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 (Section 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 (Section 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.

3.7 Severity, Capability, and Confidence Intervals (CIs)

It was shortly before Egon offered him a faculty position at University College starting 1934 that Neyman gave a paper at the Royal Statistical Society (RSS) which included a portion on confidence intervals, intending to generalize Fisher's fiducial intervals. With K. Pearson retired (he's still editing *Biometrika* but across campus with his assistant Florence David), the tension is between E. Pearson, along with remnants of K.P.'s assistants, and Fisher on the second and third floors, respectively. Egon hoped Neyman's coming on board would melt some of the ice.

Neyman's opinion was that "Fisher's work was not really understood by many statisticians . . . mainly due to Fisher's very condensed form of explaining his ideas" (C. Reid 1998, p. 115). Neyman sees himself as championing Fisher's goals by means of an approach that gets around these expository obstacles. So Neyman presents his first paper to the Royal Statistical Society (June, 1934), which includes a discussion of confidence intervals, and, as usual, comments (later published) follow. Arthur Bowley (1934), a curmudgeon on the K.P. side of the aisle, rose to thank the speaker. Rubbing his hands together in gleeful anticipation of a blow against Neyman by Fisher, he declares: "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" (p.132). Bowley was to be disappointed. 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" (Fisher 1934c, p.138). Caveats were to come later (Section 5.7). For now, Egon was relieved:

> 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! (C. Reid 1998, p. 119)

CIs, Tests, and Severity. I'm always mystified when people say they find *P*-values utterly perplexing while they regularly consume polling results in terms of confidence limits. You could substitute one for the other.

Suppose that 60% of 100 voters randomly selected from a population U claim to favor candidate Fisher. An estimate of the proportion of the population who favor Fisher, θ, at least at this point in time, is typically given by means of confidence limits. A 95% confidence interval for θ is $\overline{x} \pm 1.96\sigma_{\overline{X}}$ where $\overline{x}$ is the observed proportion and we estimate $\sigma_{\overline{X}}$ by plugging $\overline{x}$ in for θ to get $\sigma_{\overline{X}} = \sqrt{[0.60\ (0.40)/100]} = 0.048$. The 95% CI limits for $\theta = 0.6 \pm 0.09$ using the Normal approximation. The lower limit is 0.51 and the upper limit is 0.69. Often, 0.09 is reported as the *margin of error*. We could just as well have asked, having observed $\overline{x} = 0.6$,

> what value of θ would 0.6 be statistically significantly greater than at the 0.025 level, and what value of θ would 0.6 be statistically significantly less than at the 0.025 level?

The two answers would yield 0.51 and 0.69, respectively. So infer $\theta > 0.51$ and infer $\theta < 0.69$ (against their denials), each at level 0.025, for a combined error probability of 0.05.

Not only is there a duality between confidence interval estimation and tests, they were developed by Jerzy Neyman at the same time he was developing tests! The 1934 paper in the opening to this tour builds on Fisher's fiducial intervals dated in 1930, but he'd been lecturing on it in Warsaw for a few years already. Providing upper and lower confidence limits shows the range of plausible values for the parameter and avoids an "up/down" dichotomous tendency of some users of tests. Yet, for some reason, CIs are still often used in a dichotomous manner: rejecting μ values excluded from the interval, accepting (as plausible or the like) those included. There's the tendency, as well, to fix the confidence level

at a single $1 - \alpha$, usually 0.9, 0.95, or 0.99. Finally, there's the adherence to a performance rationale: the estimation method will cover the true θ 95% of the time in a series of uses. We will want a much more nuanced, inferential construal of CIs. We take some first steps toward remedying these shortcomings by relating confidence limits to tests and to severity.

To simply make these connections, return to our test T+, an IID sample from a Normal distribution, H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$. In a CI estimation procedure, an observed statistic is used to set an upper or lower (one-sided) bound, or both upper and lower (two-sided) bounds for parameter μ. Good and best properties of tests go over into good or best properties of corresponding confidence intervals. In particular, the uniformly most powerful (UMP) test T+ corresponds to a uniformly most accurate lower confidence bound (see Lehmann and Romano 2005, p. 72). The $(1 - \alpha)$ uniformly most accurate (UMA) lower confidence bound for μ, which I write as $\hat{\mu}_{1-\alpha}(\overline{X})$, corresponding to test T+ is

$$\mu > \overline{X} - c_\alpha(\sigma/\sqrt{n}),$$

where $\overline{X}$ is the sample mean, and the area to the right of c_α under the standard Normal distribution is α. That is $\Pr(Z \geq c_\alpha) = \alpha$ where Z is the standard Normal statistic. Here are some useful approximate values for c_α:

α	0.5	0.16	0.05	0.025	0.02	0.005	0.001
c_α	0	1	1.65	1.96	2	2.5	3

The Duality

"Infer: $\mu > \overline{X} - 2.5(\sigma/\sqrt{n})$" alludes to the rule for inferring; it is the CI *estimator*. Substituting $\overline{x}$ for $\overline{X}$ yields an *estimate*. Here are some abbreviations, alluding throughout to our example of a UMA estimator:

> A *generic* $1 - \alpha$ lower confidence interval estimator is $\mu > \hat{\mu}_{1-\alpha}(\overline{X}) = \mu > \overline{X} - c_\alpha(\sigma/\sqrt{n})$.
>
> A *specific* $1 - \alpha$ lower confidence interval estimate is $\mu > \hat{\mu}_{1-\alpha}(\overline{x}) = \mu > \overline{x} - c_\alpha(\sigma/\sqrt{n})$.

The corresponding value for α is close enough to 0.005 to allow $c_{0.005} = 2.5$ (it's actually closer to 0.006). The impressive thing is that, regardless of the true value of μ, these rules have high coverage probability. If, for any observed $\overline{x}$, in our example, you shout out

$$\mu > \overline{X} - 2.5(\sigma/\sqrt{n}),$$

your assertions will be correct 99.5% of the time. The specific inference results from plugging in $\overline{x}$ for $\overline{X}$. The specific 0.995 lower limit = $\hat{\mu}_{0.995}(\overline{x}) = \overline{x} - 2.5(\sigma/\sqrt{n})$, and the specific 0.995 estimate is $\mu > \hat{\mu}_{0.995}(\overline{x})$. This inference is qualified by the error probability of the method, namely the confidence level 0.995. But the upshot of this qualification is often misunderstood. Let's have a new example to show the duality between the lower confidence interval estimator $\mu > \hat{\mu}_{1-\alpha}(\overline{X})$ and the *generic* (α level) test T+ of form: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$. The "accident at the water plant" has a nice standard error of 1, but that can mislead about the role of sample size n. Let $\sigma = 1$, $n = 25$, $\sigma_{\overline{X}} = (\sigma/\sqrt{n}) = 0.2$. (Even though we'd actually have to estimate σ, the logic is the same and it's clearer.) I use $\sigma/\sqrt{n}$ rather than $\sigma_{\overline{X}}$ when a reminder of sample size seems needed.

Work backwards. Suppose we've collected the 25 samples and observed sample mean $\overline{x} = 0.6$. (The 0.6 has nothing to do with the polling example at the outset.) For what value of μ_0 would $\overline{x} = 0.6$ exceed μ_0 by $2.5\sigma_{\overline{X}}$? Since $2.5\sigma_{\overline{X}} = 0.5$, the answer is $\mu = 0.1$. If we were testing H_0: $\mu \leq 0.1$ vs. H_1: $\mu > 0.1$ at level 0.005, we'd reject with this outcome. The corresponding 0.995 lower estimate would be

$$\mu > 0.1.$$

(see Note 1).

