Is the Philosophy of Probabilism an Obstacle to Statistical Fraud Busting?

Deborah G. Mayo

2013 was the "year of celebrating statistics", but it might well have been dubbed the "year of frustration with statistics"

Well-worn criticisms of how easy it is to lie with statistics have their own slogans:

- Association is not causation.
- Statistical significance is not substantive significance.
- No evidence of risk is not evidence of no risk.
- If you torture the data enough, they will confess.

Professional manuals and treatises written for popular consumption are rife with exposés of fallacies and foibles.

- Task forces are organized, cottage industries regularly arise to expose the lack of replication and all manner of selection and publication biases.

- The rise of "big data" might have made cookbook statistics easier but these fallacies were addressed by the founders of the tools.

R. A. Fisher (birthday: Feb 17) observed that:

> [t]he political principle that anything can be proved by statistics arises from the practice of presenting only a selected subset of the data available" (Fisher 1955, 75).

- However, nowadays it's the tools that are blamed instead of the abuser.
- We don't need "scaling up" so much as scaling back many misdiagnoses of what is going wrong.
- In one sense, it's blatantly obvious: sufficient finagling may practically predetermine that a researcher's favorite hypothesis gets support, even if it's unwarranted by evidence.

2. Severity requirement

If a test procedure has little or no ability to find flaws in *H*, finding none scarcely counts in *H*'s favor.

H might be said to have "passed" the test, but it is a test that lacks stringency or severity (verification bias).

> *Severity Requirement:* If data $\boldsymbol{x}_0$ agree with a hypothesis *H,* but the method would very probably have issued so good a fit even if *H* is false, then data $\boldsymbol{x}_0$ provide poor evidence for *H*.

It is a case of bad evidence/no test (BENT).

- This seems utterly uncontroversial

- I argue that the central role of probability in statistical inference is severity—its assessment and control.

- Methods that scrutinize a test's capabilities, according to their severity, I call *error statistical*.

- Existing error probabilities (confidence levels, significance levels) *may* but need not provide severity assessments.

- The differences in justification and interpretation call for a new name: existing labels—frequentist, sampling theory, Fisherian, Neyman-Pearsonian—are too associated with hard line views.

To the error statistician, the list I began with is less a list of embarrassments than key features to be recognized by an adequate account

- Association is not causation.
- Statistical significance is not substantive significance.
- No evidence of risk is not evidence of no risk.
- If you torture the data enough, they will confess.

The criticisms are unified by a call to block the too-easy claims for evidence that are formalized in error statistical logic.

Either grant the error statistical logic or deny the criticisms.

In this sense, error statistics is self-correcting—

3. **Are philosophies about science relevant?**

They should be because these are questions about the nature of inductive-statistical evidence that we are to care about.

A critic might protest: "There's nothing philosophical about my criticism of significance tests: a small p-value is invariably, and erroneously, interpreted as giving a small probability to the null hypothesis that the observed difference is mere chance."

Really? P-values are not intended to be used this way; presupposing they should be stems from a conception of the role of probability in statistical inference—this conception is philosophical.

4. Two main views of the role of probability in inference

Probabilism. To provide a post-data assignment of degree of probability, confirmation, support or belief in a hypothesis, absolute or comparative, given data $\boldsymbol{x}_0$ (Bayesian posterior, Bayes ratio, Bayes boosts)

Performance. To ensure long-run reliability of methods, coverage probabilities, control the relative frequency of erroneous inferences in a long-run series of trials.

What happened to the goal of scrutinizing BENT science by the severity criterion?

Neither "probabilism" nor "performance" directly captures it.

Good long-run performance is a necessary not a sufficient condition for avoiding insevere tests.

The problems with selective reporting, multiple testing, stopping when the data look good are not problems about long-runs—

It's that *we cannot say about the case at hand* that it has done a good job of avoiding the sources of misinterpretation.

Probativeness: Statistical considerations arise to ensure we can control and assess how severely hypotheses have passed.

Probabilism is linked to a philosophy that says *H* is not justified unless it's true or probable (or increases probability, makes firmer).

Error statistics (*probativism*) says *H* is not justified unless something has been done to probe ways we can be wrong about *H* (C.S. Peirce).

My work is extending and reinterpreting frequentist error statistical methods to reflect the severity rationale.

