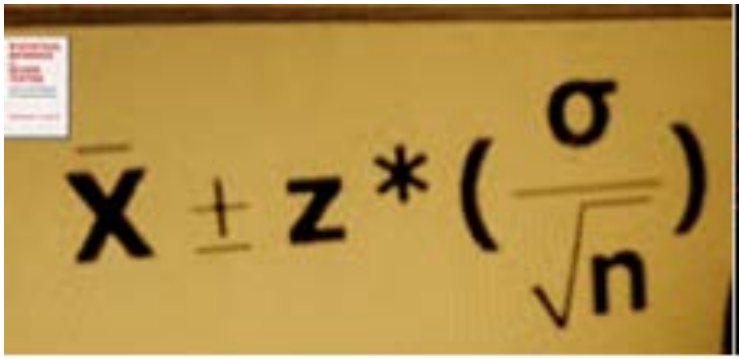
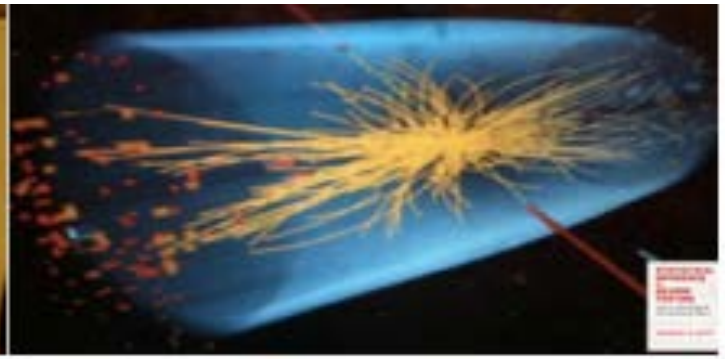
Excursion 3 Tour III
Capability and Severity: Deeper Concepts



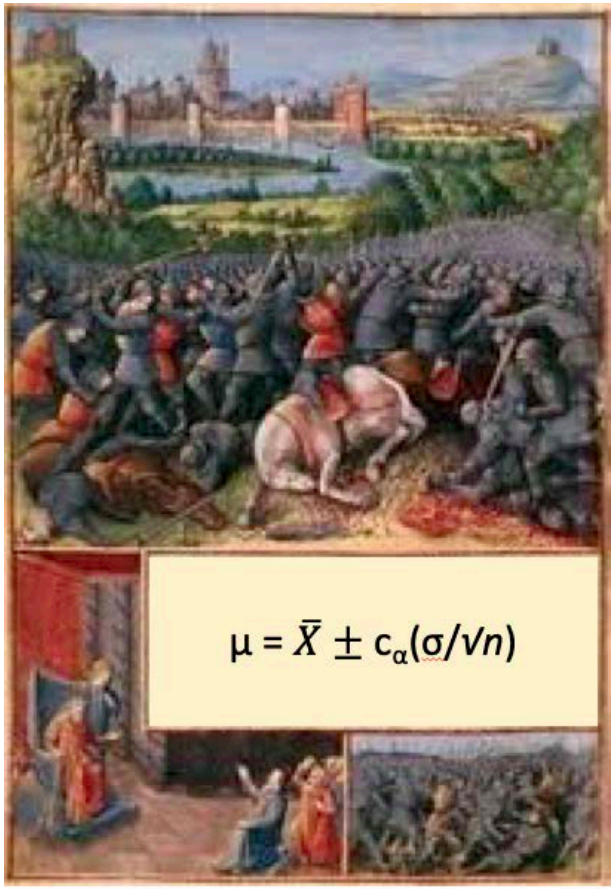
3.7



3.8

Frequentist Family Feud

A long-standing statistics war is between hypotheses tests and confidence intervals (CIs) ("New Statistics")



Historical aside…(p. 189)

"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 not at all sure that the 'confidence' is not a confidence trick." (C. Reid p. 118)

“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.”
(Fisher 1934c, p. 138)

“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)

Duality Between Tests and CIs

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

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

(we would really estimate σ)

$Pr(Z > c_\alpha) = \alpha$ where *Z* is the Standard Normal statistic.

