DAY SIX AUG 2

Excursion 3 Tour III: Capability and Severity: Deeper Concepts

From the itinerary: A long-standing family feud among frequentists is between hypotheses tests and confidence intervals (CIs), but in fact there's a clear duality between the two. The dual mission of the first stop (3.7) of this Tour is to illuminate both CIs and severity by means of this duality. A key idea is arguing from the capabilities of methods to what may be inferred. The severity analysis seamlessly blends testing and estimation. A typical inquiry first tests for the existence of a genuine effect and then estimates magnitudes of discrepancies,

or inquires if theoretical parameter values are contained within a confidence interval.

At the second stop (3.8) we reopen a highly controversial matter of interpretation that is often taken as settled. It relates to statistics and the discovery of the Higgs particle– displayed in a recently opened gallery on the "Statistical Inference in Theory Testing" level of today's Museum.

Historical aside…

It was shortly before Egon offers him a faculty position at University College starting 1934 that Neyman gave a paper at the Royal Statistical Society (RSS) that included a portion on confidence intervals, intending to generalize Fisher's Fiducial intervals.

Arthur Bowley: "I am very glad Professor Fisher is present, as it is his work that Dr Neyman has accepted and incorporated…. I am not at all sure that the 'confidence' is not a confidence trick" (C. Reid p. 118).

When it was Fisher's turn, he was full of praise. "Dr Neyman…claimed to have generalized the argument of fiducial probability, and he had every reason to be proud of the line of argument he had developed for its perfect clarity (ibid)."
Caveats were to come later (5.2).

> Fisher had on the whole approved of what Neyman had said. If the impetuous Pole had not been able to make peace between the second and third floors of University College, he had managed at least to maintain a friendly foot on each! (E. Pearson, p. 119)

In a CI estimation procedure, an observed statistic is used to set an upper or lower (1-sided) bound, or both upper and lower (2-sided) bounds for parameter μ.

Consider our test *T+*, H_0: $\mu \leq \mu_0$. against H_1: $\mu > \mu_0$.

The (1 – α) (uniformly most accurate) lower confidence bound for μ, which I write as $\hat{\mu}_{1-\alpha}(\bar{X})$, corresponding to test *T+* is

$$\mu \geq \bar{X} - c_\alpha(\sigma/\sqrt{n})$$

$Pr(Z > c_\alpha) = \alpha$ where Z is the Standard Normal statistic. Here are some useful values for c_α.

α	.5	.25	.05	.025	.02	.005	.001
c_α	0	.1	1.65	1.96	2	2.5	3

The Duality

"Infer: $\mu \geq \bar{X} - 2.5\,(\sigma/\sqrt{n})$" alludes to the rule for inferring; it is the CI estimat*or*. Substituting $\bar{x}$ for $\bar{X}$ yields an esti*mate*.

A *generic* 1-α lower confidence estimator is $\mu \geq \hat{\mu}_{1-\alpha}(\bar{X})$ = $\mu \geq \bar{X} - c_\alpha(\sigma/\sqrt{n})$.

A *specific* 1-α lower confidence estimate is $\mu \geq \hat{\mu}_{1-\alpha}(\bar{x})$ = $\mu \geq \bar{x} - c_\alpha(\sigma/\sqrt{n})$.

If, for any observed $\bar{X}$, in our example, you shout out:

$\mu \geq \bar{X} - 2(\sigma/\sqrt{n})$,

your assertions will be correct 97.5 percent of the time. The specific inference results from plugging in $\bar{x}$ for $\bar{X}$.

The specific .995 lower limit = $\hat{\mu}_{.995}(\bar{x}) = \bar{x} - 2.5(\sigma/\sqrt{n})$, and the specific .995 estimate is $\mu \geq \hat{\mu}_{.995}(\bar{x})$.

Consider our test *T+*, H_0: $\mu \leq 150$. against H_1: $\mu > 150$, σ=10, n = 100.

Work backwards. For what value of μ_0 would $\bar{x} = 152$ exceed μ_0 by $2\sigma_{\bar{X}}$?

$(\sigma/\sqrt{n}) = \sigma_{\bar{X}}$

Answer: $\mu = 150$.

If we were testing H_0: $\mu \leq 149$ vs. H_1: $\mu > 149$ at level .025, we'd reject with this outcome.

