P-values, Probability, Priors, Rabbits, Quantifauxcation, and Cargo-Cult Statistics

Philip B. Stark, www.stat.berkeley.edu/~stark, @philipbstark
Department of Statistics, University of California, Berkeley

If we are uncritical we shall always find what we want: we shall look for, and find, confirmations, and we shall look away from, and not see, whatever might be dangerous to our pet theories. In this way it is only too easy to obtain what appears to be overwhelming evidence in favor of a theory which, if approached critically, would have been refuted.
—Karl Popper

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.
—J.W. Tukey

It is inappropriate to be concerned with mice when there are tigers abroad.
— George Box

Where does probability come from?

- Rates are not probabilities
- Not all uncertainty is probability. Haphazard/random/unknown
- A coefficient in a model may not be a "real" probability, even if it's called "probability"
- A P-value may not be a relevant probability, even though it is a "probability"

What is Probability?

Axiomatic aspect and philosophical aspect.

- Kolmogorov's axioms:
  - "just math"
  - triple (S, Ω, P)
    - S a set
    - Ω a sigma-algebra on S
    - P a non-negative countably additive measure with total mass 1

- Philosophical theory that ties the math to the world
  - What does probability *mean*?
  - Standard theories
    - Equally likely outcomes
    - Frequency theory
    - Subjective theory
  - Probability models as empirical commitments
  - Probability as metaphor

How does probability enter a scientific problem?

- underlying phenomenon is random (radioactive decay)

- deliberate randomization (randomized experiments, random sampling)

- subjective probability & "pistimetry"
  - posterior distributions require prior distributions
  - prior generally matters but rarely given attention (Freedman)
  - elicitation issues
  - arguments from consistency, "Dutch book," ...
  - why should I care about your subjective probability?

- invented model that's supposed to describe the phenomenon
  - in what sense?
  - to what level of accuracy?
  - description v. prediction v. predicting effect of intervention
  - testable to desired level of accuracy?

- metaphor: phenomenon behaves "as if random"

Two very different situations:

1. Scientist creates randomness by taking a random sample, assigning subjects at random to treatment or control, etc.

2. Scientist invents (assumes) a probability model for data the world gives.

(1) allows sound inferences.

(2) is only as good as the assumptions.

Gotta check the assumptions against the world

- Empirical support?
- Plausible?
- Iffy?
- Absurd?

Cargo-Cult Science: Feynman

In the South Seas there is a cargo cult of people. During the war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to imitate things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things cargo cult science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.

Now it behooves me, of course, to tell you what they're missing. But it would he just about as difficult to explain to the South Sea Islanders how they have to arrange things so that they get some wealth in their system. It is not something simple like telling them how to improve the shapes of the earphones. But there is one feature I notice that is generally missing in Cargo Cult Science. That is the idea that we all hope you have learned in studying science in school—we never explicitly say what this is, but just hope that you catch on by all the examples of scientific investigation. It is interesting, therefore, to bring it out now and speak of it explicitly. It's a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty—a kind of leaning over backwards. For example, if you're doing an experiment, you should report everything that you think might make it invalid—not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked—to make sure the other fellow can tell they have been eliminated.

Details that could throw doubt on your interpretation must be given, if you know them. You must do the best you can—if you know anything at all wrong, or possibly wrong—to explain it. If you make a theory, for example, and advertise it, or put it out, then you must also put down all the facts that disagree with it, as well as those that agree with it. There is also a more subtle problem. When you have put a lot of ideas together to make an elaborate theory, you want to make sure, when explaining what it fits, that those things it fits are not just the things that gave you the idea for the theory; but that the finished theory makes something else come out right, in addition.

In summary, the idea is to try to give all of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another.

[] We've learned from experience that the truth will come out. Other experimenters will repeat your experiment and find out whether you were wrong or right. Nature's phenomena will agree or they'll disagree with your theory. And, although you may gain some temporary fame and excitement, you will not gain a good reputation as a scientist if you haven't tried to be very careful in this kind of work. And it's this type of integrity, this kind of care not to fool yourself, that is missing to a large extent in much of the research in cargo cult science.

The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that.

—Richard Feynman, 1974. http://calteches.library.caltech.edu/51/2/CargoCult.htm (http://calteches.library.caltech.edu/51/2/CargoCult.htm)

What's a P-value?

- A probability
- But of what?

