert in election forensics, Dr. Walter Mebane, Jr. of the University of Michigan, has found statistical evidence of vote manipulation in the 2024 U.S. election. His working report analyzing the 2024 Pennsylvania election results corroborates the findings of Election Truth Alliance's (ETA), a non-partisan nonprofit that recently shared an analysis of election results in three counties in Pennsylvania.

Dr. Mebane states in his Pennsylvania analysis that it is possible that "the election was decided or nearly decided by malevolent distortions of electors' intentions".

Mebane is recognized internationally as a leading authority on election fraud detection, and his analysis of Pennsylvania employed his independent "eforensics" model. This model has been validated in professional scientific publications and has been used to evaluate the integrity of elections in countries such as Venezuela, Turkey, and Kenya.

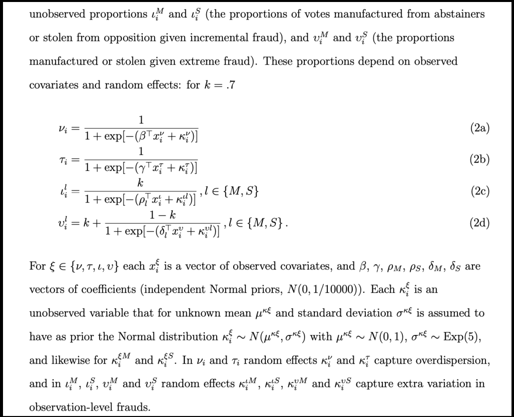


2 Model Justification and Specification

Statistical approaches based only on counts of electors and votes are challenged because neither electors' preferences, strategies nor information are observed, yet the election forensics task is to assess whether electors' intentions are accurately reflected in the election outcome. `eforensics` is based on an explicit model: functional form commitments stand in place of features it is impossible to observe. What's the problem?

The `eforensics` model assumes that if there are no frauds then each elector decides whether to vote and, if so, for whom in a way that can be represented by two binary choices governed by Bernoulli probabilities. The turnout choice is between "vote" and "abstain," and the vote choice decision is between "leader" and "opposition." Conditioning on the number of electors (N_i) at aggregation unit i, the number of votes cast is then an overdispersed binomial random variable: the turnout probability averages the electors'

3



unobserved proportions ι_i^M and ι_i^S (the proportions of votes manufactured from abstainers or stolen from opposition given incremental fraud), and υ_i^M and υ_i^S (the proportions manufactured or stolen given extreme fraud). These proportions depend on observed covariates and random effects: for $k = .7$

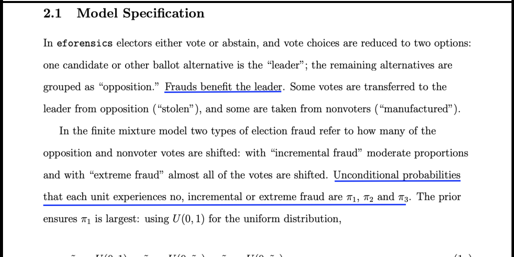
$$\nu_i = \frac{1}{1+\exp[-(\beta^\top x_i^\nu + \kappa_i^\nu)]} \tag{2a}$$

$$\tau_i = \frac{1}{1+\exp[-(\gamma^\top x_i^\tau + \kappa_i^\tau)]} \tag{2b}$$

$$\iota_i^l = \frac{k}{1+\exp[-(\rho_l^\top x_i^\iota + \kappa_i^{\iota l})]}\,, l \in \{M, S\} \tag{2c}$$

$$\upsilon_i^l = k + \frac{1-k}{1+\exp[-(\delta_l^\top x_i^\upsilon + \kappa_i^{\upsilon l})]}\,, l \in \{M, S\}\,. \tag{2d}$$

For $\xi \in \{\nu, \tau, \iota, \upsilon\}$ each x_i^ξ is a vector of observed covariates, and β, γ, ρ_M, ρ_S, δ_M, δ_S are vectors of coefficients (independent Normal priors, $N(0, 1/10000)$). Each κ_i^ξ is an unobserved variable that for unknown mean $\mu^{\kappa\xi}$ and standard deviation $\sigma^{\kappa\xi}$ is assumed to have as prior the Normal distribution $\kappa_i^\xi \sim N(\mu^{\kappa\xi}, \sigma^{\kappa\xi})$ with $\mu^{\kappa\xi} \sim N(0,1)$, $\sigma^{\kappa\xi} \sim \text{Exp}(5)$, and likewise for $\kappa_i^{\xi M}$ and $\kappa_i^{\xi S}$. In ν_i and τ_i random effects κ_i^ν and κ_i^τ capture overdispersion, and in ι_i^M, ι_i^S, υ_i^M and υ_i^S random effects $\kappa_i^{\iota M}$, $\kappa_i^{\iota S}$, $\kappa_i^{\upsilon M}$ and $\kappa_i^{\upsilon S}$ capture extra variation in observation-level frauds.



2.1 Model Specification

In `eforensics` electors either vote or abstain, and vote choices are reduced to two options: one candidate or other ballot alternative is the "leader"; the remaining alternatives are grouped as "opposition." Frauds benefit the leader. Some votes are transferred to the leader from opposition ("stolen"), and some are taken from nonvoters ("manufactured").

In the finite mixture model two types of election fraud refer to how many of the opposition and nonvoter votes are shifted: with "incremental fraud" moderate proportions and with "extreme fraud" almost all of the votes are shifted. Unconditional probabilities that each unit experiences no, incremental or extreme fraud are π_1, π_2 and π_3. The prior ensures π_1 is largest: using $U(0,1)$ for the uniform distribution,

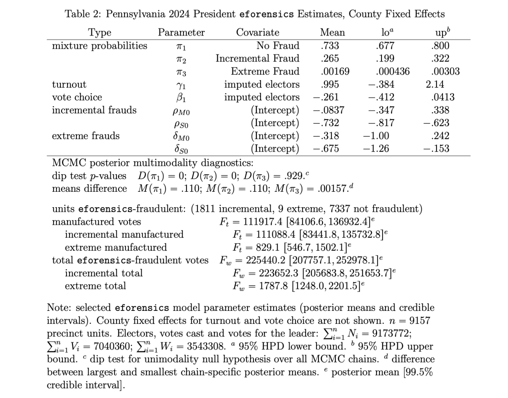


Table 2: Pennsylvania 2024 President `eforensics` Estimates, County Fixed Effects

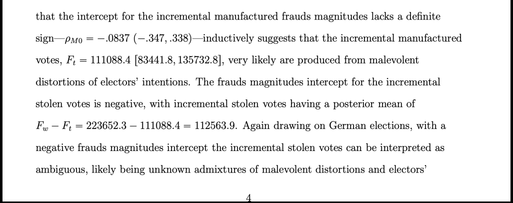
Type	Parameter	Covariate	Mean	lo[a]	up[b]
mixture probabilities	π_1	No Fraud	.733	.677	.800
	π_2	Incremental Fraud	.265	.199	.322
	π_3	Extreme Fraud	.00169	.000436	.00303
turnout	γ_1	imputed electors	.995	$-.384$	2.14
vote choice	β_1	imputed electors	$-.261$	$-.412$	.0413
incremental frauds	ρ_{M0}	(Intercept)	$-.0837$	$-.347$	.338
	ρ_{S0}	(Intercept)	$-.732$	$-.817$	$-.623$
extreme frauds	δ_{M0}	(Intercept)	$-.318$	-1.00	.242
	δ_{S0}	(Intercept)	$-.675$	-1.26	$-.153$

MCMC posterior multimodality diagnostics:
dip test p-values $D(\pi_1) = 0$; $D(\pi_2) = 0$; $D(\pi_3) = .929$.[c]
means difference $M(\pi_1) = .110$; $M(\pi_2) = .110$; $M(\pi_3) = .00157$.[d]

units `eforensics`-fraudulent: (1811 incremental, 9 extreme, 7337 not fraudulent)

manufactured votes	$F_t = 111917.4$ $[84106.6, 136932.4]$[e]
incremental manufactured	$F_t = 111088.4$ $[83441.8, 135732.8]$[e]
extreme manufactured	$F_t = 829.1$ $[546.7, 1502.1]$[e]
total `eforensics`-fraudulent votes	$F_w = 225440.2$ $[207757.1, 252978.1]$[e]
incremental total	$F_w = 223652.3$ $[205683.8, 251653.7]$[e]
extreme total	$F_w = 1787.8$ $[1248.0, 2201.5]$[e]

Note: selected `eforensics` model parameter estimates (posterior means and credible intervals). County fixed effects for turnout and vote choice are not shown. $n = 9157$ precinct units. Electors, votes cast and votes for the leader: $\sum_{i=1}^{n} N_i = 9173772$; $\sum_{i=1}^{n} V_i = 7040360$; $\sum_{i=1}^{n} W_i = 3543308$. [a] 95% HPD lower bound. [b] 95% HPD upper bound. [c] dip test for unimodality null hypothesis over all MCMC chains. [d] difference between largest and smallest chain-specific posterior means. [e] posterior mean [99.5% credible interval].



that the intercept for the incremental manufactured frauds magnitudes lacks a definite sign—$\rho_{M0} = -.0837$ $(-.347, .338)$—inductively suggests that the incremental manufactured votes, $F_t = 111088.4$ $[83441.8, 135732.8]$, very likely are produced from malevolent distortions of electors' intentions. The frauds magnitudes intercept for the incremental stolen votes is negative, with incremental stolen votes having a posterior mean of $F_w - F_t = 223652.3 - 111088.4 = 112563.9$. Again drawing on German elections, with a negative frauds magnitudes intercept the incremental stolen votes can be interpreted as ambiguous, likely being unknown admixtures of malevolent distortions and electors'

4

Model contradicts basic features of voting behavior.