Note: The severity construal blends testing and estimation, but I keep to testing talk to underscore the probative demand.

5. Optional Stopping: Capabilities of methods to probe errors are altered not just by cherry picking, multiple testing, and *ad hoc* adjustments, but also via data dependent stopping rules:

In Normal testing, 2-sided H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$

Keep sampling until H_0 is rejected at the .05 level

(i.e. keep sampling until $|\bar{X}| \geq 1.96\ \sigma/\sqrt{n}$).

Optional stopping: "Trying and trying again": having failed to rack up a 1.96 SD difference after, say, 10 trials, the researcher went on to 20, 30 and so on until finally obtaining a 1.96 SD unit difference is obtained.

With this stopping rule the actual significance level differs from, and will be greater than the .05 that would hold for *n* fixed. (nominal vs. actual significance levels).

Jimmy Savage (a subjective Bayesian) famously assured statisticians: "optional stopping is no sin" so the problem must be with significance levels (because they pick up on it).

"The likelihood principle emphasized in Bayesian statistics implies, … that the rules governing when data collection stops are irrelevant to data interpretation" (1962, p. 193).

"This irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels" (in the sense of Neyman and Pearson) (Edwards, Lindman, Savage 1963, p. 239).

6. Likelihood Principle (LP) (the formal correlate)

I claim the acceptability of methods that declare themselves free of error-probability complexities has a lot to do with the growth of fallacious statistics.

> "A likelihood ratio may be a criterion of relative fit but it "is still necessary to determine its sampling distribution in order to control the error involved in rejecting a true hypothesis, because a knowledge of [likelihoods] alone is not adequate to insure control of this error (Pearson and Neyman, 1930, p. 106).

The **key difference**: likelihoods fix the ***actual*** outcome, while error statistics considers outcomes ***other than the one observed*** in order to assess the error properties.

LP → irrelevance of, and no control over, error probabilities (Birnbaum).

> [I]t seems very strange that a frequentist could not analyze a given set of data, such as ($\boldsymbol{x}_1$,..., $\boldsymbol{x}_n$) if the stopping rule is not given. . . . [D]ata should be able to speak for itself. (Berger and Wolpert 1988, p. 78)

For our debunker, data no more speak for themselves in the case of stopping rules than they do with cherry-picking, hunting for significance and the like.

It's ironic we hear "frequentists tell us only P(data $\boldsymbol{x}_0$ | hypothesis *H*)"; in fact, we never look just at the likelihood (we reject the LP).

7. DOE: Design of Experiments

The same standpoint (that data speak for themselves) downplays the importance of design of experiments (DOE).

Lucien Le Cam: "One of the claims [of the Bayesian approach] is that the experiment matters little, what matters is the likelihood function after experimentation…. It tends to undo what classical statisticians have been preaching for many years: think about your experiment, design it as best you can to answer specific questions, take all sorts of precautions against selection bias and your subconscious prejudices". (Le Cam 1977, p. 158)

The many criticisms of statistical tests tend to overlook that the methods don't work apart from

- design of experiments
- validating model assumptions (whether its "experimental" or observational inquiry).

No methods do.

Reports are now trickling in about the consequences of ignoring frequentist principles of DOE

Stanley Young (Nat. Inst. Of Stat) "There is a relatively unknown problem with microarray experiments, in addition to the multiple testing problems.

Until relatively recently, the microarray samples were not sent through assay equipment in random order.

Essentially all the microarray data pre-2010 is unreliable.

The problem is not with p-value methodology, but with the design and conduct of the study.

"Stop Ignoring Experimental Design (or my head will explode)"

(Lambert, of a bioinformatics software Co.)

Statisticians "tell me how they are never asked to help with design before the experiment begins, only asked to clean up the mess after millions have been spent."

Fisher: "To consult the statistician after an experiment is finished is often merely to ask him to conduct a *post mortem* examination. [to] say what the experiment died of."

But one needs a rationale for the types of DOE: they must make a difference and be called for by the methods.

Probabilist accounts lack this… (LP again)

I may step on Bayesian toes…

Larry Wasserman: "Some of the greatest contributions of statistics to science involve adding additional randomness… … randomized experiments, permutation tests, cross-validation and data-splitting. These are unabashedly frequentist ideas and, while one can strain to fit them into a Bayesian framework, they don't really have a place in Bayesian inference." (Wasserman 2008, p. 465).