Now for the duality. $\overline{X}$ is not statistically significantly greater than any μ value larger than 0.1 (e.g., 0.15, 0.2, etc.) at the 0.005 level. A test of form T+ would fail to reject each of the values in the CI interval at the 0.005 level, with $\overline{x} = 0.6$. Since this is continuous, it does not matter if the cut-off is at 0.1 or greater than or equal to 0.1.[1] By contrast, if we were testing μ_0 values 0.1 or less (T+: H_0: $\mu \leq 0.1$ against H_1: $\mu > 0.1$), these nulls *would* be rejected by $\overline{x} = 0.6$ at the 0.005 level (or even lower for values less than 0.1). That is, under the supposition that the data were generated from a world where H_0: $\mu \leq 0.1$, at least 99.5% of the time a *smaller* $\overline{X}$ than what was observed (0.6) would occur:

$$\Pr(\overline{X} < 0.6; \mu = 0.1) = 0.995.$$

The probability of observing $\overline{X} \geq 0.6$ would be low, 0.005.

Severity Fact (for test T+): Taking an outcome $\overline{x}$ that just reaches the α level of significance ($\overline{x}_\alpha$)as warranting H_1: $\mu > \mu_0$ with severity $(1 - \alpha)$

[1] To avoid confusion, note the duality is altered accordingly. If we set out the test rule for T + H_0: $\mu \leq \mu_0$ vs H_1: $\mu > \mu_0$ as reject H_0: iff $\overline{X} \geq \mu_0 + c_\alpha(\sigma/\sqrt{n})$, then we do not reject H_0 iff $\overline{X} < \mu_0 + c_\alpha(\sigma/\sqrt{n})$. This is the same as $\mu_0 > \overline{X} - c_\alpha(\sigma/\sqrt{n})$, the corresponding lower CI bound. If the test rule is $\overline{X} > \mu_0 + c_\alpha(\sigma/\sqrt{n})$, the corresponding lower bound is $\mu_0 \geq \overline{X} - c_\alpha(\sigma/\sqrt{n})$.

is mathematically the same as inferring $\mu > \overline{x} - c_\alpha(\sigma/\sqrt{n})$ at level $(1 - \alpha)$.

Hence, there's an intimate mathematical relationship between severity and confidence limits. However, severity will break out of the fixed $(1 - \alpha)$ level, and will supply a non-behavioristic rationale that is now absent from confidence intervals.[2]

Severity and Capabilities of Methods

Begin with an instance of our "Fact": To take an outcome that just reaches the 0.005 significance level as warranting H_1 with severity 0.995, is the same as taking the observed $\overline{x}$ and inferring μ just exceeds the 99.5 and lower confidence bound: $\mu > 0.1$. My justification for inferring $\mu > 0.1$ (with $\overline{x} = 0.6$) is this. Suppose my inference is false. Take the smallest value that renders it false, namely $\mu = 0.1$. Were $\mu = 0.1$, then the test very probably would have resulted in a smaller observed $\overline{X}$ than I got (0.6). That is, 99.5% of the time it would have produced a result *less discordant* with claim $\mu > 0.1$ than what I observed. (For μ values less than 0.1 this probability is increased.) Given that the method was highly *in*capable of having produced a value of $\overline{X}$ as large as 0.6, if $\mu \leq 0.1$, we argue that there is an indication at least (if not full blown evidence) that $\mu > 0.1$. The severity with which $\mu > 0.1$ "passes" (or is indicated by) this test is approximately 0.995.

Some caveats: First, throughout this exercise, we are assuming these values are "audited," and the assumptions of the model permit the computations to be licit. Second, we recognize full well that we merely have a single case, and inferring a genuine experimental effect requires being able to produce such impressive results somewhat regularly. That's why I'm using the word "indication" rather than evidence. Interestingly though, you don't see the same admonition against "isolated" CIs as with tests. (Rather than repeating these auditing qualifications, I will assume the context directs the interpretation.)

Severity versus Performance. The severity interpretation differs from both the construals that are now standard in confidence interval theory: The first is the *coverage probability* construal, and the second I'm calling *rubbing-off*. The coverage probability rationale is straightforwardly performance oriented. The rationale for the rule: infer

[2] For the computations, in test T+: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$. Suppose the observed $\overline{x}$ just reaches the c_α cut-off: $\overline{x} = \mu_0 + c_\alpha\sigma_{\overline{X}}$. The $(1 - \alpha)$ CI lower bound, CI_L, is $\mu > \overline{X} - c_\alpha\sigma_{\overline{X}}$. So Pr(test T+ does not reject $H_0; \mu = \text{CI}_\text{L}) = \Pr(\overline{X} < \mu_0 + c_\alpha\sigma_{\overline{X}}; \mu = \mu_0)$. Standardize $\overline{X}$ to get Z: $Z = [(\mu_0 + c_\alpha\sigma_{\overline{X}}) - \mu_0](1/\sigma_{\overline{X}}) = c_\alpha$. So the severity for $\mu > \mu_0 = \Pr$(test T+ does not reject $H_0; \mu = \text{CI}_\text{L}) = \Pr(Z < c_\alpha) = (1 - \alpha)$.

$$\mu > \overline{X} - 2.5\sigma/\sqrt{n},$$

is simply that you will correctly cover the true value at least 99.5% of the time in repeated use (we can allow the repetitions to be actual or hypothetical):

$$\Pr(\mu > (\overline{X} - 2.5\sigma/\sqrt{n}); \mu) = 0.995.$$

Aside: The equation above is not treating μ as a random variable, although it might look that way. $\overline{X}$ is the random variable. It's the same as asserting $\Pr(\overline{X} \geq \mu + 2.5(\sigma/\sqrt{n}); \mu) = 0.005$. Is this performance-oriented interpretation really all you can say? The severe tester says no. Here's where different interpretive philosophies enter.

Cox and Hinkley (1974) do not adhere to a single choice of $1 - \alpha$. Rather, to assert a 0.995 CI estimate, they say, is to follow:

> . . . a procedure that would be wrong only in a proportion α of cases, in hypothetical repeated applications, whatever may be the true value μ. Note that this is a hypothetical statement that gives an empirical meaning, which in principle can be checked by experiment, rather than a prescription for using confidence limits. In particular, we do not recommend or intend that a fixed value α_0 should be chosen in advance and the information in the data summarized in the single assertion $[\mu > \hat{\mu}_{1-\alpha}]$. (p. 209, μ is substituted for their θ)

We have the *meaning versus application* gap again, which severity strives to close. "[W]e define procedures for assessing evidence that are calibrated by how they would perform were they used repeatedly. In that sense they do not differ from other measuring instruments" (Cox 2006a, p. 8). Yet this performance is not the immediate justification for the measurement in the case at hand. What I mean is, it's not merely that if you often use a telescope with good precision, your measurements will have a good track record – no more than with my scales (in Section 1.1). Rather, the thinking is, knowing how they would perform lets us infer how they're performing now. Good long-run properties "rub-off" in some sense on the case at hand (provided at least they are the relevant ones).

It's not so clear what's being rubbed off. You can't say the probability it's correct *in this case* is 0.995, since either it's right or not. That's why "confidence" is introduced. Some people say from the fact that the procedure is rarely wrong we may assign a low probability to its being wrong in the case at hand. First, this is dangerously equivocal, since the probability properly attaches to the method of inferring. Some espouse it as an informal use of "probability" outside of statistics, for instance, that confidence is "the degree of belief of a rational person that the confidence interval covers

the parameter" (Schweder and Hjort 2016, p. 11). They call this "epistemic probability." My main gripe is that neither epistemic probability, whatever it is, nor performance gives a report of well-testedness associated with the claim at hand.