E.g., implies that absent fraud, precinct vote shares for the winner should be unimodal in every jurisdiction.

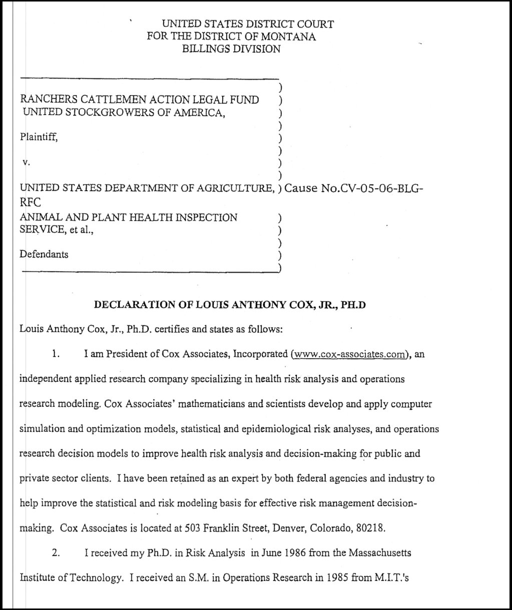
Application 2: risk of importing BSE from Canada, 2003.

5/2003: BSE-infected beef cow found in Alberta, Canada.

USDA banned imports of Canadian cattle and beef in response.

11/2003: USDA proposed allowing beef and cattle less than 30 months old to be imported, w/ restrictions on slaughter and disposal of distal ileum.

Ranchers-Cattlemen Action Legal Fund (R-CALF) sued to keep Canadian cattle out.



UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MONTANA
BILLINGS DIVISION

RANCHERS CATTLEMEN ACTION LEGAL FUND)
UNITED STOCKGROWERS OF AMERICA,)

Plaintiff,)

v.)

UNITED STATES DEPARTMENT OF AGRICULTURE,) Cause No.CV-05-06-BLG-RFC
ANIMAL AND PLANT HEALTH INSPECTION)
SERVICE, et al.,)

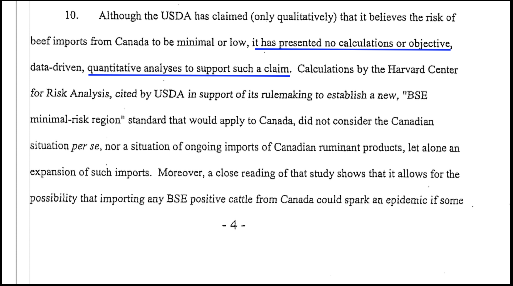
Defendants)

DECLARATION OF LOUIS ANTHONY COX, JR., PH.D

Louis Anthony Cox, Jr., Ph.D. certifies and states as follows:

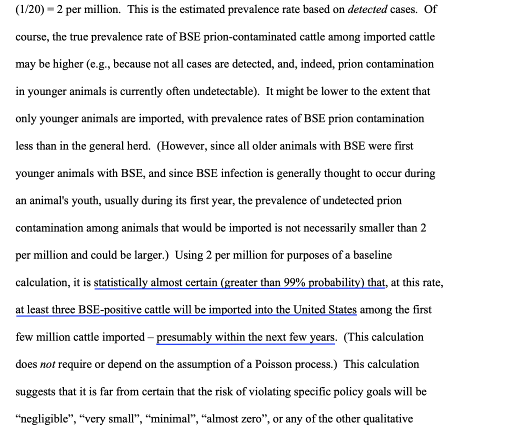
1. I am President of Cox Associates, Incorporated (www.cox-associates.com), an independent applied research company specializing in health risk analysis and operations research modeling. Cox Associates' mathematicians and scientists develop and apply computer simulation and optimization models, statistical and epidemiological risk analyses, and operations research decision models to improve health risk analysis and decision-making for public and private sector clients. I have been retained as an expert by both federal agencies and industry to help improve the statistical and risk modeling basis for effective risk management decision-making. Cox Associates is located at 503 Franklin Street, Denver, Colorado, 80218.

2. I received my Ph.D. in Risk Analysis in June 1986 from the Massachusetts Institute of Technology. I received an S.M. in Operations Research in 1985 from M.I.T.'s



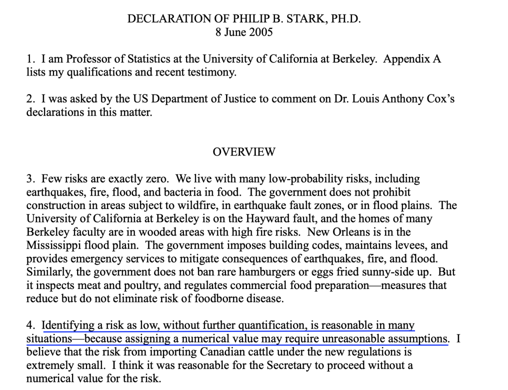
10. Although the USDA has claimed (only qualitatively) that it believes the risk of beef imports from Canada to be minimal or low, it has presented no calculations or objective, data-driven, quantitative analyses to support such a claim. Calculations by the Harvard Center for Risk Analysis, cited by USDA in support of its rulemaking to establish a new, "BSE minimal-risk region" standard that would apply to Canada, did not consider the Canadian situation *per se*, nor a situation of ongoing imports of Canadian ruminant products, let alone an expansion of such imports. Moreover, a close reading of that study shows that it allows for the possibility that importing any BSE positive cattle from Canada could spark an epidemic if some

- 4 -



(1/20) = 2 per million. This is the estimated prevalence rate based on *detected* cases. Of course, the true prevalence rate of BSE prion-contaminated cattle among imported cattle may be higher (e.g., because not all cases are detected, and, indeed, prion contamination in younger animals is currently often undetectable). It might be lower to the extent that only younger animals are imported, with prevalence rates of BSE prion contamination less than in the general herd. (However, since all older animals with BSE were first younger animals with BSE, and since BSE infection is generally thought to occur during an animal's youth, usually during its first year, the prevalence of undetected prion contamination among animals that would be imported is not necessarily smaller than 2 per million and could be larger.) Using 2 per million for purposes of a baseline calculation, it is statistically almost certain (greater than 99% probability) that, at this rate, at least three BSE-positive cattle will be imported into the United States among the first few million cattle imported – presumably within the next few years. (This calculation does *not* require or depend on the assumption of a Poisson process.) This calculation suggests that it is far from certain that the risk of violating specific policy goals will be "negligible", "very small", "minimal", "almost zero", or any of the other qualitative

- Whether cows have BSE is IID Bernoulli.
- Count of imported infected cows is Poisson.
- Whether human gets vCJD from infected beef is IID Bernoulli.
- Ignored biology & epidemiology of BSE and vCJD: feed bans, age, concentration of prions.



DECLARATION OF PHILIP B. STARK, PH.D.
8 June 2005

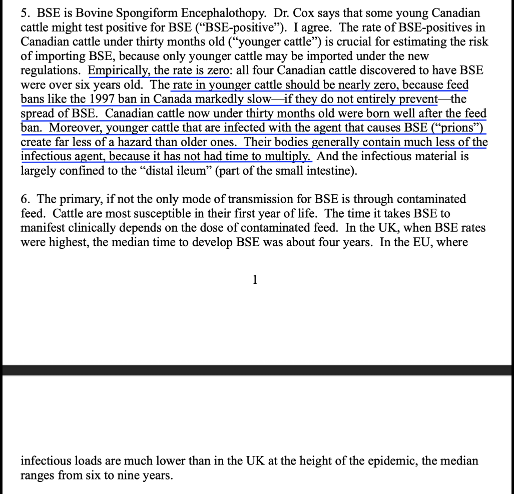
1. I am Professor of Statistics at the University of California at Berkeley. Appendix A lists my qualifications and recent testimony.

2. I was asked by the US Department of Justice to comment on Dr. Louis Anthony Cox's declarations in this matter.

OVERVIEW

3. Few risks are exactly zero. We live with many low-probability risks, including earthquakes, fire, flood, and bacteria in food. The government does not prohibit construction in areas subject to wildfire, in earthquake fault zones, or in flood plains. The University of California at Berkeley is on the Hayward fault, and the homes of many Berkeley faculty are in wooded areas with high fire risks. New Orleans is in the Mississippi flood plain. The government imposes building codes, maintains levees, and provides emergency services to mitigate consequences of earthquakes, fire, and flood. Similarly, the government does not ban rare hamburgers or eggs fried sunny-side up. But it inspects meat and poultry, and regulates commercial food preparation—measures that reduce but do not eliminate risk of foodborne disease.