P-values

- Observe data $X \sim \mathbb{P}$.
- Null hypothesis $\mathbb{P} = \mathbb{P}_0$ (or more generally, $\mathbb{P} \in \mathcal{P}_0$).
- Nested (monotone) hypothesis tests:
  - $\{A_\alpha : \alpha \in (0, 1]\}$
  - $\mathbb{P}_0\{X \notin A_\alpha\} \leq \alpha$ (or more generally, $\mathbb{P}\{X \notin A_\alpha\} \leq \alpha, \ \forall \mathbb{P} \in \mathcal{P}_0$)
  - $A_\alpha \subset A_\beta$ if $\beta < \alpha$ (Can always re-define $A_\alpha \leftarrow \cup_{\beta \geq \alpha} A_\beta$)
- If we observe $X = x$, P-value is $\sup\{\alpha : x \in A_\alpha\}$.

C.f. informal definition in terms of ”extreme” values?

- What does "more extreme" mean?

It's all about the null hypothesis

- P-values measure the strength of the evidence against the null: smaller values, stronger evidence.
- If the P-value equals p, either:

    1. the null hypothesis is false
    2. an event occurred that had probability no greater than p

- Alternative hypothesis matters for power, but not for level.
- Rejecting the null is not evidence *for* the alternative: it's evidence *against* the null.
- If the null is unreasonable, no surprise if we reject it. Null needs to make sense.
- Unreasonable null is not support for the alternative.

The Rabbit Axioms

1. For the number of rabbits in a closed system to increase, the system must contain at least two rabbits.
2. No negative rabbits.

Freedman's Rabbit-Hat Theorem

You cannot pull a rabbit from a hat unless at least one rabbit has previously been placed in the hat.

Corollary

You cannot "borrow" a rabbit from an empty hat, even with a binding promise to return the rabbit later.

Applications of the Rabbit-Hat Theorem

- Probablility doesn't come out of a calculation unless probability went into the calculation.
    - Can't turn a rate into a probability without assuming the phenomenon is random in the first place.
- Can't conclude that a process is random without making assumptions that amount to assuming that the process is random. (Something has to put the randomness rabbit into the hat.)
- Testing whether the process appears to be random using the *assumption* that it is random cannot prove that it is random. (You can't borrow a rabbit from an empty hat.)
- Posterior distributions don't exist without prior distributions.

When did the rabbit enter the hat?

Anytime you see a P-value, you should ask what the null hypothesis is.

E.g., $\mu = 0$ is not the whole null hypothesis:

- null has to completely specify (a family of possible) probability distributions of the data
- otherwise, can't set acceptance regions $\{A_\alpha\}$.

Anytime you see a posterior probability, you should ask what the prior was.

- no posterior distribution without a prior distribution.
- prior usually matters, despite claims about asymptopic results

Anytime you see a confidence interval or standard error, you should ask what was random.

- no confidence intervals or standard errors without either random sampling or stochastic errors.
- box models

Quantifauxcation
Assign a meaningless number, then pretend that since it's quantitative, it's meaningful.

Many P-values and other "probabilities" and most cost-benefit analyses are quantifauxcation.

Cargo-cult statistics

Usually involves some combination of data, pure invention, *ad hoc* models, inappropriate statistics, and logical lacunae.

Example: The 2-sample problem

- Randomization model: two lists. Are they "different"?
- t-test. Assumptions?
- Permutation distribution

Example: Effect of treatment in a randomized controlled experiment

11 pairs of rats, each pair from the same litter.

Randomly—by coin tosses—put one of each pair into "enriched" environment; other sib gets "normal" environment.

After 65 days, measure cortical mass (mg).

enriched	689	656	668	660	679	663	664	647	694	633	653
impoverished	657	623	652	654	658	646	600	640	605	635	642
difference	32	33	16	6	21	17	64	7	89	-2	11

How should we analyze the data?

Cartoon of Rosenzweig, M.R., E.L. Bennet, and M.C. Diamond, 1972. Brain changes in response to experience, *Scientific American*, *226*, 22–29 report an experiment in which 11 triples of male rats, each triple from the same litter, were assigned at random to three different environments, "enriched" (E), standard, and "impoverished." See also Bennett et al., 1969.

Informal Hypotheses

Null hypothesis: treatment has "no effect."

Alternative hypothesis: treatment increases cortical mass.

Suggests 1-sided test for an increase.