By providing several limits at different values, we get a more informative assessment, sometimes called a confidence distribution (CD). An early reference is Cox (1958). "The set of all confidence intervals at different levels of probability. . . [yields a] confidence distribution" (Cox 1958, p. 363). We'll visit others later. The severe tester still wants to nudge the CD idea; whether it's a large or small nudge is unclear because members of CD tribes are unclear. By and large, they're either a tad bit too performance oriented or too close to a form of probabilism for a severe tester. Recall I've said I don't see the severity construal out there, so I don't wish to saddle anyone with it. If that is what some CD tribes intend, great.

The severity logic is the counterfactual reasoning: Were μ less than the 0.995 lower limit, then it is very probable (> 0.995) that our procedure would yield a smaller sample mean than 0.6. This probability gives the severity. To echo Popper, $\mu > \hat{\mu}_{1-\alpha}$ is corroborated (at level 0.995) because it may be presented as a *failed attempt to falsify* it statistically. The severe testing philosophy hypothesizes that this is how humans reason. It underwrites formal error statistics as well as day-to-day reasoning.

Exhibit (vii): Capability. Let's see how severity is computed for the CI claim $(\mu > \hat{\mu}_{0.995})$ with $\overline{x} = 0.6$:

1. The particular assertion h is $\mu > 0.1 (\hat{\mu}_{0.995} = 0.1)$.
2. $\overline{x} = 0.6$ accords with h, an assertion about a positive discrepancy from 0.1.
3. Values of $\overline{X}$ less than 0.6 accord less well with h. So we want to compute the probability $(\overline{X} < 0.6)$ just at the point that makes h false: $\mu = 0.1$.
   Pr(method would yield $\overline{X} < 0.6; 0.1) = 0.995$.
4. From (3), SEV($\mu > 0.1$) = 0.995 (or we could write ≥, but our convention will be to write =).

Although we are moving between values of the parameter and values of $\overline{X}$, so long as we are careful, there is no illegitimacy. We can see that CI limits follow severity reasoning. For general lower $1 - \alpha$ limits, with small level α:

The inference of interest is h: $\mu > \hat{\mu}_{1-\alpha}$.
Since Pr(method would yield $\overline{X} < \overline{x}; \mu = \hat{\mu}_{1-\alpha}) = (1 - \alpha)$,
it follows that SEV(h) = $(1 - \alpha)$.

(Lower case *h* emphasizes these are typically members of the full alternative in a test.) Table 3.5 gives several examples.

Perhaps "capability or incapability" of the method can serve to get at what's rubbing off. The specific moral I've been leading up to can be read right off the Table, as we vary the value for α (from 0.001 to 0.84) and form the corresponding lower confidence bound from $\hat{\mu}_{0.999}$ to $\hat{\mu}_{0.16}$.

> The higher the test's capability to produce such large (or even larger) differences as we observe, under the assumption $\mu = \hat{\mu}$, the *less* severely tested is assertion $\mu > \hat{\mu}$. (See Figure 3.3.)

The third column of Table 3.5 gives the complement to the severity assessment: the capability of a more extreme result, which in this case is α: $\Pr(\overline{X} > \overline{x}; \mu = \hat{\mu}_{1-\alpha}) = \alpha$. This is the Π function – the attained sensitivity in relation to μ: $\Pi(\gamma)$ (section 3.3) – but there may be too many moving parts to see this simply right away. You can return to it later.

We do not report a single, but rather several confidence limits, and the corresponding inferences of form h_1. Take the third row. The 0.975 lower limit that would be formed from $\overline{x} = 0.6$, $\hat{\mu}_{0.975}$, is $\mu = 0.2$. The estimate takes the form $\mu > 0.2$. Moreover, the observed mean, 0.6, is statistically significantly greater than 0.2 at level 0.025. Since $\mu = 0.2$ would very probably produce $\overline{X} < 0.6$, the severe tester takes the outcome as a good indication of $\mu \geq 0.2$. I want to draw your attention to the fact that the probability of producing an $\overline{X} \geq 0.6$ ranges from 0.005 to 0.5 for values of μ between 0.1 and the observed $\overline{x} = 0.6$. It never exceeds 0.5. To see this compute $\Pr(\overline{X} \geq 0.6; \mu = \mu')$ letting μ' range from 0.1 to 0.6. We standardize $\overline{X}$ to get $Z = (\overline{X} - \mu')/(\sigma/\sqrt{n})$ which is N(0,1). To find $\Pr(\overline{X} \geq 0.6; \mu = \mu')$, compute $Z = (0.6 - \mu')/0.2$ and use the areas under the standard Normal curve to get $\Pr(Z \geq z_0)$, μ' ranging from 0.1 to 0.6.

Do you notice it is only for negative z values that the area to the right of z exceeds 0.5? The test only begins to have more than 50% capability of generating observed means as large as 0.6, when μ is larger than 0.6. An important benchmark enters. The lower 0.5 bound $\hat{\mu}_{0.5}$ is 0.6. Since a result even larger than observed is brought about 50% of the time when $\mu = 0.6$, we rightly *block* the inference to $\mu > 0.6$.

Go to the next to last row: using a lower confidence limit at level 0.31! Now nobody goes around forming confidence bounds at level 0.5, let alone 0.31, but they might not always realize that's what they're doing! We could give a performance-oriented justification: the inference to $\mu > 0.7$ from $\overline{x} = 0.6$ is an instance of a rule that errs 31% of the time. Or we could use counterfactual,

Table 3.5 Lower confidence limit with $\overline{x} = 0.6$, α ranging from 0.001 to 0.84 in T+: $\sigma = 1$, $n = 25$, $\sigma_{\overline{X}} = (\sigma/\sqrt{n}) = 0.2, \overline{x} = 0.6$

α	c_α	$\hat{\mu}_{1-\alpha}$	h_1:$\mu > \hat{\mu}_{1-\alpha}$	$\Pr(\overline{X} \geq 0.6; \mu = \hat{\mu}_{1-\alpha}) = \alpha$	SEV(h_1)
0.001	3	0	($\mu > 0$)	0.001	0.999
0.005	2.5	0.1	($\mu > 0.1$)	0.005	0.995
0.025	2	0.2	($\mu > 0.2$)	0.025	0.975
0.07	1.5	0.3	($\mu > 0.3$)	0.07	0.93
0.16	1	0.4	($\mu > 0.4$)	0.16	0.84
0.3	0.5	0.5	($\mu > 0.5$)	0.3	0.7
0.5	0	0.6	($\mu > 0.6$)	0.5	0.5
0.69	−0.5	0.7	($\mu > 0.7$)	0.69	0.31
0.84	−1	0.8	($\mu > 0.8$)	0.84	0.16

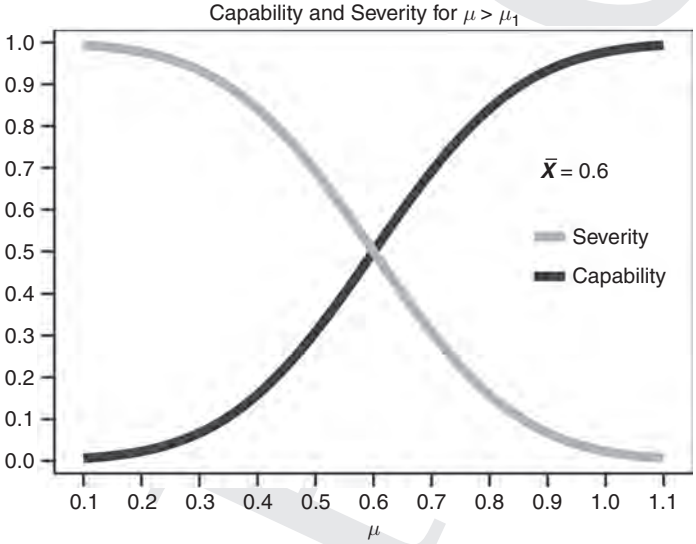

Figure 3.3 Severity for $\mu > \hat{\mu}$.

severity reasoning: even if μ were only 0.7, we'd get a larger $\overline{X}$ than we observed a whopping 69% of the time. Our observed $\overline{X}$ is terrible grounds to suppose μ must exceed 0.7. If anything, we're starting to get an indication that $\mu < 0.7$! Observe that, with larger α, the argument is more forceful by emphasizing >, rather than ≥, but it's entirely correct either way, as it is continuous.