The corresponding lower estimate would be:

$\mu > 150$.

Now for the duality

These are the μ values that would not be statistically significant at the .025 level, *had they been the ones tested in the null of T+*. 154 is not statistically significantly greater than any μ value larger than 150 at the .025 level.

> *Severity Fact (for test T+):* To take an outcome $\bar{x}$ that just reaches the α level of significance as warranting H_1: $\mu > \mu_0$ with severity $(1 - \alpha)$, is mathematically the same as inferring $\mu \geq \bar{x} - c_\alpha(\sigma/\sqrt{n})$ at level $(1 - \alpha)$.

Severity will break out of the fixed $(1 - \alpha)$ level, and will supply a non-behavioristic rationale that is now absent from confidence intervals.

Severity and Capabilities of Methods

My justification for inferring $\mu > 150$ is this. Suppose my inference is false.

Were $\mu \leq 150$, then the test very probably would have resulted in a smaller observed $\bar{X}$ than I got, 152

Premise $\Pr(\bar{X} < 152; \mu \leq 150) = .975$.

Premise: Observe: $\bar{X} \geq 152$

Data indicate $\mu > 150$

The method was highly *in*capable of having produced so large a value of $\bar{X}$ as 154, if $\mu \leq 150$, we argue that there is an indication at least (if not full blown evidence) that $\mu > 150$.

To echo Popper, ($\mu > \hat{\mu}_{1-\alpha}$) is corroborated (at level .975) because *it may be presented as a failed attempt to falsify it* statistically.

Non-rejection or non-statistically significant or moderate P-value.

Let $\bar{X}$ = 151, the test does not reject H_0.

The standard formulation of N-P (as well as Fisherian) tests stops there.

We want to be alert to a fallacious interpretation of a "negative" result: inferring there's no positive discrepancy from μ = 150.

Is there evidence of compliance? μ ≤ 150?

The data "accord with" H_0, but what if the test had little capacity to have alerted us to discrepancies from 150?

No evidence against H_0 is not evidence for it.

Condition (S-2) requires us to consider Pr($\bar{X}$ > 151; 150), which is only .16.

(saw on Day #5, depending on how far we got)

Can they say $\bar{X}$ = 151 is a good indication that μ ≤ 150.5?

No, SEV(T, $\bar{X}$ = 151, C: μ ≤ 150.5) = ∼.3.
[Z = 151 – 150.5 = .5]

But $\bar{X}$ = 151 is a good indication that μ ≤ 152
[Z = 151 – 152 = -1; Pr (Z > -1) = .84]
SEV(μ ≤ 152) = .84

It's an even better indication μ ≤ 153 (Table 3.3, p. 145)
[Z = 151 – 153 = -2; Pr (Z > -2) = .97]

With non-rejection, we seek an upper bound, and this corresponds to the upper bound of a CI

Two sided confidence interval may be written
$(\mu = \bar{X} \pm 2\sigma/\sqrt{n})$,

Upper bound is $(\mu < \bar{X} + 2\sigma/\sqrt{n})$,

If one wants to emphasize the post-data measure, one can write:

SEV($\mu < \bar{x} + \gamma\sigma_x$) to abbreviate:

The severity with which

($\mu < \bar{x} + \gamma\sigma_x$).

passes test T+

It's computed $\mathbf{Pr}(\boldsymbol{d}(\boldsymbol{X}) > \boldsymbol{d}(\boldsymbol{x}_0); \mu = \mu_0 + \gamma)$

One can consider a series of upper discrepancy bounds…
$\bar{x}$ = 151, p. 145

The first, third and fifth entries in bold correspond to the three entries of Table 3.3 (p.145)

SEV($\mu < \bar{x} + 0\sigma_x$) = .5

SEV($\mu < \bar{x} + .5\sigma_x$) = .7

SEV($\mu < \bar{x} + 1\sigma_x$) = .84

SEV($\mu < \bar{x} + 1.5\sigma_x$) = .93

SEV($\mu < \bar{x} + 1.96\sigma_x$) = .975

Note the connection with the first two entries of Table 3.3

SEV($\mu < \bar{x} - 1\sigma_x$) = .16

SEV($\mu < \bar{x} - .5\sigma_x$) = .3

But aren't I just using this as another way to say how probable each claim is?