8. Testing statistical assumptions

- All methods of statistical inferences rest on statistical models.
- Our statistical philosophy and methodology must be rich enough to include testing assumptions.
- Many Bayesian accounts assume the model, either as reflecting subjective beliefs, backgrounds.
- The most popular probabilists these days are "conventionalist" Bayesians (where priors are intended to have the least influence on the inference).

- Some call themselves "objective" (O-Bayesians) because they "only" assume the statistical model and the data.

- But error statistics does not assume the model

> An important part of frequentist theory is its ability to check model assumptions. The use of statistics whose distribution does not depend on the model assumption to be checked lets the frequentist split off this task from the primary inference of interest. (Cox and Mayo 2010, p. 302).

(The fullest discussion of error statistical testing of assumptions that I know of is in the work of my colleague Aris Spanos.)

- George Box, a hero to many modern Bayesians, issues his plea for *ecumenism* largely because he denies Bayesian updating can serve for model checking.

- Why? Because we'd have to set out the models M_1, M_2, ... M_k, and compute the posterior probability for each M_i (Box 1983, p. 73).

  "The difficulty with this approach is that by supposing all possible sets of assumptions known *a priori*, it discredits the possibility of new discovery. But new discovery is, after all, the most important object of the scientific process." (ibid., p. 74)

This open-endedness is crucial, yet is precluded if you need a probabilistic measure over an exhaustive group of hypotheses.

M_i might be given a high posterior probability *relative* to the other models considered, overlooking the M which has actually produced the data, notes 1983.

However, testing assumptions involves examining residuals from a model, which violates the Likelihood Principle (LP):

> To test assumptions one needs "to examine residuals from a model, …[S]ince the residuals are not based on sufficient statistics" this violates the Likelihood Principle. (Casella and Berger pp. 295-6).

But if you're violating the LP for model testing, why not for the primary inference (is that a coherent probabilistic methodology)?

Those much vilified (pure) significance tests are the primary means for testing model assumptions.

> H_0: the assumption(s) of statistical model M hold for data $\boldsymbol{x}_0$

Graphical analysis, non-parametric and parametric tests should be combined.

Evaluating the effect of violations on error probability assessments are important.

9. The error statistical methodology

The error statistical methodology offers an interconnected account of piecemeal steps linking:

- substantive question
- statistical inference
- generating and checking data

We don't need an exhaustive list of hypotheses to split off the problem of how well (or poorly) probed a given hypothesis is (do not need the Bayesian catchall hypothesis "everything other than *H*"

Efron: its key asset is being piecemeal

- Within a probability model we may deductively assign probabilities to an exhaustive set of events.

- But statistical hypotheses are not themselves events, may allow predicting events, explaining events.

- They may be claims about data generating mechanisms, often put in terms of parameters in models.

- We need an account that affords a genuinely inductive (not deductive) inference to such claims.

10. Statistics in the Discovery of the Higgs

- Take the case of announcing evidence for the discovery of a standard model (SM) Higgs particle based on a "5 sigma observed effect" (in July 2012).

- Physicists did not assign a high probability to

  H^*: SM Higgs exists

  (whatever it might mean

- Besides, many believe in beyond the standard model physics.

- They want to ensure that before announcing the hypothesis *H**: "a SM Higgs boson has been discovered" (with such and such properties) that

  *H** has been given a severe run for its money

  That with extremely high probability we would have observed a smaller excess of signal-like events, were we in a universe where:

  H_0: $\mu = 0$ —background only hypothesis,

  So, very probably H_0 would have survived a cluster of tests, fortified with much cross-checking T, were $\mu = 0$.

Note what's being given a high probability:

Pr(test T would produce less than 5 sigma; H_0) $\geq$ 9999997.

With probability .9999997, our methods would show that the bumps disappear (as so often occurred), *under* the assumption data are due to background H_0.

Assuming we want a posterior probability in H^* seems to be a slide from the value of knowing this probability is high for assessing the warrant for H^*

Granted, this inference relies on an implicit severity principle of evidence.