Test contenders

- 2-sample Student t-test:

$$\frac{\text{mean(treatment) - mean(control)}}{\text{pooled estimate of SD of difference of means}}$$

- 1-sample Student t-test on the differences:

$$\frac{\text{mean(differences)}}{\text{SD(differences)}/\sqrt{11}}$$

Better, since littermates are presumably more homogeneous.

- Permutation test using t-statistic of differences: same statistic, different way to calculate P-value.

Assumptions of the tests

1. 2-sample t-test:
    - masses are iid sample from normal distribution, same unknown variance, same unknown mean.
    - Tests weak null hypothesis (plus normality, independence, non-interference, etc.).
2. 1-sample t-test on the differences:
    - mass differences are iid sample from normal distribution, unknown variance, zero mean.
    - Tests weak null hypothesis (plus normality, independence, non-interference, etc.)
3. Permutation test:
    - Randomization fair, independent across pairs.
    - Tests strong null hypothesis.

Assumptions of the permutation test are true by design: That's how treatment was assigned.

If we reject the null for the 1-sample t-test, what have we learned?

> *That the data are not (statistically) consistent with the assumption that they are an IID random sample from a normal distribution with mean 0.*

So what? We never thought they were.

This is a **straw man** null hypothesis.

Making sense of probabilities in applied problems

- Reflexive way to try to represent uncertainty (post-WWII phenomenon)
- Not all uncertainty can be represented by a probability
- "Aleatory" versus "Epistemic"

- Aleatory
  - Canonical examples: coin toss, die roll, lotto, roulette
  - under some circumstances, behave "as if" random (but not perfectly)
- Epistemic: stuff we don't know
- "Pistimetry": measuring beliefs

- Le Cam's (1977) three examples of uncertainty:
  - did Eudoxus have larger feet than Euclid? (ignorance)
  - will a fair coin land "heads" the next time it is tossed? (randomness)
  - is the $10^{137} + 1$st digit of π a 7? (limited resources)

- Bayesian way of combining aleatory variability epistemic uncertainty puts beliefs on a par with an unbiased physical measurement w/ known uncertainty.
    - Claims that by introspection, can estimate without bias, with known accuracy—as if one's brain were unbiased instrument with known accuracy
    - Bacon's triumph over Aristotle should put this to rest, but empirically:
        - people are bad at making even rough quantitative estimates
        - quantitative estimates are usually biased
        - bias can be manipulated by anchoring, priming, etc.
        - people are bad at judging weights *in their hands*: biased by shape & density
        - people are bad at judging when something is random
        - people are overconfident in their estimates and predictions
        - confidence unconnected to actual accuracy.
        - anchoring effects entire disciplines (e.g., Millikan, c, Fe in spinach)
- what if I don't trust your internal scale, or your assessment of its accuracy?
- same observations that are factored in as "data" are also used to form beliefs: the "measurements" made by introspection are not independent of the data

LeCam's coin-tossing example

Toss a fair coin k times independently; X is the number of heads; θ is the chance of heads.

$$\mathbb{P}(X = k||\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Suppose prior is of the form

$$\pi(\theta) = \frac{\theta^\alpha(1-\theta)^\beta}{\int t^\alpha(1-t)^\beta dt}.$$

After tossing the coin, the posterior distribution will be of the same form.

Suppose it turns out to be

$$p(\theta) = C\theta^{100}(1-\theta)^{100}.$$

According to Bayesian inference, that is everything there is to know about θ based on prior beliefs and the experiment.

But doesn't it matter whether this is simply a prior, the posterior after 5 tosses, or the

posterior after 200 tosses?

Bayesian formalism does not distinguish between these cases.

Experiments are not the same as experiences

LeCam's (1977) shopkeeper example

As a final comment, it seemes necessary to mention that in certain respects the theory of personal probability is very similar to a theory of personal mass, which exhibits the same shortcomings.

Suppose that a store owner is asked to assign weights to the items in his store. For this purpose he can group items in sets and compare them by hand. If a set A appears to him lighter than a set B we shall say that $(A, B) \in R$. It is fairly easy to see, in the spirit of Theorem 3, that if the relation R is not compatible with an assignment of individual masses to the items and with the additivity of masses, the system is not very coherent. It is also possible to show that if there are enough items which could be indefinitely divided into 'equally weighty parts' the assignment of masses will be unique up to a multiplicative constant.

Nobody would be particularly surprised however if it turned out that ten thousand peas which were judged all alike when compared pairwise turn out to be quite different when parted into two sets of 5000.

