Pay No Attention To the Model Behind the Curtain

International Stochastics Seminar
Amirkabir University of Technology
(Tehran Polytechnic)

Philip B. Stark

8 February 2023

University of California, Berkeley

How not to do applied statistics

1. Assign a number to everything, even if it's a meaningless number.
2. Do arithmetic with the numbers as if they all represent the same thing.
3. Optional: make up uncertainties for the numbers. If you do, pretend there's no difference between different kinds of uncertainty.
4. Pick a model for the data based on how the data *look* rather than their connection to the world. Ignore the sampling design (e.g., experiment, survey, observational study).
5. Give terms in the model the names of things you would like to know. For instance, call a term "the effect of x on y," or "the probability of z."
6. Fit the model to the data. Ideally, use a method that requires high-performance computing.
7. Test hypotheses and construct confidence sets as if the model really generated the data.
8. Repeat for different models and hypotheses until you reject at least one.
9. Publish your "discovery."

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.

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.

Me

Many models are importantly wrong.

Quantifauxcation. Assign a meaningless number, then conclud that because the result 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.

Numbers, numbers everywhere, nor any a datum to drink

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

Qualitative arguments considered unscientific.

History, sociology, and econ exalt computation as scientific and objective.

Numbers, numbers everywhere, nor any a datum to drink

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

Qualitative arguments considered unscientific.

History, sociology, and econ exalt computation as scientific and objective.

> *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.*

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 a single, one-dimensional scale,

Conjoint analysis (Luce and Tukey, 1964)

To put multi-attribute things on a single preference 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)

Example: Uncertainty & The Ludic Fallacy: Nassim Taleb (2007)

"The casino is the only human venture I know where the probabilities are known, [] and almost computable." . . . [W]e automatically, spontaneously associate chance with these Platonified games. . . . Those who spend too much time with their noses glued to maps will tend to mistake the map for the territory. . . . Probability is a liberal art; it is a child of skepticism, not a tool for people with calculators on their belts to satisfy their desire to produce fancy calculations and certainties. Before Western thinking drowned in its "scientific" mentality, . . . people prompted their brain to think—not compute.

Lucien LeCam (1977) on treating all uncertainties alike

It is clear that we 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 [].

Bias can be manipulated through processes such as *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.

More than this, 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.

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), it 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), it 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 rules:

- Shoot the tallest person
  - no probability
  - you are or aren't the tallest person
- 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%.

How does probability enter a scientific problem?

- underlying physical phenomenon is random, 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* Choose a *prior distribution* 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' it is random

Simulation and probability

In physics, geophysics, climate science, sensitivity analysis, and uncertainty quantification, there's a popular impression that probabilities can be estimated in a 'neutral' or 'automatic' way by doing Monte Carlo simulations: just let the computer generate the distribution.

But Monte Carlo simulation just estimates numerical values that result from an *assumed* distribution. It is a substitute for doing an integral, not a way to uncover laws of Nature.

Doesn't tell you anything that wasn't already baked into the simulation.

Example 1: 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 help decide where to build nuclear power plants and nuclear waste disposal sites
- 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 the 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.
- have been many destructive earthquakes 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 when they do happen, but not how
- some places are more common targets than others (e.g., places near active faults)
- some places are more vulnerable to damage
- no probability *per se*

Example 2: 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 process with rate that depends parametrically on selected properties of the turbines.
  - collisions are random and independent
  - probability distribution same for all birds
  - rate follows a hierarchical Bayesian model
  - parameters for design and location
  - 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
  - birds toss coins independently

Displacement

The framing changes the subject from "how many birds does this turbine kill?" to "what are the numerical values of some coefficients in this 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.'*

Example 3: 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.

Example 4: 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

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

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	perm P	t-test P
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

Example 5: 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 (possibly) different 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

Example 6: 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

Example 7: 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

Example 8: The Rhodium Group American Climate Prospectus

Bloomberg Philanthropies, Office of Hank Paulson, Rockefeller Family Fund, Skoll Global Threats Fund, and TomKat Charitable Trust funded a study that purports to predict impacts of climate change.

> *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 factoring in the uncertainty in climate models.