Data provide good evidence for inferring H (just) to the extent that H passes severely with $\boldsymbol{x}_0$*, i.e., to the extent that H would (very probably) not have survived the test so well were H false.*

They then quantify various properties of the particle discovered (inferring ranges of magnitudes)

11. The probabilists and the p-value police

- Leading Bayesian, Dennis Lindley had a letter sent around (to ISBA members)[1]:
- Why demand such strong evidence?
- (Could only be warranted if beliefs in the Higgs extremely low or costs of error exorbitant.)
- Are they so wedded to frequentist methods? Lindley asks. "If so, has anyone tried to explain what bad science that is?"

- Other critics rushed in to examine if reports (by journalists and scientists) misinterpreted the sigma levels as posterior probability assignments to the models.
- Many critics have claimed that the .99999 was fallaciously being assigned to H^* itself—a posterior probability in H^*[1].
- What critics are doing is interpret a legitimate error probability as a posterior in H^*: SM Higgs

---

1

- One may say informally, "so probably we have experimentally demonstrated an SM-like Higgs".
- When you hear: what they really want are posterior probabilities, ask:
- How are we to interpret prior probabilities? Posterior probabilities?
- The probabilist finds himself holding an account with murky undefined terms.

12. Prior probabilities lead to more flexibility

Probabilists will rarely tell you what they're talking about.

Subjectivists: Degree of belief, opinion (actual, rational), betting behavior.

For a long time it was thought that only subjective Bayesianism had a shot at firm philosophical foundations.

Although subjective elicitation largely abandoned, we still hear throwaway lines: "all methods are subjective".

- It is one thing to say statistical models are strictly false (Box again), they are objects of belief, **and** quite another to convert the entire task to modeling beliefs.

Shift from *phenomena to epiphenomena* (Glymour 2010)

- Our statistical inference philosophy should be in sync with scientific inference more generally.

- Scientists learn about phenomena (through imperfect models), not just beliefs ….

- And they do not set sail by assigning degrees of belief to an exhaustive set of hypotheses that could explain data.

- Physicists believed in some kind of Higgs before building the big collider–very different than having evidence for a discovery.

The most popular probabilism these days: non-subjective.

Conventional (default, reference):

An undefined mathematical construct for obtaining posteriors (giving highest weight to data, or satisfying invariance, or matching frequentists, or….).

Conventional priors not intended to be beliefs (often improper).

Some suggest the reference prior is a stop-gap measure until you have degrees of belief.

A small group of Bayesians, following George Box, proposes to include the prior as part of the model and test it along with assumptions like independence, distribution …

- How can you test something if you don't have an idea of its intended meaning?
- Duhemian problems (of where to lay blame for any anomaly) loom large.
- Is it really kosher to go back and change priors after seeing the data?

I'm not saying priors never "work", when they do they occur within ordinary frequentist probability assessments (empirical Bayes, or just ordinary frequentism).

13. Error statistical scrutiny is self-correcting

- True, p-values and other error statistical methods can be misused (severity demands an assessment of discrepancies).
- I'm saying they needn't be, and when they are this is detected only thanks to the fact that *they have the means to register formally problems in the list with which I began.*
- Far from ensuring a high capacity to have alerted us to curb our enthusiasm, various gambits make it easy to send out enthusiastic signals erroneously.
- ***This is a failure for sure, but don't trade them in for methods that cannot detect failure at all.***

14. P-values can't be trusted except when used to argue that p-values can't be trusted

Fraud-busting tools rely on error statistical reasoning, and explicit error statistical methods (e.g., Simonsohn's statistical forensics that busted Smeesters in social psychology).

Ironically, critics of significance tests (and confidence intervals) often chime in with fraudbusting even if based on methods the critics themselves purport to reject?

Is there a whiff of inconsistency here? (Schizophrenia?)

If you think p-values untrustworthy, how can you take them as legitimating criticisms of fraud, especially of a career-ending sort?

15. Winning with Pseudoscience

When we hear there's statistical evidence of some unbelievable claim (distinguishing shades of grey and being politically moderate, ovulation and voting preferences) some probabilists claim—you see, if our beliefs were mixed into the interpretation of the evidence, we wouldn't be fooled.

We know these things are unbelievable.

That could work in some cases (though it still wouldn't show what they'd done wrong).

It wouldn't help with our most important problem:

How to distinguish tests of one and the same hypothesis with different methods used (e.g., one with searching, post data subgroups, etc, another without)?