In grasping the duality between tests and confidence limits we consider the *general form* of the test in question, here we considered T+. Given the general form, we imagine the test hypotheses varying, with a fixed outcome $\overline{x}$. Considering other instances of the general test T+ is a heuristic aid in interpreting confidence limits using the idea of statistical inference as severe testing. We will often allude to confidence limits to this end. However, the way the severe tester will actually use the duality is best seen as a post-data way to ask about various discrepancies indicated. For instance, in testing H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$, we may wish to ask, post-data, about a discrepancy such as $\mu > 0.2$. That is, we ask, for each of the inferences, how severely passed it is.

Granted this interval estimator has a nice pivot. If I thought the nice cases weren't the subject of so much misinterpretation, I would not start there. But there's no chance of seeing one's way into more complex cases if we are still hamstrung by the simple ones. In fact, the vast majority of criticism and proposed reforms revolve around our test T+ and two-sided variants. If you grasp the small cluster of the cases that show up in the debates, you'll be able to extend the results. The severity interpretation enables confidence intervals to get around some of their current problems. Let's visit a few of them now. (See also Excursion 4 Tour II, Excursion 5 Tour II.)

Exhibit (viii): Vacuous and Empty Confidence Intervals: Howlers and Chestnuts. *Did you hear the one about the frequentist who reports a confidence level of 0.95 despite knowing the interval must contain the true parameter value?*

Basis for the joke: it's possible that CIs wind up being vacuously true: including all possible parameter values. "Why call it a 95 percent CI if it's known to be true?" the critics ask. The obvious, performance-based, answer is that the confidence level refers to the probability the method outputs true intervals; it's not an assignment of probability to the specific interval. It's thought to be problematic only by insisting on a probabilist interpretation of the confidence level. Jose Bernardo thinks that the CI user "should be subject to some re-education using well-known, standard counterexamples. . . . conventional 0.95-confidence regions may actually consist of the whole real line" (Bernardo 2008, p. 453). Not so.

Cox and Hinkley (1974, p.226) proposed interpreting confidence intervals, or their corresponding confidence limits (lower or upper), as the set of parameter values consistent at the confidence level.

> This interpretation of confidence intervals also scotches criticisms of examples where, due to given restrictions, it can happen that a $(1 - \alpha)$ estimate contains all possible

parameter values. Although such an inference is 'trivially true,' it is scarcely vacuous in our construal. That all parameter values are consistent with the data is an informative statement about the limitations of the data to detect discrepancies at the particular level. (Cox and Mayo 2010, p. 291)

Likewise it can happen that all possible parameter points are inconsistent with the data at the $(1 - \alpha)$ level. Criticisms of "vacuous" and empty confidence intervals stem from a probabilist construal of $(1 - \alpha)$ as the degree of support, belief, or probability attached to the particular interval; but this construal isn't endorsed by CI interval methodology. There is another qualification to add: the error probability computed must be relevant. It must result from the relevant sampling distribution.

Pathological Confidence Set. Here's a famous chestnut that is redolent of Exhibit (vi) in Section 3.4 (Cox's 1958 two measuring instruments with different precisions). It is usually put in terms of a "confidence set" with $n = 2$. It could also be put in the form of a test. Either way, it is taken to question the relevance of error statistical assessments in the case at hand (e.g., Berger and Wolpert 1988, Berger 2003, p. 6). Two independent and identically distributed observations are to be made represented by random variables X_1, X_2. Each X can take either value $\psi - 1$ or $\psi + 1$ with probability of 0.5, where ψ is the unknown parameter to be estimated using the data. The data can result in both outcomes being the same, or both different.

Consider the second case: With both different, we know they will differ by 2. A possibility might be $\langle 9, 11 \rangle$. Right away, we know ψ must be 10. What luck! We know we're right to infer ψ is 10. To depict this case more generally, the two outcomes are $\boldsymbol{x}_1 = x' - 1$ and $\boldsymbol{x}_2 = x' + 1$, for some value x'.

Consider now that the first case obtains. We are not so lucky. The two outcomes are the same: $x_1 = x_2$ (maybe they're both 9 or whatever). What should we infer about the value of parameter ψ? We know ψ is either $x_1 - 1$ or $x_1 + 1$ (e.g., 8 or 10); each accords equally well with the data. Say we infer ψ is $x_1 - 1$. The method is correct with probability 0.5. Averaging over the two possibilities, the probability of an erroneous inference is 0.25. Now suppose I was lucky and observed two different outcomes. Then I know the value of ψ so it makes no sense to infer "ψ is $(x_1 + x_2)/2$" while attaching a confidence coefficient of 0.75.

You see the pattern. The example is designed so that some outcomes yield much more information than others. As with Cox's "two measuring instruments," the data have two parts: First, an indication of whether the two outcomes are the same or different; second, the observed result. Let A be an indicator of the first part: $A = 1$ if both are the same

(unlucky); $A = 2$ if the sample values differ by 2 (lucky!). The full data may be represented as $(A, \boldsymbol{x})$. The distribution of A is fixed independently of the parameter of interest: $\Pr(A = 1) = \Pr(A = 2) = 0.5$. It is an example of an *ancillary* statistic. However, learning whether $A = 1$ or $A = 2$ is very informative as to the precision achieved by the inference. Thus the relevant properties associated with the particular inference would be conditional on the value of A.

The tip-off that we're dealing with a problem case is this: The sufficient statistic S has two parts $(A, \boldsymbol{X})$, that is it has *dimension* 2. But there's only one parameter ψ. Without getting into the underlying theory, this alone indicates that S has a property known as being *incomplete*, opening the door to different P-values or confidence levels when calculated conditionally on the value of A. In particular, the marginal distribution of a P-value averaged over the two possibilities $(0.5(0) + 0.5(0.5) = 0.25)$ would be misleading for any particular set of data. Instead we condition on the value of A obtained. David Cox calls this process "*technical conditioning to induce relevance* of the frequentist probability to the inference at hand" (Cox and Mayo 2010, pp. 297–8).

Such examples have other noteworthy features: the ancillary part A gives a sneaky way of assigning a probability to "being correct" in the subset of cases given by the value of A. It's an example of what Fisher called "recognizable subsets." By careful artifice, the event that "a random variable A takes a given value a" is equivalent to "the data were generated by a hypothesized parameter value." So the probability of $A = a$ gives the probability a hypothesis is true. Aris Spanos considers these examples "rigged" for this reason, and he discusses these and several other famous pathological examples (Spanos 2012).

Even putting pathologies aside, is there any reason the frequentist wouldn't do the sensible thing and report on how well probed the inference is once A is known? No. Certainly a severe testing theorist would.

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 this system.