No. This would lead to inconsistencies (if we mean mathematical probability), but the main thing is, or so I argue, **probability gives the wrong logic for "how well-tested" (or "corroborated") a claim is**

Note: low severity is not just a little bit of evidence, but *bad evidence, no test* (BENT)

Severity vs. Rubbing–off

The severity construal is different from what I call the

> *Rubbing off construal*: The procedure is rarely wrong, therefore, the probability it is wrong in this case is low.

Still too much of a *performance* criteria, too *behavioristic*

The long-run reliability of the rule is a necessary but not a sufficient condition to infer *H* (with severity)

The reasoning instead is counterfactual:

$$H: \quad \mu \leq \bar{x} + 1.96\sigma_{\bar{x}}$$

$$(\text{i.e., } \mu \leq CI_u)$$

H passes severely because were this inference false, and the true mean $\mu > CI_u$ then, very probably, we would have observed a larger sample mean.

This is equivalently captured in **Severity Rule** (Mayo 1983, 1991, 1996, Mayo and Spanos 2006, Mayo and Cox 2006):

Test T+: Normal testing: H_0: $\mu \leq \mu_0$ vs H_1: $\mu > \mu_0$
σ is known

(FEV/SEV): If d($\boldsymbol{x}$) is <u>not</u> statistically significant, then
test T passes $\mu < \overline{\boldsymbol{x}} + k_\varepsilon \sigma / \sqrt{n}$ with severity $(1 - \varepsilon)$,
where $P(d(\mathbf{X}) > k_\varepsilon) = \varepsilon$.

- The connection with the upper confidence limit: Infer: $\mu < CI_u$
- We often return to the question of interpreting confidence levels
- Examples to follow

Higgs discovery: "5 sigma observed effect"

- One of the biggest science events of 2012-13 was the announcement on July 4, 2012 of evidence for the discovery of a Higgs-like particle based on a "5 sigma observed effect."

- Because the 5 sigma report refers to frequentist statistical tests, the discovery was immediately imbued with controversies from philosophy of statistics

- I'm an outsider to high-energy particle (HEP) physics but, apart from being fascinated by it, anyone who has come on our journey should be able to decipher the more public controversies about using P-values.

Bad Science? (O'Hagan, prompted by Lindley)

To the ISBA: "Dear Bayesians: We've heard a lot about the Higgs boson. ...Specifically, the news referred to a confidence interval with 5-sigma limits.… Five standard deviations, assuming normality, means a p-value of around 0.0000005…

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 has extremely small prior probability and/or (b) the consequences of erroneously announcing its discovery are dire in the extreme. ...

.... Are the particle physics community completely wedded to frequentist analysis? If so, has anyone tried to explain what bad science that is?"

Not bad science at all!

- HEP physicists are sophisticated with their statistical methodology: they'd seen too many bumps disappear.
- They want to ensure that before announcing the hypothesis *H**: "a new particle has been discovered" that:

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

Significance tests and cognate methods (confidence intervals) are methods of choice here for good reason

Statistical significance test in the Higgs:

(i) Null or test hypothesis: in terms of a model of the detector

μ is the "global signal strength" parameter

H_0: μ = 0 i.e., zero signal (background only hypothesis)

$$H_0: \mu = 0 \quad \text{vs.} \quad H_1: \mu > 0$$

μ = 1: Standard Model (SM) Higgs boson signal in addition to the background

(ii) Test statistic or distance statistic: **d(X)**: how many *excess events* of a given type are observed (from trillions of collisions) in comparison to what would be expected from background alone (in the form of bumps).

(iii) The P-value (or significance level) associated with $d(\mathbf{x}_0)$ is the probability of a difference as large or larger than $d(\mathbf{x}_0)$, under H_0 :