In theory one would try to reconcile these contradictory feelings but it is not clear that it could be achieved nor that it would be worth the trouble, since similar difficulties may then crop up elsewhere.

In spite of the theoretical possibility of assigning masses by hand comparison in this manner, nobody seems to claim that this is just what should be done in stores. Nobody even claims that since masses are masses there is no point in specifying whether they were obtained by hand comparison, or by using a spring scale or by using a balance. In addition the hand comparison system would lead to classify people in categories according to their ability to guess weights and according to their ability to avoid self-deceptions due to size of containers or density of the material.

The parallelism between this and the proposals of the neo-Bayesian school is quite evident. The proposal to classify people according to the sharpness of their ability for statistical guessing has already been made. For instance, Halphen could state that there is no good and bad statistics, there are good and bad statisticians.

If the process of measuring something as definite as masses by hand comparison seems rather unreliable, can one really expect a similar theory of measurement of ethereal opinions to inspire much confidence? If an indication of the process of measurement is helpful in the masses

Rates versus probabilities

- In a series of trials, if each trial has the same probability p of success, and if the trials are independent, then the rate of successes converges (in probability) to p. Law of Large Numbers
- If a finite series of trials has an empirical rate p of success, that says nothing about whether the trials are random.
- If the trials are random *and* have the same chance of success, the empirical rate is an estimate of p.
- If the trials are random *and* have the same chance of success *and* the dependence of the trials is known (e.g., the trials are independent), can quantify the uncertainty of the estimate.

Thought experiments

You are one of a group of 100 people, of whom one will die in the next year.
What's the chance it is you?

You are one of a group of 100 people, of whom one is named "Philip."
What's the chance it is you?

Why does the first invite an answer, and the second not?

Ignorance ≠ Randomness

Cargo Cult Confidence Intervals

- Have a collection of numbers, e.g., MME climate model predictions of warming
- Take mean and standard deviation.
- Report mean as the estimate; construct a confidence interval or "probability" statement from the results, generally using Gaussian critical values
- IPCC does this, as do many others.

What's wrong with it?

- No random sample; no stochastic errors.
- Even if there were a random sample, what justifies using normal theory?
- Even if random and normal, misinterprets confidence as probability. Garbled; something like Fisher's fiducial inference
- Ignores known errors in physical approximations
- Ultimately, quantifauxcation.

Random/haphazard/unpredictable/unknown

- Consider taking a sample of soup to tell whether it is too salty.
  - Stir the soup well, then take a tablespoon: random sample
  - Stick in a tablespoon without looking: haphazard sample

- Tendency to treat haphazard as random
  - random requires deliberate, precise action
  - haphazard is jusy sloppy

- Notions like probability, p-value, confidence intervals, etc., *apply only if the sample is random* (or for some kinds of measurement errors)

  - Don't apply to samples of convenience, haphazard samples, etc.
  - Don't apply to populations.

Two brief examples

- Avian / wind-turbine interactions
- Earthquake probabilities

Wind power: "avian / wind-turbine interactions"

Wind turbines kill birds, notably raptors.

- how many, and of what species?
- how concerned should we be?
- what design and siting features matter?
- how do you build/site less lethal turbines?

Measurements

Periodic on-the-ground surveys, subject to:

- censoring
- shrinkage/scavenging
- background mortality
- is this pieces of two birds, or two pieces of one bird?
- how far from the point of injury does a bird land? attribution...

Is it possible to ...

- make an unbiased estimate of mortality?
- reliably relate the mortality to individual turbines in wind farms?

Stochastic model

Common: Mixture of a point mass at zero and some distribution on the positive axis. E.g., "Zero-inflated Poisson"

Countless alternatives, e.g.:

- observe $\max\{0, \text{Poisson}(\lambda_j) - b_j\}, b_j > 0$
- observe $b_j \times \text{Poisson}(\lambda_j), b_j \in (0, 1)$.
- observe true count in area j with error ϵ_j, where $\{\epsilon_j\}$ are dependent, not identically distributed, nonzero mean

Consultant

- bird collisions random, Poisson distributed
- same for all birds
- independent across birds
- rates follow hierarchical Bayesian model that depends on covariates: properties of site and turbine design

What does this mean?

- when a bird approaches a turbine, it tosses a coin to decide whether to throw itself on the blades
- chance coin lands heads depends on site and turbine design
- all birds use the same coin for each site/design
- birds toss their coins independently

Where do the models come from?