What if the test or interval estimation procedure does not pass the audit? Consider for the moment that there has been optional stopping, or cherry picking, or multiple testing. Where these selection effects are well understood, we may adjust the error probabilities so that they do pass the audit. But what if the moves are so tortuous that we can't reliably make the adjustment? Or

perhaps we don't feel secure enough in the assumptions? Should the severity for $\mu > \mu_0$ be low or undefined?

You are free to choose either. The severe tester says SEV($\mu > \mu_0$) is low. As she sees it, having evidence requires a minimum threshold for severity, even without setting a precise number. If it's close to 0.5, it's quite awful. But if it cannot be computed, it's also awful, since the onus on the researcher is to satisfy the minimal requirement for evidence. I'll follow her: If we cannot compute the severity even approximately (which is all we care about), I'll say it's low, along with an explanation as to why: It's low because we don't have a clue how to compute it!

A probabilist, working with a single "probability pie" as it were, would take a low probability for H as giving a high probability to $\sim H$. By contrast we wish to clearly distinguish between having poor evidence for H and having good evidence for $\sim H$. Our way of dealing with bad evidence, no test (BENT) allows us to do that. Both SEV(H) and SEV($\sim H$) can be low enough to be considered lousy, even when both are computable.

Souvenir N: Rule of Thumb for SEV

> Can we assume that if SEV($\mu > \mu_0$) is a high value, $1 - \alpha$, then SEV($\mu \leq \mu_0$) is α?

Because the claims $\mu > \mu_0$ and $\mu \leq \mu_0$ form a partition of the parameter space, and because we are assuming our test has passed (or would pass) an audit, else these computations go out the window, the answer is yes.

> If SEV($\mu > \mu_0$) is high, then SEV($\mu \leq \mu_0$) is low.

The converse need not hold – given the convention we just saw in Exhibit (ix). At the very least, "low" would not exceed 0.5.

A rule of thumb (for test T+ or its dual CI):

- If we are pondering a claim that an observed difference from the null seems *large* enough to indicate $\mu > \mu'$, we want to be sure the test was highly capable of producing *less* impressive results, were $\mu = \mu'$.
- If, by contrast, the test was highly capable of producing *more* impressive results than we observed, even in a world where $\mu = \mu'$, then we block an inference to $\mu > \mu'$ (following weak severity).

This rule will be at odds with some common interpretations of tests. Bear with me. I maintain those interpretations are viewing tests through "probabilist-colored" glasses, while the correct error-statistical view is this one.

3.8 The Probability Our Results Are Statistical Fluctuations: Higgs' Discovery

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." With the March 2013 data analysis, the 5-sigma difference grew to 7 sigmas, and some of the apparent anomalies evaporated. In October 2013, the Nobel Prize in Physics was awarded jointly to François Englert and Peter W. Higgs for the "theoretical discovery of a mechanism" behind the particle experimentally discovered by the collaboration of thousands of scientists (on the ATLAS and CMS teams) at CERN's Large Hadron Collider in Switzerland. Yet before the dust had settled, the very nature and rationale of the 5-sigma discovery criterion began to be challenged among scientists and in the popular press. Because the 5-sigma standard refers to a benchmark from frequentist significance testing, the discovery was immediately imbued with controversies that, at bottom, concern statistical philosophy.

Why a 5-sigma standard? Do significance tests in high-energy particle (HEP) physics escape the misuses of *P*-values found in social and other sciences? Of course the main concern wasn't all about philosophy: they were concerned that their people were being left out of an exciting, lucrative, many-years project. But unpacking these issues is philosophical, and that is the purpose of this last stop of Excursion 3. I'm an outsider to HEP physics, but that, apart from being fascinated by it, is precisely why I have chosen to discuss it. Anyone who has come on our journey should be able to decipher the more public controversies about using *P*-values.

I'm also an outsider to the International Society of Bayesian Analysis (ISBA), but a letter was leaked to me a few days after the July 4, 2012 announcement, prompted by some grumblings raised by a leading subjective Bayesian, Dennis Lindley. The letter itself was sent around to the ISBA list by statistician Tony O'Hagan. "Dear Bayesians," the letter began. "We've heard a lot about the Higgs boson."

> 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. (O'Hagan 2012)

Neither of these seemed to be the case in his opinion: "[Is] the particle physics community completely wedded to frequentist analysis? If so, has anyone tried to explain what bad science that is?" (ibid.).

Bad science? Isn't that a little hasty? HEP physicists are sophisticated with their statistical methodology: they'd seen too many bumps disappear. They want to ensure that before announcing a new particle has been discovered that, at the very least, the results being spurious is given a run for its money. Significance tests, followed by confidence intervals, are methods of choice here for good reason. You already know that I favor moving away from traditional interpretations of statistical tests and confidence limits. But some of the criticisms, and the corresponding "reforms," reflect misunderstandings, and the knottiest of them all concerns the very meaning of the phrase (in the title of Section 3.8): "the probability our results are merely statistical fluctuations." Failing to clarify it may well impinge on the nature of future big science inquiry based on statistical models. The problem is a bit delicate, and my solution is likely to be provocative. You may reject my construal, but you'll see what it's like to switch from wearing probabilist, to severe testing, glasses.

The Higgs Results

Here's a quick sketch of the Higgs statistics. (I follow the exposition by physicist Robert Cousins (2017). See also Staley (2017). There is a general model of the detector within which researchers define a "global signal strength" parameter μ "such that H_0: $\mu = 0$ corresponds to the background-only hypothesis and $\mu = 1$ corresponds to the [Standard Model] SM Higgs boson signal in addition to the background" (ATLAS collaboration 2012c). The statistical test may be framed as a one-sided test:

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

The test statistic $d(\boldsymbol{X})$ data records how many *excess events* of a given type are "observed" (from trillions of collisions) in comparison to what would be expected from background alone, given in standard deviation or sigma units. Such excess events give a "signal-like" result in the form of bumps off a smooth curve representing the "background" alone.

The improbability of the different $d(\boldsymbol{X})$ values, its sampling distribution, is based on simulating what it would be like under H_0 fortified with much cross-checking of results. These are converted to corresponding probabilities under a standard Normal distribution. The probability of observing results as extreme as or more extreme than 5 sigmas, under H_0, is approximately 1 in 3,500,000! Alternatively, it is said that the probability that the results were just a statistical fluke (or fluctuation) is 1 in 3,500,000.

Why such an extreme evidence requirement, Lindley asked. Given how often bumps disappear, the rule for interpretation, which physicists never intended to be rigid, is something like: if $d(\boldsymbol{X}) \geq 5$ sigma, infer discovery, if $d(\boldsymbol{X}) \geq 2$ sigma, get more data.

Now "deciding to announce" the results to the world, or "get more data" are actions all right, but each corresponds to an evidential standpoint or inference: infer there's evidence of a genuine particle, and infer that spurious bumps had not been ruled out with high severity, respectively.

What "the Results" Really Are

You know from the Translation Guide (Souvenir C) that $\Pr(d(\boldsymbol{X}) \geq 5; H_0)$ is to be read Pr (the test procedure would yield $d(\boldsymbol{X}) \geq 5; H_0$). Where do we record Fisher's warning that we can only use *P*-values to legitimately indicate a genuine effect by demonstrating an *experimental phenomenon*. In good sciences and strong uses of statistics, "the results" may include demonstrating the "know-how" to generate results that rarely fail to be significant. Also important is showing the test passes an audit (it isn't guilty of selection biases, or violations of statistical model assumptions). "The results of test T" incorporates the entire display of know-how and soundness. That's what the severe tester means by Pr(test T would produce $d(\boldsymbol{X}) \geq d(x_0); H_0$). So we get:

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

There are essentially two stages of analysis. The first stage is to test for a genuine Higgs-like particle, the second, to determine its properties (production mechanism, decay mechanisms, angular distributions, etc.). Even though the SM Higgs sets the signal parameter to 1, the test is going to be used to learn about the value of any discrepancy from 0. Once the null is rejected at the first stage, the second stage essentially shifts to learning the particle's properties, and using them to seek discrepancies from a new null hypothesis: the SM Higgs.