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

"Infer: $\mu \geq \bar{X} - 2.5\,(\sigma/\sqrt{n})$" is a rule for inferring; it is the CI estimat*or*.

Substituting $\bar{x}$ for $\bar{X}$ yields an esti*mate*. (p. 191)

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 $\bar{X}$, 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 $\bar{X}$.

Consider test *T*+, H_0: μ ≤ 150 vs H_1: μ > 150, σ=10, *n* = 100.

(same as test for H_0: μ = μ_0 against H_1: μ > μ_0.)

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

(It should really be 1.96, I'm rounding to 2)

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

Answer: $\mu = 150$.

If we were testing H_0: $\mu \leq 149$ vs. H_1: $\mu > 149$ at level .025,

$\bar{x} = 152$ would lead to reject.

The lower .975 estimate would be: $\mu > 150$.

The CI contains the μ value that wouldn't be rejected were they being tested

152 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 $\overline{x}$ that just reaches the α level of significance as warranting

$$H_1: \mu > \mu_0$$

with severity (1 – α), is mathematically the same as inferring $\mu \geq \overline{x} - c_\alpha(\sigma/\sqrt{n})$ at level (1 – α).

CIs (as often used) inherit problems of behavioristic N-P tests

- Too dichotomous: in/out
- Justified in terms of long-run coverage
- All members of the CI treated on par
- Fixed confidence levels (need several benchmarks)

Move away from a purely "coverage" justification for CIs

A severity justification for inferring μ > 150 is this: Suppose my inference is false.

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

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

Premise: Observe: $\bar{X}$ ≥ 152

Data indicate μ > 150

The method was highly *in*capable of having produced so large a value of $\bar{X}$ as 154, if $\mu \leq 150$,

So 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.

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+.

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

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: $\mu \leq \bar{x} + 1.96\sigma_{\bar{x}}$
(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.

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

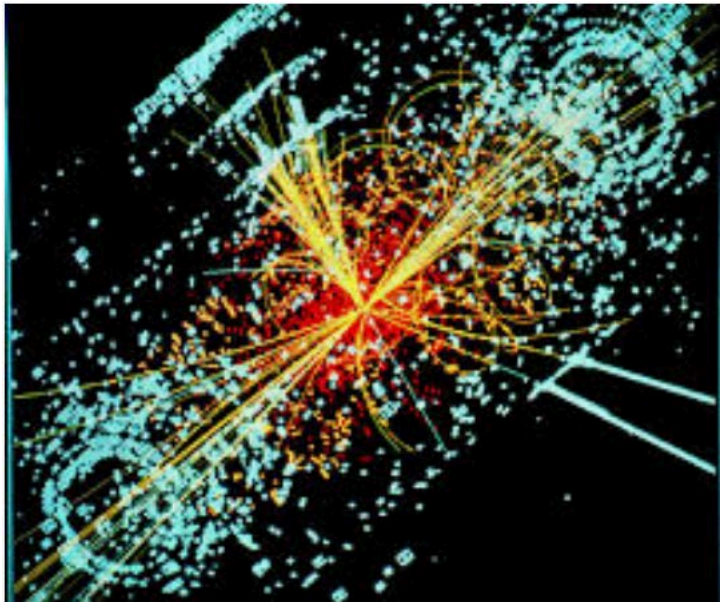
(FEV/SEV): If d($\boldsymbol{x}$) is <u>not</u> statistically significant, then test T passes $\mu < \bar{x} + k_\varepsilon \sigma / \sqrt{n}$ with severity $(1 - \varepsilon)$,

where $P(d(\mathbf{X}) > k_\varepsilon) = \varepsilon$.

(Mayo 1983, 1991, 1996, Mayo and Spanos 2006, Mayo and Cox 2006)

Higgs discovery: “5 sigma observed effect”

One of the biggest science events of 2012-13 (July 4, 2012): the discovery of a Higgs-like particle based on a “5 sigma observed effect.”



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 had 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.