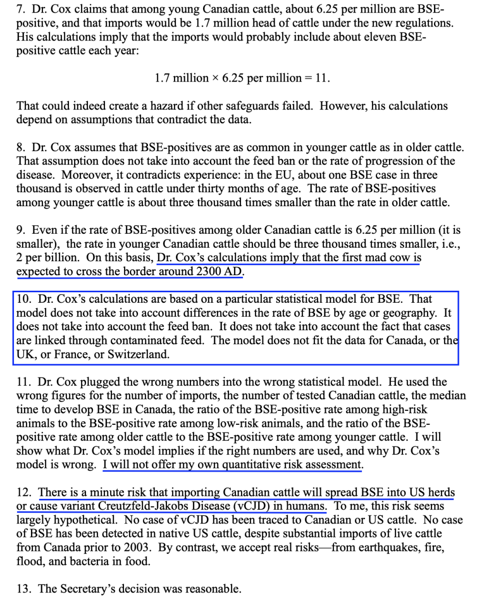
4. Identifying a risk as low, without further quantification, is reasonable in many situations—because assigning a numerical value may require unreasonable assumptions. I believe that the risk from importing Canadian cattle under the new regulations is extremely small. I think it was reasonable for the Secretary to proceed without a numerical value for the risk.



5. BSE is Bovine Spongiform Encephalothopy. Dr. Cox says that some young Canadian cattle might test positive for BSE ("BSE-positive"). I agree. The rate of BSE-positives in Canadian cattle under thirty months old ("younger cattle") is crucial for estimating the risk of importing BSE, because only younger cattle may be imported under the new regulations. Empirically, the rate is zero: all four Canadian cattle discovered to have BSE were over six years old. The rate in younger cattle should be nearly zero, because feed bans like the 1997 ban in Canada markedly slow—if they do not entirely prevent—the spread of BSE. Canadian cattle now under thirty months old were born well after the feed ban. Moreover, younger cattle that are infected with the agent that causes BSE ("prions") create far less of a hazard than older ones. Their bodies generally contain much less of the infectious agent, because it has not had time to multiply. And the infectious material is largely confined to the "distal ileum" (part of the small intestine).

6. The primary, if not the only mode of transmission for BSE is through contaminated feed. Cattle are most susceptible in their first year of life. The time it takes BSE to manifest clinically depends on the dose of contaminated feed. In the UK, when BSE rates were highest, the median time to develop BSE was about four years. In the EU, where

infectious loads are much lower than in the UK at the height of the epidemic, the median ranges from six to nine years.



7. Dr. Cox claims that among young Canadian cattle, about 6.25 per million are BSE-positive, and that imports would be 1.7 million head of cattle under the new regulations. His calculations imply that the imports would probably include about eleven BSE-positive cattle each year:

$$1.7 \text{ million} \times 6.25 \text{ per million} = 11.$$

That could indeed create a hazard if other safeguards failed. However, his calculations depend on assumptions that contradict the data.

8. Dr. Cox assumes that BSE-positives are as common in younger cattle as in older cattle. That assumption does not take into account the feed ban or the rate of progression of the disease. Moreover, it contradicts experience: in the EU, about one BSE case in three thousand is observed in cattle under thirty months of age. The rate of BSE-positives among younger cattle is about three thousand times smaller than the rate in older cattle.

9. Even if the rate of BSE-positives among older Canadian cattle is 6.25 per million (it is smaller), the rate in younger Canadian cattle should be three thousand times smaller, i.e., 2 per billion. On this basis, Dr. Cox's calculations imply that the first mad cow is expected to cross the border around 2300 AD.

10. Dr. Cox's calculations are based on a particular statistical model for BSE. That model does not take into account differences in the rate of BSE by age or geography. It does not take into account the feed ban. It does not take into account the fact that cases are linked through contaminated feed. The model does not fit the data for Canada, or the UK, or France, or Switzerland.

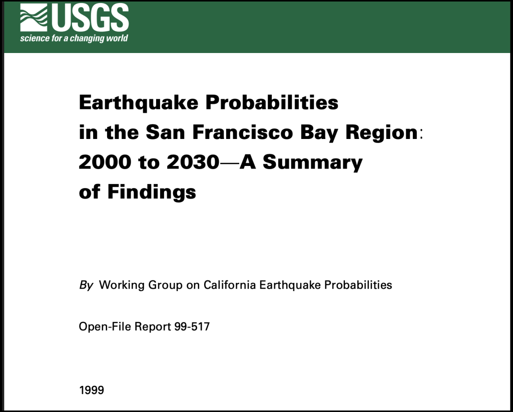
11. Dr. Cox plugged the wrong numbers into the wrong statistical model. He used the wrong figures for the number of imports, the number of tested Canadian cattle, the median time to develop BSE in Canada, the ratio of the BSE-positive rate among high-risk animals to the BSE-positive rate among low-risk animals, and the ratio of the BSE-positive rate among older cattle to the BSE-positive rate among younger cattle. I will show what Dr. Cox's model implies if the right numbers are used, and why Dr. Cox's model is wrong. I will not offer my own quantitative risk assessment.

12. There is a minute risk that importing Canadian cattle will spread BSE into US herds or cause variant Creutzfeld-Jakobs Disease (vCJD) in humans. To me, this risk seems largely hypothetical. No case of vCJD has been traced to Canadian or US cattle. No case of BSE has been detected in native US cattle, despite substantial imports of live cattle from Canada prior to 2003. By contrast, we accept real risks—from earthquakes, fire, flood, and bacteria in food.

13. The Secretary's decision was reasonable.

- R-CALF lost; imports resumed.
- No cows imported from Canada found infected in the intervening 22 years.
- Only "atypical" (spontaneous–unrelated to feed or epidemic) BSE found in US cattle since 2003.

Application 3: chance of a large earthquake in the Bay Area.



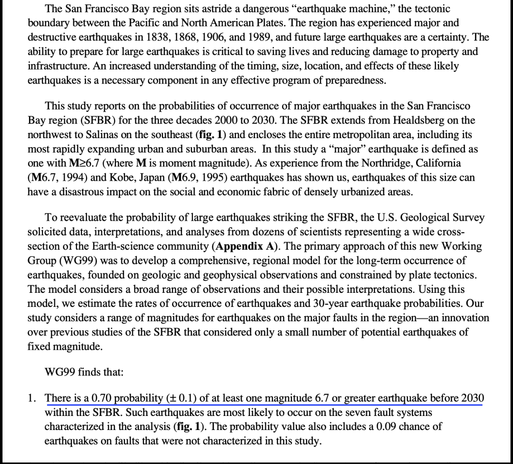
USGS
science for a changing world

Earthquake Probabilities in the San Francisco Bay Region: 2000 to 2030—A Summary of Findings

By Working Group on California Earthquake Probabilities

Open-File Report 99-517

1999



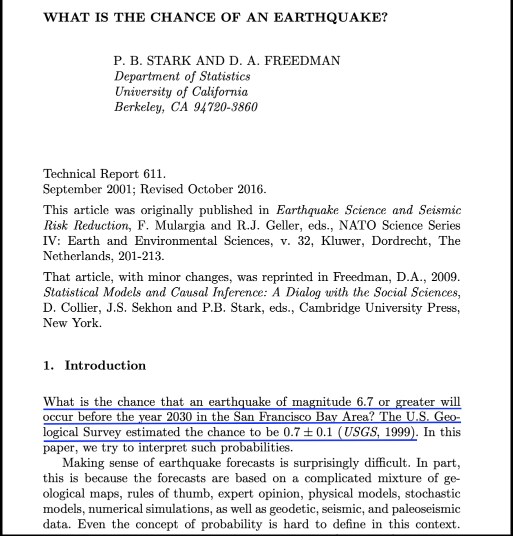
The San Francisco Bay region sits astride a dangerous "earthquake machine," the tectonic boundary between the Pacific and North American Plates. The region has experienced major and destructive earthquakes in 1838, 1868, 1906, and 1989, and future large earthquakes are a certainty. The ability to prepare for large earthquakes is critical to saving lives and reducing damage to property and infrastructure. An increased understanding of the timing, size, location, and effects of these likely earthquakes is a necessary component in any effective program of preparedness.

This study reports on the probabilities of occurrence of major earthquakes in the San Francisco Bay region (SFBR) for the three decades 2000 to 2030. The SFBR extends from Healdsberg on the northwest to Salinas on the southeast (**fig. 1**) and encloses the entire metropolitan area, including its most rapidly expanding urban and suburban areas. In this study a "major" earthquake is defined as one with **M**≥6.7 (where **M** is moment magnitude). As experience from the Northridge, California (**M**6.7, 1994) and Kobe, Japan (**M**6.9, 1995) earthquakes has shown us, earthquakes of this size can have a disastrous impact on the social and economic fabric of densely urbanized areas.