The *P*-Value Police

The July 2012 announcement gave rise to a flood of buoyant, if simplified, reports heralding the good news. This gave ample grist for the mills of *P*-value critics. Statistician Larry Wasserman playfully calls them the "P-Value Police"(2012a) such as Sir David Spiegelhalter (2012), a

professor of the Public's Understanding of Risk at the University of Cambridge. Their job was to examine if reports by journalists and scientists could be seen to be misinterpreting the sigma levels as posterior probability assignments to the various models and claims. Thumbs up or thumbs down! Thumbs up went 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 3 million would see an apparent signal this strong in a universe without a Higgs.* (2012a, emphasis added)

Now HEP physicists have a term for an apparent signal that is actually produced due to chance variability alone: a *statistical fluctuation* or *fluke*. Only one experiment in 3 million would produce so strong a background fluctuation. ATLAS (2012b) calls it the "background fluctuation probability." By contrast, Spiegethalter gave a thumbs down to:

> There is less than a one in 3 million chance that their results are a statistical fluctuation.

If they had written "would be" instead of "is" it would get thumbs up. Spiegelhalter's ratings are generally echoed by other Bayesian statisticians. According to them, the thumbs down reports are guilty of misinterpreting the *P*-value as a posterior probability on H_0.

A careful look shows this is not so. H_0 does not say the observed results are due to background alone; H_0 does not say the result is a fluke. It is just H_0: $\mu = 0$. Although if H_0 were true it *follows* that various results would occur with specified probabilities. In particular, it entails (along with the rest of the background) that large bumps are improbable.

It may in fact be seen as an ordinary error probability:

$$(1)\ \Pr(\text{test T would produce } d(\boldsymbol{X}) \geq 5;\ H_0) \leq 0.0000003.$$

The portion within the parentheses is how HEP physicists understand "a 5-sigma fluctuation." Note (1) is not a conditional probability, which involves a prior probability assignment to the null. It is not

$$\Pr(\text{test T would produce } d(\boldsymbol{X}) \geq 5 \text{ and } H_0)/\Pr(H_0).$$

Only random variables or their values are conditioned upon. This may seem to be nit-picking, and one needn't take a hard line on the use of "conditional." I mention it because it may explain part of the confusion here. The relationship between the null hypothesis and the test results is intimate: the assignment of probabilities to test outcomes or values of $d(\boldsymbol{X})$ "under the null" may be seen as a tautologous statement.

Since it's not just a single result, but also a dynamic test display, we might even want to emphasize a fortified version:

(1)* Pr(test T would display $d(\boldsymbol{X}) \geq 5$; H_0) ≤ 0.0000003.

Critics may still object that (1), even fortified as (1)*, only entitles saying:

> There is less than a one in 3 million chance of a fluctuation (at least as strong as in their results).

It does not entitle one to say:

> There is less than a one in 3 million chance that *their results* are a statistical fluctuation.

Let's compare three "ups" and three "downs" to get a sense of the distinction that leads to the brouhaha:

Ups

U-1. The probability of the background alone fluctuating up by this amount or more is about one in 3 million. (CMS 2012)
U-2. Only one experiment in 3 million would see an apparent signal this strong in a universe described in H_0.
U-3. The probability that their signal would result by a chance fluctuation was less than one chance in 3 million.

Downs

D-1. The probability their results were due to the background fluctuating up by this amount or more is about one in 3 million.
D-2. One in 3 million is the probability the signal is a false positive – a fluke produced by random statistical fluctuation.
D-3. The probability that their signal was the result of a statistical 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 fluctuation. Critics might say the objection to "this" is that the *P*-value refers to a difference as great or greater – a tail area. But if the probability of $\{d(\boldsymbol{X}) \geq d(\boldsymbol{x})\}$ is low under H_0, then Pr $(d(\boldsymbol{X}) = d(\boldsymbol{x}); H_0)$ is even lower. We've dealt with this back with Jeffreys' quip (Section 3.4). No statistical account recommends going from improbability of a point result on a continuum under H to rejecting H. The Bayesian looks to the prior

probability in H and its alternatives. The error statistician looks to the general procedure. The notation $\{d(\boldsymbol{X}) \geq d(\boldsymbol{x})\}$ is used to signal the latter.

But if we're talking about the procedure, the critic rightly points out, we are not assigning probability to these particular data or signal. 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 always treated as a generic type of event. We are never considering the probability "the background fluctuates up this much on Wednesday July 4, 2012," except as that is construed as a type of collision result at a type of detector, and so on. It's illuminating to note, at this point:

> [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's point is 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: the probability the method used would produce data displays (e.g., bumps) as impressive as these, under the assumption of H_0. 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 here. 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.

The Real Problem with D-1 through D-3. The error probabilities in U-1 through U-3 are straightforward. In the Higgs experiment, the needed computations are based on simulating relative frequencies of events where H_0: $\mu = 0$ (given a detector model). In terms of the corresponding P-value:

(1) Pr(test T would produce a P-value ≤ 0.0000003; H_0) ≤ 0.0000003.

D-1, 2, 3 are just slightly imprecise ways of expressing U-1, 2, 3. So what's the objection to D-1, 2, 3? It's the danger some find in moving from such claims to their complements. If I say there's a 0.0000003 probability their results are due to chance, some infer there's a 0.999999 (or whatever) probability their results are not due to chance – are not a false positive, are not a fluctuation. And those claims are wrong. If Pr(A; H_0) = p, for some assertion A, the probability of the complement is Pr(not-A; H_0) = $1 - p$. In particular:

(1) Pr(test T would *not* display a *P*-value ≤ 0.0000003; H_0) > 0.9999993.

There's no transposing! That is, the hypothesis after the ";" does not switch places with the event to the left of ";"! But despite how the error statistician hears D-1 through D-3, I'm prepared to grant the corresponding U claims are safer. I assure you that my destination is not merely refining statistical language, but when critics convince practitioners that they've been speaking Bayesian prose without knowing it (as in Molière), the consequences are nontrivial. I'm about to get to them.

Detaching Inferences Uses an Implicit Severity Principle

Phrases such as "the probability our results are a statistical fluctuation (or fluke) is very low" are common enough in HEP – although physicists tell me it's the science writers who reword their correct U-claims as slippery D-claims. Maybe so. But if you follow the physicist's claims through the process of experimenting and modeling, you find they are alluding to proper error probabilities. You may think they really mean an illicit posterior probability assignment to "real effect" or H_1 if you think that statistical inference takes the form of probabilism. In fact, if you're a Bayesian probabilist, and assume the statistical inference must have a posterior probability, or a ratio of posterior probabilities, you will regard U-1 through U-3 as legitimate but irrelevant to inference; and D-1 through D-3 as relevant only by misinterpreting *P*-values as giving a probability to the null hypothesis H_0.

If you are an error statistician (whether you favor a behavioral performance or a severe probing interpretation), even the correct claims U-1 through U-3 are not statistical inferences! They are the (statistical) justifications associated with implicit statistical inferences, and even though HEP practitioners are well aware of them, they should be made explicit. Such inferences can take many forms, such as those I place in brackets:

U-1. The probability of the background alone fluctuating up by this amount or more is about one in 3 million.
[Thus, our results are not due to background fluctuations.]