Simmons, Nelson, and Simonsohn (2011): Using Bayesian rather than frequentist approaches…actually increases researcher degrees of freedom."(p. 7)

Moreover, members of committees investigating questionable research find the researchers really do believe their hypotheses.

Richard Gill (in his role as statistician investigating potential fraud cases): "Integrity or fraud or just questionable research practices?"

> Probably, [X's] data has been obtained by … the usual "questionable research practices" prevalent in the field in question. Everyone does it this way, in fact, if you don't, you'd never get anything published: … effects are too small, noise is too large. People are not deliberately cheating: they honestly believe in their theories and believe the data is supporting them and are just doing the best to make this as clear as possible to everyone.

I've no reason to doubt what Gil says.

Then these inquiries are pseudoscientific and they shouldn't be doing a statistical analysis purporting to link data and scientific claims.

Using subjective Bayesian calculation, a high posterior for the hypothesis given the data may readily be had.

The onus should be on the researcher to show the methods have been subjected to scrutiny.

- *Questionable science*: A statistical (or other) inference to H^* is questionable if it stems from a method with little ability to have found flaws if they existed.

- *Pseudoscience*: A research area that regularly fails to be able to vouchsafe the capability of discerning/reporting mistakes at the levels of data, statistical model, substantive inference

It's useless to apply statistical methods if you cannot show what you're measuring is relevant to the inference you purport to draw.

Either attempt to develop some theory, or call it something else.

On this view, many of those artificial experiments on undergraduates should be scrapped

Triangulation

It's not so much replication but triangulation that's needed.

Where it's ensured that if one analysis doesn't find a mistake the others will with high probability.

Else we're hearing merely that once again I was able to construe my data as evidence of the real effect.

16. How the severity analysis avoids classic fallacies

I don't ignore misinterpretations and misuses of methods—a main focus of my work in philosophy of statistics)

Fallacies of Rejection: Statistical vs. Substantive Significance

i. **Take statistical significance as evidence of substantive theory that explains the effect.**
ii. Infer a discrepancy from the null beyond what the test warrants.

(i) Handled easily with severity: since flaws in the substantive alternative H^* have not been probed by the test, the inference from a statistically significant result to H^* fails to pass with severity.

> Merely refuting the null hypothesis is too weak to corroborate substantive *H**, "we have to have 'Popperian risk', 'severe test' [as in Mayo], or what philosopher Wesley Salmon called a *highly improbable coincidence*" (Meehl and Waller 2002, p. 184).

So-called NHSTs that allow this exist only as abuses of tests: they are not licensed by any legitimate test (apparently allowed in psychology).

(ii) Infer a discrepancy from the null beyond what the test warrants

Finding a statistically significant effect, $d(\mathbf{x}_0) > c_\alpha$ (cut-off for rejection) need not be indicative of large or meaningful effect sizes — especially with *n* sufficiently large: large *n* problem

- The severity assessment makes short work of this: an α-significant difference is indicative of *less* of a discrepancy from the null with large n than if it resulted from a smaller sample size.

  (What's more indicative of a large effect (fire), a fire alarm that goes off with burnt toast or one so insensitive that it doesn't go off unless the house is fully ablaze? The larger sample size is like the one that goes off with burnt toast.)

The *"Jeffrey-Good-Lindley" paradox*

- With *n* sufficiently large, a statistically significant result can correspond to a high posterior probability to a null hypothesis.
- If you give it a lump of prior (spiked prior).
- But why the spiked prior?

  (And why compare error probabilities and posteriors?)

Fallacy of Non-Significant results: The test might not be sensitive enough

- One of the slogans at the start is an instance of this:

No evidence against the null is not evidence for the null

- True, negative results do not warrant 0 discrepancy from the null, but we can use severity to set an upper bound (as with confidence intervals), to infer the discrepancy from 0 isn't as great as γ.

What counts as substantively important, of course, is context dependent.

16. We've tried: Many say they've tried to get people to interpret error probabilities correctly, it didn't work so we should ban significance tests altogether.

(It's not a matter of being a fan of significance tests, they are of a piece with confidence intervals, and other error probability methods.

Many CI reformers use tests in the dichotomous fashion they supposedly deplore).