To reevaluate the probability of large earthquakes striking the SFBR, the U.S. Geological Survey solicited data, interpretations, and analyses from dozens of scientists representing a wide cross-section of the Earth-science community (**Appendix A**). The primary approach of this new Working Group (WG99) was to develop a comprehensive, regional model for the long-term occurrence of earthquakes, founded on geologic and geophysical observations and constrained by plate tectonics. The model considers a broad range of observations and their possible interpretations. Using this model, we estimate the rates of occurrence of earthquakes and 30-year earthquake probabilities. Our study considers a range of magnitudes for earthquakes on the major faults in the region—an innovation over previous studies of the SFBR that considered only a small number of potential earthquakes of fixed magnitude.

WG99 finds that:

1. There is a 0.70 probability (± 0.1) of at least one magnitude 6.7 or greater earthquake before 2030 within the SFBR. Such earthquakes are most likely to occur on the seven fault systems characterized in the analysis (**fig. 1**). The probability value also includes a 0.09 chance of earthquakes on faults that were not characterized in this study.



WHAT IS THE CHANCE OF AN EARTHQUAKE?

P. B. STARK AND D. A. FREEDMAN
Department of Statistics
University of California
Berkeley, CA 94720-3860

Technical Report 611.
September 2001; Revised October 2016.

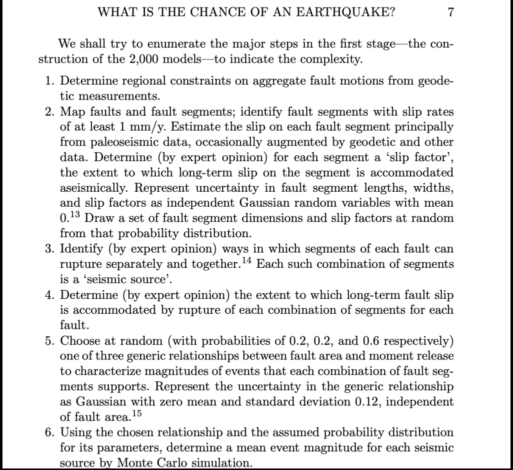
This article was originally published in *Earthquake Science and Seismic Risk Reduction*, F. Mulargia and R.J. Geller, eds., NATO Science Series IV: Earth and Environmental Sciences, v. 32, Kluwer, Dordrecht, The Netherlands, 201-213.

That article, with minor changes, was reprinted in Freedman, D.A., 2009. *Statistical Models and Causal Inference: A Dialog with the Social Sciences*, D. Collier, J.S. Sekhon and P.B. Stark, eds., Cambridge University Press, New York.

1. Introduction

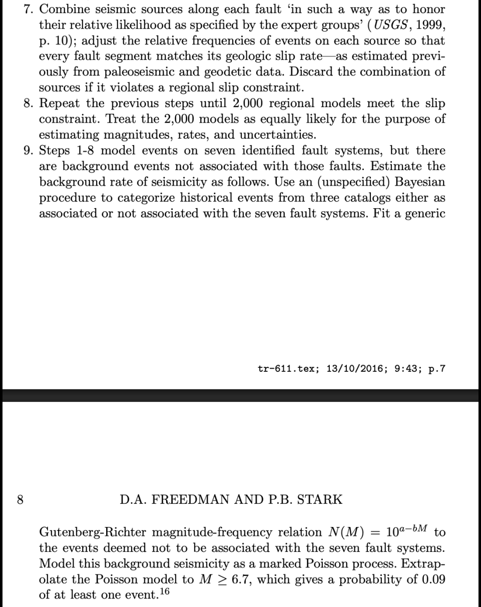
What is the chance that an earthquake of magnitude 6.7 or greater will occur before the year 2030 in the San Francisco Bay Area? The U.S. Geological Survey estimated the chance to be 0.7 ± 0.1 (*USGS*, 1999). In this paper, we try to interpret such probabilities.

Making sense of earthquake forecasts is surprisingly difficult. In part, this is because the forecasts are based on a complicated mixture of geological maps, rules of thumb, expert opinion, physical models, stochastic models, numerical simulations, as well as geodetic, seismic, and paleoseismic data. Even the concept of probability is hard to define in this context.



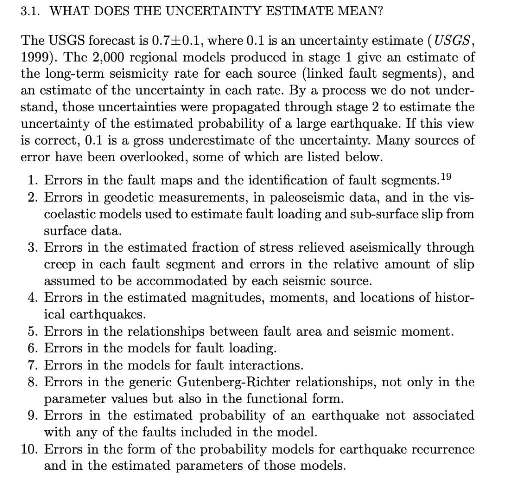
We shall try to enumerate the major steps in the first stage—the construction of the 2,000 models—to indicate the complexity.

1. Determine regional constraints on aggregate fault motions from geodetic measurements.
2. Map faults and fault segments; identify fault segments with slip rates of at least 1 mm/y. Estimate the slip on each fault segment principally from paleoseismic data, occasionally augmented by geodetic and other data. Determine (by expert opinion) for each segment a 'slip factor', the extent to which long-term slip on the segment is accommodated aseismically. Represent uncertainty in fault segment lengths, widths, and slip factors as independent Gaussian random variables with mean 0.[13] Draw a set of fault segment dimensions and slip factors at random from that probability distribution.
3. Identify (by expert opinion) ways in which segments of each fault can rupture separately and together.[14] Each such combination of segments is a 'seismic source'.
4. Determine (by expert opinion) the extent to which long-term fault slip is accommodated by rupture of each combination of segments for each fault.
5. Choose at random (with probabilities of 0.2, 0.2, and 0.6 respectively) one of three generic relationships between fault area and moment release to characterize magnitudes of events that each combination of fault segments supports. Represent the uncertainty in the generic relationship as Gaussian with zero mean and standard deviation 0.12, independent of fault area.[15]
6. Using the chosen relationship and the assumed probability distribution for its parameters, determine a mean event magnitude for each seismic source by Monte Carlo simulation.



7. Combine seismic sources along each fault 'in such a way as to honor their relative likelihood as specified by the expert groups' (*USGS*, 1999, p. 10); adjust the relative frequencies of events on each source so that every fault segment matches its geologic slip rate—as estimated previously from paleoseismic and geodetic data. Discard the combination of sources if it violates a regional slip constraint.
8. Repeat the previous steps until 2,000 regional models meet the slip constraint. Treat the 2,000 models as equally likely for the purpose of estimating magnitudes, rates, and uncertainties.
9. Steps 1-8 model events on seven identified fault systems, but there are background events not associated with those faults. Estimate the background rate of seismicity as follows. Use an (unspecified) Bayesian procedure to categorize historical events from three catalogs either as associated or not associated with the seven fault systems. Fit a generic

Gutenberg-Richter magnitude-frequency relation $N(M) = 10^{a-bM}$ to the events deemed not to be associated with the seven fault systems. Model this background seismicity as a marked Poisson process. Extrapolate the Poisson model to $M \geq 6.7$, which gives a probability of 0.09 of at least one event.[16]



3.1. WHAT DOES THE UNCERTAINTY ESTIMATE MEAN?

The USGS forecast is 0.7±0.1, where 0.1 is an uncertainty estimate (*USGS*, 1999). The 2,000 regional models produced in stage 1 give an estimate of the long-term seismicity rate for each source (linked fault segments), and an estimate of the uncertainty in each rate. By a process we do not understand, those uncertainties were propagated through stage 2 to estimate the uncertainty of the estimated probability of a large earthquake. If this view is correct, 0.1 is a gross underestimate of the uncertainty. Many sources of error have been overlooked, some of which are listed below.

1. Errors in the fault maps and the identification of fault segments.[19]
2. Errors in geodetic measurements, in paleoseismic data, and in the viscoelastic models used to estimate fault loading and sub-surface slip from surface data.
3. Errors in the estimated fraction of stress relieved aseismically through creep in each fault segment and errors in the relative amount of slip assumed to be accommodated by each seismic source.
4. Errors in the estimated magnitudes, moments, and locations of historical earthquakes.
5. Errors in the relationships between fault area and seismic moment.
6. Errors in the models for fault loading.
7. Errors in the models for fault interactions.
8. Errors in the generic Gutenberg-Richter relationships, not only in the parameter values but also in the functional form.
9. Errors in the estimated probability of an earthquake not associated with any of the faults included in the model.
10. Errors in the form of the probability models for earthquake recurrence and in the estimated parameters of those models.