ASA 2016 Guide: Principle #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. (Wasserstein and Lazar 2016, p. 131)

*full list, note 4 pp 215-16

Hypotheses vs Events

- Statistical hypotheses assign probabilities to data or events $Pr(\boldsymbol{x}_0;H_1)$, but it's rare to assign frequentist probabilities to hypotheses
- The inference is qualified by probabilistic properties of the method (methodological probabilities-Popper)

Hypotheses

- A coin tossing (or lady tasting tea) trial is Bernoulli with Pr(heads) on each trial = .5.
- The deflection of light due to the sun λ is 1.75 degrees
- IQ is more variable in men than women
- Covid recovery time is shortened in those given treatment R

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: μ = 0 vs. H_1: μ > 0

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

(ii) Test 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).

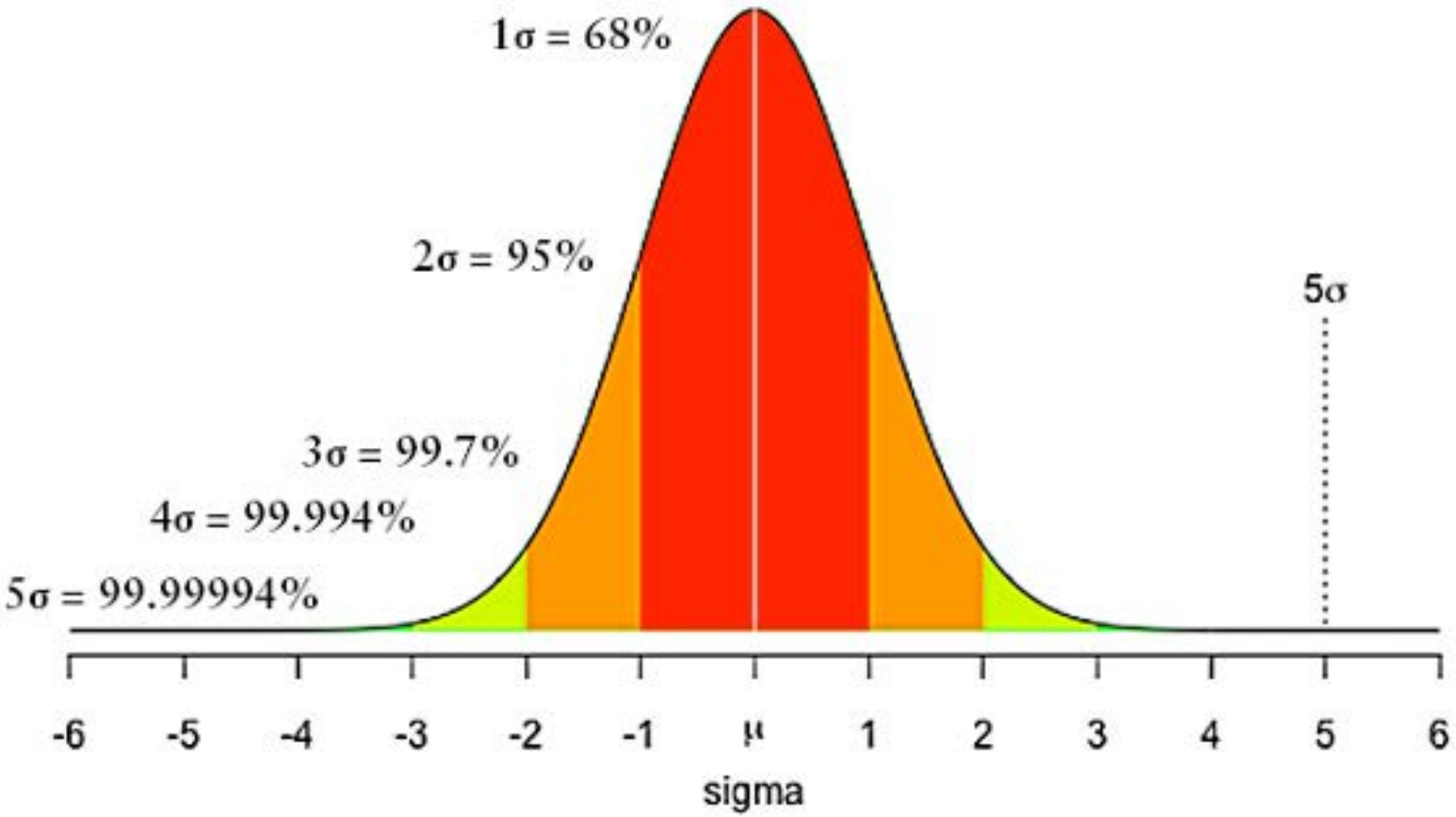
(iii) The P-value (or significance level) associated with $d(\mathbf{x}_0)$: the probability of an excess at least as large as $d(\mathbf{x}_0)$, under H_0:

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

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

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

The computations are based on simulating what it would be like were H_0: μ = 0 (signal strength = 0)


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)

Translation Guide (Souvenir (C) Excursion 1, p. 52).

Pr(d(***X***)> 5; H_0) is to be read Pr(the test procedure would yield d(***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.
"the results" may include demonstrating the "know how" to generate results that rarely fail to be significant.

The P-Value Police (SIST p. 204)

When the July 2012 report came out, some graded the different interpretations of the P-value report: thumbs up or down e.g., Sir David Spiegelhalter (Professor of public Understanding of Risk, Cambridge)

Thumbs up, to the ATLAS group report:

> "A statistical combination of these channels and others puts the significance of the signa 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 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. (p. 205)

Critics allege the “thumbs down” construals misinterpret P-value as a posterior probability on H_0.

There’s disagreement; problem is delicate (p. 203).

It may be seen as an ordinary error probability.

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

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

(SIST p. 205)
(Note: not strictly conditional probs)

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.

(SIST 208-9)

Thumbs down cases allude to “this” signal or “these” data are due to chance or are a fluke

True, but that’s how frequentists give probabilities to general events, whether they have occurred, or it’s 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.).

- For an error statistician: the probability that the results in front of us are a mere statistical fluctuation, refers to a methodological probability

To 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 the P-value police 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.

Some admissions

But U-type statements are preferable because of a tendency to misinterpret the complements: (p. 207)

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-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.]

They move in stages from indications, to evidence, to discovery–implicitly assuming something along the lines of:

> ***Severity Principle Popperian:*** (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.

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 null hypothesis of no Higgs (or Higgs-like) boson was definitively rejected upon the announcement of the observation of a new boson by both ATLAS and CMS on July 4, 2012. The confidence intervals for signal strength θ . . . were in reasonable agreement with the predictions for the SM Higgs boson. Subsequently, much of the focus shifted to measurements of … production and decay mechanisms. For measurements of continuous parameters, … the tests … use the frequentist duality … between interval estimation and hypothesis testing. One constructs (approximate) confidence intervals and regions for parameters … and checks whether the predicted values for the SM Higgs boson are within the confidence regions. (Cousins 2017, p. 414)

Now the corresponding null hypothesis, call it H_0^2, is the SM Higgs boson

$$H_0^2\text{: SM Higgs boson: } \mu = 1$$

and discrepancies from it are probed and estimated with confidence intervals. The most important role for statistical significance tests is actually when results are insignificant, or the *P*-values are not small: *negative* results. They afford a standard for blocking inferences that would be made too readily. In this episode, they arose to

(a) block precipitously declaring evidence of a new particle;
(b) rule out values of various parameters, e.g., spin values that would preclude its being "Higgs-like," and various mass ranges of the particle.

" Search for . . ." (2017, p. 412). They are regarded as important and informative

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

Interpreting "A small P-value indicates it's improbable that the results are due to chance alone (as described in H_0)".

(1) 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:

$$p = \Pr(H_0 \mid x).$$

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.

Could Bayesians be illicitly sliding? (p. 215)

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*

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

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).

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).

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.

Live Exhibit (ix). What Should We Say When Severity Is Not Calculable?

- In developing a system like severity, at times a conventional decision must be made.
- However, the reader can choose a different path and still work within

Other issues:

Souvenir N (p. 201) (negations)

Excursion 3 Tour II (chestnuts and howlers, p. 165-)