At first I thought, they tried…

But when you look at their textbooks and papers you see:

- p-values do not give posterior probabilities in hypotheses, which is what you want. (So they may be used only by misinterpretation.)
- They're perfectly fine if the model is infallibly given and you only care about good performance in the long run.
- They're perfectly fine if the significance level matches your degrees of belief in the hypothesis (as with certain "uninformative" priors).
- They're fine if you deny any background knowledge should enter inference…

And these are the middle-of-the roaders…

17. A new paradigm: "Science-wise rates" (FDRs):
Combines probabilism and performance in ways that risk entrenching just about every fallacy in the books.

A: finding a statistically significant result at the .05 level

$$P(H_0|A) = \frac{P(A|H_0)P(H_0)}{P(A|H_0)P(H_0) + P(A|H_1)P(H_1)} = \frac{\alpha(1-\pi)}{\alpha(1-\pi)+\gamma\pi}.$$

If we

- imagine two point hypotheses H_0 and H_1 – H_1 identified with some "meaningful" effect, *H*,* all else ignored,
- assume P(*H**) is very small (.1),

- permit a dichotomous "thumbs up-down" pronouncement,
- from a single (just) .05 significant result (ignoring magnitudes),
- allow the ratio of type 1 error probability to the power against H_1 to supply a "likelihood ratio".

The unsurprising result is that most "positive results" are false.

But their computations might at best hold for crude screening exercises (e.g., for associations between genes and disease).

18. Concluding remarks: Need for new foundations

I worry that just when there's a move to raise consciousness about questionable statistics, some may buy into methods that sacrifice the very thing that is doing the work in criticism

Error statistical methods include all of resampling statistics, causal modeling, model specification…

Some positive signs

1. Fortunately, some of the worst offenders are now taking serious steps to demand authors report how flexible they've been, register trials, etc.

2. I used "probabilism" in my title because we are starting to see some error statistical Bayesians who also reject "probabilisms".

In an attempted meeting of the minds (a Bayesian and an error statistician) they suggest that:

Gelman and Shalizi (2013):

> Implicit in the best Bayesian practice is a stance that has much in common with the error-statistical approach of Mayo (1996), despite the latters frequentist orientation. Indeed crucial parts of Bayesian data analysis, such as model checking, can be understood as 'error probes' in Mayo's sense. (p. 10).

The idea of error statistical foundations for Bayesian tools is not as preposterous as it may seem.

The concept of severe testing is sufficiently general to apply to any of the methods now in use.

Any inference can be said to be (un)warranted just to the extent that it has (not) withstood severe testing.

"There have been technical advances, now we need an advance in philosophy…" (Gelman and Shalizi)

I agree, and to promote debunking and fraudbusting, it should be error statistical.

Overview:

Non-fraudulent uses of statistics demands an account capable of registering how various gambits alter the error probing capacities of methods. This turns on error statistical considerations that are absent in accounts that fall under the umbrella of "probabilism". Examples of such gambits are: stopping rules (optional stopping), data-dependent selections, flawed randomization, and violated statistical model assumptions. If little has been done to rule out flaws in construing the data as evidence for claim *H*, then *H* "passes" a test that lacks severity. Methods that scrutinize a method's capabilities, according to their severity, I call *error statistical*. Using probability to control and assess severity differs from the goals of probabilism and that of long-run performance. Assuming probabilism often leads to presupposing that p-values, confidence levels and other error statistical properties are misinterpreted. Reinterpreting them as degrees of belief or plausibility begs the question against error statistical goals at the heart of debunking. Such twists turn out to license, rather than hold accountable, cases of questionable science (she showed the greatest integrity in promoting her beliefs in *H*), while some of the most promising research (e.g., discovery of the Higgs boson) is misunderstood and/or dubbed pseudoscientific. Attempts to assess "science-wise rates of false discoveries" (unintentionally) institutionalizes howlers, abuses, and cookbook statistics.

References and Selected Background

Berger, J. & Wolpert, R. (1988). *The Likelihood Principle*. (2nd ed). Vol. 6. Lecture Notes-Monograph Series. Hayward, Calif.: IMS.

Birnbaum, A. (1962). On the Foundations of Statistical Inference. In S. Kotz & N. Johnson (Eds.), *Breakthroughs in Statistics*, (1, 478–518). Springer Series in Statistics. New York: Springer-Verlag.

Birnbaum, A. (1969). Concepts of Statistical Evidence. In S. Morgenbesser, P. Suppes, & M. G. White (Eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, (112–143). New York: St. Martin's Press.