Rates are not probabilities: Hydrology, Klemeš (1989)

The automatic identification of past frequencies with present probabilities is the greatest plague of contemporary statistical and stochastic hydrology. It has become so deeply engrained that it prevents hydrologists from seeing the fundamental difference between the two concepts. It is often difficult to put across the fact that whereas a histogram of frequencies for given quantities . . . can be constructed for any function whether it has been generated by deterministic or random mechanism, it can be interpreted as a probability distribution only in the latter case. . . . Ergo, automatically to interpret past frequencies as present probabilities means a priori to deny the possibility of any signal in the geophysical history; this certainly is not science but sterile scholasticism.

The point then arises, why are these unreasonable assumptions made if it is obvious that probabilistic statements based on them may be grossly misleading, especially when they relate to physically extreme conditions where errors can have catastrophic consequences? The answer seems to be that they provide the only conceptual framework that makes it possible to make probabilistic statements, i.e. they must be used if the objective is to make such probabilistic statements.

- In a sequence of random trials with chance p of success in each, the empirical rate of success is an unbiased estimate of p. Under additional conditions (pairwise independence or exchangeability, for instance), the rate converges almost surely to p.
- *But the fact that something has a rate doesn't mean a random process produced it.*

- In a sequence of random trials with chance p of success in each, the empirical rate of success is an unbiased estimate of p. Under additional conditions (pairwise independence or exchangeability, for instance), the rate converges almost surely to p.

- *But the fact that something has a rate doesn't mean a random process produced it.*

Two thought experiments:

1. You are in a group of 100 people. You are told that one person in the group will die next year. What is the chance it is you?

- In a sequence of random trials with chance p of success in each, the empirical rate of success is an unbiased estimate of p. Under additional conditions (pairwise independence or exchangeability, for instance), the rate converges almost surely to p.

- *But the fact that something has a rate doesn't mean a random process produced it.*

Two thought experiments:

1. You are in a group of 100 people. You are told that one person in the group will die next year. What is the chance it is you?

2. You are in a group of 100 people. You are told that one of them is named Philip. What is the chance it is you?

Both involve a rate of 1% in a group.

Probability is in the method of selection, not the existence of a rate

Possible selection rules:

- Draw lots and shoot whoever gets the short straw
  - *might be* reasonably modeled as random and uniform
  - if so, the probability you die is 1%.

Probability is in the method of selection, not the existence of a rate

Possible selection rules:

- Draw lots and shoot whoever gets the short straw
  - *might be* reasonably modeled as random and uniform
  - if so, the probability you die is 1%.
- Shoot the tallest person
  - no probability
  - you are or aren't the tallest person

How does probability enter a scientific problem?

- underlying physical phenomenon is random (per theory), e.g. radioactive decay and other quantum effects

How does probability enter a scientific problem?

- underlying physical phenomenon is random (per theory), e.g. radioactive decay and other quantum effects
- scientist deliberately introduces randomness, e.g., by randomizing treatment assignments or drawing a random sample

How does probability enter a scientific problem?

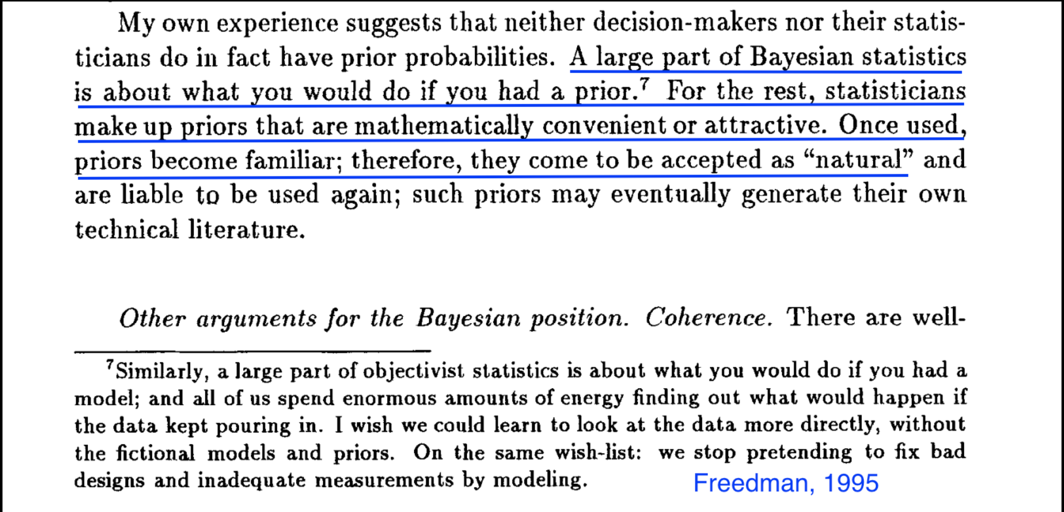
- underlying physical phenomenon is random (per theory), e.g. radioactive decay and other quantum effects
- scientist deliberately introduces randomness, e.g., by randomizing treatment assignments or drawing a random sample
- *subjective prior probability* for unknowns

How does probability enter a scientific problem?

- underlying physical phenomenon is random (per theory), e.g. radioactive decay and other quantum effects
- scientist deliberately introduces randomness, e.g., by randomizing treatment assignments or drawing a random sample
- *subjective prior probability* for unknowns
- *probability model* that is supposed to describe a phenomenon, e.g., a regression model, a Gaussian process model, or a stochastic PDE.
    - In what sense, to what level of accuracy, and for what purpose?

How does probability enter a scientific problem?

- underlying physical phenomenon is random (per theory), e.g. radioactive decay and other quantum effects
- scientist deliberately introduces randomness, e.g., by randomizing treatment assignments or drawing a random sample
- *subjective prior probability* for unknowns
- *probability model* that is supposed to describe a phenomenon, e.g., a regression model, a Gaussian process model, or a stochastic PDE.
    - In what sense, to what level of accuracy, and for what purpose?
- *metaphor*: claim the phenomenon in question behaves ‘as if’ random



My own experience suggests that neither decision-makers nor their statisticians do in fact have prior probabilities. A large part of Bayesian statistics is about what you would do if you had a prior.[7] For the rest, statisticians make up priors that are mathematically convenient or attractive. Once used, priors become familiar; therefore, they come to be accepted as "natural" and are liable to be used again; such priors may eventually generate their own technical literature.

Other arguments for the Bayesian position. Coherence. There are well-

[7] Similarly, a large part of objectivist statistics is about what you would do if you had a model; and all of us spend enormous amounts of energy finding out what would happen if the data kept pouring in. I wish we could learn to look at the data more directly, without the fictional models and priors. On the same wish-list: we stop pretending to fix bad designs and inadequate measurements by modeling.

Freedman, 1995

Simulation and probability

In physics, geophysics, climate science, sensitivity analysis, and uncertainty quantification, popular belief that probabilities can be estimated in a ‘neutral’ or ‘automatic’ way by Monte Carlo simulation: just let the computer generate the distribution.

- Simulation estimates numerical values from an *assumed* distribution.
- Simulations are transformations of assumptions–not new information.
- Substitute for an integral, not a way to uncover laws of Nature.

Application 4: Probabilistic Seismic Hazard Assessment (PSHA)

Cornell (1968):

> *In this paper a method is developed to produce [various characteristics of ground motion] and their average return period . . . The minimum data needed are only the seismologist's best estimates of the average activity levels of the various potential sources of earthquakes . . . The technique to be developed provides the method for integrating the individual influences of potential earthquake sources, near and far, more active or less, into the probability distribution of maximum annual intensity (or peak-ground acceleration, etc.). The average return period follows directly.*
>
> *. . .*
>
> *In general the size and location of a future earthquake are uncertain. They shall be treated therefore as random variables.*

- basis of seismic building codes in many countries; used to pick sites for nuclear power plants and nuclear waste disposal
- claims to find the probability of a given level of ground shaking
- models earthquakes as occurring at random in space, time and with random magnitude (marked point process)
- models ground motion assumed to be random w/ known distribution given quake occurs
- treats Gutenberg-Richter (G-R) law, the historical spatial distribution of seismicity, and ground acceleration given the distance and magnitude of an earthquake as probability distributions
- that earthquakes occur at random is an *assumption*, not a matter of physics–the physics is not understood

- hinges on *metaphor* that earthquakes occur *as if* in a casino game
  - as if there is a special deck of cards
  - game involves dealing one card per time period
  - If the card is blank, no earthquake.
  - If the card is 8, magnitude 8 earthquake. Etc.
- tens of thousands of journal pages arguing about how many cards of each kind are in the deck, how well the deck is shuffled, whether after each draw you replace the card in the deck and shuffle again before dealing the next card, whether you add high-numbered cards to the deck if no high card has been drawn in a while, etc.
- many destructive earthquakes occurred where PSHA says risk is small

A different metaphor: earthquakes are like terrorist bombings

- don’t know when or where they’re going to happen or how big they will be
- do know they could hurt people
- some places are more common targets than others (e.g., places near active faults)
- some places are more vulnerable to damage
- but no probability *per se*

Application 5: Avian-Turbine Interactions.