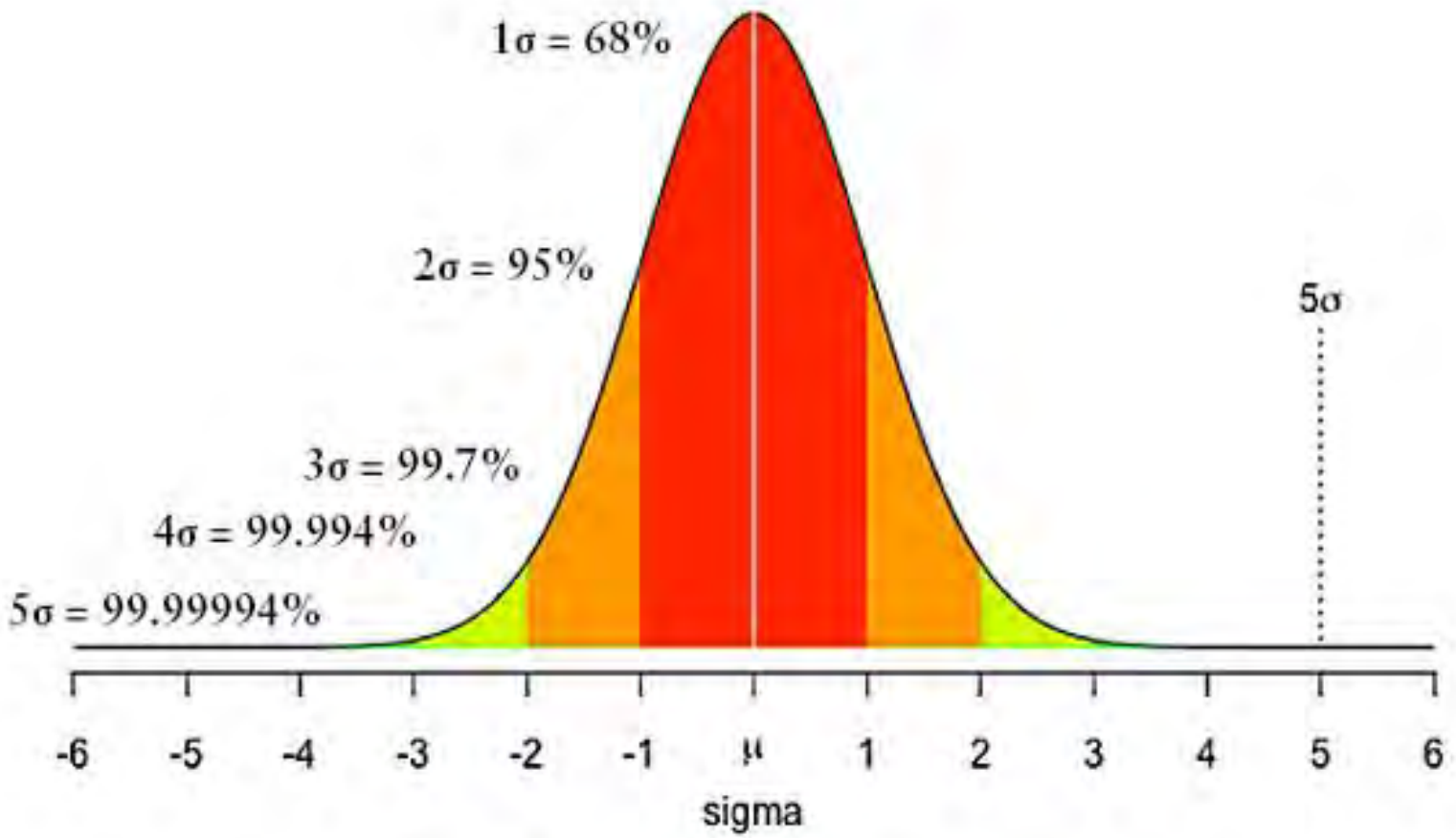
$$P\text{-value}=\Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0); H_0)$$

Usually the P-value is sufficiently small if $\sim$.05, .01, .001

$$\Pr(d(\mathbf{X}) > 5; H_0) = .0000003$$

The probability of observing results as or more extreme as 5 sigmas, under H_0, is approximately 1 in 3,500,000.

The actual computations are based on simulating what it would be like were H_0: $\mu = 0$ (signal strength = 0), fortified with much cross-checking of results


1σ = 68%
2σ = 95%
3σ = 99.7%
4σ = 99.994%
5σ = 99.99994%
5σ
-6
-5
-4
-3
-2
-1
μ
1
2
3
4
5
6
sigma

What "the Results" Really Are (p. 204)

From Translation Guide (Souvenir (C) Excursion 1, p. 52). Pr(d($\boldsymbol{X}$)> 5; H_0) is to be read Pr(the test procedure would yield d($\boldsymbol{X}$) > 5; H_0).