Box, G. E. P. (1983). An apology for ecumenism in statistics. In G.E.P. Box, T. Leonard, & D. F.J. Wu (Eds.), *Scientific Inference, Data Analysis, and Robustness,* (51-84). Academic Press.

Cassella, G. & Berger, R. (2002). *Statistical Inference*, (2nd ed.). Belmont, CA: Duxbury Press.

Cox, D. R., & Mayo, D. G. (2010). Objectivity and Conditionality in Frequentist Inference. In D. G. Mayo & A. Spanos (Eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, (276–304). Cambridge: CUP.

Edwards, W., Lindman, H. & Savage, L. J. (1963). Bayesian Statistical Inference for Psychological Research. *Psych. Rev.* 70 (3), 193–242.

Fisher, R. A. (1938). Presidential Address to the First Indian Statistical Congress, 1938. *Sankhya* 4, 14-17.

Fisher, R. A. (1947). *The Design of Experiments* (4th ed.). Edinburgh: Oliver and Boyd.

Fisher, R. A. (1955). Statistical Methods and Scientific Induction. *J. Roy. Stat. Soc., B (Meth.)* 17(1), 69–78.

Gelman, A. (2011). Induction and Deduction in Bayesian Data Analysis. *Rationality, Markets and Morals: Studies at the Intersection of Philosophy and Economics*, *2* (Special Topic: Statistical Science and Philosophy of Science), 67–78.

Gelman, A., & Shalizi, C. (2013). Philosophy and the Practice of Bayesian Statistics and Rejoinder. *Brit. J. Math. & Stat. Psych.* 66 (1), 8–38; 76-80.

Gill, R. (2012/updated 2013). Integrity or fraud or just questionable research practices? *http://www.math.leidenuniv.nl/~gill/Integrity.pdf*

Goodman S. & Greenland S. (2007). Assessing the unreliability of the medical literature: A response to "Why most published research findings are false". JHU, Dept. Biostatistics, Paper 135, 1-25. Available: http://www.bepress.com/jhubiostat/paper135.

Ioannidis J.P.A. (2005). Why most published research findings are false. *PLoS Med*. 2 (8), e124. doi: 10.1371/journal.pmed.0020124.

Lambert, C. (2010) Stop Ignoring Experimental Design (or my head will explode). *Golden Helix* blog post (September 29). http://blog.goldenhelix.com/?p=322.

LeCam, L. (1977). A Note on Metastatistics or "An Essay Toward Stating a Problem in the Doctrine of Chances". *Synthese* 36 (1), 133–160.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: UCP.

Mayo, D. G. (2004). An Error-statistical Philosophy of Evidence. In M. Taper & S. Lele (Eds.), *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*, (79–97). Chicago: UCP.

Mayo, D. G. (2008). How to Discount Double-Counting When It Counts: Some Clarifications. *Brit. J. Phil. Sci.* 59 (4), 857–879.

Mayo, D. G. (2010). Frequentist Statistics as a Theory of Inductive Inference In D. G. Mayo & A. Spanos (Eds.), *Error and Inference*, (247–274). Cambridge: CUP.

Mayo, D. G. (2010). An Error in the Argument From Conditionality and Sufficiency to the Likelihood Principle. In D. G. Mayo & A. Spanos (Eds.), *Error and Inference*, (305–314). Cambridge: CUP.

Mayo, D. G. (2011). Statistical Science and Philosophy of Science: Where Do/Should They Meet in 2011 (and Beyond)? *RMM* 2 (Special Topic: Statistical Science and Philosophy of Science), 79-102.

Mayo, D. G. (2012). Statistical Science Meets Philosophy of Science Part 2: Shallow Versus Deep Explorations. *RMM* 3 (Special Topic: Statistical Science and Philosophy of Science), 71-107.

Mayo, D. G. (2013). The Error-statistical Philosophy and the Practice of Bayesian Statistics: Comments on Gelman and Shalizi: "Philosophy and the Practice of Bayesian Statistics". *Brit. J. Math. & Stat. Psych.* 66 (1), 57–64.

Mayo, D. G. (2013). Discussion: Bayesian Methods: Applied? Yes. Philosophical Defense? In Flux. *The Amer. Stat.* 67(1), 11–15.