Wind turbine generators occasionally kill birds, including raptors.

- How many?
- What species?
- What design and siting features of the wind turbines matter?
- Can you design turbines or wind farms in a way that reduces avian mortality?
- What design changes would help?

- Raptors are rare; raptor collisions with wind turbines are rarer.
- Data: look for pieces of birds near the turbines.
  - background mortality
  - find bird fragments, not birds
  - carcasses decompose
  - scavengers
  - birds may land far from the turbine they hit.

- Consultant modelled bird collisions as zero-inflated Poisson regression.
  - collisions are random and independent
  - probability same for all birds
  - rate follows a hierarchical Bayesian model
  - parameters for design and location of turbine
  - additional smoothing to make parameters identifiable
  - estimated the coefficients
- According to the model, when a bird approaches a turbine, it tosses a biased coin.
  - heads, the bird hits turbine; tails not
  - for each turbine location and design, every bird uses a coin with the same chance of heads
  - P(heads) is a known parametric function of turbine design/site features
  - birds toss coins independently

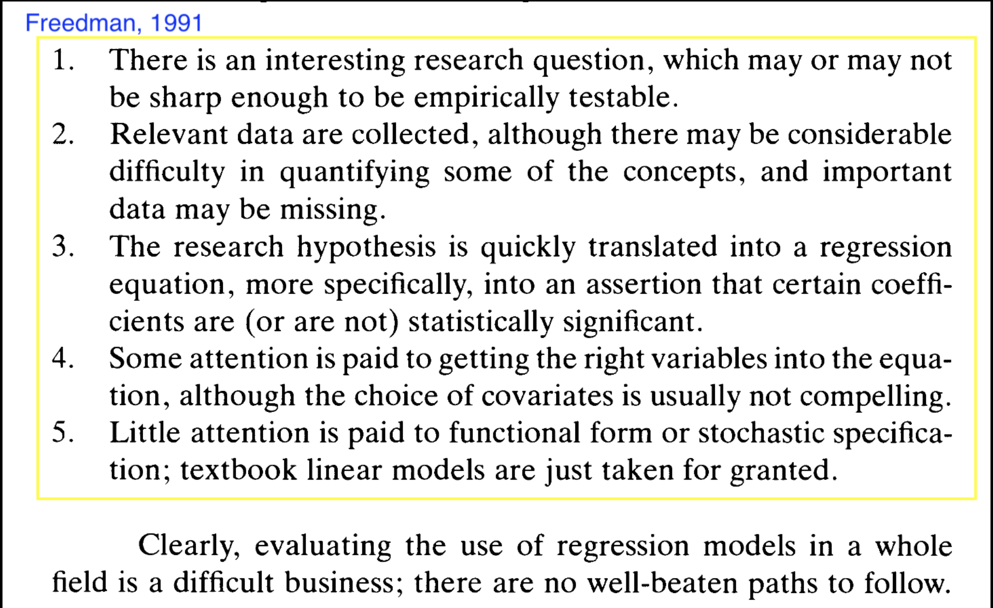
Displacement

Changes subject from “how many birds does this turbine kill?” to “what are the numerical values of some coefficients in a zero-inflated Poisson regression model?”

Type III error: testing a statistical model with little connection to the scientific question.

Rayner (2012, p. 120):

> *Displacement is the term that I use to describe the process by which* ***an object or activity, such as a computer model, designed to inform management of a real-world phenomenon actually becomes the object of management****. Displacement is more subtle than diversion in that it does not merely distract attention away from an area that might otherwise generate uncomfortable knowledge by pointing in another direction, which is the mechanism of distraction, but substitutes a more manageable surrogate. The inspiration for recognizing displacement can be traced to A. N. Whitehead’s fallacy of misplaced concreteness, ‘the accidental error of mistaking the abstract for the concrete.’*

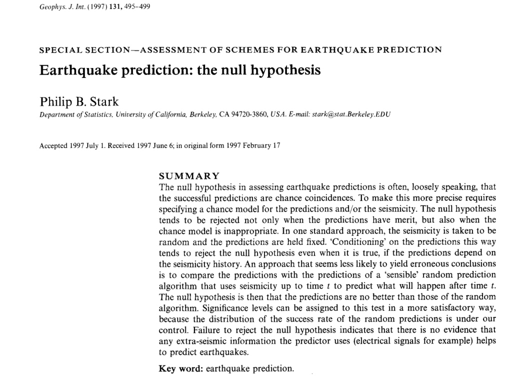


Freedman, 1991

1. There is an interesting research question, which may or may not be sharp enough to be empirically testable.
2. Relevant data are collected, although there may be considerable difficulty in quantifying some of the concepts, and important data may be missing.
3. The research hypothesis is quickly translated into a regression equation, more specifically, into an assertion that certain coefficients are (or are not) statistically significant.
4. Some attention is paid to getting the right variables into the equation, although the choice of covariates is usually not compelling.
5. Little attention is paid to functional form or stochastic specification; textbook linear models are just taken for granted.

Clearly, evaluating the use of regression models in a whole field is a difficult business; there are no well-beaten paths to follow.

Application 6: Testing earthquake predictions



Geophys. J. Int. (1997) **131,** 495–499

SPECIAL SECTION—ASSESSMENT OF SCHEMES FOR EARTHQUAKE PREDICTION

Earthquake prediction: the null hypothesis

Philip B. Stark

Department of Statistics, University of California, Berkeley, CA 94720-3860, *USA. E-mail: stark@stat.Berkeley.EDU*

Accepted 1997 July 1. Received 1997 June 6; in original form 1997 February 17

SUMMARY

The null hypothesis in assessing earthquake predictions is often, loosely speaking, that the successful predictions are chance coincidences. To make this more precise requires specifying a chance model for the predictions and/or the seismicity. The null hypothesis tends to be rejected not only when the predictions have merit, but also when the chance model is inappropriate. In one standard approach, the seismicity is taken to be random and the predictions are held fixed. 'Conditioning' on the predictions this way tends to reject the null hypothesis even when it is true, if the predictions depend on the seismicity history. An approach that seems less likely to yield erroneous conclusions is to compare the predictions with the predictions of a 'sensible' random prediction algorithm that uses seismicity up to time *t* to predict what will happen after time *t*. The null hypothesis is then that the predictions are no better than those of the random algorithm. Significance levels can be assigned to this test in a more satisfactory way, because the distribution of the success rate of the random predictions is under our control. Failure to reject the null hypothesis indicates that there is no evidence that any extra-seismic information the predictor uses (electrical signals for example) helps to predict earthquakes.

Key word: earthquake prediction.

- are predictions better for real seismicity than for simulations?
- what stochastic model for seismicity?
- basic feature of seismicity is *clustering* in time and space
- if seismicity were different, predictions would be different: conditioning on predictions while randomizing times answers the wrong question.
- Type III error

Application 7: Student's T-test in RCTs

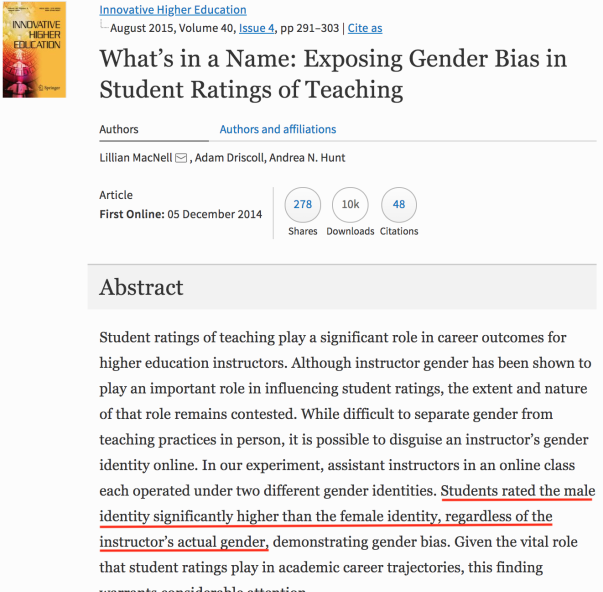
Scientific null: treatment does not affect the outcome, either subject-by-subject (the *strong null*) or on average (the *weak null*).

Statistical null: all N responses are IID Gaussian (same mean & variance).

- In experiment, the treatment and control groups are dependent: random partition of single group.
- In statistical null, the groups are independent.
- In the experiment, the only source of randomness is the random allocation to treatment or control, and nothing is known about distribution of responses.
- In the statistical null, subjects' responses are random and Gaussian.

Type III error (but asymptotically makes the same decisions in some situations).