U-2. Only one experiment in 3 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 in 3.5 million.

[Thus the signal was not due to chance.]

The formal statistics moves from

(1) Pr(test T produces $d(\boldsymbol{X}) \geq 5$; H_0) < 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. Admittedly, moving from (1) to inferring (2) relies on the implicit assumption of error statistical testing, the severity principle. I deliberately phrase it in many ways. Here's yet another, in a Popperian spirit:

> *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.

What *is* the probability that H_0 would have "survived" (and not been falsified) at the 5-sigma level? It is the probability of the complement of the event $\{d(\boldsymbol{X}) \geq 5\}$, namely, $\{d(\boldsymbol{X}) < 5\}$ under H_0. Its probability is correspondingly 1 – 0.0000003. So the overall argument starting from a fortified premise goes like this:

> (1)* With probability 0.9999997, 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 modeled by H_0.

They did not disappear, they grew (from 5 to 7 sigma). So,

(2a) infer there's evidence of H_1: non-fluke, or (2b) infer H^*: a Higgs (or a Higgs-like) particle.

There's always the error statistical qualification of the inference in (2), given by the relevant methodological probability. Here it is a report of the stringency or severity of the test that the claim has passed, as given in (1)*: 0.9999997. We might even dub it the severity coefficient. Without making the underlying principle of testing explicit, some critics assume the argument is all about the reported *P*-value. It's a mere stepping stone to an inductive inference that is detached.

Members of a strict (N-P) behavioristic tribe might reason as follows: If you follow the rule of behavior: Interpret 5-sigma bumps as a real effect (a discrepancy from 0), you'd erroneously interpret data with probability less than 0.0000003 – a very low *error probability*. Doubtless, HEP physicists are keen to avoid repeating such mistakes as apparently finding particles that move faster than light, only to discover some problem with the electrical wiring (Reich 2012). I claim the specific evidential warrant for the 5-sigma Higgs inferences aren't low long-run errors, but being able to detach an inference based on a stringent test or a *strong* argument from coincidence.[3]

Learning How Fluctuations Behave: The Game of Bump-Hunting

Dennis Overbye (2013) wrote an article in the *New York Times*: "Chasing the Higgs," based on his interviews with spokespeople Fabiola Gianotti (ATLAS) and Guido Tonelli (CMS). It's altogether common, Tonelli explains, that the bumps they find are "random flukes" – spuriously significant results – "So 'we crosscheck everything' and 'try to kill' any anomaly that might be merely random."

> One bump on physicists' charts . . . was disappearing. But another was blooming like the shy girl at a dance. . . . nobody could remember exactly when she had come in. But she was the one who would marry the prince . . . It continued to grow over the fall until it had reached the 3-sigma level – the chances of being a fluke [spurious significance] were less than 1 in 740, enough for physicists to admit it to the realm of "evidence" of something, but not yet a discovery. (Overbye 2013)

What's one difference between HEP physics and fields where most results are claimed to be false? HEP physicists don't publish on the basis of a single, isolated (nominal) *P*-value. That doesn't mean promising effects don't disappear. "'We've made many discoveries,' Dr. Tonelli said, 'most of them false'" (ibid.).

Look Elsewhere Effect (LEE). The null hypothesis is formulated to correspond to regions where an excess or bump is found. Not knowing the mass region in advance means "the local *p*-value did not include the fact that 'pure chance' had lots of opportunities . . . to provide an unlikely occurrence" (Cousins 2017, p. 424). So here a nominal (they call it local) *P*-value is assessed at a particular, data-determined, mass. But the probability of so impressive a difference anywhere in a mass range – the global

[3] The inference to (2) is a bit stronger than merely falsifying the null because certain properties of the particle must be shown at the second stage.

P-value – would be greater than the local one. "The original concept of '5σ' in HEP was therefore mainly motivated as a (fairly crude) way to account for a multiple trials factor ... known as the 'Look Elsewhere Effect'" (ibid. p. 425). HEP physicists often report both local and global *P*-values.

Background information enters, not via prior probabilities of the particles' existence, but as to how researchers might be led astray. "If they were flukes, more data would make them fade into the statistical background ... If not, the bumps would grow in slow motion into a bona fide discovery" (Overbye 2013). So, they give the bump a hard time, they stress test, look at multiple decay channels, and they hide the details of the area they found it from the other team. When two independent experiments find the same particle signal at the same mass, it helps to overcome the worry of multiple testing, strengthening an argument from coincidence.

Once the null is rejected, the job shifts to testing if various parameters agree with the SM predictions.

> 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.

While the popular press highlighted the great success for the SM, the HEP physicists, at both stages, were vigorously, desperately seeking to uncover BSM (Beyond the Standard Model) physics.

Once again, the background knowledge of fluke behavior was central to curbing their enthusiasm about bumps that hinted at discrepancies with the new null: H_0^2: $\mu = 1$. Even though July 2012 data gave evidence of the existence of a Higgs-like particle – where calling it "Higgs-like" still kept the door open for an anomaly with the "plain vanilla" particle of the SM – they also showed some hints of such an anomaly.

Matt Strassler, who, like many, is longing to find evidence for BSM physics, was forced to concede: "The excess (in favor of BSM properties) has become a bit smaller each time ... That's an unfortunate sign, if one is hoping the excess isn't just a statistical fluke" (2013a). Or they'd see the bump at ATLAS ... and not CMS. *"Taking all of the LHC's data, and not cherry picking* ... there's nothing here that you can call 'evidence'" for the much sought BSM (Strassler 2013b). They do not say the cherry-picked results 'give evidence, but disbelief in BSM physics lead us to discount it,' as Royall's Likelihoodist may opt to. They say: "There's nothing here that you can call evidence."

Considering the frequent statistical fluctuations, and the hot competition between the ATLAS and CMS to be first, a tool for when to "curb their enthusiasm" is exactly what was wanted. So, this negative role of significance tests is crucial for denying BSM anomalies are real, and setting upper bounds for these discrepancies with the SM Higgs. Since each test has its own test statistic, I'll use $g(\boldsymbol{x})$ rather than $d(\boldsymbol{x})$.

> *Severity Principle (for non-significance):* Data provide evidence to rule out a discrepancy $\delta^\star$ to the extent that a larger $g(\boldsymbol{x}_0)$ would very probably have resulted if δ were as great as $\delta^\star$.

This can equivalently be seen as inferring confidence bounds or applying FEV. The particular value of $\delta^\star$ isn't so important at this stage. What happens with negative results here is that the indicated discrepancies get smaller and smaller as do the bumps, and just vanish. These were not genuine effects, even though there's no falsification of BSM.

Negative results in HEP physics are scarcely the stuff of file drawers, a serious worry leading to publication bias in many fields. Cousins tells of the wealth of papers that begin "Search for ..." (2017, p. 412). They are regarded as important and informative – if only in ruling out avenues for

theory development. There's another idea for domains confronted with biases against publishing negative results.