Mayo, D. G. (In press). On the Birnbaum Argument for the Strong Likelihood Principle (with discussion). *Statistical Science.*

Mayo, D. G. & Cox, D. R. (2010). Frequentist Statistics as a Theory of Inductive Inference In D. G. Mayo & A. Spanos (Eds.), *Error and Inference*, (1-27). Cambridge: CUP. This paper appeared in *The Second Erich L. Lehmann Symposium: Optimality*, 2006, Lecture Notes-Monograph Series, (49, 1-27). IMS.

Mayo, D. G. & Cox, D. R. (2011). Statistical Scientist Meets a Philosopher of Science: A Conversation. *RMM* 2 (Special Topic: Statistical Science and Philosophy of Science), 103-114

Mayo, D. G., & Kruse, M. 2001. Principles of Inference and Their Consequences. In D. Corfield & J. Williamson (Eds.), *Foundations of Bayesianism*, (24, 381–403). Applied Logic. Dordrecht: Kluwer.

Mayo, D. G. & Spanos, A. (2004). Methodology in Practice: Statistical Misspecification Testing. *Phil. Sci.* 71 (5), 1007–1025.

Mayo, D. G. & Spanos, A. 2006. Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. *Brit. J. Phil. Sci.* 57 (2), 323–357.

Mayo, D. G. & Spanos, A. (Eds.) (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge: CUP

Mayo, D. G. and Spanos, A. (2010). Introduction and Background: Part I: Central Goals, Themes, and Questions; Part II The Error-Statistical Philosophy. In D. G. Mayo & A. Spanos (Eds.), *Error and Inference*, (1-14; 15-27). Cambridge: CUP.

Mayo, D. G. & Spanos, A. (2011). Error Statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of Statistics*, (7, 152–198). *Handbook of the Philosophy of Science*. The Netherlands: Elsevier.

O'Hagan, T. (2012). Higgs Boson (Question from Lindley). *International Society for Bayesian Analysis* blog post (July 10). http://bayesian.org/forums/news/3648.

Owhadi, H., Scovel, C. & Sullivan, T. (2013). When Bayesian Inference Shatters. arXiv:1308.6306 [math.ST].

Pearson, E.S. & Neyman, J. (1930). On the problem of two samples. In *J. Neyman and E.S. Pearson, 1967, Joint Statistical Papers*, (99-115). Cambridge: CUP.

Simmons, J., Nelson, L. & Simonsohn, U., (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psych. Sci.* 22(11), 1359-1366. *(*Available at SSRN: http://ssrn.com/abstract=1850704).

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York: Springer.

Wasserman, L. 2008. Comment on an article by Gelman. *Baysian Analysis* 3 (3), 463-466.

Young, S. & Karr, A. (2011). Deming, Data and Observational Studies. *Significance* 8 (3), 116–120.

Young, S. (2013). Better P-values Through Randomization in Microarrays. *Error Statistics* blog post (June 19). http://errorstatistics.com/2013/06/19/stanley-young-better-p-values-through-randomization-in-microarrays/.

---

[1] *Dear Bayesians,*

A question from Dennis Lindley prompts me to consult this list in search of answers.

We've heard a lot about the Higgs boson. The news reports say that the LHC needed convincing evidence before they would announce that a particle had been found .. Specifically, the news referred to a confidence interval with 5-sigma limits.

Now this appears to correspond to a frequentist significance test with an extreme significance level. Five standard deviations, assuming normality, means a p-value of around 0.0000005…

1. Why such an extreme evidence requirement? We know from a Bayesian perspective that this only makes sense if (a) the existence of the Higgs boson (or some other particle sharing some of its properties) has extremely small prior probability and/or (b) the consequences of erroneously announcing its discovery are dire in the extreme. …

2. Rather than ad hoc justification of a p-value, it is of course better to do a proper Bayesian analysis. Are the particle physics community completely wedded to frequentist analysis? If so, has anyone tried to explain what bad science that is?

3. We know that given enough data it is nearly always possible for a significance test to reject the null hypothesis at arbitrarily low p-values, simply because the parameter will never be exactly equal to its null value. …

If anyone has any answers to these or related questions, I'd be interested

to know and will be sure to pass them on to Dennis.

Regards,

Tony

—-Professor A O'Hagan Email: a.ohagan@sheffield.ac.ukDepartment of Probability and StatisticsUniversity of Sheffield

•