Application 8: Gender bias in teaching evaluations



Innovative Higher Education

August 2015, Volume 40, Issue 4, pp 291–303 | Cite as

What's in a Name: Exposing Gender Bias in Student Ratings of Teaching

Authors Authors and affiliations

Lillian MacNell, Adam Driscoll, Andrea N. Hunt

Article

First Online: 05 December 2014

278 Shares 10k Downloads 48 Citations

Abstract

Student ratings of teaching play a significant role in career outcomes for higher education instructors. Although instructor gender has been shown to play an important role in influencing student ratings, the extent and nature of that role remains contested. While difficult to separate gender from teaching practices in person, it is possible to disguise an instructor's gender identity online. In our experiment, assistant instructors in an online class each operated under two different gender identities. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender, demonstrating gender bias. Given the vital role that student ratings play in academic career trajectories, this finding warrants considerable attention.

MacNell, Driscoll, & Hunt, 2014

NC State online course.

Students randomized into 6 groups, 2 taught by primary prof, 4 by GSIs.

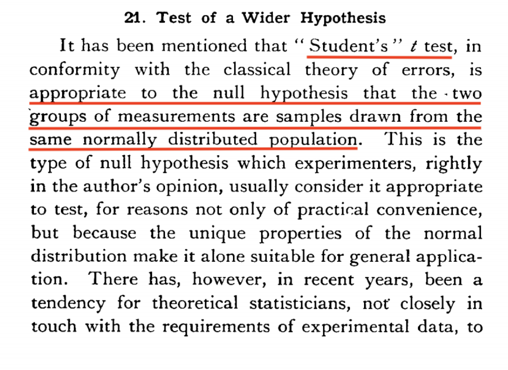
2 GSIs: 1 male, 1 female.

GSIs used actual names in 1 section, swapped names in 1 section.

Ratings on 5-point scale

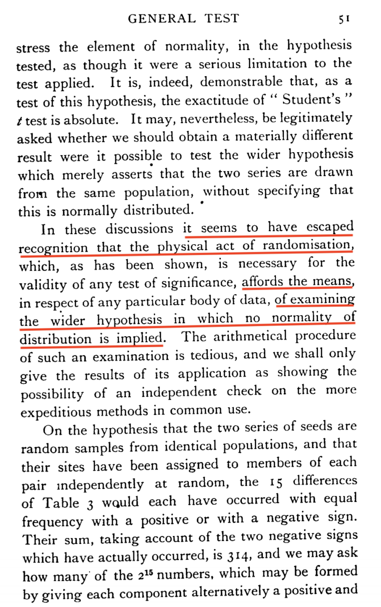
Characteristic	M - F	permutation t-test	parametric t-test
Overall	0.47	0.12	0.128
Professional	0.61	0.07	0.124
Respectful	0.61	0.06	0.124
Caring	0.52	0.10	0.071
Enthusiastic	0.57	0.06	0.112
Communicate	0.57	0.07	NA
Helpful	0.46	0.17	0.049
Feedback	0.47	0.16	0.054
Prompt	0.80	0.01	0.191
Consistent	0.46	0.21	0.045
Fair	0.76	0.01	0.188
Responsive	0.22	0.48	0.013
Praise	0.67	0.01	0.153
Knowledge	0.35	0.29	0.038
Clear	0.41	0.29	NA

R.A. Fisher, 1935



21. Test of a Wider Hypothesis

It has been mentioned that "Student's" *t* test, in conformity with the classical theory of errors, is appropriate to the null hypothesis that the two groups of measurements are samples drawn from the same normally distributed population. This is the type of null hypothesis which experimenters, rightly in the author's opinion, usually consider it appropriate to test, for reasons not only of practical convenience, but because the unique properties of the normal distribution make it alone suitable for general application. There has, however, in recent years, been a tendency for theoretical statisticians, not closely in touch with the requirements of experimental data, to



stress the element of normality, in the hypothesis tested, as though it were a serious limitation to the test applied. It is, indeed, demonstrable that, as a test of this hypothesis, the exactitude of "Student's" *t* test is absolute. It may, nevertheless, be legitimately asked whether we should obtain a materially different result were it possible to test the wider hypothesis which merely asserts that the two series are drawn from the same population, without specifying that this is normally distributed.

In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. The arithmetical procedure of such an examination is tedious, and we shall only give the results of its application as showing the possibility of an independent check on the more expeditious methods in common use.

On the hypothesis that the two series of seeds are random samples from identical populations, and that their sites have been assigned to members of each pair independently at random, the 15 differences of Table 3 would each have occurred with equal frequency with a positive or with a negative sign. Their sum, taking account of the two negative signs which have actually occurred, is 314, and we may ask how many of the 2^{15} numbers, which may be formed by giving each component alternatively a positive and

Scientific null: students assigned at random, blocked design

Statistical null for Student's T-test: student responses are IID gaussian.

Type III error

Illustration: ignoring stratification

- Two centers, A and B.
- 4 units per center, randomized 2 to treatment and 2 to control
- Response is a for control in A, $a+1$ for treatment in A. Ditto for B.

Illustration: ignoring stratification

- Two centers, A and B.
- 4 units per center, randomized 2 to treatment and 2 to control
- Response is a for control in A, $a + 1$ for treatment in A. Ditto for B.
- Permutation P value is $1/\binom{4}{2}^2 = 1/36 \approx 0.029$

Illustration: ignoring stratification

- Two centers, A and B.
- 4 units per center, randomized 2 to treatment and 2 to control
- Response is a for control in A, $a+1$ for treatment in A. Ditto for B.
- Permutation P value is $1/\binom{4}{2}^2 = 1/36 \approx 0.029$
- Student's T statistic for $b - a = 10$ is $1/\sqrt(100/6) \approx 0.2449$; one-sided P-value 0.41

Application 9: Blair-Loy et al. (2017) interruptions of academic job talks

Do academic audiences interrupt female speakers more often than they interrupt male speakers?

- 119 job talks from two engineering schools
- fit a zero-inflated negative binomial regression model with coefficients for gender, speaker's years since PhD, the proportion of faculty in the department who are female, and a dummy variable for university, and a dummy variable for department (CS, EE, or ME).
- statistical null hypothesis: the coefficient of gender in the "positive" model is zero

The standard choices for modeling count data are a Poisson model, negative binomial model, or a zero-inflated version of either of these models [55]. We prefer a zero-inflated, negative binomial (ZINB) model for this analysis . . .

We now estimate ZINB models to address our first research question: ***do women get more questions than men during the job talk?*** *(emphasis added)*

ZINB model:

- in each talk, a biased coin is tossed
  - heads, no questions
  - tails, toss another biased coin repeatedly, independently, until it lands heads for the kth time.
  - number of questions is number of tosses it takes to get kth head on 2nd coin.
- probabilities that each coin lands heads and k depend parametrically on the covariates
- Scientific null: gender has no effect on the number of questions
- Statistical null: ZINB model is true, and coefficient of gender in the “positive” part of the ZINB model is zero.
- Type III error; displacement

Application 10: soccer penalty cards Silberzahn et al. (2018)

- 29 teams comprising 61 “analysts” attempting to answer the same question from the same data: are soccer referees more likely to give penalties (“red cards”) to dark-skin-toned players than to light-skin-toned players.
- teams used a wide variety of models and came to different conclusions: 20 found a “statistically significant positive effect”
- great example of reproducible research: data, models, and algorithms were made available to the public

The teams used models and tests including:

Least-squares regression, with or without robust standard errors or clustered standard errors, with or without weights; multiple linear regression; GLMs, GLMMs, with or without a logit link; negative binomial regression, with or without a logit link; multilevel regression; hierarchical log-linear models; linear probability models; logistic regression; Bayesian logistic regression, mixed-model logistic regression; multilevel logistic regression; multilevel Bayesian logistic regression, multilevel logistic binomial regression; clustered robust binomial logistic regression; Dirichlet-process Bayesian clustering; Poisson regression; hierarchical Poisson regression; zero-inflated Poisson regression; Poisson multilevel modelling; cross-classified multilevel negative binomial regression; hierarchical generalized linear modeling with Poisson sampling; Tobit regression; Spearman correlation

- chose 21 distinct subsets of the 14 available covariates
- “round-robin” peer feedback on each team’s initial work, after which the approaches and models were revised
- 2nd period of discussion and revision

- scientific null: skin tone does not affect whether referees give penalty flags
- statistical null: red cards are issued according to a parametric probability model, and the coefficient of skin tone in that model is zero
- Type III error

Unsurprising that the teams came to differing conclusions

Application 11: IPCC Climate Models

. . . quantified measures of uncertainty in a finding expressed probabilistically (based on statistical analysis of observations or model results or expert judgment).

. . . Depending on the nature of the evidence evaluated, teams have the option to quantify the uncertainty in the finding probabilistically. In most cases, level of confidence. . .