> **Fisher's Testing Principle**: If you know how to bring about results that rarely fail to be statistically significant, there's evidence of a genuine experimental effect.

In good sciences and strong uses of statistics, "the results" include demonstrating the "know how" to generate results that rarely fail to be significant, and showing the test passes an audit (it isn't guilty of selection biases, or violations of statistical model assumptions).

The P-Value Police (SIST p. 204)

When the July 2012 report came out, a number of people set out to grade the different interpretations of the P-value report:

Larry Wasserman ("Normal Deviate" on his blog) called them the "P-Value Police".

- Job: to examine if reports by journalists and scientists could by any stretch of the imagination be seen to have misinterpreted the sigma levels as posterior probability assignments to the various models and claims.

Thumbs up or down

Thumbs up, to the ATLAS group report:

> "A statistical combination of these channels and others puts the significance of the signal at 5 sigma, meaning that *only one experiment in three million would see an apparent signal this strong in a universe without a Higgs."*

Thumbs down to reports such as:

> "There is less than a one in 3.5 million chance that their results are a statistical fluke."

statistical fluctuation or fluke: an apparent signal that is actually produced due to chance variability alone.

Critics (Spiegelhalter) allege they are misinterpreting the P-value as a posterior probability on H_0.

Not so.

H_0 does not say the observed results are due to background alone, or are flukes,

$$H_0: \mu = 0$$

Although if H_0 were true **it follows that** various results would occur with specified probabilities.

(In particular, it entails that large bumps are improbable.)

In fact it is an ordinary error probability.

(1) Pr(Test T would produce d(**X**) $\geq$ 5; H_0) $\leq$ .0000003

(1)* Pr(Test T would produce d(**X**) < 5; H_0) $\leq$ .9999997

(SIST p. 205)

True, the inference actually detached goes beyond a P-value report.

(2)There is strong evidence for

*H**: a Higgs (or a Higgs-like) particle.

Inferring (2) relies on an implicit principle of evidence.

SEV Principle for statistical significance:

If Pr(Test *T* would produce d(**X**) < d($\mathbf{x}_0$); H_0) is very high, then $\mu \geq \mu_0$ passes the test with high severity…

(1)* Pr(Test *T* would produce d(**X**) < 5; H_0) > .9999997

- *With probability .9999997, the bumps would be smaller, would behave like flukes, disappear with more data, not be produced at both CMS and ATLAS, in a world given by H_0.*
- *They didn't disappear, they grew*

(2) So, *H*: a Higgs (or a Higgs-like) particle.*

Goes beyond long-run *performance*: Interpret 5 sigma bumps as a real effect (a discrepancy from 0), you'd erroneously interpret data with probability less than .0000003

An *error probability*

The warrant isn't low long-run error (in a case like this) but detaching an inference based on a severity argument.

Qualifying claims by how well they have been probed (precision, accuracy).

SIST p. 206 3 ups and downs (here are 2)

Ups

U-1. The probability of the background alone fluctuating up by this amount or more is about one in three million.

U-3. The probability that their signal would result by a chance fluctuation was less than one chance in 3.5 million

Downs

D-1. The probability their results were due to the background fluctuating up by this amount or more is about 1 in 3 million.

D-3. The probability that their signal was a result of a chance fluctuation was less than one chance in 3 million.

The difference is that the thumbs down allude to “this” signal

or “these” data are due to chance or is a fluke.

True, but that’s the way frequentists always give probabilities to general events, whether they have occurred, or we are contemplating a hypothetical excess of 5 sigma that might occur.

It’s illuminating to note, at this point that

> [t]he key distinction between Bayesian and sampling theory statistics is the issue of what is to be regarded as random and what is to be regarded as fixed. To a Bayesian, parameters are random and data, once observed, are fixed…(Kadane 2011, p. 437)

- Kadane is right that "[t]o a sampling theorist, data are random even after being observed, but parameters are fixed" (ibid.).

- When an error statistician speaks of the probability that the results standing before us are a mere statistical fluctuation, she is referring to a methodological probability

- If you're a Bayesian probabilist D-1 through D-3 appear to be assigning a probability to a hypothesis (about the parameter) because, since the data are known, only the parameter remains unknown

- But they’re to be scrutinizing a non-Bayesian procedure.
- Whichever approach you favor, my point is that they’re talking past each other.
- To get beyond this particular battle, this has to be recognized.