Back to O'Hagan and a 2015/2016 Update

O'Hagan published a digest of responses a few days later. When it was clear his letter had not met with altogether enthusiastic responses, he backed off, admitting that he had only been being provocative with the earlier letter. Still, he declares, the Higgs researchers would have been better off avoiding the "ad hoc" 5 sigma by doing a proper (subjective) Bayesian analysis. "They would surely be willing to [announce SM Higgs discovery] if they were, for instance, 99.99 percent certain" [SM Higgs] existed. Wouldn't it be better to report

$$\Pr(\text{SM Higgs}|\text{data}) = 0.9999?$$

Actually, no. Not if it's taken as a formal probability rather than a chosen way to abbreviate: the reality of the SM Higgs has passed a severe test. Physicists believed in a Higgs particle before building the big billion-dollar collider. Given the perfect predictive success of the SM, and its simplicity, such beliefs would meet the familiar standards for plausibility. But that's very different from having evidence for a discovery, or information about the characteristics of the particle. Many aver they didn't expect it to have so small a mass, 125 GeV. In fact, given the unhappy consequences some find with this low mass, some researchers may well have gone back and changed their prior probabilities to arrive at something more sensible (more "natural" in the parlance of HEP). Yet, their strong argument from coincidence via significance tests prevented the effect from going away.

O'Hagan/Lindley admit that a subjective Bayesian model for the Higgs would require prior probabilities to scads of high dimensional "nuisance" parameters of the background and the signal; it would demand multivariate priors, correlations between parameters, joint priors, and the ever worrisome Bayesian catchall factor: $\Pr(\text{data}|\text{not-}H^*)$. Lindley's idea of subjectively eliciting beliefs from HEP physicists is rather unrealistic here.

Now for the update. When the collider restarted in 2015, it had far greater collider energies than before. On December 15, 2015 something exciting happened: "ATLAS and CMS both reported a small 'bump' in their data" at a much higher energy level than the Higgs: 750 GeV (compared to 125 GeV) (Cartlidge 2016). "As this unexpected bump

could be the first hint of a new massive particle that is not predicted by the Standard Model of particle physics, the data generated hundreds of theory papers that attempt to explain the signal" (ibid.). I believe it was 500.

> The significance reported by CMS is still far below physicists' threshold for a discovery: 5 sigma, or a chance of around 3 in 10 million that the signal is a statistical fluke. (Castelvecchi and Gibney 2016)

We might replace "the signal" with "a signal like this" to avoid criticism. While more stringent than the usual requirement, the "we're not that impressed" stance kicks in. It's not so very rare for even more impressive results to occur by background alone. As the data come in, the significance levels will either grow or wane with the bumps:

> Physicists say that by June, or August [2016] at the latest, CMS and ATLAS should have enough data to either make a statistical fluctuation go away – if that's what the excess is – or confirm a discovery. (Castelvecchi and Gibney 2016)

Could the Bayesian model wind up in the same place? Not if Lindley/O'Hagan's subjective model merely keeps updating beliefs in the already expected parameters. According to Savage, "The probability of 'something else' . . . is definitely very small" (Savage 1962, p. 80). It would seem to require a long string of anomalies before the catchall is made sufficiently probable to start seeking new physics. Would they come up with a particle like the one they were now in a frenzy to explain? Maybe, but it would be a far less efficient way for discovery than the simple significance tests.

I would have liked to report a more exciting ending for our tour. The promising bump or "resonance" disappeared as more data became available, drowning out the significant indications seen in April. Its reality was falsified.

Souvenir O: Interpreting Probable Flukes

There are three ways to construe a claim of the form: A small *P*-value indicates it's improbable that the results are statistical flukes.

(1) The person is using an informal notion of probability, common in English. They mean a small *P*-value gives grounds (or is evidence) of a genuine discrepancy from the null. Under this reading there is no fallacy. Having inferred H^*: Higgs particle, one may say informally, "so probably we have experimentally demonstrated the Higgs," or "probably, the Higgs exists."

"So probably" H_1 is merely qualifying the grounds upon which we assert evidence for H_1.

(2) An ordinary error probability is meant. When particle physicists associate a 5-sigma result with claims like "it's highly improbable our results are a statistical fluke," the reference for "our results" includes: the overall display of bumps, with significance growing with more and better data, along with satisfactory crosschecks. Under this reading, again, there is no fallacy.

To turn the tables on the Bayesians a bit, maybe they're illicitly sliding from what may be inferred from an entirely legitimate high probability. The reasoning is this: With probability 0.9999997, our methods would show that the bumps disappear, under the assumption the data are due to background H_0. The bumps don't disappear but grow. Thus, infer H^*: real particle with thus and so properties. Granted, unless you're careful about forming probabilistic complements, it's safer to adhere to the claims along the lines of U-1 through U-3. But why not be careful in negating D claims? An interesting phrase ATLAS sometimes uses is in terms of "the background fluctuation probability": "This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with ... the Standard Model Higgs boson" (2012b, p.1).

(3) The person is interpreting the *P*-value as a posterior probability of null hypothesis H_0 based on a prior probability distribution: $p = \Pr(H_0|\boldsymbol{x})$. 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. It would be a kind of "confirmation bias" whereby one insists on finding a sentence among many that could be misinterpreted Bayesianly.

ASA 2016 Guide: Principle 2 reminds practitioners that *P*-values aren't Bayesian posterior probabilities, but it slides into questioning an interpretation sometimes used by practitioners – including Higgs researchers:

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)[4]

[4] 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.

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).

Some might allege that I'm encouraging a construal of *P*-values that physicists have bent over backwards to avoid! I admitted at the outset that "the problem is a bit delicate, and my solution is likely to be provocative." My question is whether it is legitimate to criticize frequentist measures from a perspective that assumes a very different role for probability. Let's continue with the ASA statement under principle 2:

> Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself. (Wasserstein and Lazar 2016, p. 131)

Start from the very last point: what does it mean, that it's not "about the explanation"? I think they mean it's not a posterior probability on a hypothesis, and that's correct. The *P*-value is a methodological probability that can be used to quantify "how well probed" rather than "how probable." Significance tests can be the basis for, among other things, falsifying a proposed explanation of results, such as that they're "merely a statistical fluctuation." So the statistical inference that emerges is surely a statement about the explanation. Even proclamations issued by high priests – especially where there are different axes to grind – should be taken with severe grains of salt.

As for my provocative interpretation of "probable fluctuations," physicists might aver, as does Cousins, that it's the science writers who take liberties with the physicists' careful U-type statements, turning them into D-type statements. There's evidence for that, but I think physicists may be reacting to criticisms based on how things look from Bayesian probabilists' eyes. For a Bayesian, once the data are known, they are fixed; what's

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.

These principles are of minimal help when it comes to understanding and using *P*-values. The first thing that jumps out is the absence of any mention of *P*-values as error probabilities. (Fisher-N-P Incompatibilist tribes might say "they're not!" In tension with this is the true claim (under #4) that cherry picking results in spurious *P*-values; p. 132.) The ASA effort has merit, and should be extended and deepened.

random is an agent's beliefs or uncertainties on what's unknown – namely the hypothesis. For the severe tester, considering the probability of $\{d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0)\}$ is scarcely irrelevant once $d(\boldsymbol{x}_0)$ is known. It's the way to determine, following the severe testing principles, whether the null hypothesis can be falsified. ATLAS reports, on the basis of the *P*-value display, that "these results provide conclusive evidence for the discovery of a new particle with mass [approximately 125 GeV]" (ATLAS collaboration 2012b, p. 15).

Rather than seek a high probability that a suggested new particle is real; the scientist wants to find out if it disappears in a few months. As with GTR (Section 3.1), at no point does it seem we want to give a high formal posterior probability to a model or theory. We'd rather vouchsafe some portion, say the SM model with the Higgs particle, and let new data reveal, perhaps entirely unexpected, ways to extend the model further. The open-endedness of science must be captured in an adequate statistical account. Most importantly, the 5-sigma report, or corresponding *P*-value, strictly speaking, *is not the statistical inference*. Severe testing premises – or something like them – are needed to move from statistical data plus background (theoretical and empirical) to detach inferences with lift-off.