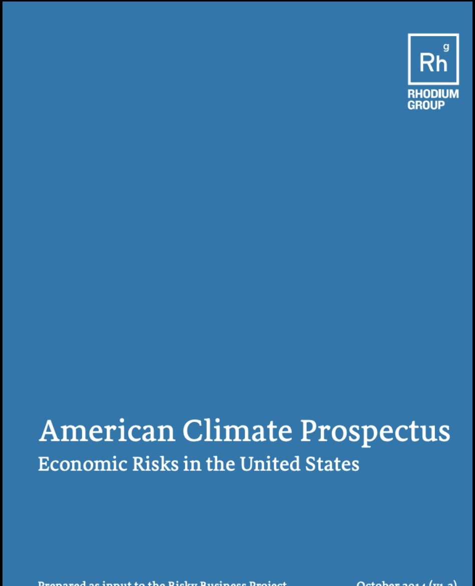
. . . Because risk is a function of probability and consequence, information on the tails of the distribution of outcomes can be especially important. . . Author teams are therefore encouraged to provide information on the tails of distributions of key variables. . .

- Subjective probability assessments (even by experts) generally untethered to reality
- subjective confidence is unrelated to accuracy.
- mixing measurement errors with subjective probabilities doesn’t work
- climate parameters have unknown values, not probability distributions.

‘Multi-model ensemble approach’: bogus confidence intervals

- take a “found” group of models, compute mean and SD of their predictions
- treat the mean as the expected value of the outcome and SD as the standard error of the natural process that is generating climate
- compute a normal confidence interval from the mean and standard deviation
- treat the confidence interval as a prediction interval

Application 12: The Rhodium Group American Climate Prospectus



Rh g
RHODIUM GROUP

American Climate Prospectus

Economic Risks in the United States

Prepared as input to the Risky Business Project October 2014 (v1.2)

Bloomberg Philanthropies, Office of Hank Paulson, Rockefeller Family Fund, Skoll Global Threats Fund, TomKat Charitable Trust.

> *In this climate prospectus, we aim to provide decision-makers in business and government with the facts about the economic risks and opportunities climate change poses in the United States.*

- estimates the effects of climate change on mortality, crop yields, energy use, the labor force, and crime, *at the level of individual counties in the United States through the year 2099*.

- predicts that violent crime will increase just about everywhere, with different increases in different counties.
    - In some places, on hot days there is on average more crime than on cool days
    - Fit a regression model to the increase.
    - Assume that the fitted regression model is a *response schedule*, i.e., how Nature generates crime rates from temperature.
    - Input average temperature change predicted by a climate model; out comes the average increase in crime rate.

Even if you knew exactly what the temperature and humidity would be in every cubic centimeter of the atmosphere every millisecond of every day, you would have no idea what the crime rate in the U.S. would be next year, much less in 2099, much less at the level of individual counties.

And that is before considering the uncertainty in climate models.

What to do instead. Freedman:

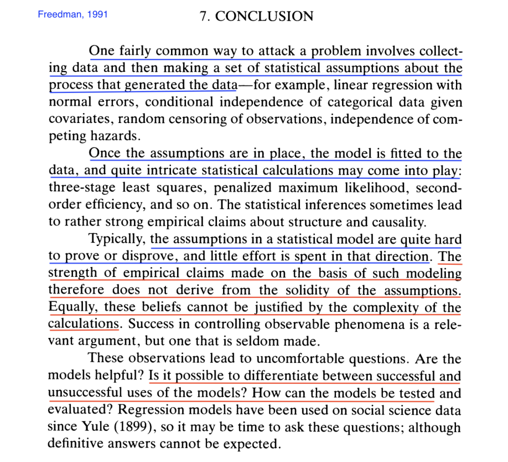
Regression models often seem to be used to compensate for problems in measurement, data collection, and study design. By the time the models are deployed, the scientific position is nearly hopeless. Reliance on models in such cases is Panglossian . . .

Given the limits to present knowledge, I doubt that models can be rescued by technical fixes. Arguments about the theoretical merit of regression or the asymptotic behavior of specification tests for picking one version of a model over another seem like the arguments about how to build desalination plants with cold fusion as the energy source. The concept may be admirable, the technical details may be fascinating, but thirsty people should look elsewhere . . .

Causal inference from observational data presents may difficulties, especially when underlying mechanisms are poorly understood.

There is a natural desire to substitute intellectual capital for labor, and an equally natural preference for system and rigor over methods that seem more haphazard. These are possible explanations for the current popularity of statistical models.

Indeed, far-reaching claims have been made for the superiority of a quantitative template that depends on modeling — by those who manage to ignore the far-reaching assumptions behind the models. However, the assumptions often turn out to be unsupported by the data. If so, the rigor of advanced quantitative methods is a matter of appearance rather than substance.



Freedman, 1991

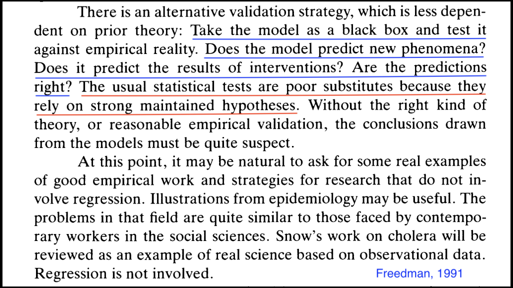
7. CONCLUSION

One fairly common way to attack a problem involves collecting data and then making a set of statistical assumptions about the process that generated the data—for example, linear regression with normal errors, conditional independence of categorical data given covariates, random censoring of observations, independence of competing hazards.

Once the assumptions are in place, the model is fitted to the data, and quite intricate statistical calculations may come into play: three-stage least squares, penalized maximum likelihood, second-order efficiency, and so on. The statistical inferences sometimes lead to rather strong empirical claims about structure and causality.

Typically, the assumptions in a statistical model are quite hard to prove or disprove, and little effort is spent in that direction. The strength of empirical claims made on the basis of such modeling therefore does not derive from the solidity of the assumptions. Equally, these beliefs cannot be justified by the complexity of the calculations. Success in controlling observable phenomena is a relevant argument, but one that is seldom made.

These observations lead to uncomfortable questions. Are the models helpful? Is it possible to differentiate between successful and unsuccessful uses of the models? How can the models be tested and evaluated? Regression models have been used on social science data since Yule (1899), so it may be time to ask these questions; although definitive answers cannot be expected.



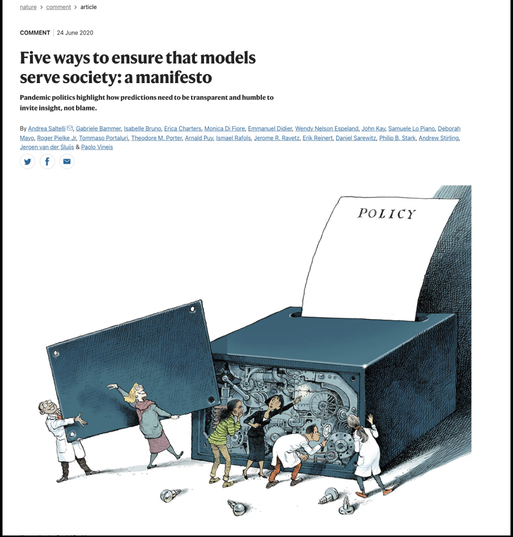
There is an alternative validation strategy, which is less dependent on prior theory: Take the model as a black box and test it against empirical reality. Does the model predict new phenomena? Does it predict the results of interventions? Are the predictions right? The usual statistical tests are poor substitutes because they rely on strong maintained hypotheses. Without the right kind of theory, or reasonable empirical validation, the conclusions drawn from the models must be quite suspect.

At this point, it may be natural to ask for some real examples of good empirical work and strategies for research that do not involve regression. Illustrations from epidemiology may be useful. The problems in that field are quite similar to those faced by contemporary workers in the social sciences. Snow's work on cholera will be reviewed as an example of real science based on observational data. Regression is not involved.

Freedman, 1991

John Tukey:

> *The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.*



nature > comment > article

COMMENT | 24 June 2020

Five ways to ensure that models serve society: a manifesto

Pandemic politics highlight how predictions need to be transparent and humble to invite insight, not blame.

By Andrea Saltelli, Gabriele Bammer, Isabelle Bruno, Erica Charters, Monica Di Fiore, Emmanuel Didier, Wendy Nelson Espeland, John Kay, Samuele Lo Piano, Deborah Mayo, Roger Pielke Jr, Tommaso Portaluri, Theodore M. Porter, Arnald Puy, Ismael Rafols, Jerome R. Ravetz, Erik Reinert, Daniel Sarewitz, Philip B. Stark, Andrew Stirling, Jeroen van der Sluijs & Paolo Vineis

- mind the assumptions
- mind the hubris
- mind the framing
- mind the consequences
- mind the unknowns
- questions, not answers

- Be honest–starting with yourself.
- If you don't know the subject matter, you will probably answer the wrong question. Learn the subject.
- Think about where the data come from and how they happened to become your sample.
- Enumerate the assumptions. Check those you can; flag those you can't. Which are plausible? Which are plainly false? How much might it matter?
- Why is what you did the right thing to have done?
- Emphasize design-based inference.
- "But mom, *everybody's* doing it" isn't scientific justification.
- The most important work is often not the most technical, but it requires the most patience: a technical tour-de-force is typically less useful than curiosity, skeptical persistence, and shoe leather.