Those who think we want a posterior probability in H^* might be sliding from what may be inferred from this legitimate high probability:

Pr(test T would not reach 5 sigma; H_0) > .9999997

With probability .9999997, our methods would show that the bumps disappear, *under* the assumption data are due to background H_0.

Most HEP physicists believe in Beyond Standard Model physics (BSM) but to their dismay, they find themselves unable to reject the SM null (bumps keep disappearing)

U-1 through U-3 are not statistical inferences!

They are the (statistical) justifications associated with statistical inferences

U-1. The probability of the background alone fluctuating up by this amount or more is about one in three million.

[Thus, our results are not due to background fluctuations.]

U-2. Only one experiment in three million would see an apparent signal this strong in a universe [where H_0 is adequate].

[Thus H_0 is not adequate.]

U-3. The probability that their signal would result by a chance fluctuation was less than one chance in 3.5 million.

[Thus the signal was not due to chance.]

The formal statistics moves from

(1) Pr(test T produces d($\boldsymbol{X}$) < 5; H_0) < .0000003.

to

(2) There is strong evidence for

(first) (2a) a genuine (non-fluke) discrepancy from H_0.

(later) (2b)H^*: a Higgs (or a Higgs-like) particle.

They move in stages from indications, to evidence, to discovery–they assume

Severity Principle: (from low P-value) Data provide evidence for a genuine discrepancy from H_0 (just) to the extent that H_0 would (very probably) have survived, were H_0 a reasonably adequate description of the process generating the data.

(1)* With probability .9999997, the bumps would be smaller, would behave like statistical fluctuations: disappear with more data, wouldn't be produced at both CMS and ATLAS, in a world adequately modelled by H_0.

<u>They didn't disappear, they grew (from 5 to 7 sigma)</u>
So, (2a) infer there's evidence of H_1: non-fluke, or (2b) infer H^*: a Higgs (or a Higgs-like) particle.

Look Elsewhere Effect (LEE) (p. 210)

Lindley/O'Hagan: "Why such an extreme evidence requirement?"

Their report is of a **nominal** (or local) P-value: the P-value at a particular, data-determined, mass.

- The probability of so impressive a difference anywhere in a mass range would be greater than the local one.

- Requiring a P-value of at least 5 sigma, is akin to adjusting for multiple trials or look elsewhere effect LEE.

This leads to THE key issue of controversy in the philosophy of statistics: whether to take account of selection effects

In the second part of the inquiry, having found the Higgs particle, they're blocking inferences to BSM (p. 211)

Here they use SEV for non-significance or setting upper bounds to discrepancies

Souvenir O (p. 214) ***Interpreting Probable Flukes***

There are three ways to construe a claim of form: A small P-value indicates it's improbable that the results are due to chance alone (as described in H_0).

- The person is using an informal notion of probability, common in English. …Under this reading there is no fallacy. Having inferred H^*: Higg's particle, one may say informally, "so probably we have experimentally demonstrated the Higgs".
- "So probably" H_1 is *merely qualifying the grounds upon which we assert evidence for* H_1.

(2) An ordinary error probability is meant: "the results" in "it's highly improbable our results are a statistical fluke" include: the overall display of bumps, with significance growing with more and better data, Under this reading, again, there is no fallacy.

(3) The person interpreting the p-value as a posterior probability of null hypothesis H_0 based on a prior probability distribution: p = Pr(H_0 |x).
Under this reading there is a fallacy.
Unless the P-value tester has explicitly introduced a prior, it would be "ungenerous" to twist probabilistic assertions into posterior probabilities.

ASA 2016 Guide: Principle #2

> *P*-values do not measure (a) the probability that the studied hypothesis is true, or (b) the probability that the data were produced by random chance alone.
> (Wasserstein and Lazar 2016, p. 131)

I insert the (a), (b), absent from the original principle #2, because while (a) is true, phrases along the lines of (b) should not be equated to (a).

Even proclamations issued by high priests–especially where there are different axes to grind–should be taken with severe grains of salt.

The ASA 2016 Guide's Six Principles:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

ASA acknowledges the statistics wars & lack of agreement

We return to Tour II It's the Methods, Stupid on Monday.