- Why random?
- Why Poisson?
- Why independent from site to site? From period to period? From bird to bird? From encounter to encounter?
- Why doesn't chance of detection depend on size, coloration, groundcover, …?
- Why do different observers miss carcasses at the same rate?
- What about background mortality?

Complications at Altamont

Earthquake probabilities

The PSHA equation

Model earthquake occurrence as a marked stochastic process with known parameters.

Model ground motion in a given place as a stochastic process, given the quake location and magnitude.

Then,

> *probability of a given level of ground movement in a given place is the integral (over space and magnitude) of the conditional probability of that level of movement given that there's an event of a particular magnitude in a particular place, times the probability that there's an event of a particular magnitude in that place*

- That earthquakes occur at random is an *assumption* not based in theory or observation.
- involves taking rates as probabilities
    - Standard argument:
        - M = 8 events happen about once a century.
        - Therefore, the chance is about 1% per year.

Earthquake casinos

- Models amount to saying there's an "earthquake deck"
- Turn over one card per period. If the card has a number, that's the size quake you get.
- Journals and journals full of arguments about how many "8"s in the deck, whether the deck is fully shuffled, whether cards are replaced and re-shuffled after dealing, etc.

But this is just a metaphor!

Earthquake terrorism

- Why not say earthquakes are like terrorist bombings?
  - don't know where or when
  - know they will be large enough to kill
  - know some places are "likely targets"
  - but no probabilities
- What advantage is there to the casino metaphor?

Rabbits and Earthquake Casinos

What would make the casino metaphor apt?

1. The physics of earthquakes might be stochastic. But it isn't.
2. A stochastic model might provide a compact, accurate description of earthquake phenomenology. But it doesn't.
3. A stochastic model might be useful for predicting future seismicity. But it isn't (Poisson, Gamma renewal, ETAS)

3 of the most destructive recent earthquakes were in regions seismic hazard maps showed to be relatively safe (2008 Wenchuan M7.9, 2010 Haiti M7.1, & 2011 Tohoku M9) Stein, Geller, & Liu, 2012 (http://web.missouri.edu/~lium/pdfs/Papers/seth2012-tecto-hazardmap.pdf)

See also Mulargia, Geller, & Stark, 2017 (http://www.sciencedirect.com/science/article/pii/S0031920116303016)

What good are the numbers?

Further reading

- Freedman, D.A., 1995, Some issues in the foundations of Statistics, *Foundations of Science*, *1*, 19–39.

- LeCam, L., 1977. A note on metastatistics or 'an essay towards stating a problem in the doctrine of chances', *Synthese*, *36*, 133–160.

- Mulargia, F., R.J. Geller, and P.B. Stark, 2017. Why is Probabilistic Seismic Hazard Analysis (PSHA) still used?, *Physics of the Earth and Planetary Interiors*, *264*, 63–75.

- Stark, P.B. and D.A. Freedman, 2003. What is the Chance of an Earthquake? 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. Preprint: https://www.stat.berkeley.edu/~stark/Preprints/611.pdf (https://www.stat.berkeley.edu/~stark/Preprints/611.pdf)

- Stark, P.B. and L. Tenorio, 2010. A Primer of Frequentist and Bayesian Inference in Inverse Problems. In Large Scale Inverse Problems and Quantification of Uncertainty, Biegler, L., G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders and K. Willcox, eds. John Wiley and Sons, NY. Preprint: https://www.stat.berkeley.edu/~stark/Preprints/freqBayes09.pdf (https://www.stat.berkeley.edu/~stark/Preprints/freqBayes09.pdf)

- Stark, P.B., 2015. Constraints versus priors. SIAM/ASA Journal on Uncertainty Quantification, 3(1), 586–598. doi:10.1137/130920721, Reprint: http://epubs.siam.org/doi/10.1137/130920721 (http://epubs.siam.org/doi/10.1137/130920721), Preprint: https://www.stat.berkeley.edu/~stark/Preprints/constraintsPriors15.pdf (https://www.stat.berkeley.edu/~stark/Preprints/constraintsPriors15.pdf)

- Stark, P.B., 2016. Pay no attention to the model behind the curtain. https://www.stat.berkeley.edu/~stark/Preprints/eucCurtain15.pdf (https://www.stat.berkeley.edu/~stark/Preprints/eucCurtain15.pdf)